Detecting and Mitigating of biasness in Machine Translation

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

RISHABH JAIN

2K22/AFI/15

Under the supervision of

Ms. GARIMA CHHIKARA

Computer Science and Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

RISHABH JAIN, Roll No's – 2K22/AFI/15 students of M.Tech (COMPUTER SCIENCE AND ENGINEERING),hereby declare that the project Dissertation titled "Detecting and Mitigating of biasness in Machine Translation" which is submitted by us to the DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi Rishabh Jain

Date: 31.05.2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Detecting and Mitigating of biasness in Machine Translation" which is submitted by RISHABH JAIN, Roll No – 2K22/AFI/15, COMPUTER SCIENCE AND ENGINEERING ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Ms. GARIMA CHHIKARA

Date: 31.05.2024

SUPERVISOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

We wish to express our sincerest gratitude to Ms. GARIMA CHHIKARA for his continuous guidance and mentorship that he provided us during the project. She showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. She was always ready to help us and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi Rishabh Jain

Date: 31.05.2024

Abstract

Machine translation (MT) systems have become indispensable tools for cross-lingual communication, yet their ability to accurately convey gender-specific nuances remains a significant challenge. This study provides a comprehensive evaluation of popular MT systems like Google Translate and Bing Microsoft Translator, focusing on the Hindi-to-English language pair, where grammatical gender plays a crucial role.

We investigate the systems' performance in translating gendered vocabulary, encompassing personality adjectives, professions, nouns, and pronouns. Analyzing the prevalence of female, male, and neutral forms in the translated output, we assess the systems' capacity to preserve the original gender intent and identify potential biases.

Our research emphasizes professional contexts, where gender bias can have profound implications for individuals and society. We scrutinize how MT systems handle gendered language nuances in these settings, examining whether they inadvertently memorialize stereotypes or introduce discriminatory language.

Through rigorous statistical analysis of translation outputs, we aim to quantify bias in these systems. By identifying specific areas where biases occur, we can provide valuable insights to guide the development of more equitable and inclusive translation technologies.

This study contributes to a deeper understanding of the capabilities and limitations of current MT systems in accurately representing gender-specific language constructs. Ultimately, our goal is to foster more nuanced and sensitive cross-lingual communication by highlighting areas for improvement and promoting the development of MT tools that respect and preserve the diversity of gendered language.

Contents

List of Tables

List of Figures

List of Symbols

Chapter 1

INTRODUCTION

1.1 Overview

In an increasingly interconnected world, machine translation (MT) has completely revolutionized cross-lingual communication, empowering individuals and organizations to effortlessly transcend language barriers. Fueled by advanced algorithms and vast linguistic databases, MT systems have democratized translation, making it accessible to a broader audience. This accessibility is absolutely crucial for various applications, from global business expansion to personal communication across cultures. Machine translation Systems struggles with gender representation, especially when translating between languages with differing gender systems. This often leads to unintentional gender bias in translations, particularly in the choice of professions, pronouns, and verbs.

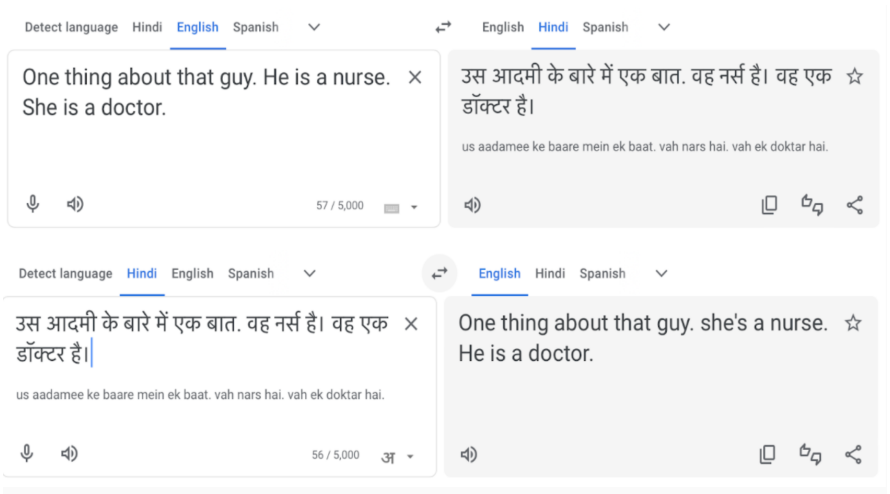

Figure 1.1: Revealing Bias: How Machine Translation Perpetuates Gender Stereotypes in Occupations (Hindi-English).

The potential for gender bias in MT systems has been highlighted in recent studies. Even prominent tools like Google Translate exhibit a tendency to default to masculine forms, especially in fields with perceived gender imbalances (as shown in Figure 1.1). This bias can emerge even when translating gender-neutral source texts, suggesting that the underlying algorithms and training data may inherently contain societal biases.

This research aims to systematically evaluate gender bias in MT systems, specifically focusing on the Hindi-to-English language pair. We will analyze translations of professional texts, examining the frequency with which certain professions are rendered

neutrally versus being assigned a male or female gender.Additionally, we are conducting an in-depth analysis of the connection between professions and gender. This will provide crucial insights into how MT systems handle gender-specific language in professional contexts, where biases can have significant consequences. Our research aims to develop more equitable and inclusive Hindi-to-English machine translation (MT) tools. By analyzing the limitations of current MT systems and applying bias mitigation strategies, we aim to create efficient translations that promote fairness, gender sensitivity, and respectful treatment for all users.

1.2 Motivation and objectives

The escalating reliance on machine translation (MT) systems across different sectors demands a thorough investigation into its capacity to perpetuate and exacerbate detrimental gender biases. Despite significant advancements in MT technology, effectively translating gender remains a formidable obstacle, frequently leading to the unintentional reinforcement of societal stereotypes and discriminatory behaviors. This has tangible implications, including the misrepresentation of individuals and the obstruction of initiatives aimed at achieving gender equality.

Language is a fascinating and complex skill we use daily, often without realizing its importance in communication. Understanding meanings in natural languages is not as simple as following the logical rules of programming languages. Natural language processing (NLP), a branch of artificial intelligence (AI), strives to make natural languages understandable to computers. Similarly, translating between different natural languages falls under the scope of machine translation (MT). Machine learning (ML) offers tools for analyzing data and building models by detecting patterns within the data. More specifically, deep learning (DL) leverages neural networks to improve learning tasks' performance compared to traditional statistical models, especially in sequence-to-sequence problems like translation tasks. Neural machine translation (NMT), a modern approach in MT, utilizes deep neural networks to learn patterns between the source and target languages for text translation.

A downside of models trained on human-generated corpora is that they can learn social biases present in the data. This is evident when training word embeddings, vector representations of words, with news sets and crowd-sourced evaluation to quantify biases such as gender bias in these representations. These biases can affect downstream applications and risk being amplified. The aim of this work is to examine gender bias in machine translation and explore the impact of debiasing such systems. While some previous studies have detected gender bias in MT, this study is among the first to propose debiasing techniques for this application. Therefore, defining an appropriate framework to evaluate these debiasing techniques is necessary.

There is a notable gap in understanding the specific nuances and challenges associated with the Hindi-to-English language pair. The unique contrast between Hindi's predominantly gender-neutral structure and English's grammatical gender system makes this language pair a compelling area for further investigation. By delving into the complexities of gender bias in this context, this research seeks to develop targeted strategies to mitigate these biases and create more equitable and inclusive MT tools. Ultimately, this work aspires to develop MT systems that not only facilitate effective cross-lingual communication but also actively promote fairness and gender sensitivity, thus contributing to a more inclusive digital landscape.

Chapter 2

LITERATURE REVIEW

For a thorough review of work done in machine translation system, we will start with some surveys conducted and delve into the results. Subsequently, we will explore recent studies and publications that address gender bias in machine translation across various domains. Finally, we will conclude this section by discussing the different methods and techniques employed to address this issue.

2.1 Introduction

Many studies have investigated gender bias in machine translation across various languages using diverse methodologies. Gender bias research has a long history, starting with investigations into biases related to motherhood and gender stereotypes in the workplace, and extending into almost all aspects of life. Nasrina Siddiqi discusses the daily challenges women face on social networks due to stereotypes. Bolukbasi[1] used principal component analysis to examine bias in textual data, while Kurita[2] assessed bias in BERT using the Word Embedding Association Test (WEAT) as a baseline, calculating the mean of the log probability bias score for each attribute. Recently, Gupta[3] evaluated gender bias in Hindi-English machine translation. However, there is limited research comparing gender bias across different domains.

2.2 Surveys and Case Studies: Gender Bias

To gain a better understanding of efforts in debiasing gender roles or addressing gender bias, we reviewed several surveys and case studies, learning about the current focus of recent research. Friedman research provides a comprehensive analysis of gender gaps in society by consolidating data from sources like the World Values Survey and tweets from 99 countries[4]. The study captures various statistics in politics, economy, and education and explores the correlations between linguistic gender bias and gender valuations. Bias, though it can refer to racial, religious, or gender-related issues, typically arises from stereotyping.

An older case study on Google Translate highlights stereotypical biases in gender roles that reflect real-world perceptions of men and women. Translating sentences using English as an intermediary, Google Translate shows an imbalance by associating stronger adjectives with men and weaker adjectives with women. Similarly, translations tend to add gendered pronouns in professions. For further details, see Prates[5].

A survey by Kalyan[6] examines the increasing use of transformer models in NLP

tasks, discussing pre-trained models and various techniques and architectures employed for bias mitigation. Another survey by Blodgett[7] conducts a review of publications on bias in natural language processing systems, classifying these papers according to their motivational tasks. It shows that the motivation and work on gender bias in NLP tasks are often ambiguous and inconsistent.

