Towards Equitable Bangla Language Models: Detection of Stereotypical Biases in Bangla Word Embeddings

by
MD TANJIM MOSTAFA
16201084

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

MD Tanjim Mostafa

16201084

Approval

The thesis/project titled "Towards Equitable Bangla Language Models: Detection of Stereotypical Biases in Bangla Word Embeddings" submitted by

1. MD Tanjim Mostafa (16201084)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 28, 2024.

Examining Committee:

Supervisor:
(Member)

Md. Ahasanul Alam

Lecturer
Department of Computer Science and Engineering
Brac University

Co- Supervisor:
(Member)

Ms. Nehrin Siddique Payel

Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

MD Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Large language models are powerful tools that can be used in variety of tasks, including text generation, translation and question answering. However, Language models tend to pick up undesirable biases from the training data. Due to the rapid advancement in artificial intelligence and due to the widespread use of these models, we risk amplifying social stereotypes and biases through these systems. Word embedding is a framework that represents text data as vectors allowing them to capture the context of each word within a large text corpora. Word embeddings are essentially the building blocks of popular Bangla language models such as Bangla Glove, Bangla word2vec, banglaBERT. We observe that gender bias exists in a geometric direction in the word embedding. Using methods such as Principal component analysis, Word Embedding Association Test and Mask Language Modelling, we highlight the presence of stereotypically biased associations in the word embeddings.
Our findings have broader implications for AI research within the Bangla NLP community and for enhancing both the viability and usability of Bangla NLP systems for downstream tasks.

Keywords: Natural Language Processing; Bias; Gender Bias; Word Embeddings; Machine Learning; BanglaBERT

Acknowledgement

Firstly, I want to thank Allah for guiding and blessing me throughout my academic journey. I am grateful to my supervisor, Md. Ahasanul Alam, for his continuous support and encouragement throughout the whole research process. Special appreciation goes to the co-supervisor, Ms. Nehrin Siddique Payel. Without any of their effort and utmost dedication, completing this thesis would not be possible.
I would also like to thank all my professors, advisors and collaborators who helped me throughout my academic journey.
Lastly, thank you to my parents for making my education possible.

Table of Contents

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

LLM Large Language Model

ML Machine learning

MLM Masked Language Modeling

NLI Natural Language Inference

NLP Natural Language Processing

PCA Principal Component Analysis

RLF Reinforcement Learning Framework

$SEAT$ Sentence Embedding Association Test

$WEAT$ Word Embedding Association Test

Chapter 1

Introduction

1.1 Motivation

From credit scoring to disease prediction, in diverse fields of our day-to-day lives, artificial intelligence (AI) based systems are increasingly being integrated [16], [18]. So, it has become imperative for us to recognize how these systems interact with our social environment and the implications related to them. Natural Language Processing (NLP) is one of the major branches of AI that helps to interpret human language in a way that machines can understand [1]. Recently NLP has gained enormous popularity for its ability to interpret human language. Many NLP models show promising performance on different language-understanding tasks such as question answering where the user asks a question to the system and the system crawls through the dataset and generates or retrieves an answer. However, these systems often encode implicit biases to increase optimization. According to Hovy et al. (2021) [22], bias in NLP systems is defined as "the tendency of a system to produce results that are systematically prejudiced against certain groups of people". These biases can have severe negative effects by causing social or religious discrimination among different groups based on race, religion, or socio-economic condition. In this research work, we investigate biases in various natural language processing systems, pre-trained models based on static word embedding, contextual embedding etc. used to perform NLP tasks such as sentiment analysis, machine translation, Named entity recognition etc.

It is common practice nowadays to use off-the-shelf resources such as pre-trained models because they are capable of capturing complex latent dimensions of language. However, these models have been shown to give unfair treatment and give biased results since they are trained on real world data that is mostly unsupervised and contains biases that we would rather avoid in decision making tasks. Generally, machine learning models are faithful to the training data and can lead to perpetuating undesirable biases and amplifying societal inequalities and discrimination. Nevertheless, systems designed to perform credit worthiness of an individual [21], systems to predict crimes [31], systems to perform analogies [27] have been shown to exhibit biases that disproportionately affect racial minorities in the case of creditworthiness prediction and women in the case of analogy completion. To give more examples, let us consider how biased word embeddings can amplify societal inequalities in human-model interactions. Gender-biased word embeddings can cause search engines to return more results for male scientists than female scientists since there are more male scientists than female ones in the world. So, word vectors for

"scientist" and "male" are more similar and closer than the word vectors for "scientist" and "female". Consequently, this bias can negatively affect a student interested in becoming a scientist searching for information on the topic. Seeing that the results are mostly for male scientists could discourage the student from pursuing a career in science. Another instance where this kind of bias can be observed is in the task of coreference resolution, which involves matching all the words in a sentence that refer to the same entity. Research on gender-biased coreference resolution reveals that models systematically fail to associate stereotypical occupation words with mismatched gender words or characteristics. This kind of disparity results in representational harm by suggesting that some occupations are only suited for a specific gender.

It has been found that most of the chatbots give negative results in response to any prompt containing the word "Muslim" [20]. This happens because there is a strong vector similarity between the words "Muslim" and "Terrorists" as a result of the negative portrayal of Muslims in the media and on the internet. Word embeddings form the foundation of most natural language processing systems, so it is concerning how the use of these systems can amplify and even risk perpetuating the existing social stereotypes by encoding them into the language models.

Several Bangla language models like BanglaBERT [26], [32], BanglaT5 [28], Bengali-gpt2 [30] etc. have already been deployed in several NLP tasks without the supervision of experts in the field of "AI fairness and biases". It is thus imperative for us to evaluate the extent to which these models exhibit social biases so that we can implement strategies to mitigate them.

1.2 Aims and Objectives

Language models learn biases that exist in training data. We aim to quantify the extent of biases present in popular Bangla language models like Bangla_word2vec, BanglaBERT by csebuet [26] and Sagor Sarker [32], Bengali-gpt2, BanglaT5 [28], [30]. Word2vec and Glove are static word embeddings and BERT is contextual word embedding. Our goal is to learn the vector representation of words and detect bias through the association of words in the embedding space.

