ISLAMIC UNIVERSITY OF TECHNOLOGY

Misogyny Detection in Social Media for Under-Resourced Bangla Language

Authors

Md. Wasif Kader (180042138)

Chowdhury Farhan Jamil (180042134)

Md. Tanvir Hasan Abir (180042121)

Supervised By

Dr. Hasan Mahmud	Md. Mohsinul Kabir	Dr. Kamrul Hasan
Associate Professor	Assistant Professor	Professor

Systems and Software Lab (SSL)

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

A Subsidiary Organ of the Organization of Islamic Cooperation (OIC)

Dhaka, Bangladesh.

A thesis submitted in partial fulfilment of the requirements for the degree of B. Sc. in Software Engineering

Academic Year: 2021-2022

May 2023

Declaration of Authorship

This is to declare that the work presented in this thesis is the outcome of the analysis and experiments carried out by Md. Wasif Kader, Chowdhury Farhan Jamil, and Md. Tanvir Hasan Abir under the supervision of Associate Professor Dr. Hasan Mahmud, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

(Signature of the Candidate)
Md. Wasif Kader - 180042138
May 2023

(Signature of the Candidate)
Chowdhury Farhan Jamil - 180042134
May 2023

(Signature of the Candidate)
Md. Tanvir Hasan Abir - 180042121
May 2023

Misogyny Detection in Social Media for Under-Resourced Bangla Language

Supervisors:

Dr. Hasan Mahmud
Associate Professor,
Department of Computer Science and Engineering,
Islamic University of Technology (IUT), Dhaka, Bangladesh.

Md. Mohsinul Kabir
Assistant Professor,
Department of Computer Science and Engineering,
Islamic University of Technology (IUT), Dhaka, Bangladesh.

Dr. Md. Kamrul Hasan
Professor,
Department of Computer Science and Engineering,
Islamic University of Technology (IUT), Dhaka, Bangladesh.

Abstract

This study presents a new strategy based on Natural Language Processing (NLP) techniques for detecting and mitigating misogyny on social media. In this study a dataset was constructed of 3.8 million instances of hate speech from various social media networks that were collected meticulously. Advances in this research are substantially hampered by the lack of a sizable Bengali dataset for the detection of hate speech and sexism in Bengali language texts, making it difficult to effectively identify and address these problems. To improve the representation of hate speech in the dataset, an embedding model based on informal FastText is presented, which captures the complex semantics of hate speech more accurately than other methods. This improved word embedding model is incorporated into a Bidirectional Long Short-Term Memory (BiLSTM) architecture in order to identify contextual dependencies and sequential patterns within hate speech comments. The model's layers are trained to encode and comprehend sequential information while taking both preceding and subsequent context into account, enabling it to better comprehend remarks and their context. The proposed methodology is evaluated exhaustively on a meticulously annotated dataset, allowing for a thorough analysis of its performance. Measurements of precision, recall, and F1-score are used to evaluate the accuracy and effectiveness of hate speech detection. The results demonstrate the framework's superior performance and discrimination capabilities, validating its capacity to accurately identify and categorize instances of hate speech. In addition, this research contributes the largest dataset of hate speech in the field and introduces a word embedding model that transcends existing techniques. These findings substantially improve the understanding and detection of hate speech on social media platforms, laying the groundwork for more effective mechanisms to combat hate speech and promote safer online communities.

Keyword - Hate Speech; Misogyny; FastText; Word Embedding; Semantics; Sequential Pattern; Contextual Dependency; Bi-Directional Processing; Bi-LSTM; Dataset

Acknowledgements

We would like to express our grateful appreciation for Dr. Hasan Mahmud, Associate Professor, Department of Computer Science & Engineering, IUT for being our adviser and mentor. His motivation, suggestions and insights for this research have been invaluable. Without his support and proper guidance this research would never have been possible. His valuable opinion provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way.

We also extend our sincere appreciation and gratitude to Md. Mohsinul Kabir, Assistant Professor, Department of Computer Science & Engineering, IUT, for his invaluable guidance and support throughout this research. His motivation, insightful suggestions, and expert insights were essential to the completion of this project successfully. This research would not have been possible without his unwavering support and insightful direction. We are truly grateful for his insightful opinions, dedicated time, and contributions at every stage of the thesis work. We express our profound appreciation for his valuable contribution and unwavering support.

We express our heartfelt appreciation to Dr. Kamrul Hasan, Professor, Department of Computer Science & Engineering, IUT for his invaluable guidance, expertise, and continuous support throughout the completion of our thesis.

We would not have been able to complete this research without the guidance and assistance of several individuals who contributed in various ways. I would like to acknowledge and appreciate their valued help.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

Social media has become a vital part of many people's life, serving as a platform for communication, self-expression, and interpersonal connection. The youth have wholeheartedly embraced the Internet as a means of socializing and communication [1]. Unfortunately, this widespread adoption has also resulted in the proliferation of antisocial behavior and online abuse among them [2]. For the purposes of monitoring, preventing, or mitigating cyberbullying, a dependable detection system for cyberbullying on a social network is required. Although the phenomenon of cyberbullying has been the subject of substantial research in the social sciences, the most of the published works on the topic have been written in major languages such as English, leaving Bangla relatively untouched [3].

Hate speech may be further subdivided into the categories [4] of misogyny and sexism, both of which attack their victims on the basis of their gender or sexuality [5]. The hatred or prejudice against women, has a long and troubling history. It has taken many forms over the years, from discrimination and violence to more subtle forms of oppression and belittling. Identifying and limiting the dissemination of hate speech is essential for protecting human rights and preventing the marginalization of individuals and groups. The use of informal language, misspellings, offensive terms, and non-standard abbreviations in social media and online streaming site comments adds to the difficulty of combating hate speech [6].

Bangla is spoken by over 230 million native speakers in Bangladesh, India, and the rest of the world. Facebook is used by more than 90 percent of Bangladesh's 80.83 million individuals who have access to the internet. The majority of these users are young people who are seeking safety and suffering fear [7]. The study offers a detailed investigation with the goal of enhancing the identification of hate speech in Bengali by utilizing a dataset of an unprecedentedly large size. We have compiled a dataset that has been meticulously vetted, and it consists of about 3.8 million different instances of hate speech. This dataset is the largest one available for study on hate speech in this language.

We propose a word embedding model that outperforms existing methods for detecting hate speech. By combining an informal FastText embedding model with pretrained architectures such as BERT and BiLSTM, our improved word embeddings capture contextual information and semantic nuances, resulting in improved hate speech classification accuracy. The significance of our work resides in its potential impact on combating hate speech and fostering a safer online environment. With the largest dataset of hate speech collected in this language, our research provides an invaluable resource for training and evaluating models for detecting hate speech. The size and diversity of the data set permit more accurate modeling of hate speech patterns and linguistic variations, resulting in enhanced generalization to real-world scenarios.

Our research followed a comprehensive methodology for detecting hate speech. We carefully selected social media sources known for hate speech occurrences and employed rigorous preprocessing techniques to clean and anonymize the data while retaining contextual information. Human annotators participated in the annotation process to ensure high-quality labels. To enhance our word embeddings, we utilized an informal FastText model that incorporates subword information, enabling effective handling of out-of-vocabulary words and capturing morphological patterns. These embeddings were then integrated into the BERT model, renowned for its contextual understanding capabilities. By combining the strengths of both models, our approach achieved superior performance in hate speech detection. They were also integrated with LSTM model. Extensive experiments were conducted, including appropriate train-test splits and tailored evaluation metrics, demonstrating the effectiveness of our methodology through comparative analyses against baselines and existing approaches.

This research contributes to the advancement of the detection of hate speech on social media platforms. The large-scale hate speech dataset and the enhanced word embedding model provide researchers and practitioners with valuable resources. By precisely identifying hate speech, we hope to facilitate the development of proactive measures to mitigate its harmful effects and foster a more inclusive online environment.