From these observations, we conclude that bias in MT systems stems from inherent bias in training data or real-world examples used. Although studies on linguistic gender bias detection and mitigation are increasing, the term itself remains unclear as part of the motivation in studies, leading to a disconnect. Most linguistic biases are studied only in English, causing varied gendered outputs when embeddings are created in MT models from one language to another.

2.3 Recent Studies

Several advanced transformer language models are emerging, excelling on benchmark NLP tasks and achieving better translations, leading to improved embeddings for bias mitigation. Reviewing the surveys raises questions: Is gender bias domain-specific? Does the stereotype shift with the domain? Domains can refer to professions, industries, education, etc. For example, a national survey might link women with the role of 'homemaker,' while the hospitality or wellness industry might describe women with adjectives like 'strong' and 'driven.' Which domains are most impacted by gender role bias? Can language models translating highly gendered languages exacerbate bias if embeddings remain unchecked? Is debiasing embeddings the only strategy for bias mitigation? Is evaluating domain-specific biases more effective? These questions warrant further exploration.

Dacon and Liu[8] investigate 'Does Gender matter in News?' revealing significant representation disparities in news articles. Their experiments captured bias based on power, influence, career, and family. Predictably, adjectives associated with men related to influence and decisions, while 'female' resonated with family, home, and weddings. Fu[9] used a game language model to quantify gender bias in sports journalism, concluding that higher-ranked male athletes were asked more game-related questions than their female counterparts. Analyzing interviews with 167 male and 143 female tennis players, they highlighted stark bias in press conference queries. Another study on women in sports journalism observed the marginalization of females in certain countries, surveying journalists from over 700 newspapers in Australia, the UK, and the US. They found less coverage for women's sports, acknowledging a significant gender disparity. Observations included a low percentage of female journalists (just over 5%) and a higher number of articles about men in sports compared to women. In articles covering both genders, women were more often discussed in domestic roles.

Current research focuses on whether fairness measurement can adapt to domains, highlighting misrepresented or underrepresented categories, adjectives, and stereotypes specific to domains. A study on Wikipedia corpora used the WEAT metric to evaluate bias across domain embeddings in English texts, identifying bias categories using cluster embeddings (Chaloner and Maldonado[10]). Further research by Saunders and Byrne[11] addresses gender bias as a domain-specific MT problem, comparing bias in eight languages using a transformer model, with English as the base. Entertainment also showcases gender bias. In a video (Johnson[12]), actress Scarlett Johansson discusses the types of questions she faces compared to her male counterparts while filming Marvel's Avengers, illustrating that even when portraying strong female characters, the more "interesting" questions are

directed towards male actors.

Having understood the domain research, we now turn to gender fairness issues in language translation. It is crucial to address these issues for several reasons: Most language research has been conducted in English. How do we measure bias in highly gendered languages like Hindi and Spanish? What about less widely spoken languages or older languages needing preservation, like Sanskrit? Once detected, can bias be controlled in these languages with appropriate models? A survey on Arabic highlights similar challenges in machine translation. According to Darwish[13], dialectal variations in Arabic present challenges for many NLP applications, often necessitating rule-based systems. Efforts have been made to detect gender bias in Hindi across various domains. Kapoor, Bhuptani, and Agneswaran[14] used the Bechdel test on Hindi movies to identify gendered content and biases. Madaan[15] studied similar biases and stereotypes in Hindi films and went further by constructing knowledge graphs to mitigate these biases. Pujari[16] attempted to debias Hindi using an SVM classifier. A recent paper by Gupta, Ramesh, and Singh[3] addresses de-biasing gendered words in Hindi, discussing the gendered nature of the language, which can cause embeddings to detect biased translations when using a non-gendered or low-order language like English.

Additionally, various model architectures have been employed to analyze and mitigate gender bias in language. Sutskevar[17] utilized LSTM on the WMT'14 dataset for English to French, experimenting with sequence-to-sequence learning models and achieving improved results for longer sentences. Vaswani[18] introduced the transformer, an attention-based model that enabled inputs to interact over longer sequences with parallelization, reducing training time. In 2020, Dinan[19] discussed using multiple classifiers for male, female, and neutral gender categories with a pretrained transformer model by Vaswani[18], employing a bi-encoder architecture trained with cross-entropy to rank the classes.

Detection and mitigation techniques include Gonen and Webster[20] introducing a novel approach to detect gender bias using perturbations with BERT, which automatically detected gender differences when translating sentences in gendered languages. Wong[21] suggested altering the data by introducing false and/or substitute data, conducting experiments on datasets of English and Spanish to observe how data augmentation impacts gender bias and BLEU score, finding that data augmentation could mitigate gender bias to some extent. Word embeddings remain a popular method to detect gender bias, with debiasing them being one approach to mitigate these biases.

We explored various studies and papers to comprehend the prevalence of gender bias across different domains and languages. We posed inquiries about the potential challenges in language translation and examined some solutions. Specifically, we delved into research concerning gender biases and stereotypes in Hindi. Subsequently, we examined recent studies aiming to mitigate biases in gendered languages like Hindi. Following this, we delved into recent language models and methods for bias mitigation, ranging from SVMs and classifiers to transformer architectures and word embeddings. We compared several approaches and highlighted metrics more suited for gender-word ratio detection.

Chapter 3

METHODOLOGY

Detection and mitigation of gender bias in machine translation for Hindi to English is the main objective of this thesis.

3.1 Neural Machine Translation: An Overview

Neural network and word embedding concepts are introduced in this section. The attention mechanism, recurrent neural network (RNN) encoder-decoder, and the Transformer architecture for machine translation techniques are explored. Common metrics for evaluation are also discussed.

3.1.1 Neural Networks (NN)

Prior to exploring neural machine translation (NMT) models, it is essential to establish a foundation in the concepts of neural networks and word embeddings, as they are fundamental components of numerous natural language processing tasks, including NMT. Neural networks, drawing inspiration from the neuronal structure of the human brain, comprise interconnected computational units that process multiple inputs to generate a singular output, which can subsequently be relayed as input to other units within the network. These networks represent a class of machine learning algorithms designed to predict outputs based on multiple input variables. Unlike linear models that seek to identify linear relationships between features, neural networks are characterized by their non-linear nature, affording them greater flexibility and adaptability. By incorporating multiple layers of computational units, neural networks can effectively model and potentially solve an extensive array of complex problems.

A feedforward neural network with a single hidden layer is characterized by the following elements:

- Input vector: $\mathbf{x} = (x_1, x_2, ..., x_n)$
- Hidden vector: $\mathbf{h} = (h_1, h_2, ..., h_m)$
- Output vector: $\mathbf{y} = (y_1, y_2, ..., y_l)$
- Weight matrix W: Connects input nodes to hidden nodes.
- Weight matrix U: Connects hidden nodes to output nodes.

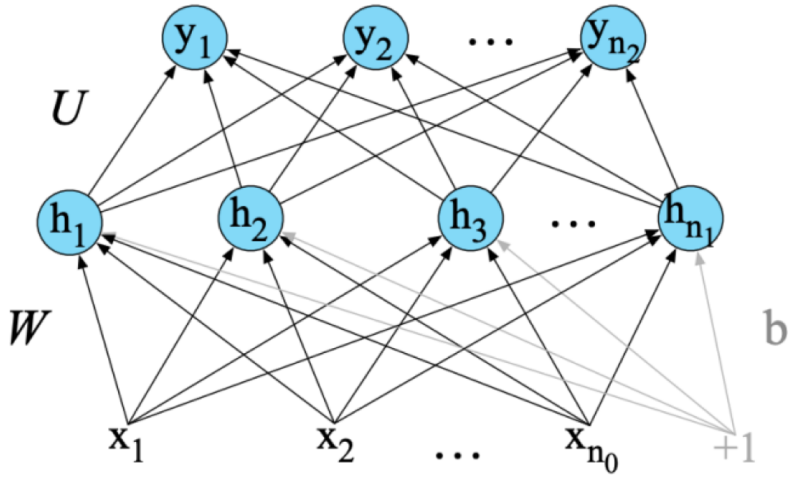

Figure 3.1: Feedforward Neural Network

In a typical neural network architecture, each node in a layer is fully interconnected with all nodes in the preceding and subsequent layers. Bias units, which are beneficial when all input values are zero, can be incorporated into the input layer and any hidden layers.

Eq. 1 illustrates the calculation of each hidden node's output. This is achieved by multiplying the input vector $\mathbf{x}$ by the weight matrix W, incorporating a bias term $\mathbf{b}$, and applying an activation function g (such as sigmoid, tanh, or ReLU) to yield the hidden output $\mathbf{h}$:

$$\mathbf{h} = g(W\mathbf{x} + \mathbf{b}) \quad \text{(Eq. 1)}$$

Subsequently, for each output node, an intermediate output $\mathbf{z}$ is obtained by multiplying the hidden vector $\mathbf{h}$ with the weight matrix U (Eq. 2). This intermediate output is then transformed into a probability distribution $\mathbf{y}$ through the application of the softmax function (Eq. 3), ensuring that all output values are within the range of 0 to 1 and their sum equals 1:

$$\mathbf{z} = U\mathbf{h} \quad \text{(Eq. 2)}$$

$$\mathbf{y} = \text{softmax}(\mathbf{z}) \quad \text{(Eq. 3)}$$

In neural networks, the softmax function plays a vital role in converting the intermediate output into a format suitable for representing the likelihood of different outcomes in classification tasks. By normalizing the values, it allows the model to produce a probability distribution that reflects the confidence or certainty associated with each possible class. However,architectures themselves can also be enhanced for better performance. Deep learning utilizes deeper networks created by stacking multiple hidden layers together in neural architectures. In natural language processing (NLP), besides feedforward neural networks, other neural frameworks such as RNN(Recurrent neural networks), LSTM(Long Short-Term Memory) networks, and CNN(Convolutional neural networks) are frequently employed. Notably, the encoder-decoder architecture,particularly with recurrent neural networks, is commonly used in machine translation.