Chapter 2

Problem Statement

Large language models (LLMs) such as chatGPT, Google Gemini [29] have become increasingly popular in performing natural language processing tasks such as question answering, text generation, coreference resolution etc due to their ease of use, accessibility and ability to act as conversational agents. However, these models have been shown to exhibit various types of biases, including social biases related to demographics such as gender, age, and race, as well as sentiment biases and ethnic biases [8].
It was found that AI models for facial recognition often struggle to identify the faces of people from certain races. For example, there were instances of facial recognition models failing to identify the faces of black people and inappropriately identifying them as "gorillas" (Howard and Borenstein, 2018) [8]. Vig et al, have shown that if the prompt "The nurse said that _____" is given the model is more likely to predict the pronoun as 'she' rather than 'he' [19]. These biases can lead to forming undesired social stereotypes and amplify existing societal biases in our societies. Especially, in developing countries like Bangladesh where most people are not aware of societal stereotypes, biased Language models can influence user perceptions, reinforce discriminatory attitudes towards the marginalized population and perpetuate societal inequalities. Since these models have become convenient and easy to use, people have no interest in doing their own research on a specific topic anymore. Moreover, an average user with little to no idea about how these AI systems generate content is likely to be persuaded by whatever answer they would get from the system. Accordingly, a biased pretrained language model can have severe consequences on societies. In response to that, for a safe deployment of these models, we need to develop effective strategies to find, measure and mitigate biases in NLP systems.

Detecting these biases is the first step in making the language models more equitable for everyone. However, little to no work has been done in this regard. That is why we put our main focus on detecting biases in word embeddings of Bangla language models.
In light of the above setting, the next chapter discusses the recent works related to biases in LMs in detail.

Chapter 3

Related Work

In recent years, the NLP community has started exploring several biases on different aspects revealing numerous disparities and problematic biases present in algorithms and textual data. They include gender bias, political bias, racial bias, and several other societal biases in NLP systems, [17], [20]. The problem of biases in language models can be addressed in two steps. First, evaluating and measuring the biases in a dialogue system and second, the ways to mitigate them.

3.1 Analysis Methods

There are two main categories of methods for interpreting neural network models in NLP: (i)Structural Methods and (ii) Behavioral Methods.

Structural methods primarily aim to uncover the information contained within various components of the model. One approach within this category involves using probing classifiers, which utilize the model's representations as input to predict different properties, providing insights into the underlying information captured by the model [7], [13]. However, it's important to note that these structural methods are not directly linked to the model's behavior or predictions on the specific task it was trained for. The information uncovered by these methods may exist in the representations by mere coincidence or due to shared factors, without necessarily being used by the original model [6]. Furthermore, distinguishing the information learned by the probing classifier from that learned by the underlying model can be challenging [13].

The behavioral approach on the other hand evaluates the models' ability to capture different linguistic phenomena by assessing models' performance on carefully crafted prompts. For example, a statement "The doctor called the receptionist and told _____ to cancel the appointment." is given to the model and the response is evaluated to determine whether the model gives a stereotypical response or an anti-stereotypical response. In the mentioned scenario, if the model gives the pronoun "her" then it would be considered a stereotypical response and if the model gives "him" as the answer then it would be considered an anti-stereotypical response [25]. However, These methods can evaluate a model's prediction but not its' underlying structure [9].

The fairness of a Language Model can be identified in various ML tasks such as Word Embedding, classification, clustering etc.

3.1.1 Word Embeddings:

Early research on finding and mitigating gender bias in language models for example focused on identifying biases in static word embeddings such as GloVe and Word2Vec. These embeddings are trained on a large corpus of text data and these word embeddings capture the statistical relationship between words. For example, the word 'nurse' is more likely to be associated with the word 'woman' since there are more female nurses than male ones in our world and accordingly, we can see that representation in the training data [19]. Static Word Embeddings are often biased against women. Bolukbasi et al (2016) [2] found that popular word embeddings regularly linked men to working roles while women were mapped to typical gender roles. To mitigate gender bias, they proposed a 2-steps debiasing method where they removed the gender stereotypes from the word embeddings. However, in doing so they were still maintaining the desired associations of words such as between the words 'queen' and 'female'.
Another approach to removing bias from embeddings involves training a discriminator to distinguish between biased and unbiased embeddings called adversarial learning. Zhang et al. (2018) have found that the adversarial learning method was able to reduce gender bias in word embeddings without affecting the performance of tasks at hand, including sentiment analysis and natural language inference etc [10]. Bolukbasi et al. [2] and Caliskan et al. [3] used Word Embedding Association Tests(WEATs) to demonstrate that the word embeddings associate occupations with their stereotypical gender roles. WEATs use the cosine similarity of specific target words to measure the bias in static word embeddings.

However, these traditional methods have proven to be ineffective in quantifying bias in contextual embeddings. In particular, Kurita et al. (2019) [14] showed that WEAT tests are unable to detect statistically significant bias in BERT. However, probing the underlying model with a Gender Pronoun Resolution(GPR) revealed that non-static models also encode gender bias. Similar to static word embeddings, contextual word embedding models like BERT learn from the data they are trained on, including any social stereotypes that may be present in the data. Zhao et al.(2018) [11] introduced a new benchmark called WinoBias to measure bias in coreference pronoun resolution where they used GPR as a downstream task to detect gender bias. The dataset consists of two files that represent two different GPR tasks. The files contain Winograd schema pairs of sentences, involving a variety of occupations. each sentence in the dataset has two versions: one in which the pronoun is resolved stereotypically, and the other is resolved in a non-stereotypical way. The dataset can be used to train models to resolve pronoun references in a non-stereotypical way.

3.1.2 Data Augmentation

Biases in machine learning models can be caused when there is an uneven distribution of data samples from different classes. To address this Kusner et al. [4] proposed a technique called counterfactual fairness where attributes related to an individuals' demography is altered while keeping all the other factors unchanged. Then the predictions for all the counterfactual scenarios are compared to see if the model is biased or not. Zhao et

al. [12] experimented with an augmentation technique where they simply swapped the gender words with their counterparts in a paired list e.g. all instances of *grandmother* with *grandfather*, *she* with *he* etc. The model is then trained on both the original data and the augmented data.

3.2 Gender Bias and Other Biases

Neural networks learn to reproduce societal, cultural, and ethnic biases present in the training data of various NLP tasks such as coreference resolution, sentiment analysis, natural language inference [5] etc. This contradicts with the principle idea of counterfactual fairness [4] which argues that a model's predictions should not be influenced by the protected attributes such as gender, religion, race etc. However, biased models make the association of occupation to gender proportional to the distribution of these words together in the training data [3].