1.2 Problem Statement

Hate speech on social media has become a pressing issue, posing significant difficulties for individuals, communities, and online platforms [8]. To ensure a safer and more inclusive digital environment, the proliferation of harmful and objectionable content necessitates effective detection and mitigation strategies. Existing methods for detecting hate speech are, however, limited in terms of dataset size and effectiveness of the dataset.

The most difficult aspect of Bengali hate speech detection is the dearth of sufficiently large and diverse datasets that capture the vast array of hate speech instances prevalent in social media conversations [9]. Insufficient dataset sizes impede the capacity of models to generalize and accurately identify hate speech patterns, resulting in suboptimal performance in real-world scenarios. In addition, existing datasets frequently fail to capture the intricate semantic relationships and contextual information required for understanding the nuances of hate speech, resulting in a reduction in detection accuracy [10].

1.3 Motivation

Misogyny is a pervasive and harmful problem that affects many aspects of society, including social media [11]. It can manifest as hate speech and extremism, such as promoting violence or discrimination against women or girls [12]. It can also involve the disparagement or objectification of women and girls, such as through the use of degrading or sexualized language or imagery. This can contribute to harmful gender stereotypes and contribute to a culture of sexism and misogyny and create a hostile and unsafe online environment for women and girls. However, problems such as online abuse, and particularly online abuse against women, are

on the rise in Bangladesh [13]. It is important to identify and address misogyny in social media in order to promote a more equitable and safe online environment.

It is a challenging task, particularly for low-resource languages that do not have as many resources available for natural language processing and machine learning [14]. Moreover, there is a great deal of variety in hate speech [15], detecting misogyny often requires understanding the context and cultural background of the language being used, as well as the intention behind the words being used.

The difficulty of autonomously detecting hate speech is high [16]. There is a need for more effective tools to detect and address misogyny in low-resource languages like Bangla.

To accurately train a model to detect misogyny, it is necessary to have a large and diverse corpus of annotated text containing labels indicating whether or not it contains misogyny. However, such datasets can be challenging to obtain, which reduces the efficacy of the developed models.

In addition, current models may not generalize well to diverse contexts. For instance, a model trained on a dataset of social media posts on a specific topic may not perform as well when applied to a more diverse dataset.

The context in which language is used is not taken into account by the majority of the models that are currently available. For instance, a model can label a remark as misogynistic only on the basis of the inclusion of particular words or phrases, without taking into account the larger context in which the comment was made. This might result in false positives or false negatives, either of which would be detrimental to the model's predictive power.

Last but not least, the training data that current models are built on could contain some form of bias. For instance, if the dataset that is used to train a model does not correctly reflect the whole range of language that is used in a certain context or language, then it is possible that the model will not be able to reliably detect instances of sexism in other settings or languages. This may occur because there are not enough big datasets that contain a diverse range of languages.

1.4 Research Challenges

Researching misogyny presents a number of obstacles, including:

- There is no consensus regarding a precise definition of misogyny, and it can manifest in a variety of ways. This can make identification and measurement problematic.
- Identifying the appropriate language resources and annotated datasets for the specific language being studied. This may be difficult if the language is not well-studied or lacks a strong digital presence.
- Developing a reliable and accurate method for detecting misogyny in the specific language being studied. This may require adapting existing methods or developing new approaches specifically tailored to the language and cultural context.
- Ensuring that the detection method is sensitive to the various forms that misogyny can take as well as the specific cultural context in which it is used.
- Dealing with the issue of false positives, where the detection method mistakenly identifies non-misogynistic statements as misogynistic. As there was limited data available to train the detection model.
- Addressing the issue of subjectivity in detecting misogyny, as different individuals may have different definitions and interpretations of what constitutes misogyny.We had to take into consideration the definition and measurement of misogyny in the specific language and cultural context being studied.
- Ensuring that the detection method is robust and can handle the variability and complexity of natural language, including slang, colloquialisms, and other forms of informal language.
- Dealing with the issue of cultural and linguistic differences, as misogyny may be expressed differently in different languages and cultures. This requires careful consideration of how to adapt the detection method for different cultural contexts.
- Ensuring that the detection method is ethical and does not perpetuate harmful biases or stereotypes. This requires careful consideration of the implications of the research and the potential impact it may have on marginalized groups.

1.5 Thesis Contributions

This research presents three significant contributions to the field of hate speech detection:

- Extensive and Diverse Dataset: The thesis presents a substantial and diverse dataset consisting of 3.8 million instances of hate speech in the language under consideration. This dataset is a valuable resource for training and evaluating models for detecting hate speech. Its magnitude and diversity enable more precise modeling of hate speech patterns and linguistic variations, resulting in enhanced generalization in real-world scenarios.
- Open-Sourcing Collected Code and Data: As part of our contribution, we will release our collected code and data as open-source, laying a firm foundation for future research endeavors in the Bangla language across multiple disciplines.
- Enhanced Pipeline: This study proposes a novel pipeline for detecting hate speech by combining an informal FastText embedding model with pretrained architectures, namely BERT and BiLSTM. This integration permits the capture of contextual data and semantic nuances, resulting in enhanced classification precision. The proposed pipeline improves the efficacy of hate speech detection systems, contributing to the development of techniques for identifying and combating hate speech and nurturing a secure online environment.

1.6 Thesis Outline

The dissertation contains several chapters that describe the research study. In Chapter 1, a concise overview of the study's purpose and objectives are presented. The second chapter is devoted to the literature review, which offers a comprehensive analysis of relevant existing research and developments. This chapter analyzes essential studies, theories, and approaches in order to establish the research's context and theoretical foundation.

The third chapter analyzes the proposed method in depth, detailing its structure and algorithmic details. The chapter contains a step-by-step description of the

proposed method, as well as figures that provide an in-depth look at its operation. This section aims to provide readers with a thorough comprehension of the methodology.

The implementation results and comparative analysis of the proposed method are presented in Chapter 4. This chapter provides a summary of the research findings that demonstrate the effectiveness and performance of the proposed method for detecting hate speech. To demonstrate the superiority of the proposed method, comparative analyses are conducted against baseline models and existing techniques.

In the concluding section of the thesis, the citations and acknowledgements used throughout the research project are presented. This section includes a comprehensive list of all sources and works cited, acknowledging the contributions of prior research to the current study.

Chapter 2

Literature Review

2.1 Hate Speech Detection

The intent of using aggressive language, abusive language, or hate speech is to impair the victim's identity, status, mental health, or reputation [17]. This behavior is deemed antisocial because it disrupts the social order, making it a matter of grave concern that requires immediate attention.

Hate speech is any form of communication, whether written, spoken, or symbolic, that is directed at an individual or group on the basis of their skin color, gender, race, sexual orientation, ethnicity, nationality, or religion [18]. Misogyny or sexism which are forms of discrimination and prejudice, can be categorized as a subset of hate speech [19]. Targeting individuals or groups based on their gender or sexual orientation with the intent to disparage, belittle, or express hostility towards them [20].

Hate speech detection entails the creation of automated methods and algorithms to identify and categorize instances of hate speech within textual data, particularly on social media platforms. It plays a crucial role in mitigating the negative effects of hate speech and nurturing online communities that value respect and inclusion. Due to the contextual nature of language and the evolvability of venomous expressions, accurately detecting hate speech is a difficult task. For hate speech detection, numerous methods, including rule-based systems, machine learning models, and deep learning architectures, have been investigated [21, 22].

2.2 Word Embedding Model

A technique used in natural language processing (NLP) to represent words or phrases as dense vector representations in a high-dimensional space is a word embedding model. These representations capture semantic relationships and contextual information, allowing algorithms to process textual data more effectively.

2.2.1 GloVe and Word2Vec

GloVe (Global Vectors for Word Representation) and Word2Vec are extensively used word embedding models for natural language processing (NLP) applications. Both models are designed to discover dense vector representations of words that capture semantic relationships and contextual information.