3.1.2 Word Embeddings in Natural Language Processing

The foundation of word embeddings, and vector semantics in general, is the distributional hypothesis in linguistics, which proposes that words used and occurring in similar contexts tend to possess similar meanings. This correlation between distributional and semantic similarities is leveraged to map words to high-dimensional vectors. For instance, semantically related words like "dog" and "cat," as well as morphologically related words like "dog" and "dogs," often appear in similar contexts and thus would exhibit similar representations of vector, being positioned close to each other in the vector. Assigning a numeric vector to every word in a dictionary allows for easy quantification of similarity between words using distance functions like the cosine similarity metric.

One method for constructing these word vectors involves creating co-occurrence matrices (either term-document or term-term) based on word frequencies within a document collection. In a term-document matrix, each word is represented by a vector of length d (the size of the document collection), whereas in a term-term matrix, the vector length is $|V|$ (the vocabulary size). Algorithms such as PPMI (Positive Pointwise Mutual Information) and tf-idf (term frequency-inverse document frequency) are subsequently employed to transform raw counts into word vectors. However, due to the infrequent co-occurrence of words with all other words, co-occurrence matrices often yield long and sparse vectors (mostly 0). To address this, the word2vec software package was introduced, facilitating the creation of short, dense vectors known as embeddings, which offer a more efficient and effective way to capture semantic relationships between words.

Figure 3.2: Two-dimensional Visualization of Word Embeddings

The word2vec algorithm, instead of calculating word frequency, concentrates on predicting word co-occurrences. The skip-gram model, when combined with negative sampling, exemplifies this approach. Within a text, words in the context window of a given word are considered positive examples of environments in which the target word occurs. For each positive target-context pair, negative sampling algorithms randomly select k

words from the vocabulary that do not co-occur with the target word to generate negative examples.

A binary classifier is then trained using logistic regression to differentiate between positive and negative examples derived from a corpus. The loss function applicable to a single target word w is illustrated in Eq. 4:

$$L = -[\log P(+|w, c) + \sum_{i=1}^{k} \log(1 - P(-|w, n_i))] \quad (Eq.4),$$

In Eq. 4, the term (w, c) signifies a target-context word pair drawn from the positive examples, k denotes the quantity of negative samples generated for each positive sample, and n_i represents an individual negative sample. The weights learned during the training process ultimately serve as word embeddings.

Word2vec embeddings are limited by their static representation of each word, leading to inadequate modeling of words with multiple meanings (polysemy) in different contexts. Contextualized embeddings like BERT and ELMo address this by generating dynamic word representations.

ELMo pre-trains a bidirectional LSTM model on a large text corpus, predicting both the preceding and following words. It learns a task-specific function of hidden states from the entire input sentence, producing contextualized word representations. This approach allows ELMo to capture nuances in word meaning across contexts.

BERT, using a stacked Transformer architecture with self-attention, predicts words based on their surrounding context. Similar to ELMo, BERT is pre-trained on a large corpus and then fine-tuned for specific tasks. Due to their superior performance, neural machine translation (NMT) often favors dynamic contextualized embeddings like ELMo and BERT for encoding and decoding words.

3.1.3 Encoder-Decoder Models for Neural Machine Translation

At the sentence level, machine translation involves transforming a sequence of tokens from a source language into a corresponding sequence in the target language. The encoder-decoder network, a prominent architecture within the broader class of sequence-to-sequence neural network models, is frequently employed for this task.

This framework comprises three fundamental components:

Component	**Description**	**Input**	**Output**
Encoder	Processes an input sequence of tokens from the source language and encodes them into contextualized representations.	$x = (x_1, ..., x_n)$	$h = (he_1, ..., he_n)$
Context Vector	Derived from all encoder hidden states, it encapsulates the semantic essence of the source text.	$he_1, ..., he_n$	c
Decoder	Utilizes the context vector to generate a variable-length sequence of vector representations, which are then converted into tokens in the target language.	c	$y = (y_1, ..., y_m)$

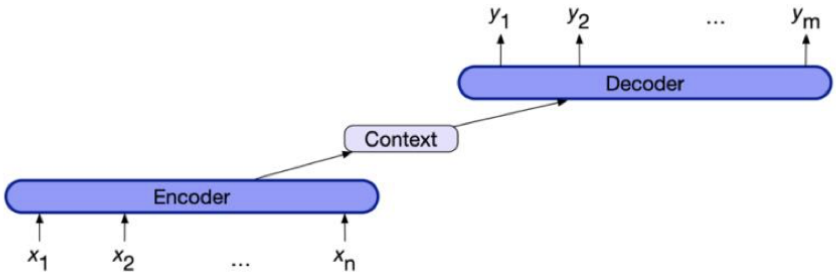

Figure 3.3: Encoder-Decoder Model Architecture: Contextualization and Generation Process

While specific implementations of encoder-decoder models may vary in their underlying language models (e.g., employing different types of neural networks like RNN, LSTM, or CNN), the fundamental architecture remains consistent across different variations.

Encoder-decoder models are trained on a parallel corpus containing aligned sentence pairs in the source and target languages. These sentence pairs are combined with a special token acting as a separator. The model's encoder and decoder components are trained in tandem to maximize the conditional log-likelihood of generating the target sentence given the source sentence.

In order to translate an input sentence, the algorithm iteratively predicts an output word at each decoding step. This is achieved by initially calculating the probability distribution across all potential outputs (i.e., every word within the vocabulary). While a greedy approach might always select the word with the highest probability at each step, this locally optimal decision may not result in the optimal translation for the entire sentence. The most accurate translation could include words that, at first glance, appear less likely.

Beam search offers an alternative to the greedy method in selecting the best translation. Instead of choosing a single most likely token at each step, beam search maintains a list of k potential translations, where k represents the width of the search beam. The algorithm then selects the top k most probable words based on the probability distribution for the next position in the sequence. To explore these potential translations, the algorithm employs separate decoders to expand each of the k hypotheses, resulting in $k \times |V|$ total candidates, where $|V|$ denotes the vocabulary size. The probability of each candidate is then assessed based on the conditional probability $P(y_i|x, y_{<i})$, considering the input sequence and the previously generated words. Finally, the search space is pruned by keeping only the k highest-scoring hypotheses, ensuring that the number of potential translations under consideration remains manageable.

The decoding process concludes when a termination symbol is generated for each of the k potential translations. Following this, for each hypothesis y a score is calculated using Eq. 5. Importantly, this score is adjusted by normalizing the negative log probability with the length of y, thus mitigating the inherent preference for shorter sequences in the raw probability. Depending on the specific application, the output will either be the highest-scoring translation or a collection of the top-scoring translations.

$$\text{score}(y) = -\log P(y|x) = \sum_{i=1}^{t} \log P(y_i|y_1, \ldots, y_{i-1}, x) (Eq.5)$$

A recurrent neural network (RNN) is a popular choice for the underlying model of both encoder and decoder in neural machine translation (NMT) due to its ability to handle sequential data and capture long-distance dependencies. An RNN-based encoder-decoder framework typically consists of two RNNs, one for encoding the source language input and another for decoding into the target language.

As the encoder RNN processes the input sequence $x = (x_1, x_2, ..., x_n)$, where each x_i is a word embedding, it maintains a hidden state vector h^e that is continually updated according to the equation:

$$h_t^e = f(h_{t-1}^e, x_t) \qquad (Eq.6)$$

Here, f is a non-linear activation function (e.g., sigmoid or LSTM), and h_{t-1}^e represents the cumulative hidden state from the previous time step. Upon reaching the end-of-sentence symbol, the final hidden state of the encoder, known as the context vector c, encapsulates the contextual information of the entire input sentence.

This context vector is then passed to the decoder RNN, which initializes its hidden state h_0^d with c. At each time step t in the decoding phase, the decoder updates its hidden state using:

$$h_t^d = f(h_{t-1}^d, y_{t-1}, c) \qquad (Eq.7)$$

and computes a probability distribution using:

$$P(y_t|y_{t-1}, y_{t-2}, ..., y_1, c) = g(h_t^d, y_{t-1}, c) \qquad (Eq.8)$$

where g is an activation function like softmax. This probability distribution is used to generate the output token. The context vector c remains a parameter to both f and g throughout the decoding process, ensuring its influence is not diluted. The decoding process continues until the end-of-sentence token is generated.

3.1.4 Attention Mechanism

Recurrent Neural Networks (RNNs) are a popular choice for encoding and decoding in machine translation due to their ability to handle sequential data and capture dependencies between distant words in a sentence. A typical RNN-based encoder-decoder model uses two RNNs: one to encode the source language input into a context vector and another to decode this vector into the target language.

The encoder RNN reads the input sequence word by word, updating its hidden state at each step. This final hidden state, called the context vector, summarizes the meaning of the entire input sentence. The decoder RNN then uses this context vector to initialize its hidden state and generate the target language output word by word.

While RNNs are effective, they can struggle to represent the beginning of long sentences due to the fixed-length context vector. This is because the context vector is typically derived solely from the final hidden state of the encoder. This issue has led to the development of attention mechanisms, which we will discuss in the next section.