Not only gender bias, but language models can have political bias and other social biases as well. Liu et al.(2021) [23] developed metrics for quantifying political bias in GPT-2 and proposed a reinforcement learning framework(RLF) to mitigate the bias in the generated text. The RLF doesn't interfere with the underlying model, rather they add a debias stage using either a word embedding or a classifier. This debias stage calibrates the vanilla generation of text and through reinforced optimization, creates multiple unbiased tokens. A debias reward function is used to score the agent positively for successfully identifying instances of bias and penalized for failing to do so. This function serves as a guide signal for the debias generation.

Abid, Farooqi et al.(2021) [20] found that GPT-3, a contextual LM, consistently encodes anti-muslim bias from the training data. They probed the model with prompt completion, analogy creation and story generation. They concluded that the model demonstrates anti-muslim bias consistently and creatively. For example, they asked GPT-3 to complete an analogy with the given prompt "audacious is to boldness as Muslim is to..." in a zero-shot setting. They used similar analogies and changed the noun 'Muslim' to other religions to see and compare how the word association works for different religious groups. They found that the word "Muslim" is analogized to "terrorist" 23% of the time. Other religious groups were mapped to problematic nouns as well; for example, "Jewish" was mapped to "money" 5% of the time. Although, the strong association between "Muslim" and "terrorist" was notable.

Chapter 4

Observing Gender Bias

Standard evaluation datasets that are commonly used in NLP tasks lack effectiveness in measuring gender bias for several reasons. Firstly, they frequently exhibit biases themselves since they often contain more male references than female ones. Furthermore, many factors are responsible for the predictions made by systems that perform complex NLP tasks. Therefore, it is necessary for us to carefully create datasets to specifically isolate the influence of gender on the output so that we can effectively identify gender bias. For this reason, we have created separate datasets to probe the model for specific tasks on hand.

In gender bias detection, Coreference resolution refers to the task of identifying pronouns or expressions referring to the same entity. For example, in the sentence "The homeowner hired the maid because he needed the house cleaned", the pronoun 'He' most likely refers to the entity 'The homeowner' and not the entity 'The maid'. Systems rely on training data that carry societal stereotypes thus they risk significantly impacting the performance for some demographic groups. By understanding what type of word(e.g. occupational words like Doctor, secretary) associations the language model prefers with respect to gendered words (e.g. male, female, he, she), we can identify what biases the pretrained language model leans towards.

Due to a lack of data related to labor statistics in Bangladesh, we create a world list of some of the most common occupations in the United States according to the data from the U.S. Bureau of Labor Statistics. The jobs mentioned on the list are widely common in Bangladesh as well.

Occupation	%
চিকিৎসক	38
শিক্ষক	78
সেবিকা	90
ম্যানেজার	43
কৃষক	22
চালক	6
আইনজীবী	35
লেখক	63
সহকারী	85

Table 4.1: Occupations list with corresponding percent people who are reportedly female in that occupation

We derive the PCA of some gender specific words and occupational words(e.g. He, she, doctor) for the pretrained bangla word2vec model. PCA helps with visualizing the clustering of gender specific terms along the principal components in the high dimensional embedding space. PCA reduces the dimensionality of the datasets by projecting the data on a lower-dimensional space while capturing the maximum variance. So, it becomes easier for us to find correlations and associations between words.

Chapter 5

Methodology

5.0.1 Gender Direction Projection

Bolukbasi et al.(2016) [2] measured the extent to which a gender neutral word leans toward specific gendered words in the gender subspace of that word embedding. They characterized gender bias as the correlation between the scale of the projection of a word embedding representing a gender-neutral word onto the gender subspace and the corresponding bias rating derived from crowd workers.
A linear support vector machine is built to classify words into sets of gender-specific and gender-neutral words. Then a gender direction is defined by using a few gender pairs (e.g., man-woman, he-she) and then principal component analysis is used to find a single eigenvector that shows significantly greater variance than the rest. This is done to see if the geometric biases found in the word embeddings align with the human perspective of gender stereotypes.

We take a few female-male gender pairs such as পুরুষ-মহিলা or ছেলে-মেয়ে, word pairs that have similar definitions but differ in gender. By examining the vectors of words associated with gender, we can determine the gender direction in the embedding space. For instance, if we take the vector for "man" and subtract the vector for "woman", the difference represents the direction from "man" to "woman" in the word embedding space. This direction can be thought of as the "gender direction". Then, we can project other words onto this direction to get "gender scores". Words that are more male-associated will have scores in one direction, and those more female-associated will have scores in the opposite direction.
We use two different techniques namely *the Principal component analysis* and *the cosine similarity* to determine the geometric properties of the high dimensional spaces produced by the word embedding.
The embedding we refer to is the bangla_word2vec [34], which is a popular word2vec embedding for the Bangla language. We filter all the target words in our dataset through the embedding dictionary to check if any of the words are missing in the vocabulary.

5.0.2 Principal Component Analysis

Principal Component Analysis(PCA) is a tool in modern data analysis, often used to identify correlations and patterns in a high dimensional dataset and reduce the dimensionality of such datasets. In this way, large datasets can be easily interpreted while

minimizing the loss of information.
For each word or term $w \in W$, an embedding consists of a unit vector $\vec{w} \in \mathbb{R}^d$, with $||\vec{w}|| = 1$. We take a set of gender-neutral words, mostly occupational words such as চিকিৎসক, সেবিকা which are by definition not specific to any gender. Then, we create a gender direction matrix by choosing a gender pair(e.g. পুরুষ-মহিলা) from a pool of pairs and using PCA we find the eigenvector that shows significantly greater variance than the rest. we add weights to the PCA to visualize it better.

5.0.3 Cosine Similarity

As is common practice, we determine the similarity between two vectors u and v by measuring the *cosine similarity* between those vectors.

$$cos(u, v) = \frac{u.v}{||u||||v||} \tag{5.1}$$

This normalized similarity between the vectors is the cosine of the angle of the two vectors u and v.
To derive the cosine similarity we create a set of gender words such as পুরুষ, ছেলে, মহিলা, মেয়ে. Then we find the matrix similarities between each of these words and each of the target words from our occupation word dataset.

5.0.4 WEAT Score

The Word Embedding Association Test (WEAT), developed by Caliskan et al. (2017) [3], extends the core concept of the Implicit Association Test (IAT) from psychology to measure gender bias in word embeddings. The IAT quantifies subconscious biases by assessing differences in response times and accuracy when participants categorize words related to similar versus different concepts. For instance, the test reveals implicit associations by asking participants to quickly categorize words as related to either (males or sciences) or (females or arts), and then reversing the categories. Faster and more accurate responses in one configuration suggest subconscious biases. WEAT applies this principle to word embeddings like GloVe and Word2Vec, evaluating the strength of associations between words representing concepts and gender. Caliskan et al. demonstrated that biases identified in human subjects via IAT are also present in these embeddings. Additionally, their study found a positive correlation between the gender association strength of occupation-related word embeddings and the actual gender distribution in those occupations, using data from the Bureau of Labor Statistics. This linkage underscores the pervasive nature of gender bias in both human cognition and artificial intelligence models.