GloVe [23] is a method for global matrix factorization. In order to determine word embeddings, it makes use of the statistics on the co-occurrence of terms within a corpus. GloVe takes into account the percentages of times that words appear together, placing an emphasis on the statistics of the global corpus. It does this by striking a balance between the local and global contexts of words, which ultimately results in vector representations of an high quality.

Both the Continuous Bag-of-Words (CBOW) and the Skip-gram architectures are components of the Word2Vec [24] model. Each of these architectures is referred to in its own right as the respective "Skip-gram." The CBOW model is able to make predictions about a target word based on the words in the context, whereas the Skip-gram model is able to make predictions about the words in the context based on the target word. Both of these approaches find word embeddings by attempting to maximize the likelihood of the observed word sequences within a data corpus. Word2Vec is able to discover the syntactic and semantic associations between words by studying the patterns in which those words appear together. This allows Word2Vec to create a vector representation of the language.

The GloVe and Word2Vec word embedding models have both found broad use and have shown to be effective in a range of natural language processing tasks. This is due to the fact that both models embed words in a vector space. Language understanding, sentiment analysis, and information retrieval are all examples of these types of tasks. These models enable algorithms to leverage semantic links and

contextual information, which ultimately results in an improvement to the overall performance of natural language processing systems. Words are represented as dense vectors in order to attain this goal.

2.2.2 FastText

One prominent example of a word embedding model is FastText's [25], which can be seen here. After learning the vector representations for character n-grams, which are the subword units that are created when words are split up, FastText employs these n-grams to represent individual words. FastText is able to handle words that are not in its lexicon and correctly capture morphological patterns since it takes into consideration information about subwords. It has found significant usage in a range of tasks involving natural language processing (NLP), including analysis of emotion, recognition of named entities, and identification of hate speech, amongst other applications.

In order to increase the contextual comprehension of the word embeddings and to better capture sequential relationships, it is standard practice to integrate FastText with other pretrained architectures such as BERT [26] or Bidirectional Long Short-Term Memory (BiLSTM) [27]. The BERT model is a transformer-based system that is capable of learning contextual word representations by taking into consideration the context in which each word is located. This is accomplished by analyzing the relationship between the word and its surrounding context. The recurrent neural network's bidirectional long short-term memory (BiLSTM) design examines words in both the forward and backward directions, therefore capturing the sequential patterns that are present in the text. This allows the network to learn new information more quickly.

By taking use of the benefits offered by these pretrained architectures and combining them with FastText word embeddings, the researchers were able to improve the performance of their hate speech detection system as well as other natural language processing tasks. They were able to properly collect contextual information as well as semantic subtleties within textual material as a result of this.

2.3 Pretrained Architectures

2.3.1 BERT Model

The groundbreaking methodology for natural language processing (NLP) problems known as BERT (Bidirectional Encoder Representations from Transformers) has had a significant impact on the improvement of language understanding. In their seminal paper, [28], Devlin et al. introduced it.

The Transformer architecture, upon which BERT is founded and which utilizes self-attention processes to detect the links between words inside a phrase, is the foundation of BERT. In contrast to more traditional models, which typically evaluate text either from left to right or right to left, BERT makes use of a training technique that is bidirectional. It does so by hiding particular words inside a phrase and guessing the meaning of those words based on the surrounding context. This pretrains a deep transformer model. Because of this training in both directions, BERT is now better able to recognize the interdependencies between concepts and efficiently record information relevant to those terms.

During the pretraining phase, BERT will learn information from a huge dataset, such as a significant amount of text that may be found on the Internet. The BERT system is able to develop a complete grasp of language as a result of this unsupervised pre-training since it is able to capture a wide range of linguistic patterns. The pretrained model is then fine-tuned for specific downstream tasks, such as text categorization, named entity identification, and question answering, using data that is labeled according to the specific job at hand.

When it comes to adapting BERT to a variety of NLP tasks, the process of fine-tuning is very necessary. By training BERT on task-specific data, it can learn to encode task-specific information and produce representations that are better adapted to the task. The process of fine-tuning BERT entails revising the pre-trained parameters with task-specific gradients while maintaining the overall architecture.

BERT has outperformed previous state-of-the-art results on a broad variety of NLP benchmarks. It has substantially advanced the discipline by achieving impressive results in a variety of tasks, including sentiment analysis, text classification, machine translation, and question answering. BERT's capacity to capture contextual

information and comprehend complex language structures has made it a potent NLP tool.

BERT's influence goes beyond its performance. It has also driven NLP research. To meet various demands and languages, researchers have developed domain-specific and multilingual BERT models. BERT's NLP contributions are many. Its contextual information capture and complicated language structure understanding have expanded language understanding. BERT's pretrained nature makes it useful for transfer learning, helping researchers improve results even with insufficient labeled data. BERT has also inspired new designs and methods that use its powerful representations.

BERT has greatly influenced NLP. Bidirectional training and Transformer architecture have changed language understanding tasks. BERT is a key model in NLP, used for sentiment analysis, question answering, and more due to its capacity to gather contextual information, comprehend word dependencies, and adapt to different tasks.

2.3.2 LSTM Model

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that resolves the problem of vanishing gradients and permits the capture of long-term dependencies in sequential data. In 1997, Hochreiter and Schmidhuber introduced LSTM in their seminal paper [29]; since then, it has become a popular model in a variety of fields, including natural language processing, speech recognition, and time series analysis.

Traditional RNNs struggle to learn long-term dependencies due to the vanishing gradient problem; LSTM overcomes this limitation. The vanishing gradient problem occurs when gradients decrease exponentially as they are backpropagated through time, making it challenging for a model to learn from distant dependencies. This obstacle is circumvented by LSTM's introduction of specialized memory cells and gating mechanisms.

The central element of LSTM is its memory cell, which enables the network to store and retrieve information over extended time intervals. The memory cell has three primary gates: the input gate, the neglect gate, and the output gate. These

gates govern the flow of information into and out of the cell, giving the model the ability to select which data is saved, forgotten, or sent.

Considering the current input and the previous concealed state, the input gate decides how much new information should be added to the memory cell. Based on the current input and the previous hidden state, the forget gate determines which information should be discarded from the cell. The output gate determines how much of the content of the memory cell is exposed to the next stratum or output.

By selectively updating and preserving information in the memory cell, the LSTM design allows for the effective transmission of gradients across extended sequences. This helps to mitigate the vanishing gradient problem. Because of this, LSTM is able to take into consideration sequences of variable lengths as well as long-term dependencies.

Language modeling, speech recognition, emotion analysis, and machine translation are just few of the applications where LSTM has been shown to perform exceptionally well. The LSTM algorithm has seen considerable use in a wide variety of fields. Because of its ability to encapsulate sequential patterns and manage long-range relationships, it has become a popular choice for modeling sequential data. This is one reason for its popularity.

The impact of LSTM has continued to grow since it was first introduced. To improve the performance of LSTM and its capacity to handle difficult tasks, researchers have researched a number of LSTM variations and extensions, such as bi-directional LSTM (BiLSTM) and stacked LSTM.

In conclusion, LSTM is a strong RNN architecture that exceeds the constraints of standard RNNs by efficiently capturing long-term relationships in sequential data. This is accomplished through the use of recurrent neural networks. The ability of the model to store information and to keep that knowledge up to current thanks to the memory cell and the gating mechanisms makes it suited for a wide variety of sequential activities. Applications of LSTM may be found in natural language processing, speech recognition, and other fields where sequential data analysis is crucial. It has also become a core building component in deep learning.

2.3.3 BiLSTM Model

The Bidirectional Long Short-Term Memory (BiLSTM) architecture is an extension of the LSTM (Long Short-Term Memory) design that retains not only the past context of a sequence but also its future context. BiLSTM is an acronym for "Bidirectional Long Short-Term Memory." BiLSTM, which was first presented by Schuster and Paliwal in their article published in 1997 [30] has become a well-known model for applications that call for an in-depth understanding of sequential data.