To derive the context matrix c_t at each time step t, the decoder calculates an attention vector α_t, which quantifies the relevance of each encoder hidden state, ensuring $c_t = H^e\alpha_t$. This process involves the following steps:

1. **Similarity Assessment:** For each encoder hidden state h_i^e, its similarity with the preceding decoder hidden state h_{t-1}^d is evaluated by computing an attention score $score(h_{t-1}^d, h_i^e)$.
    - The scoring function can be a simple dot product between h_i^e and h_{t-1}^d, provided they share the same dimensions.
    - Alternatively, a more complex scoring function like the bilinear function can be employed, incorporating a learnable parameter W_s to enhance expressiveness and accommodate vectors of differing dimensions.
2. **Normalization:** Having computed a score for each vector in H^e, these scores are normalized into a probability distribution using the softmax function. This yields the proportional relevance of h_i^e to h_t^d:

$$\alpha_i^t = \text{softmax}(score(h_{t-1}^d, h_i^e) \forall i \in e) = \frac{\exp(score(h_{t-1}^d, h_i^e))}{\sum_k \exp(score(h_{t-1}^d, h_k^e))} \qquad (Eq.9)$$

3. **Weighted Average Computation:** The weighted average of all encoder hidden states is calculated, with the weights determined by the attention vector α_t. This weighted average then serves as the context vector for the current time step t:

$$c_t = \sum_i a_i^t h_i^e \qquad (Eq.10)$$

3.1.5 Transformer

The requirement for recurrent neural networks to process the entire input sequentially to generate hidden states is a limitation of RNN-based encoder-decoder models. This method is not only time and memory-intensive but can also result in the loss of crucial information over long sequences of recurrent connections. A solution is provided by the Transformer model, introduced in 2017, using a non-recurrent, highly parallelized architecture relying solely on the attention mechanism. When first introduced, this approach not only increased efficiency in terms of time and space but also set a new benchmark for BLEU scores. The Transformer employs self-attention within its encoder and decoder, in addition to incorporating components such as feedforward neural networks and the encoder-decoder attention mechanism.

The relationship between a target position and other positions within the same sequence is calculated by self-attention, unlike recurrent methods, to create a representation of the sequence. The model, by focusing on relevant contextual information for each input token, can better capture long-distance dependencies and provide a more refined representation. The self-attention mechanism can efficiently perform information extraction and inference for large contexts using parallel computation because the computations for each position are independent of others.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Eq. 11)}$$

The association of each input word's embedding with three weight matrices—Query, Key, and Value vectors—facilitates learning during training. Using the formula of Eq. 11,

the outputs of a self-attention layer can be computed, where d_k represents the dimension of the query and key vectors.

$$\text{MultiHead}(Q, K, V) = W^O(head_1 \oplus head_2 ... \oplus head_h) \quad \text{(Eq. 12)}$$

Multi-head self-attention, consisting of multiple self-attention layers operating in parallel, is used by the Transformer to capture both local context and long-range dependencies. Learning different contextual aspects is the focus of these layers, each having distinct parameters. Using the formula of Eq. 12, the outputs from these parallel layers are concatenated and then reduced to the original output dimension for the subsequent layer, where head_i is the output of the i-th head (computed with Eq. 11) and W^O is the weight matrix projecting the concatenated output to the original dimension.

Additional feedforward layers, residual connections, and normalization layers are then passed through by the self-attention layer outputs. The model can be built by stacking these components, forming a transformer block. Positional encoding, used by the Transformer to integrate information about the relative positions of words with their embeddings, is necessary since the order of tokens is not inherently preserved, and this information is then passed through the transformer blocks.

Attention mechanisms between its encoder and decoder are used by the Transformer, similar to RNN encoder-decoder models, to focus on relevant parts of the input sequence. Stacked self-attention layers, along with residual connections, layer normalization, and feedforward networks, are also utilized by the decoder.

3.2 Assessing Gender Bias in Machine Translation

3.2.1 Dataset

The dataset utilized in this research consists of 3888 sentences, sourced from the `https://github.com/argentina-res/gender` and generated using Large Language Models (LLMs) [ChatGPT]. Each sentence is categorized as either male (1821 sentences), female (1814 sentences), or neutral (253 sentences), ensuring a balanced distribution of gender representations. This equitable dataset is crucial for an unbiased evaluation of gender bias, as each sentence is associated with only one gender. Furthermore, the dataset encompasses Hindi translations of these sentences, each retaining its corresponding gender label.

3.2.2 Description of the MT Systems

Machine translation is an automated process that leverages statistical or neural machine learning models to transform text from one language to another. Two prominent examples are:

1. **Google Translator (GT):** Introduced in 2003 as a statistical MT system and later transitioning to a neural one in 2016, GT provides sentence-level translations. Since 2018, GT has offered alternative translations for ambiguous or under-specified English words in some languages, with male-female alternatives. However, this feature is currently unavailable for Hindi.

2. **Bing Microsoft Translator (BMT):** Owned by Microsoft, BMT is another machine translation system that initially used a statistical approach before shifting to

a neural one. Unlike GT, BMT does not present alternative translations within the translation box itself.

Following the generation of translations using both GT and BMT, an analysis was performed to ascertain the frequency of female, male, and neutral forms in the translated text. The investigation specifically targeted gendered terms such as personality adjectives, professions, and nouns.

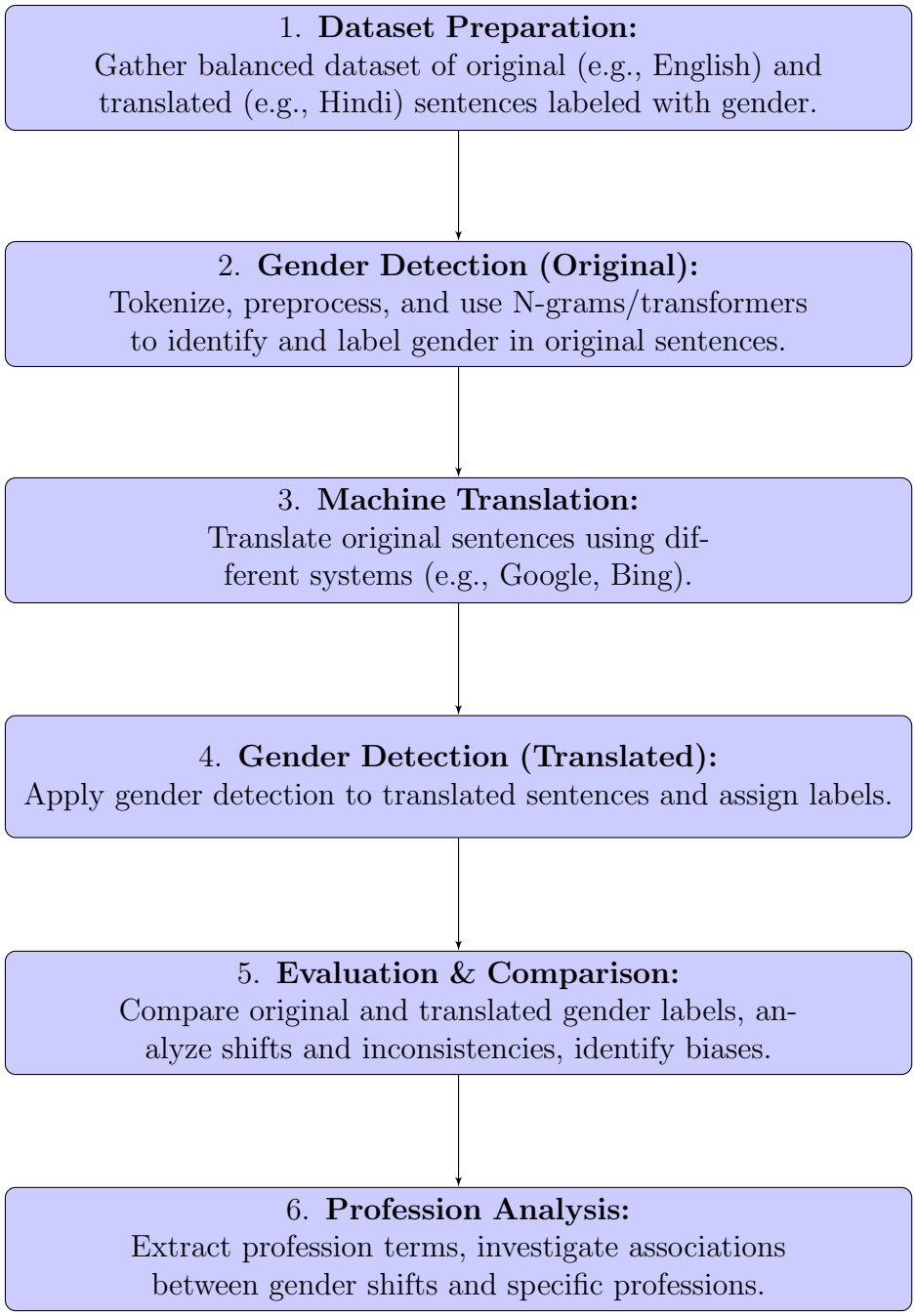

Figure 3.4: Methodology for analyzing gender bias in machine translation.

3.2.3 Bias Statement

Bias is the inclination to favor or discriminate against particular groups. In the realm of machine translation, bias can manifest as unjust or erroneous portrayals of diverse genders or identities. Biased translations have the potential to perpetuate harmful stereotypes and marginalize certain groups, excluding them from meaningful discourse. Tackling bias in machine translation is paramount for fostering more inclusive and equitable representations that serve all individuals, while promoting diversity and respect.

3.2.4 Detection of gender

Identifying gender within a text can be achieved through the utilization of N-grams (Unigrams, Bi-grams, Tri-grams) and Transformers. However, this method may yield similar results for both the original English sentences and their translations, as the sentences contain only one gender. The process commences with the tokenization of text during preprocessing, where it is segmented into individual words (tokens) for both the English and translated versions. Subsequently, all tokens are converted to lowercase to ensure case-insensitivity in the analysis. This stage may also involve stemming or lemmatization, techniques employed to standardize words to their base form.

The gender detection algorithm then identifies words indicative of gender, such as "he," "his," "him," and "himself" for male, and "she," "her," "hers," and "herself" for female, cross-referencing them with predefined gender-specific lists. For the neutral gender, terms like "they," "we," "our," and "themselves" are considered. Based on the matches found in these lists, a gender label is assigned, categorizing the noun as "male," "female," or "neutral."