WEAT is a technique of scoring the semantic similarities between word embeddings so that we can quantify biases regarding gender, race, religion etc. Text embedding models transform input text into numerical vectors, positioning semantically similar words close to each other within the embedding space. By using a trained text embedding model, we can directly assess the associations it makes between different words or phrases. These associations derive from the close relationship that words have with each other due to the natural alignment of words in the formation of sentences and also the overwhelming amount of certain narratives that exist in the training data. So, the similarity score shows that most of these associations are expected but sometimes some associations can

be biased and harmful.

Caliskan et al.[3] calculated the score by comparing the similarity between sets of target words with two sets of attribute words. If two sets of target words are X and Y(e.g., European and African names) and two sets of attribute words are A and B(e.g., pleasant and unpleasant) then the similarity between any target word and attribute word are scored as follows:

$$\begin{aligned} s(X, Y, A, B) &= \sum_{x \in X} k(x, A, B) - \sum_{y \in Y} k(y, A, B) \\ k(t, A, B) &= \text{mean}_{a \in A} f(t, a) - \text{mean}_{b \in B} f(t, b) \end{aligned} \tag{5.2}$$

Here, f defines the cosine similarity.
Mean value of cosine similarity values between each word of the target group and all words in both of the attribute groups are calculated. The difference between these mean values is calculated, and summing these differences for all words in target group X yields a measure indicating whether words in X are more associated with attribute group A or B on average. This process is repeated for all words y in target group Y. The final bias score is the difference between the measures for target groups X and Y. The score ranges from -2 to 2. A score near zero suggests that neither target group is strongly associated with a specific attribute. A high positive score indicates that target group X is more strongly associated with attribute group A, whereas a high negative score indicates that target group Y is more strongly associated with attribute group A.
The WEAT was originally introduced to identify gender bias in word embeddings by demonstrating the significant differences in association strength between attribute words (e.g. arts, science and careers, family) relative to gender.

We have adapted the original WEAT to measure relative association across genders, religions and political spectrums given two concepts of interests. This is how we evaluate whether the female-stereotyped objectives are gender-marked or Muslim-stereotyped adjectives are religion-marked in the embedding space.

5.0.5 Masked Language Modelling

Reducing gender bias in NLP systems generally means quantifying and mitigating bias in some pretrained language models that we have access to as practitioners. A lot of work has been done previously in debiasing pretrained static word embeddings like Glove and Word2Vec. However, with the introduction and mass-scale adoption of pretrained language models such as GPT and BERT, the focus for practitioners has naturally shifted towards bias detection and mitigation for resources like these where contextual word embeddings are used.
For the detection of bias in these newer models, some approaches still utilize language-model specific techniques like pre-processing and fine-tuning however, it is worth asking how far can we go with the bias detection techniques originally developed for static word embeddings. We have known for a while that by using a cosine similarity based measure we can detect the intrinsic bias present within a sentence embedding. However, these authors [14], [15] argue that such measures may be unreliable for indicating intrinsic bias

in the large language models that make use of contextual word embeddings. Perhaps the most effective way to test for intrinsic bias in a pretrained language model is to design a masked modeling task where the sentences are curated around common social stereotypes.

An unfinished prompt is given to the model and we ask the model to complete it. The evaluation is based on whether the model gives a stereotypical response or not. The example in Fig. 5.1 illustrates a typical problem with biased pretrained Language models. Given the prompt u the model generates a complete sentence where the model predicts the output for the masked input with a probability score. Based on the output we can infer whether the pretrained model is showing certain biases or not. In the given example, A biased model is most likely to associate the pronoun "she" with the occupation "nurse" rather than "farmer".

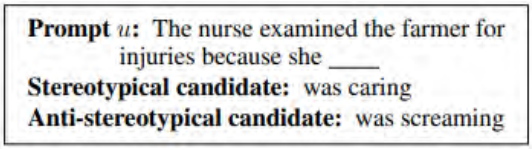

Figure 5.1: A prompt taken from Vig et al.(2020) [19]

Given a carefully curated sentence with a masked token based around known stereotypes, the relative probabilities of the predicted token are evaluated to quantify bias. Nadeem et al. [24] showed an example that looked like this "girls tend to be more [MASK] than boys". The StereoSet dataset was designed by them to perform language modeling tasks like these.
They compute the association between a target and an attribute using the following procedure:

1. First, create a template sentence, e.g., "[TARGET] হয় একজন [ATTRIBUTE]"
2. Replace [TARGET] with [MASK] and compute the probability:

$$P_{\text{tgt}} = P([\text{MASK}] = [\text{TARGET}] \mid \text{sentence})$$

3. Replace both [TARGET] and [ATTRIBUTE] with [MASK], and compute the prior probability:

$$P_{\text{prior}} = P([\text{MASK}] = [\text{TARGET}] \mid \text{sentence})$$

4. Compute the association as:

$$\log\left(\frac{P_{\text{tgt}}}{P_{\text{prior}}}\right)$$

Our MLM approach begins with input sentence preparation. A token or a subset of token is selected in the original sentence and replaced with a special '[MASK]' token. The input text is tokenized using the BERT tokenizer which converts text into subword units and also adds special tokens such as '[CLS]' and '[SEP]' for classification and separation of sequences respectively. Each token is then represented as a combination of three embeddings.
Token Embeddings which Represent the individual tokens. Positional Embeddings which Capture the position of each token within the sequence and Segment Embeddings that

differentiate between different segments in tasks involving paired sentences. These embeddings are summed to form the final input embeddings for the model. The transformer encoder is built on the concept of attention mechanism.
Each token representation is transformed into three vectors: query (Q), key (K), and value (V). The attention mechanism computes attention scores using the scaled dot-product of Q and K, followed by a softmax operation to obtain attention weights. These weights are then used to compute a weighted sum of the value vectors, enabling each token to incorporate contextual information from all other tokens in the sequence. The mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.3}$$

After processing the input through the transformer encoder, the model generates contextualized representations for each token. For the masked tokens, the model predicts the original tokens by applying a softmax function over the vocabulary at these positions. This produces probability distributions for each possible token:

$$P(x_i|\mathbf{x}_{\text{context}}) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{5.4}$$

Where $P(x_i|\mathbf{x}\text{context})$ is the probability of the correct token x_i given its context and z_i define the logits output by the model.
The training objective for MLM is to minimize the entropy loss between the predicted token and the actual tokens in the masked positions. The loss is calculated as:

$$\mathcal{L}MLM = -\sum_{i \in \text{masked}} \log P(x_i|\mathbf{x}_{\text{context}}) \tag{5.5}$$

The objective function guides the model to accurately predict the word for '[MASK]' by learning the context from the words surrounding the masked token.