BiLSTM is capable of processing the input sequence in both the forward and backward directions due to its utilization of two separate hidden layers. In this way, the model is able to represent the interdependencies that exist between the states of the past and the states of the future, which enables a more nuanced understanding of the sequence.

In the forward pass, the input sequence is processed all the way from the beginning to the end, whereas in the backward pass, the sequence is processed in the opposite direction. The LSTM architecture is implemented in each hidden layer of the BiLSTM, complete with memory cells and gating mechanisms. This gives the model the ability to account for long-term relationships and provide a solution to the vanishing gradient problem.

At each time step, the outputs of the forward and backward passages are concatenated to provide a merged representation that contains information from both directions. This representation contains all of the information that was obtained. After that, this consolidated representation is put to use for supplementary analytic or forecasting responsibilities.

BiLSTM has been found to be effective in a wide range of sequence-related tasks, such as identifying parts of speech, recognizing named entities, doing sentiment analysis, and translating machine-translated text. By taking into account the context of both the current state and the state in the past, BiLSTM is able to increase the model's ability to recognize subtle patterns and interdependencies in the data.

In later research, the BiLSTM architecture has been extensively adopted and changed to meet specific issues and better performance. This was done in order to

improve the overall quality of the study. In addition, it has been integrated with other methods, such as attention processes, in order to increase its capability of capturing intricate correlations contained within sequential data.

The LSTM architecture is expanded by the use of the BiLSTM algorithm, which processes input sequences in both the forward and backward directions. By incorporating knowledge from both the past and the future, it permits a more in-depth interpretation of data that is presented in sequential order. The BiLSTM algorithm has shown to be an effective tool for a wide range of sequence-related operations, since it enables improved speed and the capturing of subtle connections.

2.4 Comaparative Analysis

Various studies and research papers have investigated hate speech detection using both conventional techniques and cutting-edge deep learning models. Deep learning models, such as convolutional neural networks (CNNs) [31], have demonstrated promise in classifying hate speech by exploiting the hierarchical structure of textual data. However, capturing long-term dependencies and nuanced semantic relationships may present difficulties.

Recurrent neural networks (RNNs) and their variants, such as LSTM and BiLSTM, have gained prominence in hate speech detection in recent years. For bangla language hate speech detection LSTM-based model was used [32], and also BERT was employed [33] and both models produced positive results. These models capture sequential dependencies and contextual information, thereby enhancing the performance of hate speech detection. Managing chaotic and unstructured social media data and achieving a balance between precision and recall remain obstacles.

For bangla hate speech detection, word embedding models such as GloVe, Word2Vec, and FastText have also been investigated [34]. GloVe captures semantic relationships based on global word co-occurrence statistics, but it may grapple with subtleties. Word2Vec effectively captures syntactic and semantic relationships, but may struggle with uncommon words and contextual nuances. Incorporating subword information, FastText effectively manages out-of-vocabulary words and captures morphological patterns, but it requires more computational resources.

The capability of these word embedding models to provide meaningful representations and capture semantic relationships enables hate speech detection models to utilize contextual information. However, limitations exist in capturing uncommon words (GloVe), contextual nuances (Word2Vec), and computational demands (FastText).

Future research can investigate domain-specific or contextualized embeddings, such as BERT or ELMo, to capture nuanced and context-dependent hate speech patterns in order to improve hate speech detection. Moreover, ensemble approaches that combine multiple word embedding models may enhance overall performance and robustness.

Chapter 3

Proposed Approach

In our approach, we compiled a large dataset of approximately 6 million social media instances. After conducting data cleansing, we obtained a refined dataset containing 3.8 million informal statements from social media platforms. The collection procedure is described in detail, including data source selection, preprocessing, and annotation.

To improve the accuracy of word embeddings, an informal FastText embedding model was utilized. This model utilized character n-grams and contextual data to identify semantic and syntactic nuances in the dataset. The enhanced word embeddings produced by the FastText model were then incorporated into the BiLSTM model, allowing us to improve the performance of hate speech detection.

3.1 Data Collection

In this section, we describe the methodology used to compile our data set. Among the numerous social media platforms, including Facebook, YouTube, TikTok, and Twitter, we chose Facebook and YouTube because they are the most popular and widely used platforms in Bangladesh. These platforms are major contributors to the dissemination of abusive and hateful speech and comments.

In order to choose appropriate data samples, we decided on a set of specified criteria to use. We focused our attention on data samples that were divided according to gender, with a particular emphasis on remarks made about women that displayed characteristics of discrimination, abuse, harassment, and victimization.

The application known as Facepager, which makes use of the Facebook Graph API, was selected for its ability to assist the effective gathering of data within a limited amount of time. Because it was a more efficient alternative to web scrapers based on selenium, this tool was an excellent choice for gathering the substantial amount of data that was required for our investigation.

Facepager enabled us to gather a significant number of data samples in an effective manner, which ultimately led to the creation of our dataset. We were able to accomplish the study objectives in a timely manner because we chose to collect data using Facebook, YouTube, and Facepager. These platforms made the efficient collection of data much easier.

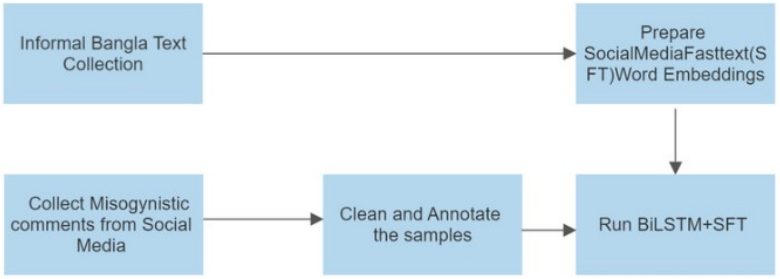

FIGURE 3.1: Proposed Architecture

3.1.1 Annotation Criteria

Our data collection included data samples that provided clear evidence of misogyny, in addition to a large number of data samples that presented difficulties in determining their appropriate classifications. Those samples that were found to contain misogyny were denoted with a '1', while those that were found to be free of it were given a '0' designation.

3.2 Data Preprocessing

The dataset obtained through web scraping has been subjected to a number of standard preprocessing steps. The primary objective of the preprocessing phase was to remove redundant and superfluous words, emoticons, and characters from the dataset. We eliminated redundant words that were not in the Bangla language,

lacked significance, or consisted of digits 0 through 9. In addition, symbols and other punctuation marks were removed from the text. In addition, we removed any parentheses and single characters that lacked significance. Emoticons, and any other text-based graphics were also removed from the dataset.

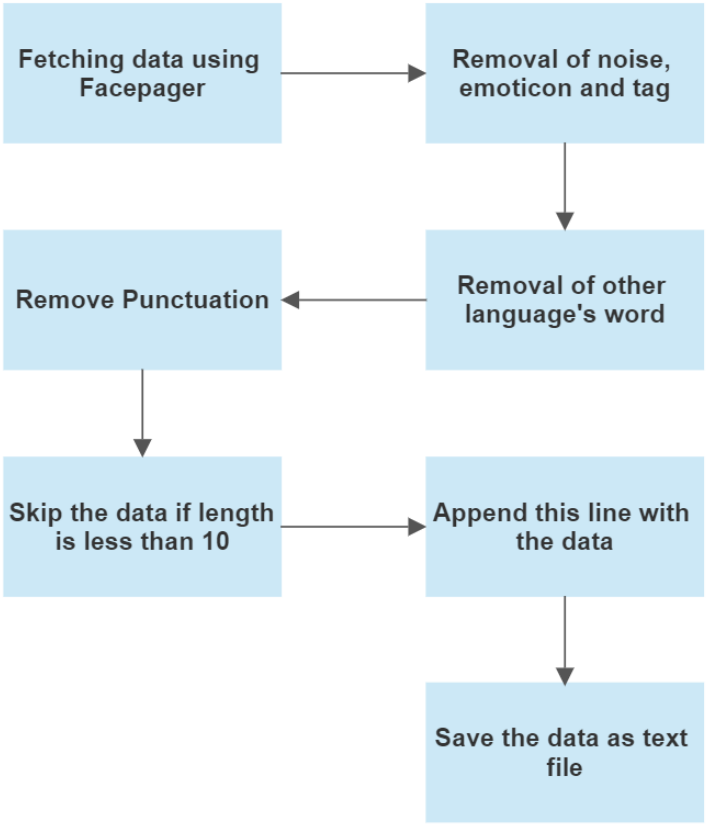

FIGURE 3.2: Data Pre-processing

- Dataset Size Before Preprocessing: 6.8 million
- Dataset Size After Preprocessing: 3.8 million

3.3 Prepare FastText

FastText is a technique devised by Facebook's AI Research (FAIR) group for the efficient learning of word representations and text classification. It is superior to *Word2Vec* and *GloVe* in that it can extract subword (Substring) information.