3.2.5 Evaluation

To assess any alterations in gender representation within the translated data, the original text is contrasted with the translations produced by Google Translator and Microsoft Bing Translator. Through various comparisons, such as Original vs. Google Translator, Original vs. Bing Microsoft Translator, and Original vs. both translators, patterns and relationships between professions and gender shifts during translation are identified. The Natural Language Toolkit (NLTK) is employed to extract profession-related terms from the dataset. This analysis aims to shed light on how machine translation systems handle gender-specific language and its implications for the portrayal of professions in translated text.

3.3 Mitigation Techniques for Gender Bias

In their 2019 study, "Evaluating Gender Bias in Machine Translation," Stanovsky[22] introduced WinoMT, the first challenge set for examining gender bias in MT systems. WinoMT is a synthetic English dataset with sentences featuring non-stereotypical gender roles (e.g., "One thing about that guy. He is a nurse. She is a doctor."). Their findings revealed that popular MT systems exhibit significant gender bias across multiple languages. Specifically, when translating from English to gendered languages, these systems often incorrectly assign female gender to traditionally male-associated professions like "doctor."

To delve deeper into the vulnerability of MT models to bias, Stanovsky[22] employed a "fighting bias with bias" strategy. This approach involved adding gender-associated adjectives (e.g., "handsome," "pretty") to occupation words in the WinoMT dataset. In the previous example, the sentence would be altered to "One thing about that guy. She is a nurse. He is a doctor." While this technique enhanced translation accuracy in certain languages (such as Russian, which experienced an 11.2% accuracy improvement), it has its limitations. It is challenging to apply this method broadly to other contexts and it depends on a precise coreference system to correctly resolve pronouns.

Beyond this experiment, the field of machine translation has been investigating various methods to address gender bias. While traditional methods typically focus on reducing bias in word embeddings and other pre-trained models, recent research has been exploring a potentially more efficient approach that involves fine-tuning MT models rather than fully retraining them.

3.3.1 Word Embedding

In their seminal 2019 paper, "Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques," Escudé Font and Costa-jussà[23] explored the potential of leveraging word embeddings to mitigate gender bias in neural machine translation (NMT). Their methodology involved integrating variations of Global Vectors (GloVe) embeddings into the encoder and decoder components of an OpenNMT Transformer architecture. Specifically, they experimented with the original GloVe embeddings, hard-debiased GloVe embeddings (created through a debiasing process), and Gender-Neutral GloVe (GN-GloVe) embeddings.

The researchers trained their models on a corpus of over 16 million English-Spanish sentence pairs derived from diverse sources such as the United Nations, Europarl, CommonCrawl, and the Workshop on Machine Translation (WMT) datasets. To assess the effectiveness of the different embedding variations, they used the standard newstest2013 benchmark, a set of 3,000 sentences provided by WMT. In addition, to evaluate the impact on gender bias, they developed a custom test set focusing on the translation of professions and pronouns in non-stereotypical contexts.

The results revealed that the model employing GN-GloVe embeddings in both the encoder and decoder outperformed the baseline model (without pre-trained embeddings) by 0.98 BLEU points on the newstest2013 benchmark. Notably, the use of hard-debiased GloVe embeddings in both encoder and decoder led to the most significant reduction in gender bias across various scenarios. This was particularly evident in the translation of occupational terms in feminine contexts, especially for technical roles like "criminal investigator," "heating mechanic," and "refrigeration mechanic."

However, like previous work in this domain, the study's scope was limited to professional occupations and the English-Spanish language pair, which may hinder the generalizability of the findings to other domains or language pairs with different linguistic properties. Nonetheless, the research provides valuable insights into the potential of debiased word embeddings as a promising avenue for mitigating gender bias in machine translation systems.

3.3.2 Domain Adaptation Techniques

In their 2020 study, Saunders and Byrne[11] introduced an innovative method for mitigating gender bias in machine translation (MT) by reframing it as a domain adaptation problem. They proposed fine-tuning existing models on a small, gender-balanced dataset and a counterfactual set with reversed gender roles, rather than relying on synthetic data or debiased word embeddings. This approach proved more efficient and less disruptive than retraining models from scratch. By learning from both balanced and intentionally unbalanced examples, the model developed a more nuanced understanding of gender representation, leading to improved performance in mitigating gender bias.

To create their datasets, the researchers leveraged a corpus of sentences featuring occupations and pronouns to identify potential biases. A handcrafted set of 388 sentences, based on the template "The [Profession] finished his, her work," was manually translated into German, Spanish, and Hebrew to ensure diverse linguistic representation. This targeted approach allowed them to focus on specific areas where bias might manifest.

For training the general-purpose models, large bilingual corpora from sources like WMT19, the United Nations, and TED talks were employed. These datasets, although containing a degree of gender bias, provided a broader linguistic context for the fine-tuning process. The researchers then meticulously compared the outcomes of fine-tuning on both the counterfactual and handcrafted profession data with the baseline results from the WinoMT study, a seminal work in gender bias evaluation.

Remarkably, fine-tuning on the occupation-focused handcrafted set proved more effective than the counterfactual set, despite its limited size. This targeted approach not only required significantly less computational resources but also yielded substantial improvements in gender-related metrics. While a slight decrease in the BLEU score, a measure of general translation quality, was observed, this was effectively mitigated by incorporating regularized training and lattice rescoring techniques.

Saunders and Byrne's[11] research significantly contributes to the ongoing efforts to address gender bias in machine translation. Their findings highlight the potential of domain adaptation and fine-tuning as efficient and effective strategies. By focusing on a smaller, carefully curated dataset, they were able to achieve substantial bias reduction without compromising overall translation quality, paving the way for more equitable and inclusive MT systems.

3.3.3 Cross-Lingual Pivoting Approach to Addressing Gender Bias in Machine Translation

The cross-lingual pivoting technique has emerged as a promising approach in various natural language processing (NLP) tasks, leveraging the power of a shared "pivot" language to bridge linguistic gaps and improve performance across different languages. **Applications of Cross-Lingual Pivoting Technique:**

- **Machine Translation (MT):** Webster and Pitler (2020) introduced a novel approach to mitigate gender bias in MT by generating gender labels through cross-lingual pivoting. They aligned English and non-English Wikipedia pages, identifying corresponding sentences and labeling ambiguous pronouns based on gendered counterparts in English. This led to the creation of a large, gender-balanced dataset used to fine-tune a BERT language model, enhancing its ability to predict pronoun gender.

- **Cross-lingual Instruction Tuning:** Zhang[24] proposed using a high-resource language as a pivot to enhance instruction tuning in low-resource languages. The model was trained to process instructions in the pivot language and generate responses in the target language, improving instruction-following abilities in low-resource languages.

- **Cross-lingual Image Captioning:** The paper "CROSS2STRA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment" (ACL Anthology, 2023) proposed a framework to address the challenges of cross-lingual image captioning in unpaired settings. The approach involved two steps: image-to-pivot captioning and pivot-to-target translation, aligning the two subtasks using the pivot language.

- **Cross-lingual Entity Linking:** In "Pivot-based Candidate Retrieval for Cross-lingual Entity Linking," researchers tackled finding referents in a target-language knowledge base for mentions in a source-language text. They proposed a pivot-based approach using resources in closely related languages to improve performance in low-resource languages.

Cross-lingual pivoting has emerged as a versatile technique with applications in various NLP tasks. By leveraging the shared knowledge and resources of a pivot language, researchers have been able to overcome language barriers, improve model performance, and enhance the capabilities of NLP systems across different languages and modalities.

Chapter 4

RESULTS and DISCUSSION

The following section is devoted to showcasing the outcomes and subsequent assessment of the research. An in-depth examination and illustrative instances will be offered. Notably, all evaluations were carried out in April 2024.

Data Gender	Value Count
Male	1821
Female	1814
Neutral	253

Table 4.1: Gender Distribution in Dataset

Table 4.1 visually depicts the frequency of each gender, facilitating effortless comparison and examination of the gender composition within the dataset.