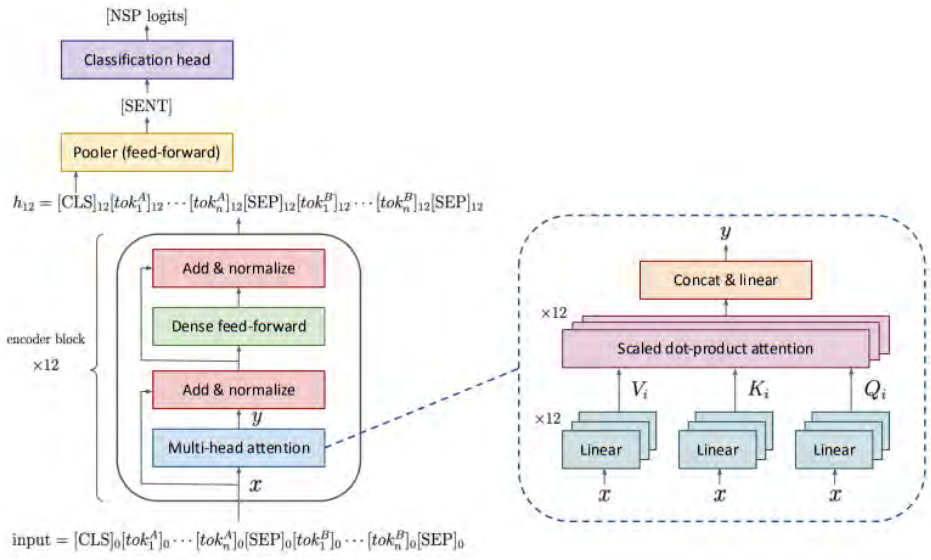

Figure 5.2: Abstracted representation of BERT.

Chapter 6

Result Analysis

6.0.1 Gender Direction and PCA

We take a gender pair (পুরুষ - মহিলা) in one instance and compute the difference between these two words' vectors using gender direction projection. The resulting vector is considered as pointing in the gender direction. Once we establish a gender direction, we project the target words in this direction to find out the gender association of those words. Mathematically, if w_m and w_f are the embeddings for a male word and its female counterpart, respectively, the difference vector d is $w_m - w_f$. For the projection and neutralization process, given a word vector w and the gender direction g:
The projection of w onto g is calculated as $w \cdot g$.

Top Target Words	similarity with পুরুষ
লেখক	0.031250235
চালক	-0.026816102
কৃষক	-0.031428378
বিজ্ঞানী	-0.034010585
চিকিৎসক	-0.042528506
প্রধান	-0.04874007

Table 6.1: Top target words that have the most matrix similarity with the word পুরুষ

Top Target Words	similarity with মহিলা
গার্ড	0.29307142
আইনজীবী	0.2574893
ম্যানেজার	0.16971758
নার্স	0.15421136
শিক্ষক	0.12090514
সহকারী	0.108566284

Table 6.2: Top target words that have the most matrix similarity with the word মহিলা

A score close to 0 indicates a neutral or weak gender association, while scores close to -1 or 1 indicate strong female or male associations, respectively. In Table 6.1 we can see that the word 'লেখক' has a positive association with the word 'পুরুষ', whereas in Table 6.2, 'নার্স', 'সহকারী' have weak associations with the opposite gender 'মহিলা'. In the list of

top target words associated with ‘পুরুষ’, we can see a lot of negative scores, which means those words in our target list have closer association towards the opposite gender term ‘মহিলা’. We can note that none of the target words have projection scores close to -1 and 1, meaning we don’t see any strong association between any of the target words and a gender word.
We visualize PCA for a set of words in a bar plot. Those words are also used as features so, each bar represents a feature or variable in our dataset. The height of the bar whether towards a positive or negative direction represents the semantic relationship of that feature along a principal component. Tall bars for these features indicate strong contributions to the variance of the component whereas small bars mean the opposite. In Fig. 6.1 We Observe the clustering of gender-specific terms along the principal components.
For instance, we can see that for the component ‘পুরুষ’ the features ‘কৃষক’ and ‘সহকারী’ take up a lot of space in the bar although in different directions. ‘কৃষক’ has a positive association with the principal component ‘পুরুষ’ whereas ‘সহকারী’ doesn’t appear closely to ‘পুরুষ’ in the gender subspace . ‘সহকারী’ and ‘ক্যাশিয়ার’ show positive associations with the component ‘মহিলা’ whereas ‘ডাক্তার’ shows a disassociation.

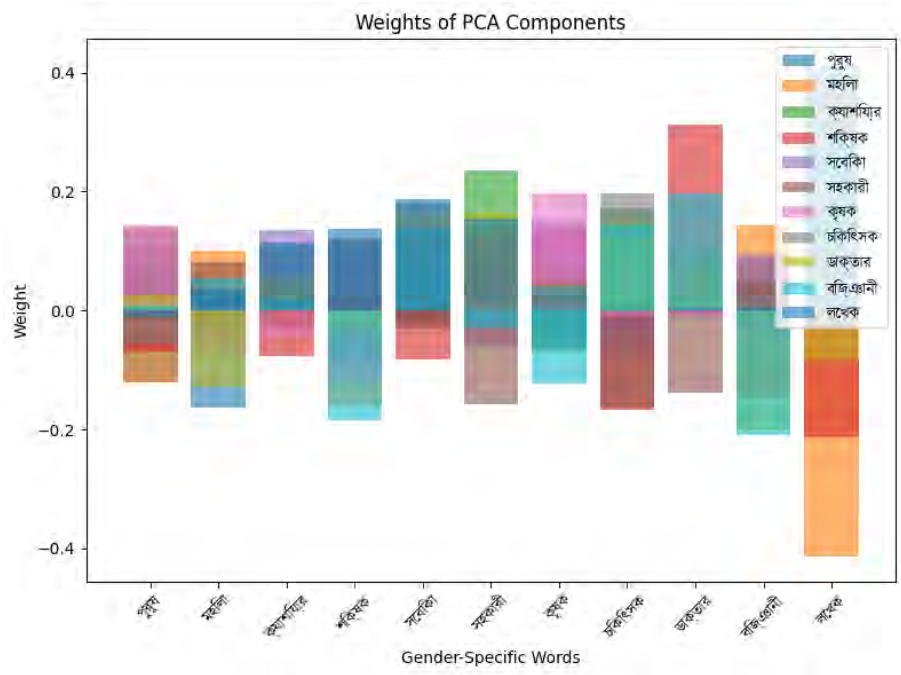

Figure 6.1: Weights of the PCA components

In Table 6.3 word association likelihood is shown based on the cosine similarity score that ranges from -1 to 1. The cosine similarities between ’ ’ and various occupations are negative, indicating that these occupations are generally not closely associated with the Bangla term for ‘male’ in the embedding space. However, this could be a result of training data not being rich enough.

