B. Sc. Information Systems
Berlin School of Economics and Law
Department 1: Business and Economics

Bachelor's Thesis

# Detecting Gender Bias in English-German Translations using Natural Language Processing

Khanh Linh Pham

Supervisors: Prof. Dr. Diana Hristova, Prof. Dr. Markus Schaal

Semester: Summer Semester 2025
Matrikel-Nr.: 77211916753
Email: klpham04@gmail.com

**Date: xx.xx.2025**

**Abstract**

XX

**Sperrvermerk**

XX

Berlin, den 01. Januar 2099

. . . . . .
*(Unterschrift des Verfassers)*

# Contents

**List of Figures** **vi**

**List of Tables** **vii**

**1 Introduction** **1**

1.1 Motivation . . . 1
1.1.1 Social and Ethical Importance of Addressing Gender Bias . . . 1
1.1.2 Why Detection Systems Are Needed . . . 2
1.2 Problem Statement and Research Questions . . . 2
1.3 Scope . . . 3
1.4 Limitations . . . 3
1.5 Overview of Chapters . . . 4

**2 Theoretical Background and Related Work** **5**

2.1 Definitions . . . 5
2.1.1 Natural Language Processing and Machine Translation . . . 5
2.1.2 Bias in Machine Translation Systems . . . 6
2.1.3 Manifestiations of Gender Bias . . . 6
2.1.4 What Gender Bias Means in Machine Translation . . . 8
2.2 Related Works . . . 9
2.2.1 Literature Search Process . . . 9
2.2.2 Linguistic Challenges in English-German Translation . . . 11
2.2.3 German Gender-Fair Language . . . 11
2.2.4 Foundational studies . . . 12
2.2.5 Why it matters . . . 13
2.3 Research Gaps . . . 15
2.4 Approach and Justification of the Technical Setup . . . 15
2.4.1 Binary Classification in NLP . . . 15
2.4.2 Transformer Architecture . . . 15
2.4.3 Language model: BERT . . . 16

2.4.4 Multilingual BERT . . . . . 20
2.4.5 Interactive Demo . . . . . 21

**3 Methodology** **22**
3.1 Goal of the project . . . . . 22
3.2 Workflow . . . . . 22
3.3 Dataset Handling . . . . . 22
3.3.1 Final Dataset Composition . . . . . 24
3.4 Data Pre-processing . . . . . 25
3.5 Feature extraction . . . . . 25
3.6 Model Selection and Training . . . . . 25
3.6.1 Hyperparameters . . . . . 25
3.7 Evaluation . . . . . 25
3.8 Demo Application Design . . . . . 25

**4 Implementation** **26**
4.1 Project Structure . . . . . 26
4.2 Environment Setup . . . . . 26
4.3 Core components and their interaction . . . . . 26
4.4 Demo Result . . . . . 26

**Bibliography** **27**
1 Prompt and Output for Pre-training/Fine-tuning Explanation . . . . . 31

# List of Figures


2.1 Frequency of different types of gender-inclusive language. Source: Waldendorf (2024) p. 367. . . . . 12
2.2 Study design. Post-editing of an MT output into both feminine and masculine gender. Source: Savoldi et al. (2024) p. 18048 . . . . . 14
2.3 Transformer encoder-decoder architecture. The encoder (left) processes input tokens $x_1, \ldots, x_m$ through: (1) a self-attention layer for contextual relationships, (2) a feed-forward network for feature transformation, and (3) residual connections with layer normalization. The decoder (right) generates outputs by attending to both the encoder's representations and its previous outputs ($y_0$ to $y_{n-1}$), producing the next-token probability distribution. Figure and description adapted from Xiao and Zhu (2023), p. 6. . . . . 17
2.4 BERT's encoder-only architecture. Figure by Smith (2024). . . . . 18

3.1 Workflow of the project. . . . . 23

1 Google Translate assigns stereotypical genders to occupational roles. . . . . 35
2 DeepL shows a similar bias in the same sentence, highlighting consistent patterns across MT tools. . . . . 35
3 Gender-specific translation by Google Translate for ambiguous pronouns. . 35

# List of Tables

2.1 Key concepts relevant to this thesis . . . . . 10

3.1 Overview of available EN-DE datasets based on past works. . . . . 24
3.2 Sample entries from the final training dataset. . . . . 25

# 1 Introduction

Machine Translation (MT) helps millions of people communicate across languages, in daily life and in areas like healthcare, law, and business (Kappl 2025). Services like Google Translate handle over 200 million users every day (Prates et al. 2019; Shrestha and Das 2022). It is a fast-growing market. A report by SkyQuest (2025) valued it at 980 million USD in 2023, with projections reaching 2.78 billion USD. New and more advanced translation models keep appearing, and many of them are free to use. As a result, MT tools are now used to translate large volumes of content across domains.

With this widespread use, the output of MT systems increasingly shapes how people receive and interpret information. But automatic translations are not neutral. There is growing concern about the social effects of biased translations. One key issue is gender bias. MT systems are often trained on large datasets that reflect social norms and stereotypes. If the data contains gender bias, the system will likely reproduce it (Cho et al. 2019; Soundararajan and Delany 2024; Smacchia et al. 2024).