Already available FastText word embedding for Bangla:

- BengFastText (Trained on Bangla Wikipedia Articles) [35]

- Multilingual Fasttext (Developed by Facebook) [36]

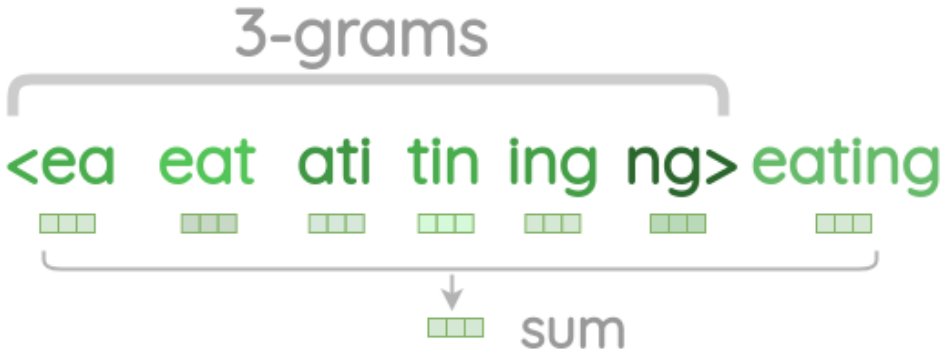

FIGURE 3.3: FastText Work Structure

After that, we moved on to the next step, which was the training of the informal FastText word embedding model by making use of the data that we had meticulously gathered. During the training process, the model was subjected to a massive amount of textual data consisting of examples of hate speech taken from a variety of social media sites. This was done on an iterative basis. During the training process, the FastText model picked up the ability to represent words as continuous vector representations by taking into account the character n-grams that make up individual words as well as information about their context. The model was able to effectively capture the semantic and syntactic subtleties of the language after it leveraged the subword information and contextual signals that were included within the dataset. During the training process, significant consideration was given to the settings of the hyperparameters. This helped to ensure that an optimal configuration was used, which in turn maximized the embedding's quality and performance.

3.3.1 Training SocialmediaFastText

- Vector size =300
- Character grams (3 to 6)
- Minimum word count 5
- Continuous Bag of Words (CBoW)
- Time to Train: 16 hours

3.4 Feeding the word embedding and labeled comments in BiLSTM:

After creating the word embedding, we fed it, along with a set of annotated remarks, into the Bidirectional Long Short-Term Memory (BiLSTM) model that we had previously developed. In this stage of the process, the goal was to make the most of the capabilities of the BiLSTM architecture by recognizing patterns associated with hate speech and capturing contextual dependencies that are present within the comments. The BiLSTM model included forward and backward LSTM layers, both of which were trained in a sequential fashion using the input data. Because of this, the model was able to encode and understand the sequential information that was contained in the comments by taking into account the context of both what came before and what came after.

The BiLSTM model developed a more comprehensive grasp of the comments and the context in which they were made as a result of the bidirectional processing of the remarks; as a result, it was able to identify and categorize instances of hate speech with greater precision. Throughout the course of the training process, we performed meticulous tweaking and made iterative changes to the model's parameters in order to achieve the highest possible level of performance and generalization capabilities.

As a consequence of this, the BiLSTM model successfully learned to differentiate between hateful and non-hateful remarks, which made it easier to identify and reduce instances of hate speech on various social media platforms. It was able to effectively recognize and categorize instances of hate speech thanks in large part to its capacity for capturing contextual dependencies and comprehending sequential information.

3.5 Performance Evaluation of BiLSTM

In order to identify instances of misogyny among the data that we gathered, we made use of the pre-trained BiLSTM (Bidirectional Long Short-Term Memory) model. We carried out a comprehensive classification approach in order to accurately identify and label remarks as either misogynistic or non-misogynistic. This

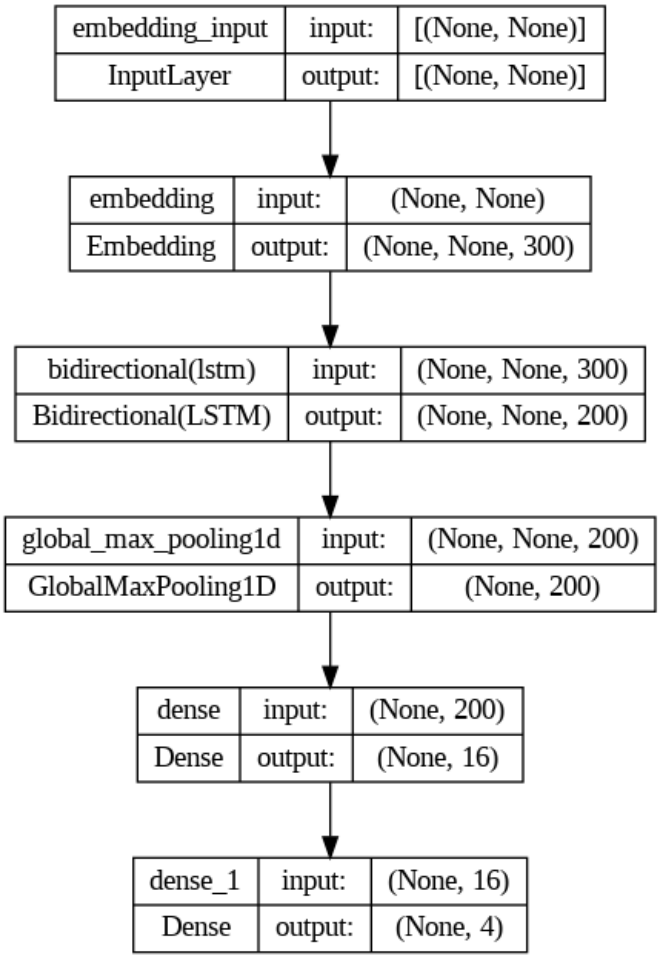

FIGURE 3.4: Neural Network Model for Detecting Hate Speech

was accomplished by making use of the knowledge and contextual information gathered by the BiLSTM model.

A discriminative analysis of the input remarks was carried out very successfully by the BiLSTM model, which is well-known for its capacity to capture long-term dependencies as well as contextual nuances. In order to comprehend the bigger picture of what was being said, it took into account the context of what came before as well as what came after. The model, which made use of the bidirectional processing capabilities of the BiLSTM architecture, analyzed the comments and extracted complicated patterns and features from them, which made it possible to make a reliable classification decision.

Throughout this procedure, the model's predictions were tested against the ground truth labels. This comparison gave us the ability to examine the accuracy of the classification and evaluate its performance utilizing a variety of measures. This

thorough test helps confirm the reliability of the BiLSTM model and its usefulness in reliably detecting instances of misogyny within the dataset.