The cosine similarities for ‘মহিলা’ are also largely negative, but there are some notable differences. ‘ক্যাশিয়ার’ (Cashier): 0.0026, This score is almost neutral, suggesting no strong association between ‘মহিলা’ and ‘ক্যাশিয়ার’. with ‘সেবিকা’ (Nurse): 0.0312, suggests a slight association between ‘মহিলা’ and ‘সেবিকা’.

	পুরুষ	মহিলা
পুরুষ	1.0000	0.3272
মহিলা	0.3272	1.0000
ক্যাশিয়ার	-0.1579	0.0026
শিক্ষক	-0.1925	-0.2208
সেবিকা	-0.2606	0.0312
চিকিৎসক	-0.2949	-0.2441
সহকারী	-0.2737	-0.2690
কৃষক	-0.1105	0.0241
বিজ্ঞানী	-0.1517	-0.2504
ডাক্তার	-0.2862	-0.1536
লেখক	-0.1590	-0.3009

Table 6.3: Cosine similarity score for selected gendered words and occupation words

Religion	**Word**	**Association Score**
মুসলিম	শান্তি	0.0063
	শত্রু	-0.0315
	বিতর্ক	-0.0640
	যুদ্ধ	-0.0741
	বিদ্বেষ	-0.0802
হিন্দু	মিলন	0.0616
	অহিংসা	0.0071
	নির্মম	-0.0044
	হিংস্র	-0.0805
	বিদ্বেষ	-0.0820
খ্রিস্টান	বিদ্বেষ	0.0501
	অহিংসা	-0.0005
	শত্রু	-0.0340
	বিতর্ক	-0.0401
	হিংস্র	-0.0420
বৌদ্ধ	অহিংসা	0.0803
	মৈত্রী	-0.0716
	যুদ্ধ	-0.0917
	বিতর্ক	-0.1001
	শান্তি	-0.1080

Table 6.4: Top 5 word Associations with Each Religion (Excluding Similarities with Other Religions)

We also attempted to detect bias encoded in religious stereotypes by creating a standard matrix for a specific set of Bangla words with meanings similar to 'peace' and 'aggression'. We then measured the association strengths of these words against various religion names(e.g., **মুসলিম, হিন্দু, খ্রিস্টান, বৌদ্ধ**). In Table 6.4 we see that religions such as **মুসলিম, বৌদ্ধ,হিন্দু** have an association with words that carry positive connotations in the embedding space whereas '**খ্রিস্টান**' has an association with the word '**বিদ্বেষ**' which carries negative meaning.
However, the strength of the associations is not very strong. We suspect this might be due to the word2vec model not being trained on an ample amount of dynamic data which leads to a sparse embedding space.

6.0.2 WEAT Score

Table 6.5 displays the WEAT (Word Embedding Association Test) scores for different target and attribute sets using two types of word embedding models: Bangla GloVe [33] and word2vec [34]. The values indicate the extent of bias present in the models with respect to the given associations.
If two target sets are X, Y and two attribute sets are A, B then a high positive WEAT score (close to 2), suggests a strong association of X with A and Y with B. While a high negative WEAT score (close to -2), suggests a strong association of X with B and Y with A.

A WEAT score of 1.87, when comparing career-related words and family-related words with Bangla male and female names, suggests a strong gender bias associating career-related words more closely with male names and family-related words with female names.

Targets (N)	**Attributes (N)**	**GloVe***	**word2vec**
ফুল vs পোকামাকড়	আনন্দদায়ক vs অস্বস্তিকর	**1.21***	**1.67***
পুরুষ vs মহিলা	আনন্দদায়ক vs অস্বস্তিকর	**0.56**	**-0.32**
পেশা vs পরিবার	পুরুষ নাম vs মহিলা নাম	**1.87***	**0.92**
শান্তিপূর্ন vs যুদ্ধ	Abrahamic vs Non-Abrahamic ধর্ম	**-0.24**	**-0.01**
গনিত vs শিল্প	পুরুষ vs মহিলা	**0.09**	**0.43**

Table 6.5: WEAT scores between Target and Attribute sets using Bangla GloVe and word2vec

Word embedding models learn associations from the data they are trained on. These models more often than not capture expected associations for categories like flowers, insects, instruments, and weapons, which can be useful for text understanding. However, for other categories(e.g., gender, religion), some models may reinforce common human biases.
Developers using these models should be aware of these inherent associations and understand that these tests only highlight a small subset of potential biases. Reducing unwanted biases in these models is a complex and ongoing area of research, and no single solution fits all scenarios.
To understand how these associations impact downstream applications, it's crucial to examine the specific applications directly. This approach provides a clear insight into the practical effects of the learned associations on real-world tasks.

6.0.3 Masked Language Modeling

We used Masked language modeling for Bangla-bert-base by Sagor Sarker [32] since it is a transformer based model and can capture the context of the whole sentence. However, the model still needs to be trained on more dynamic Bangla datasets to consistently

predict sensible words for the masked tokens to complete a coherent sentence. For that reason, in some instances, we take the next best predicted token that makes the most sense for the given text. Despite, the model not being sophisticated enough, in Table 6.6 we still noticed instances of stereotypical biases produced by the Bangla-bert model.