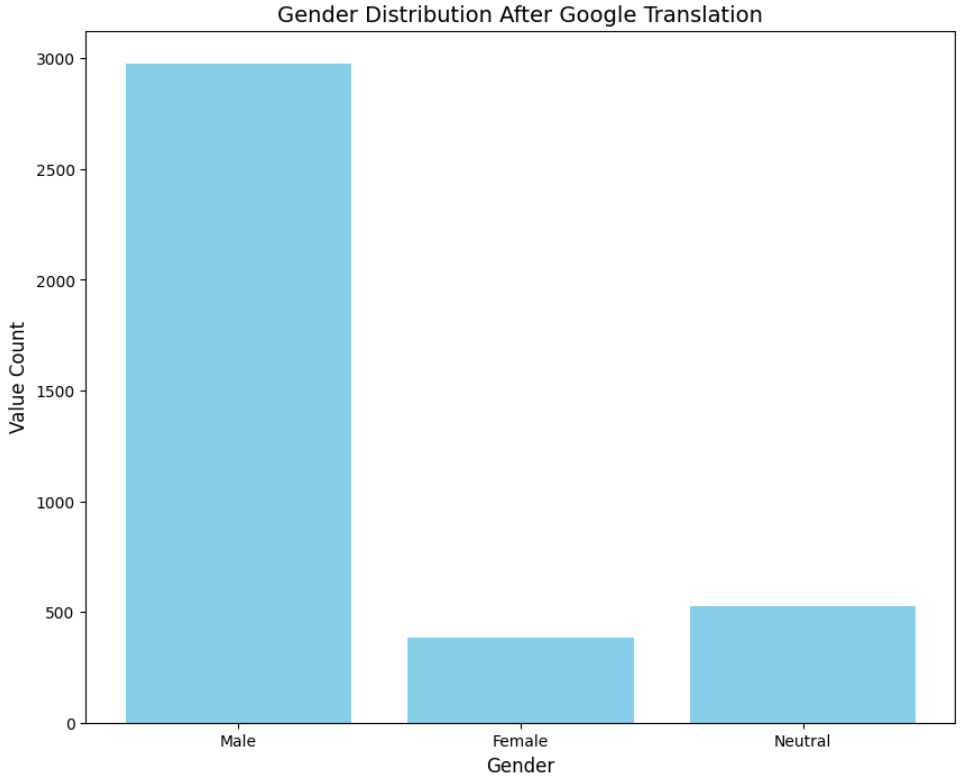

Figure 4.1: Gender Distribution after Google Translation

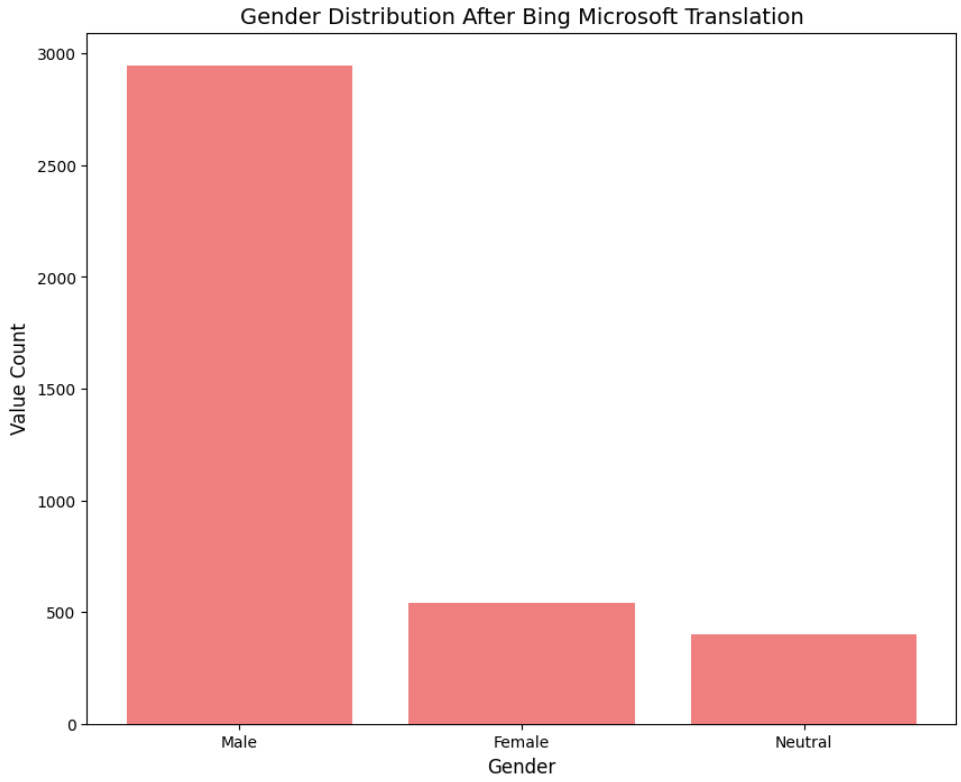

Figure 4.2: Gender Distribution after Bing Microsoft Translation

Figure 4.1 illustrates the distribution of gender labels assigned by Google Translate across the translated sentences. This offers a glimpse into how the algorithm interprets and classifies data related to gender. Contrasting this distribution with that of the original dataset allows us to identify potential inconsistencies or agreements in gender assignment between the source text and its Google-translated counterpart. Figure 4.2 depicts the gender distribution of sentences following translation by Bing Microsoft Translator.

Figure 4.3 illustrates the gender distribution across sentences in both the original dataset and the translated version using Google Translator. The table provides counts for various gender combinations. MF indicates instances where the original gender was male but was classified as female after translation. This comparison highlights any changes in gender categorization between the original and translated texts, shedding light on the translation's impact on gender assignment. Google Translator often fails to accurately translate gender-specific terms, leading to misclassification. For instance, neutral roles may be incorrectly translated as predominantly male or vice versa. Google Translator may reinforce gender biases and stereotypes by assigning incorrect genders to certain professions in specific languages. For example, it might assign a male gender to gender-neutral professions like Mechanic, Developer, Doctor, and Guard in some English contexts. Conversely, it might assign a female gender to professions like Nurse, Veterinarian, and Teacher when translating from Hindi to English, where gendered pronouns are common. Figures 4.3 and 4.4 display the gender distribution in the original dataset and the translated versions using both Bing Microsoft Translator and Google Translator.

Figure 4.5 illustrates the gender distribution and comparison among the original dataset, the Google translated dataset, and the Bing Microsoft Translator (BMT) translated dataset. Each row corresponds to a unique combination of genders, with accompa-

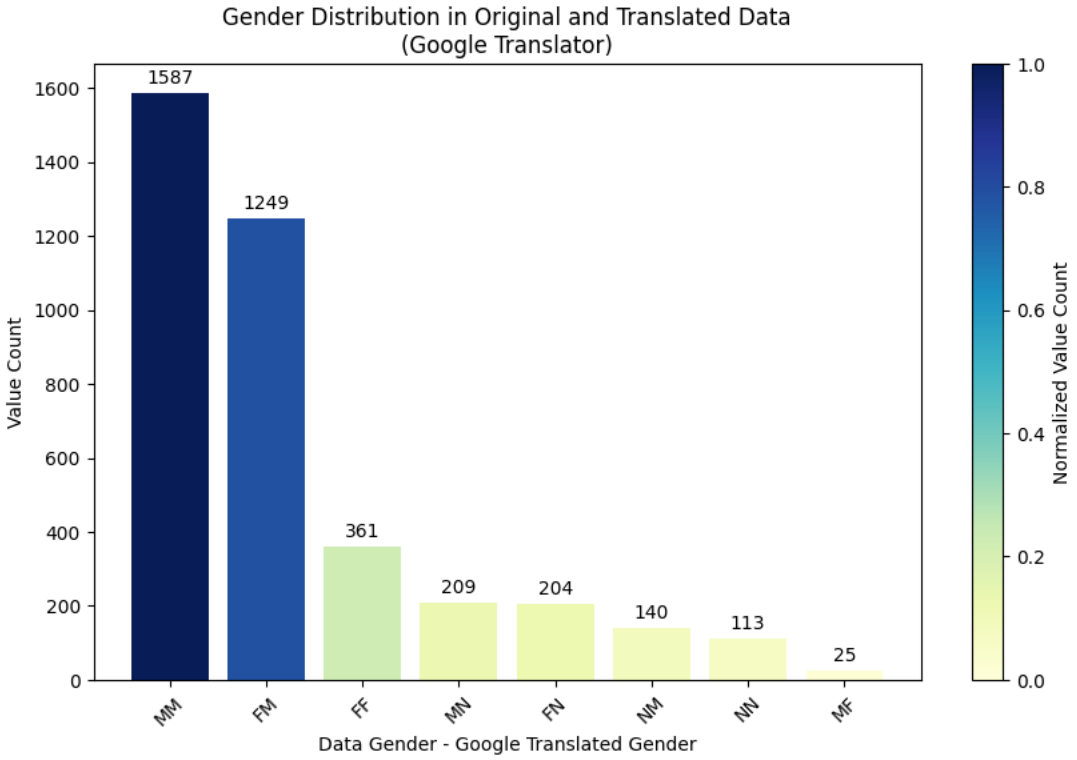

Figure 4.3: Comparison of Gender Representation in Original Text and Google Translated Text.

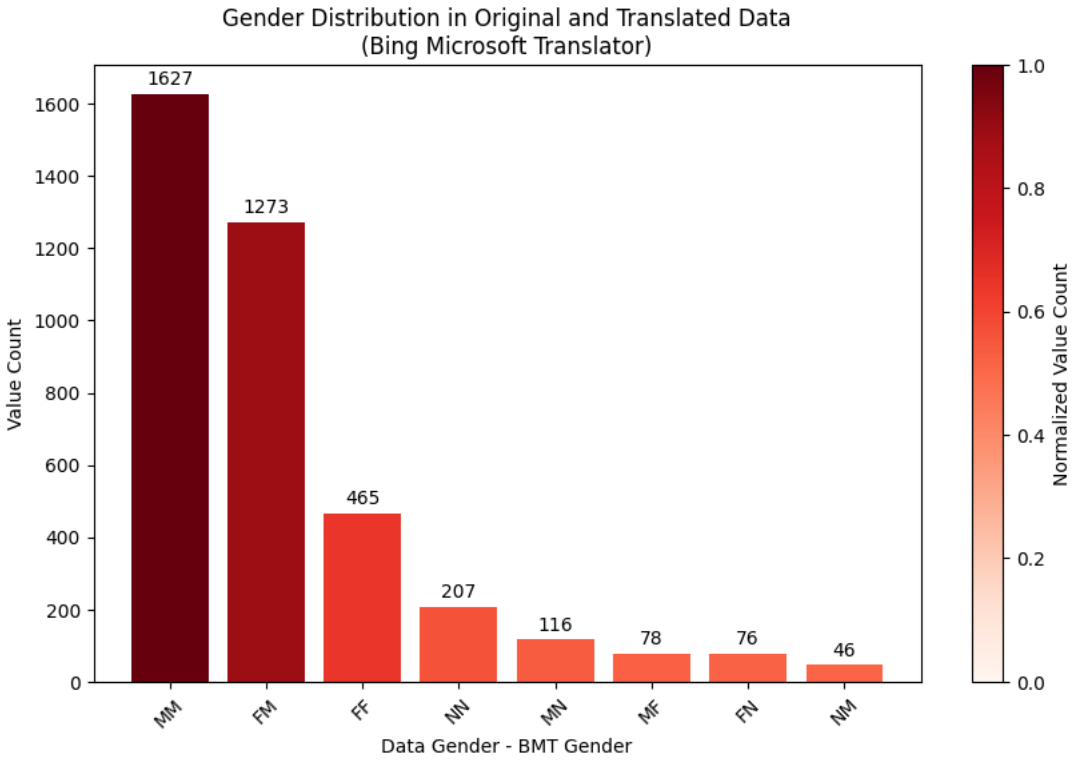

Figure 4.4: Comparison of Gender Representation in Original Text and Microsoft Translated Text.