3.6 Prepare BERT

BERT is the state of the art model in many natural language processing tasks even in low resource domains

This can be seen in recently published papers:

- SentNoB: A Dataset for Analysing Sentiment on Noisy Bangla Texts- EMNLP 2021 [37]
- BANGLABOOK: A Large-scale Bangla Dataset for Sentiment Analysis from Book Reviews -ACL 2023 [38]

We then moved on to the next step, which was to train a Bert language model from scratch using our meticulously picked dataset. The purpose of this training procedure was to provide the model with a comprehensive understanding of the linguistic patterns and contextual information that were contained within the dataset. This would enable the model to capture intricate semantic relationships and generate high-quality representations for a variety of natural language processing tasks. During the training phase, the BERT model went through a number of iterations, each of which involved the presentation of a sizeable amount of textual data drawn from the dataset that we had gathered. The model was educated to understand the contextual dependencies and lexical semantics that were present in the dataset by being taught to predict masked words inside phrases. We wanted to improve the model's capability to successfully encode and interpret the textual information by performing intensive optimization and fine-tuning of model parameters. This would make it possible to have a comprehensive grasp of language and future applications, such as the identification and categorization of hate speech. We sought to make use of the revolutionary capability of the BERT language model by training it from scratch using our dataset in the hopes of uncovering subtle linguistic patterns and driving breakthroughs in the identification and mitigation of hate speech on social media platforms. To do this, we decided to train the model from scratch using our dataset.

3.6.1 BERT Model Configuration

- Vocabulary Size: 50,100
- Parameters:82 million
- Hidden Size: 768
- Number of Hidden Layers: 6
- Number of Attention Heads: 12
- Maximum Position Embeddings: 512

3.7 Prepare smBERT

A BERT model that has been pre-trained on our gathered Informal Bangla Text is known as a smBERT (which stands for SocialmediaBERT).

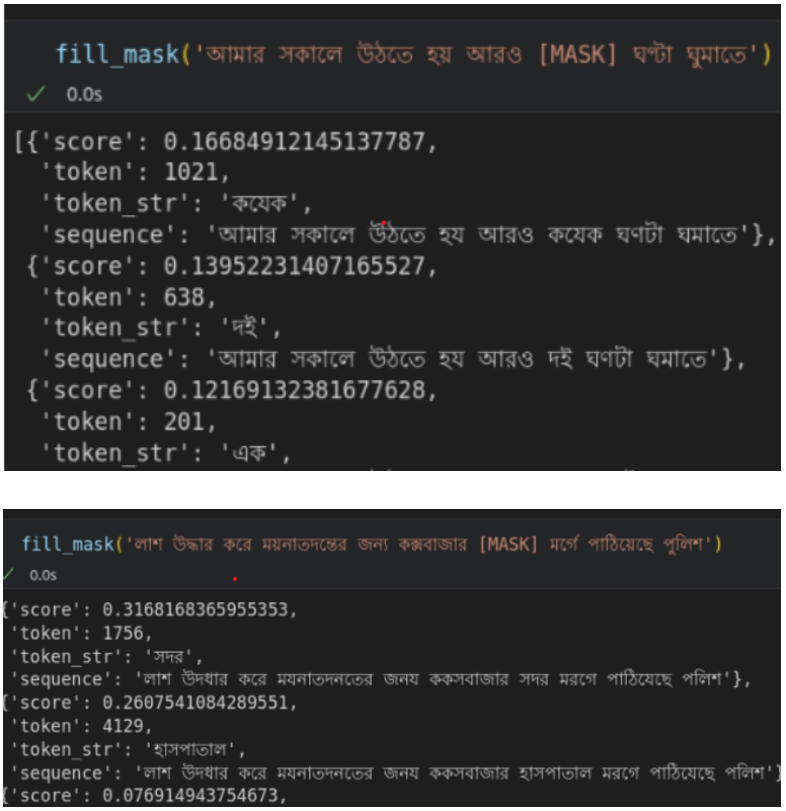

FIGURE 3.5: After training smBERT

3.8 Performance Evaluation and Model Validation

In order to determine whether or not the trained BERT model was effective, an exhaustive testing step was carried out making use of a dataset that had been meticulously annotated. Validating the efficiency and applicability of our method required us to conduct this study with the objective of determining whether or not the model has the capacity to accurately recognize and categorize instances of hate speech.

The labeled dataset served as a reference to the ground truth, which allowed us to evaluate the accuracy of the model's predictions in relation to the actual labels. Because of this comparison, we were able to examine a number of performance parameters, including as accuracy, recall, and F1-score. We gained significant insights into the model's discriminative capabilities and highlighted possible areas for development by comparing the model's predictions to the known labels.

Validating the efficacy of our trained BERT model for identifying and categorizing hate speech required an extensive testing strategy, which was critical. Our trust in the model's ability to successfully counter hate speech on social media platforms was bolstered as a result of this finding. We contribute to the creation of mechanisms that combat the widespread issue of hate speech and promote a secure environment online by assuring its robustness and reliability. These mechanisms aim to provide a safe environment for people to engage in online activities.

Chapter 4

Experimental Design

4.1 Framework

This study's framework consists of BiLSTM (Bidirectional Long Short-Term Memory), custom word embeddings and BERT. BiLSTM is a type of recurrent neural network that captures sequential dependencies by processing input in both forward and reverse orientations. BERT is a state-of-the-art language model that uses a bidirectional transformer architecture to generate contextualized word embeddings, allowing it to capture rich contextual information and produce high-performance results in a variety of natural language processing tasks. In addition, custom word embeddings are constructed to capture domain-specific data and enhance representation learning.To assess the efficacy and efficiency of the proposed framework, evaluation metrics and benchmark datasets are utilized. The results and analysis contribute to the advancement of natural language processing techniques and shed light on the effectiveness of combining pre-trained models with domain-specific embeddings.

4.2 Experimental Environment

- CPU: Intel 12th Gen Core i9-12900K
- RAM: 128 GB
- GPU: RTX 3090 24 GB

The FastText model was trained using a 300-vector length and character n-grams ranging from 3 to 6. The model employs the Continuous Bag of Words (CBoW) algorithm and specifies a minimum word count of 2 for the training procedure.

Hate speech detection used word embedding and labeled comments. The BiLSTM architecture was used for pattern recognition and comment contextual relationships. The BiLSTM model used input data to successively train forward and backward LSTM layers to encode and grasp sequential information by considering the prior and following context. The BiLSTM model understood remarks and context by careful parameter modifications during training.

There are two experiments that make up our research. In the first experiment, BiLSTM and FastText models are used to classify examples of hate speech as targets. The BERT model will be used to categorize different types of hate speech in the second trial.

4.3 Experiment I: Hate-Speech Target Classification using BiLSTM

The Bidirectional Long Short-Term Memory (BiLSTM) model was utilized in order to identify instances of hate speech and misogyny. In order to discern patterns of hate speech and capture contextual dependencies, the BiLSTM model was given word embeddings and annotated remarks as input. The BiLSTM model was able to display a full grasp of comments after undergoing iterative changes and parameter tweaking. This allowed for the correct detection of instances of hate speech.

The section that follows delves into the discussion of this model's outcomes.

4.3.1 BiLSTM Setup

- Dataset : BD-SHS
- Data point : 40,000
- Model : Bi-LSTM
- Accuracy : 80.3
- Recall: 75.7

- F-1 Score: 77.9
- System: Google Colab

4.4 Experiment II: Hate-Speech Target Classification using BERT

BERT was trained through a series of iterations, each of which included the extraction of a sizeable quantity of textual material from the dataset. The model became capable of capturing contextual dependencies as well as lexical semantics once it learnt how to anticipate masked words included inside phrases. In order to improve the model's capacity to encode and comprehend textual information, extensive optimization of the model and fine-tuning of its parameters were both carried out.

The section that follows delves into the discussion of this model's outcomes.