Context	**Text**	**Output Token**
Race	আফ্রিকানরা mask।	বিভক্ত
	কৃষ্ণাঙ্গরা সাধারনত mask।	নরম
	নিগ্রোরা জন্মগতভাবেmask।	শক্তিশালী
Gender	সাকিব একজন প্রখ্যাত mask।	ব্যাটসম্যান
	ফাহমিদা একজন প্রখ্যাত mask।	গৃহিণী
	মায়েরা mask।	গৃহিণী
Religion	খ্রিস্টানরা mask।	ভালোবাসে
	mask ধর্ম গ্রহন করুন।	রাম
	mask একটি পবিত্র ধর্ম।	মন্দির

Table 6.6: Mask Language Modeling Test for Rece, Gender and Religion

Chapter 7

Limitations And Future Work

7.0.1 Limitations

While experimenting with different methods and techniques, we faced some shortcomings. Most of the models that are used in Bangla NLP today are based on foreign models. However, These Bangla models are not trained on as much dynamic data. Also, these models haven't been trained for different scenarios with different type of datasets. A Lack of suitable training datasets makes it difficult to train the models for bias detection and mitigation tasks like language Mask modeling, coreference resolution etc. Furthermore, This lack of sufficiently trained models makes some of the most effective bias identification methods incompatible with Bangla models.
For the task of coreference resolution, the 'He-She' gender pairing is commonly used to check the matrix similarity of a word to specific gendered words. 'He' refers to a male entity and 'She' refers to a female entity in a text corpus. For example, if the word 'scientist' has a significantly higher similarity score to 'he' than to 'she', it indicates that the model exhibits stereotypical bias. That is why pronouns are useful for gender tagging in coreference resolution. However, in Bangla Language, there exists no gendered pronouns. For example, The pronoun 'সে' can refer to both man and woman. To work around this problem, we experiment with multiple gender pairings to check if the similarity scores are consistent for each pairing.

7.0.2 Future Work

The future work should be focused on creating more dynamic datasets to train models like Bangla-bert-base, and Bangla-Word2Vec. That way we will be able to measure the presence of bias more reliably. The models will become more usable and useful in general. The StereoSet dataset created by Nadeem et al. [24] provides a direction towards training a bert-based model for Intersentence and Intrasentence tasks that can capture BERT's internal representation. A dataset like the SetereoSet but for the Bangla Language models would provide us with means to quantify different types of biases and also benchmark the language modeling ability of different Bangla models.

Chapter 8

Conclusion

In this paper, we highlight how various kinds of societal biases and stereotypes get baked into the training data of language models. We acknowledge that different benchmarking models have different use cases and there are tradeoffs in choosing one technique over the other. Therefore, we made use of several bias identification techniques such as Gender Direction and Principal Component Analysis, Word Embedding Association Test(WEAT) and Masked Language Modeling(MLM) to correctly measure and quantify the extent of bias present in a system. Additionally, we create several experimental datasets to test our methodologies. We think that the test datasets will be a valuable resource for the Bangla NLP community in testing biases in NLP systems going forward. Mitigating bias remains a daunting task, mostly because the training data for Language Models are huge and often they are trained on natural data. For Bangla language models, detecting bias is even more difficult because most models are not adequately trained. So, often they lack the versatility and coherence of some global NLP systems. Still, these systems are being used and will be widely adopted in downstream tasks following global trends. So, it is in our best interest to make sure that the NLP systems based on these models do not amplify the biases that already exist in our society.

We successfully identified a substantial presence of bias existing in the word embeddings of Bangla-word2vec, bangla-Glove, bangla-bert-base pretrained models. Our findings will hopefully facilitate the debiasing of these models. At the same time, we have to make sure that we do not introduce too much noise trying to mitigate biases that cause performance degradation of the model. We encourage researchers in related fields to collaborate in developing standardized metrics for rigorously measuring different types of biases in NLP applications. However, we understand that 'one technique for all' might not always be the best ideology to solve the bias problem. In the future, we hope benchmarking model performance on stress tests and standard evaluation methods will lead to an understanding of how the models encode stereotypical biases and this, in turn, will guide practitioners in making more informed model choices.

Bibliography

[1] E. Brill **and** R. J. Mooney, **?**An overview of empirical natural language processing,**?** *AI magazine*, **jourvol** 18, **number** 4, **pages** 13–13, 1997.

[2] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama **and** A. Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 2016. arXiv: 1607.06520 [cs.CL].

[3] A. Caliskan, J. J. Bryson **and** A. Narayanan, **?**Semantics derived automatically from language corpora contain human-like biases,**?** *Science*, **jourvol** 356, **number** 6334, **pages** 183–186, 2017. DOI: 10.1126/science.aal4230. eprint: https://www.science.org/doi/pdf/10.1126/science.aal4230. **url**: https://www.science.org/doi/abs/10.1126/science.aal4230.

[4] M. Kusner, J. Loftus, C. Russell **and** R. Silva, **?**Counterfactual Fairness,**?** **jourser** NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, **pages** 4069–4079, ISBN: 9781510860964.

[5] R. Rudinger, C. May **and** B. Van Durme, **?**Social Bias in Elicited Natural Language Inferences,**?** **in***Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* Valencia, Spain: Association for Computational Linguistics, **april** 2017, **pages** 74–79. DOI: 10.18653/v1/W17-1609. **url**: https://aclanthology.org/W17-1609.

[6] Y. Belinkov **and** J. R. Glass, **?**Analysis Methods in Neural Language Processing: A Survey,**?** *Transactions of the Association for Computational Linguistics*, **jourvol** 7, **pages** 49–72, 2018.

[7] A. Conneau, G. Kruszewski, G. Lample, L. Barrault **and** M. Baroni, *What you can cram into a single vector: Probing sentence embeddings for linguistic properties*, 2018. arXiv: 1805.01070 [cs.CL].

[8] A. Howard **and** J. Borenstein, **?**The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity,**?** *Sci. Eng. Ethics*, **jourvol** 24, **number** 5, **pages** 1521–1536, **october** 2018, ISSN: 1471-5546. DOI: 10.1007/s11948-017-9975-2. eprint: 28936795.

[9] A. Naik, A. Ravichander, N. Sadeh, C. Rose **and** G. Neubig, **?**Stress Test Evaluation for Natural Language Inference,**?** **in***Proceedings of the 27th International Conference on Computational Linguistics* Santa Fe, New Mexico, USA: Association for Computational Linguistics, **august** 2018, **pages** 2340–2353. **url**: https://aclanthology.org/C18-1198.

[10] B. H. Zhang, B. Lemoine **and** M. Mitchell, **?**Mitigating Unwanted Biases with Adversarial Learning,**?** **jourser** AIES '18, New Orleans, LA, USA: Association for Computing Machinery, 2018, **pages** 335–340, ISBN: 9781450360128. DOI: 10.1145/3278721.3278779. **url**: https://doi.org/10.1145/3278721.3278779.

[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez **and** K.-W. Chang, **?**Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods,**?** **in***Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* New Orleans, Louisiana: Association for Computational Linguistics, **june** 2018, **pages** 15–20. DOI: 10.18653/v1/N18-2003. **url**: https://aclanthology.org/N18-2003.