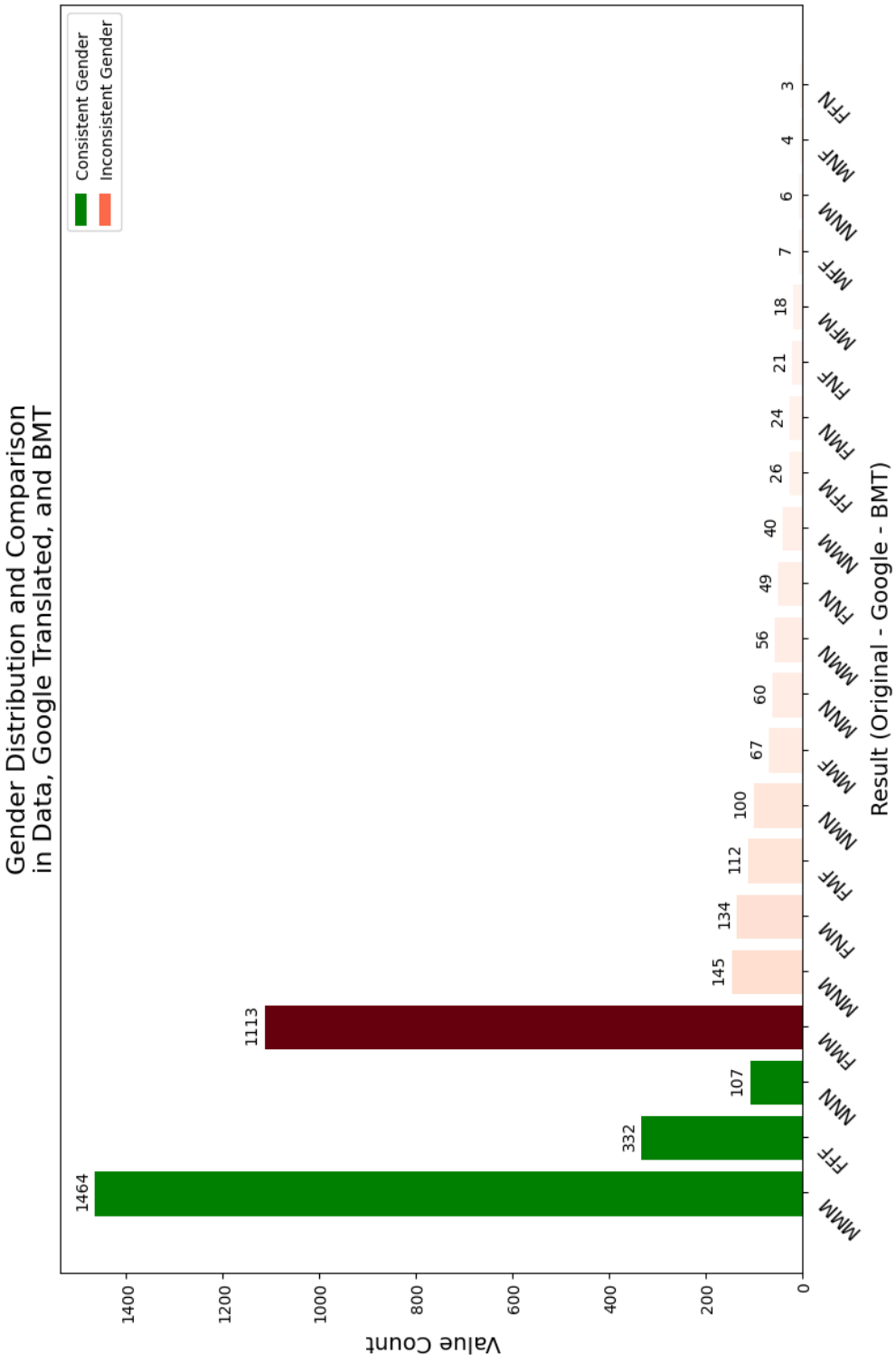

Figure 4.5: Comparison of Gender Distributions in Original Data, Google Translated Data, and BMT Translated Data.

nying counts. The first character denotes the gender in the original dataset, the second character signifies the gender in the Google-translated dataset, and the third character indicates the gender in the BMT-translated dataset. The counts in the table highlight the differences or similarities in gender assignment between the translation models. For example, the combination male-male-male, represented by the code 'MMM,' appears 1,464 times, showing how often a sentence is classified as male. Such a comparative analysis provides insights into the accuracy of gender translation by different models and highlights any discrepancies or biases introduced during the translation process.

Original	Hindi	Google Translated	Bing Microsoft Translator
The guard appreciated getting treatment from the nurse because he did a good job.(M)	गार्ड ने नर्स से इलाज करने की सराहना की क्योंकि उसने अच्छा काम किया।(N)	The guard appreciated treating the nurse as she did a good job.(F)	The guard appreciated the treatment from the nurse because she did a good job.(F)
The developer argued with the designer because she did not like the design.(F)	डेवलपर ने डिजाइनर के साथ तर्क दिया क्योंकि उसे डिजाइन पसंद नहीं था।(N)	The developer argued with the designer because he did not like the design.(M)	The developer argued with the designer because he didn't like the design.(M)
The physician told the baker that she tried the best.(F)	चिकित्सक ने बेकर को बताया कि उसने सबसे अच्छी कोशिश की।(N)	The doctor told Baker that he tried his best.(M)	The therapist told Baker that he tried the best.(M)
The laborer raced against the hairdresser because she thought it would be easy.(F)	मजदूर ने हेयरड्रेसर के खिलाफ दौड़ लगाई क्योंकि उसे लगा कि यह आसान होगा।(N)	The labourer ran against the hairdresser because he thought it would be easy.(M)	The worker raced against the hairdresser because he thought it would be easy.(M)

Figure 4.6: Example Sentences and Their Translation

Figure 4.6 demonstrates the manifestation and possible modification of gender bias in machine translation. It showcases four English sentences, their Hindi translations, and the subsequent translations back into English using Google Translate and Bing Microsoft Translator. Each sentence includes a profession and a pronoun, with the gender implied by the pronoun noted in parentheses (M for male, F for female, N for neutral).

The research investigated the gender translation accuracy of two prominent translation models across various professions. Despite their sophisticated algorithms, the models tend to alter gender from female to male or vice versa during the translation process based on the profession. These findings indicate that improvements are necessary in automated translation systems. Enhancing gender translation accuracy is vital for ensuring fair representation and inclusivity in different linguistic contexts.

The figures 4.7 and 4.8 present confusion matrices illustrating the performance of Google Translate (GT) and Bing Microsoft Translator (BMT) in classifying gender in translated text. Each matrix compares the original gender of a word in the source text to the gender predicted by the MT system.

Figure 4.7 (Google Translate):

- The diagonal elements (N-N, F-F, M-M) represent correct gender classifications.
- Off-diagonal elements indicate misclassifications.
- GT struggles most with predicting the 'F' gender, often misclassifying it as 'M'.

Figure 4.8 (Bing Microsoft Translator):

- Similar to GT, BMT also performs best when the original and predicted genders match (diagonal elements).
- However, BMT seems to have a slightly better performance in predicting 'F' gender compared to GT.