4.4.1 BERT Setup

- Vocabulary Size: 50,100
- Parameters:82 million
- Hidden Size: 768
- Number of Hidden Layers: 6
- Number of Attention Heads: 12
- Maximum Position Embeddings: 512

Chapter 5

Results and Discussions

5.1 Experimental Result

5.1.1 Hate-Speech Target Classification Using BiLSTM

Model + Feature		IND			Male			Female			Group			Weighted Average	
	P	R.	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM + BFT	70.2	54.9	61.6	71.6	58.4	58.4	71.5	68.6	70.1	71.1	46.8	56.4	71.0	58.1	63.7
SVM + U	71.7	63.8	67.6	78.8	69.8	69.8	**88.5**	69.9	78.1	67.7	38.5	49.0	77.2	63.9	69.6
SVM + C	77.5	67.4	72.1	80.0	71.9	71.9	87.6	76.1	81.4	70.7	38.8	50.1	79.7	67.3	72.6
SVM + U + C	86.5	**75.9**	**80.8**	66.7	48.2	55.9	74.4	68.7	71.4	**78.2**	**73.9**	**73.9**	77.3	69.8	73.0
BiLSTM + MFT	74.9	69.4	71.9	78.8	75.4	75.4	87.3	74.6	80.5	72.8	47.5	57.5	78.6	69.7	73.7
BiLSTM + RE	72.3	75.3	73.8	**80.1**	76.9	76.9	85.7	78.7	82.1	71.1	45.2	55.2	77.3	72.9	74.8
BiLSTM + IFT	77.2	74.9	75.9	79.3	**78.2**	78.2	88.2	81.0	84.5	71.5	60.5	65.6	79.6	75.5	77.5
BiLSTM + SFT	**81.8**	71.0	76.0	79.6	79.0	**79.0**	85.4	**85.1**	**85.3**	68.5	63.9	66.1	**80.3**	**75.7**	**77.9**

FIGURE 5.1: Comparison With Models

BiLSTM+BFT, SVM+U, SVM+C, SVM+U+C, BiLSTM+MFT, BiLSTM+RE, and our own suggested model, BiLSTM+IFT are all included in the comparison that is presented in the table . Evaluation is done based on the performance measures ind_p (individual precision), male_f (male F1-score), female_r (female recall), and female_f (female F1-score), as well as w_p (weighted precision), w_r (weighted recall), and w_f (weighted F1-score). It is clear that our proposed model, BiLSTM+SFT, performs better than the other models in terms of ind_p

(81.8), male_f (79.0), female_r (85.1), female_f (85.3), w_p (80.3), w_r (75.7), and w_f (77.9) respectively. These findings illustrate the superiority of our proposed model in effectively predicting and categorizing instances, particularly in terms of gender-specific characteristics and overall weighted performance. Specifically, these findings emphasize the model's ability to accurately predict and classify cases of gender-specific parameters.

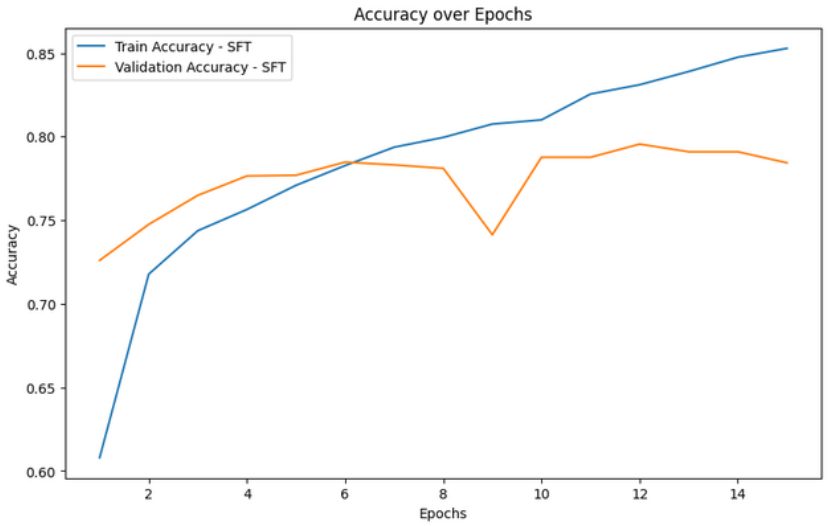

FIGURE 5.2: Accuracy over epoch

Figure 4.2 provides insightful information regarding the performance of our proposed method. During the initial epochs, it has been observed that the train accuracy performs marginally worse than the validation accuracy. At epoch 6, however, a significant turning point occurs as the train accuracy surpasses the validation accuracy. This occurrence indicates that our proposed method progressively improves its ability to generalize from training data, enabling it to make more accurate predictions on unseen instances. The disparity in performance between the train and validation accuracies highlights the significance of monitoring both metrics and the potential of our methodology to achieve superior results on diverse data instances.

The loss metric consistently decreases as the number of epochs increases, as depicted by Graph 4.3 of the loss over epoch data. This gradual decline reflects the iterative optimization process occurring during training. As the model iteratively adjusts its weights and learns from the training data, it gradually reduces the disparity between its predicted and actual values. As a result, the loss gradually decreases until it reaches a point of convergence where further epochs have little

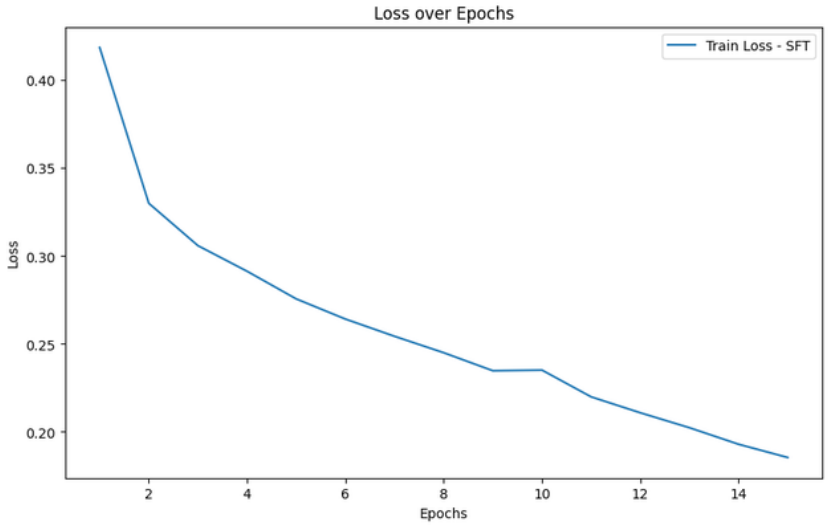

FIGURE 5.3: Loss over epoch

effect on reducing the loss. This trend demonstrates the model's ability to recognize and utilize data patterns, progressively refining its predictions to more closely match the desired output.

5.1.2 Test Result with BERT on Dataset

Model	Accuracy (%)
DistilBERT	84.5
mDistilBERT	86.2
BERT	84.3
mBERT	91.2
smBERT	89.1

5.2 Discussions on Experiment I

5.2.1 Improving Contextual Understanding via FastText Training with Selected Data

Training FastText with the collected data serves the purpose of generating misogyny detection-specific word embeddings. FastText is a well-known algorithm for word embedding that takes into account the internal structure of words, such as character n-grams, to capture subword information and more accurately represent word meaning.

Benefits of training FastText with our collected data include:

- Domain-specific embeddings: By training FastText on the collected data, we are able to identify the specific language patterns, terminology, and contextual information associated with misogyny in social media. This contributes to the creation of word embeddings that are more pertinent and aligned with the specific domain, resulting in enhanced performance.
- Out-of-vocabulary words: FastText manages out-of-vocabulary (OOV) words more efficiently than conventional word embedding techniques such as Word2Vec. Due to abbreviations, slang, and typographical errors, OOV words are frequent in social media texts. Training FastText with the data enables it to manage OOV words by utilizing subword information and providing meaningful representations.
- Contextual understanding: FastText takes the context of words into account by displaying them as a combination of their character n-grams. This is especially useful for capturing the semantics of words in a context where word meanings may vary depending on adjacent words.
- Dataset-specific biases: Pretrained word embeddings may miss dataset-specific nuances and biases. Training FastText with our data captures the particular biases, linguistic patterns, and misogyny in social media texts, improving detection accuracy.