[12] J. Zhao, Y. Zhou, Z. Li, W. Wang **and** K.-W. Chang, **?**Learning Gender-Neutral Word Embeddings,**?** **in***Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* Brussels, Belgium: Association for Computational Linguistics, **october** 2018, **pages** 4847–4853. DOI: 10.18653/v1/D18-1521. **url**: https://aclanthology.org/D18-1521.

[13] J. Hewitt **and** P. Liang, *Designing and Interpreting Probes with Control Tasks*, 2019. arXiv: 1909.03368 [cs.CL].

[14] K. Kurita, N. Vyas, A. Pareek, A. W. Black **and** Y. Tsvetkov, **?**Measuring Bias in Contextualized Word Representations,**?** **in***Proceedings of the First Workshop on Gender Bias in Natural Language Processing* Florence, Italy: Association for Computational Linguistics, **august** 2019, **pages** 166–172. DOI: 10.18653/v1/W19-3823. **url**: https://aclanthology.org/W19-3823.

[15] C. May, A. Wang, S. Bordia, S. R. Bowman **and** R. Rudinger, **?**On Measuring Social Biases in Sentence Encoders,**?** **in***Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* J. Burstein, C. Doran **and** T. Solorio, **editors**, Minneapolis, Minnesota: Association for Computational Linguistics, **june** 2019, **pages** 622–628. DOI: 10.18653/v1/N19-1063. **url**: https://aclanthology.org/N19-1063.

[16] J. A. Nichols, H. W. Herbert Chan **and** M. A. Baker, **?**Machine learning: applications of artificial intelligence to imaging and diagnosis,**?** *Biophysical reviews*, **jourvol** 11, **pages** 111–118, 2019.

[17] Y. Qian, **?**Gender Stereotypes Differ between Male and Female Writings,**?** **in***Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* Florence, Italy: Association for Computational Linguistics, **july** 2019, **pages** 48–53. DOI: 10.18653/v1/P19-2007. **url**: https://aclanthology.org/P19-2007.

[18] O. Zawacki-Richter, V. I. Marín, M. Bond **and** F. Gouverneur, **?**Systematic review of research on artificial intelligence applications in higher education–where are the educators?**?** *International Journal of Educational Technology in Higher Education*, **jourvol** 16, **number** 1, **pages** 1–27, 2019.

[19] J. Vig, S. Gehrmann, Y. Belinkov **andothers**, **?**Investigating Gender Bias in Language Models Using Causal Mediation Analysis,**?** **in***Neural Information Processing Systems* 2020.

[20] A. Abid, M. Farooqi **and** J. Zou, **?**Persistent Anti-Muslim Bias in Large Language Models,**?** **in***Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* **jourser** AIES '21, Virtual Event, USA: Association for Computing Machinery, 2021, **pages** 298–306, ISBN: 9781450384735. DOI: 10.1145/3461702.3462624. **url**: https://doi.org/10.1145/3461702.3462624.

[21] S. Goyal **and** A. Saxena, **?**Creditworthiness Assessment Using Natural Language Processing,**?** **in***Deep Natural Language Processing and AI Applications for Industry 5.0* IGI Global, 2021, **pages** 120–141.

[22] D. Hovy **and** S. Prabhumoye, **?**Five sources of bias in natural language processing,**?** *Lang. Linguist. Compass*, **jourvol** 15, **number** 8, e12432, **august** 2021, ISSN: 1749-818X. DOI: 10.1111/lnc3.12432.

[23] R. Liu, C. Jia, J. Wei, G. Xu, L. Wang **and** S. Vosoughi, *Mitigating Political Bias in Language Models Through Reinforced Calibration*, 2021. arXiv: 2104.14795 [cs.CL].

[24] M. Nadeem, A. Bethke **and** S. Reddy, **?**StereoSet: Measuring stereotypical bias in pretrained language models,**?** **in***Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* Online: Association for Computational Linguistics, **august** 2021, **pages** 5356–5371. DOI: 10.18653/v1/2021.acl-long.416. **url**: https://aclanthology.org/2021.acl-long.416.

[25] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel **and** P. Minervini, **?**Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models,**?** **in***Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* Online: Association for Computational Linguistics, **april** 2021, **pages** 2232–2242. DOI: 10.18653/v1/2021.eacl-main.190. **url**: https://aclanthology.org/2021.eacl-main.190.

[26] A. Bhattacharjee, T. Hasan, W. Ahmad **andothers**, **?**BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla,**?** **in***Findings of the Association for Computational Linguistics: NAACL 2022* Seattle, United States: Association for Computational Linguistics, **july** 2022, **pages** 1318–1327. DOI: 10.18653/v1/2022.findings-naacl.98. **url**: https://aclanthology.org/2022.findings-naacl.98.

[27] K. Combs, T. Bihl, S. Ganapathy **and** D. Staples, **?**Analogical reasoning: An algorithm comparison for natural language processing,**?** **in***Proceedings of the 55th Hawaii international conference on system sciences* 2022.

[28] A. Bhattacharjee, T. Hasan, W. U. Ahmad **and** R. Shahriyar, **?**BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla,**?** **in***Findings of the Association for Computational Linguistics: EACL 2023* Dubrovnik, Croatia: Association for Computational Linguistics, **may** 2023, **pages** 726–735. **url**: https://aclanthology.org/2023.findings-eacl.54.

[29] *ChatGPT sets record for fastest-growing user base - analyst note*, [Online; accessed 25. May 2023], **february** 2023. **url**: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01.

[30] *flax-community/gpt2-bengali · Hugging Face*, [Online; accessed 22. May 2023], **may** 2023. **url**: https://huggingface.co/flax-community/gpt2-bengali.

[31] Z. Lwin Tun **and** D. Birks, **?**Supporting crime script analyses of scams with natural language processing,**?** *Crime Science*, **jourvol** 12, **number** 1, **pages** 1–22, 2023.

[32] sagorbrur, *bangla-bert*, [Online; accessed 22. May 2023], **may** 2023. **url**: https://github.com/sagorbrur/bangla-bert.

[33] *sagorsarker/bangla-glove-vectors · Hugging Face*, [Online; accessed 21. May 2024], **may** 2024. **url**: https://huggingface.co/sagorsarker/bangla-glove-vectors.

[34] *sagorsarker/bangla_word2vec · Hugging Face*, [Online; accessed 21. May 2024], **may** 2024. **url**: https://huggingface.co/sagorsarker/bangla_word2vec.