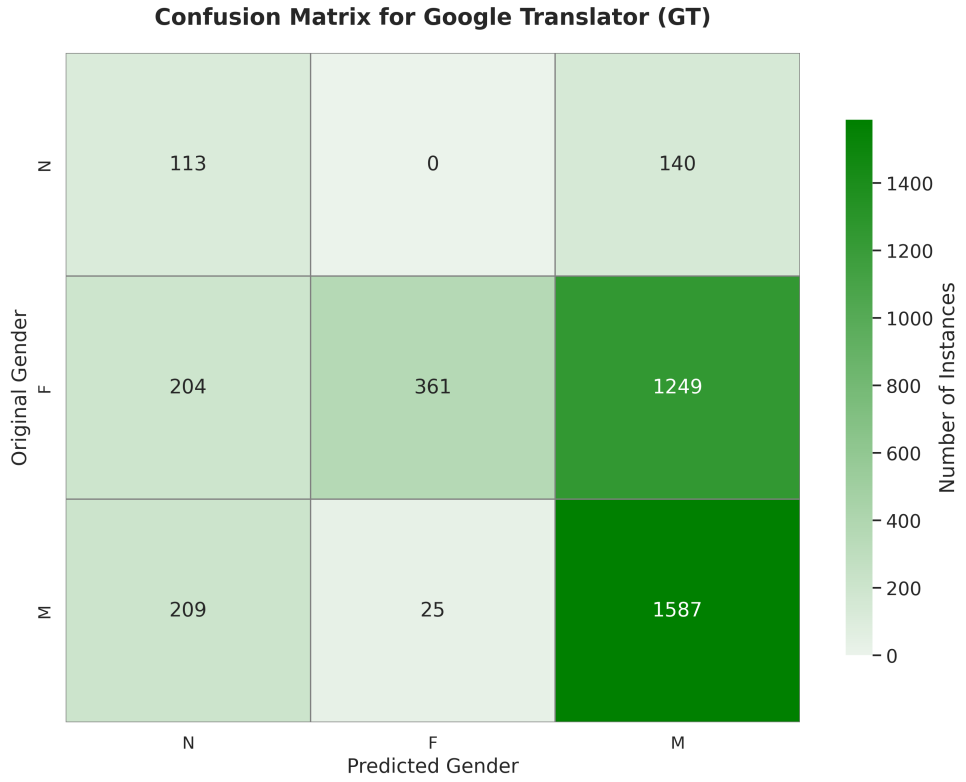

Figure 4.7: Google Translate's Gender Classification: A Confusion Matrix

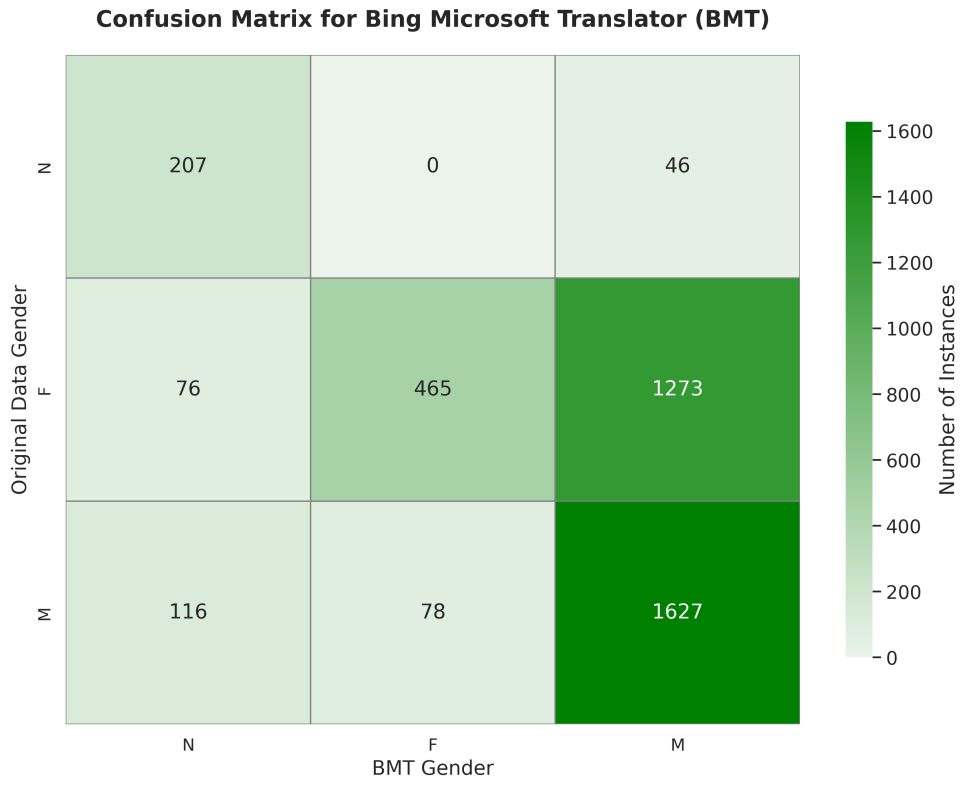

Figure 4.8: Microsoft Translate's Gender Classification: A Confusion Matrix

- The most frequent error for BMT is misclassifying 'F' as 'M', similar to GT.

Overall, both figures highlight the challenges faced by MT systems in accurately translating gender, especially for the female gender, potentially reflecting biases in the training data or the underlying algorithms.

In this research, accuracy is defined as the lack of bias in the translated text when compared to the original data. To compute the accuracy, the number of correctly translated sentences is totaled and then divided by the overall number of sentences in the dataset, which is 3888 in this instance.

For the Google translation system and the Bing Microsoft translation system, the accuracy is determined as follows:

$$\text{Accuracy (Google)} = \frac{1587 + 361 + 113}{3888} \approx 53.03\%$$

$$\text{Accuracy (Bing)} = \frac{1627 + 465 + 207}{3888} \approx 59.12\%$$

When comparing the two, it is noted that the Bing Microsoft translation system attains slightly higher accuracy than Google. However, it is crucial to acknowledge that both systems demonstrate comparable performance in translating sentences accurately without introducing bias.

Chapter 5

CONCLUSION AND FUTURE SCOPE

This study examined how well two prominent translation models handle gender translation across various professions. Despite their advanced algorithms, both systems struggled to accurately translate professions and frequently misinterpreted gender. Comparing Bing Microsoft and Google translation systems, Bing Microsoft showed slightly higher accuracy, but both performed similarly in avoiding bias. Future research aims to widen evaluation to more language pairs and a diverse vocabulary. Additionally, creating a challenge set focusing on gender-related linguistic phenomena is planned for automatic translation system evaluation. The ultimate goal is a cutting-edge machine translation system to scrutinize machine bias impact on translation outputs and explore mitigation strategies. Various approaches, including those for other languages, will be explored to minimize biases in machine translation. Implementing gender tagging at the sentence or word level could help mitigate bias, enhancing translator performance and reliability.

Research on mitigating bias in Hindi-to-English machine translation is in its developing stages and it is presenting numerous opportunities for advancements in the future. For further exploration here are some areas:

1. **Expanding Bias Mitigation:**Future studies should also address caste, religion, and regional biases present in Indian languages.

2. **Developing New Metrics:**We need precise bias evaluation metrics for Hindi that consider the language's nuances, grammar, and cultural context for better bias detection and mitigation.

3. **Incorporating Cultural Sensitivity:**Integration of cultural knowledge into machine translation models could prevent misinterpretations and offensive translations due to cultural disparities.

4. **Improving Data Diversity:**Training models on diverse datasets representative of various groups and perspectives could mitigate biases and ensure equitable representation.

5. **User Feedback and Evaluation:**Involving human users in the evaluation process can offer valuable insights into the effectiveness of bias mitigation strategies and areas for enhancement.

By addressing these challenges and opportunities, researchers and developers can foster more inclusive and equitable machine translation systems, accurately reflecting the diverse linguistic and cultural landscape of India.

Appendix A

Appendix Title

Bibliography

[1] T. Mewa, *'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings' by Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai (2016)*, may 30 2020, https://cis.pubpub.org/pub/debiasing-word-embeddings-2016.

[2] 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*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 166–172. [Online]. Available: https://aclanthology.org/W19-3823

[3] K. Ramesh, G. Gupta, and S. Singh, "Evaluating gender bias in Hindi-English machine translation," in *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, M. Costa-jussa, H. Gonen, C. Hardmeier, and K. Webster, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 16–23. [Online]. Available: https://aclanthology.org/2021.gebnlp-1.3

[4] S. Friedman, S. Schmer-Galunder, A. Chen, and J. Rye, "Relating word embedding gender biases to gender gaps: A cross-cultural analysis," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 18–24. [Online]. Available: https://aclanthology.org/W19-3803

[5] M. O. R. Prates, P. H. C. Avelar, and L. Lamb, "Assessing gender bias in machine translation – a case study with google translate," 2019.

[6] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus : A survey of transformer-based pretrained models in natural language processing," 2021.

[7] S. L. Blodgett, S. Barocas, H. D. I. au2, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in nlp," 2020.

[8] J. Dacon and H. Liu, "Does gender matter in the news? detecting and examining gender bias in news articles," in *Companion Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 385–392. [Online]. Available: https://doi.org/10.1145/3442442.3452325

[9] L. Fu, C. Danescu-Niculescu-Mizil, and L. Lee, "Tie-breaker: Using language models to quantify gender bias in sports journalism," 2016.

[10] K. Chaloner and A. Maldonado, "Measuring gender bias in word embeddings across domains and discovering new gender bias word categories," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–32. [Online]. Available: https://aclanthology.org/W19-3804

[11] D. Saunders and B. Byrne, "Reducing gender bias in neural machine translation as a domain adaptation problem," 2020.

[12] A. Johnson, "Scarlett johansson shutting down sexist comments for 5 min straight," https://www.youtube.com/watch?v=YGqQk12jBoA, 2020, [Online video].

[13] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, S. R. El-Beltagy, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, W. El-Hajj, M. Jarrar, and H. Mubarak, "A panoramic survey of natural language processing in the arab world," 2021.

[14] H. Kapoor, P. H. Bhuptani, and A. Agneswaran, "The bechdel in india: gendered depictions in contemporary hindi cinema," *Journal of Gender Studies*, vol. 26, no. 2, pp. 212–226, 2017.

[15] N. Madaan, S. Mehta, T. Agrawaal, V. Malhotra, A. Aggarwal, Y. Gupta, and M. Saxena, "Analyze, detect and remove gender stereotyping from bollywood movies," in *Conference on Fairness, Accountability, and Transparency*. PMLR, January 2018, pp. 92–105.

[16] A. K. Pujari, A. Mittal, A. Padhi, A. Jain, M. K. Jadon, and V. Kumar, "Debiasing gender biased hindi words with word-embedding," *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:211104718

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[19] E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams, "Multi-dimensional gender bias classification," 2020.

[20] H. Gonen and K. Webster, "Automatically identifying gender issues in machine translation using perturbations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1991–1995. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.180

[21] A. Wong, "Mitigating gender bias in neural machine translation using counterfactual data," Master's thesis, The Graduate Center, City University of New York, New York, NY, 9 2020, available at https://academicworks.cuny.edu/gc_etds/3990.

[22] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," 2019.

[23] J. Escudé Font and M. R. Costa-jussà, "Equalizing gender bias in neural machine translation with word embeddings techniques," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 147–154. [Online]. Available: https://aclanthology.org/W19-3821

[24] Z. Zhang, D.-H. Lee, Y. Fang, W. Yu, M. Jia, M. Jiang, and F. Barbieri, "Plug: Leveraging pivot language in cross-lingual instruction tuning," 2024.