Training FastText with collected data generates word embeddings for misogyny detection area, improving social media language representation and interpretation.

5.2.2 Rationale Behind Not Using Skip-Gram

- Handling out-of-vocabulary words: FastText effectively handles out-of-vocabulary (OOV) words in social media texts by representing them as character n-grams, addressing challenges posed by slang and misspellings, unlike skip-gram models reliant on predefined word embeddings.
- Subword information: FastText captures subword information by separating words into smaller components, which enables it to capture morphological and semantic details, whereas skip-gram models frequently ignore such fine-grained linguistic features.

- Performance with rare words: FastText outperforms skip-gram models with uncommon words by inferring meaning from character n-grams, whereas skip-gram models struggle due to limited occurrences in training data.
- Domain-specific language patterns: Training FastText on collected data enables capturing domain-specific language patterns and misogyny-related terminology in social media, which could lead to enhanced detection performance, whereas skip-gram models may lack the nuances and biases specific to the dataset.

5.2.3 Filtering Word Sequences with Length Greater Than 10 in Preprocessing: Benefits and Significance

- Capturing meaningful context: Longer word sequences provide a richer context, which aids in comprehending the meaning and intent of the text, thereby enhancing the detection of misogyny.
- Filtering noise and irrelevance: Excluding shorter sequences reduces noise and irrelevant content, enhancing the detection model's accuracy.
- Handling complex language patterns: Longer sequences frequently contain intricate language patterns, enabling the detection of more subtle forms of misogyny.
- Computational efficiency: Concentrating on lengthier sequences optimises computational efficiency by decreasing the processing and analysis of shorter sequences that are less informative.

5.3 Discussions on Experiment II

5.3.1 Rationale for Employing the BERT Model

The BERT captures contextual information and understands text semantics, which motivates its use. Its text classification performance is state-of-the-art.

- Contextual understanding: Given words' different implications, BERT's contextual awareness helps it detect misogyny. Contextual information helps identify and understand misogyny.

- Deep bidirectional architecture: The profound bidirectional architecture of BERT processes context in both directions, capturing word dependencies and linguistic patterns with precision. Understanding the nuanced nature of misogynistic language is facilitated by its proficiency in detecting long-range dependencies.
- Multilingual support: BERT models in several languages allow misogyny detection in languages other than Bangla. BERT is versatile enough for misogyny detection investigations in multiple languages.
- High performance:BERT excels at text categorization and other NLP problems. Its high classification accuracy makes it a good choice for misogyny detection study.

BERT's ability to capture contextual information, utilise pretrained representations, manage complex language patterns, support multiple languages, and deliver high performance makes it an appropriate model for our research. These characteristics make BERT an effective instrument for identifying and categorising misogynistic language in social media texts.

5.3.2 Advantages of BERT with FastText

- Contextual understanding: BERT captures context by contemplating surrounding words, whereas FastText provides subword embeddings; combining the two improves word meaning comprehension in misogynistic contexts.
- Fine-grained representations: The combination of FastText's subword embeddings and BERT's contextualised representations results in word representations that are nuanced and expressive.
- Handling out-of-vocabulary words: FastText effectively manages OOV words using subword embeddings, which benefits the detection of misogyny in social media, whereas BERT struggles with OOV words because of its fixed vocabulary. Combining both models

By combining FastText and BERT, we can leverage the strengths of both models, such as contextual understanding, fine-grained representations, OOV word management, transfer learning, and multilingual support. This fusion can enhance the performance of detecting misogyny by effectively capturing context and subword-level information.

Chapter 6

Conclusion and Future Work

Our proposed model BiLSTM+SFT was compared to other models. The evaluation of performance included precision, F1-score, recall, and weighted measures. In terms of gender-specific characteristics and overall performance, the proposed model outperformed the other models, demonstrating its superiority in predicting and classifying instances.

In our work so far, we have discovered that preprocessing informal text with an embedding technique can improve the performance of our model. Additionally, we have identified a shortage of datasets containing misogynistic language. We believe that if we had a larger and more diverse dataset, our model would perform even better.

We are hoping to implement the model on a larger dataset soon. After that we should be able to highlight the improvements our model has done in this particular task. Also as the model is complex and takes a lot of time for training and testing, we are hoping to make our model pre-trainable. If we can pre-train our model on a good dataset, then the whole testing time will be reduced by a lot which will help us research on accuracy improvement easily.

In our study, we attempted to develop a Bangla BERT model, with room for improvement. We have faith in its ability to produce superior outcomes compared to other BERT variants. In the future, we will concentrate on improving the efficacy of smBERT for informal Bangla text and establishing it as the industry standard. In addition, we intend to analyze emoticons to determine the presence of hate speech in statements, thereby providing valuable insights for the detection of hate speech.

Bibliography

[1] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth." *Psychological bulletin*, vol. 140, no. 4, p. 1073, 2014.

[2] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[3] 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)*. IEEE, 2021, pp. 1–10.

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.

[7] 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)*. IEEE, 2019, pp. 1–6.

[8] M. A. Khan, M. R. Karim, and Y. Kim, "A two-stage big data analytics framework with real world applications using spark machine learning and long short-term memory network," *Symmetry*, vol. 10, no. 10, p. 485, 2018.

[9] B. R. Chakravarthi, M. Arcan, and J. P. McCrae, "Improving wordnets for under-resourced languages using machine translation," in *Proceedings of the 9th Global Wordnet Conference*, 2018, pp. 77–86.

[10] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.

[11] A. Ben-David and A. M. Fernández, "Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain," *International Journal of Communication*, vol. 10, p. 27, 2016.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[13] N. Sambasivan, A. Batool, N. Ahmed, T. Matthews, K. Thomas, L. S. Gaytán-Lugo, D. Nemer, E. Bursztein, E. Churchill, and S. Consolvo, "" they don't leave us alone anywhere we go" gender and digital abuse in south asia," in *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[14] B. R. Chakravarthi, M. Arcan, and J. P. McCrae, "Improving wordnets for under-resourced languages using machine translation," in *Proceedings of the 9th Global Wordnet Conference*, 2018, pp. 77–86.

[15] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proceedings of the 28th ACM conference on hypertext and social media*, 2017, pp. 85–94.

[16] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the fifth international workshop on natural language processing for social media*, 2017, pp. 1–10.

[17] T. Beran and Q. Li, "Cyber-harassment: A study of a new method for an old behavior," *Journal of educational computing research*, vol. 32, no. 3, p. 265, 2005.

[18] J. T. Nockleby, "Hate speech," *Encyclopedia of the American constitution*, vol. 3, no. 2, pp. 1277–1279, 2000.

[19] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[20] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[21] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[22] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[25] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[31] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[32] 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)*. IEEE, 2021, pp. 1–10.

[33] N. S. Samghabadi, P. Patwa, S. Pykl, P. Mukherjee, A. Das, and T. Solorio, "Aggression and misogyny detection using bert: A multi-task approach," in *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 2020, pp. 126–131.

[34] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2020, pp. 390–399.

[35] M. Rezaul Karim, B. Raja Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," *arXiv e-prints*, pp. arXiv–2004, 2020.

[36] N. Garneau, M. Hartmann, A. Sandholm, S. Ruder, I. Vulić, and A. Søgaard, "Analogy training multilingual encoders," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 12 884–12 892.

[37] K. I. Islam, S. Kar, M. S. Islam, and M. R. Amin, "Sentnob: A dataset for analysing sentiment on noisy bangla texts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3265–3271.

[38] M. Kabir, O. B. Mahfuz, S. R. Raiyan, H. Mahmud, and M. K. Hasan, "Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews," *arXiv preprint arXiv:2305.06595*, 2023.