A common case is the use of gendered terms in translations of gender-neutral input. For example, the English sentence "The nurse is hard-working" does not say anything about gender. But a translation system may render it in German as "Die Krankenschwester ist fleißig," which uses the explicitly feminine term *Krankenschwester*. Similarly, "The surgeon is hard-working" may become "Der Chirurg ist fleißig," using the masculine form *Chirurg*. These choices add gendered assumptions that were not present in the original. Such patterns are not just technical side effects. They can reinforce stereotypes, especially when they appear in job ads, reports, or other public texts.

## 1.1 Motivation

### 1.1.1 Social and Ethical Importance of Addressing Gender Bias

Academia has come to the consensus that MT systems do default to male pronouns when gender in the source sentence is ambiguous (Prates et al. 2019; Cho et al. 2019; Rescigno and Monti 2023). In addition, translations often reflect traditional roles, like associating "nurse" with women and "surgeon" with men. This can affect people's perceptions of jobs

and reinforce gender roles.

When used in formal contexts like job descriptions or reference letters, biased translations can shape how a candidate is perceived. If a system always assigns male pronouns to leadership roles and female terms to caregiving roles, it may disadvantage those who do not match those stereotypes (Bolukbasi et al. 2016). This is not just a personal issue. It can reduce diversity and go against international standards. Organizations like the United Nations, UNESCO, and the European Union stress the importance of gender equality and inclusive language, making gender equality one of the 17 Sustainable Development Goals for 2030 (Sczesny et al. 2016; United Nations 2023).

Language also shapes thought. Research shows that readers often interpret masculine forms as male-specific, even if they are supposed to be generic (Sczesny et al. 2016). Inclusive forms are more common in official documents, less so in everyday language. However, exposure matters. Frequent use of fair language makes it feel more normal. Detecting and addressing bias in MT can support this shift.

### 1.1.2 Why Detection Systems Are Needed

Current research on this topic tends to focus more on the quantitative measurement of gender bias (Rescigno and Monti 2023; Barclay and Sami 2024; Smacchia et al. 2024). Common methods include counting gendered forms in outputs and comparing them to demographic baselines or human expectations (Rescigno and Monti 2023; Prates et al. 2019; Savoldi et al. 2024). These are useful, but they do not help users identify specific biased translations in real-time. Evaluations are not enough for accountability.

Other domains, like facial recognition, have already seen progress in active bias detection. For example, Schwemmer et al. (2020) showed that systems tend to label women more accurately if they match stereotypical appearances (e.g., long hair). Some models even linked female images to words like “kitchen” or “cake” based on bias patterns in training data. For MT, a detection layer is still missing. Without such tools, biased translations are likely to spread unnoticed. A detection system could flag potential bias in real time, improving transparency and encouraging more careful use.

## 1.2 Problem Statement and Research Questions

**DRAFT NEED TO REWRITE AFTER IMPLEMENTATION** This thesis focuses on gender bias in English-to-German (EN-DE) MT. This language pair is widely used in research, with many open datasets and high-quality models available. It also involves a

grammatical shift: English has limited gender marking, while German assigns gender to many nouns and pronouns. This structural difference makes gender bias more visible and easier to study in the translation outputs.

The core problem boils down to the significant bias towards the masculine form in EN-DE MTs, sometimes consituting 93-96% of translations for isolated words (Lardelli et al. 2024). These outputs often reflect social stereotypes rather than objective translations, yet current systems offer no mechanism to detect or signal when such bias occurs (Rescigno and Monti 2023). To address this, this thesis deploys a blackbox approach to explore how fine-tuning a pre-trained multilingual BERT model can help detect gender bias in MT outputs. The model takes an input sentence and its corresponding German translation and predicts whether the translation introduces gender bias.

The translation system used is Opus-MT, an open-source neural MT model. It is widely used in research, supports EN-DE translation, and is trained on real-world corpora, making it suitable for studying translation bias (Tiedemann and Thottingal 2020). Translations are then passed through BERT, trained on a dataset I have constructed by combining and adapting several existing datasets from other researchers. The classifier is lightweight and efficient, aiming for transparent behavior and easy integration into other tools (Devlin et al. 2019). The final tool highlights biased parts in a simple web demo. The goal is not a perfect classifier but a working prototype that shows how such detection could be integrated into translation workflows.

The main research question is therefore: **"How can a NLP-based binary classification model detect gender bias in English-German translations?"**.

## 1.3 Scope

**WRITE AFTER IMPLEMENTATION PART** This thesis focuses only on EN-DE MT. Other language pairs are out of scope.

## 1.4 Limitations

**WRITE AFTER IMPLEMENTATION PART** It becomes especially difficult to detect when sentences contain multiple subjects, indirect references, or ambiguous pronouns. For example, as Barclay and Sami (2024) explain, the sentence “He went to see her mother” clearly implies three people, while “He went to see his mother” could refer to either two or three. These types of structures introduce ambiguity that makes annotation and evaluation

much harder. Creating a dataset that captures such linguistic complexity would require significant effort and careful control of variables. One broader limitation in building datasets for complex scenarios with multiple subjects is the difficulty of isolating the influence of each gendered entity (Lardelli et al. 2024). When working with natural language sources, it becomes hard to tell what caused the bias in the translation. Because of this, the focus of this thesis is on simpler sentence structures with a single subject. This makes it easier to identify and explain bias patterns. It also fits the intended use case: translating business texts like job advertisements or reports, which rarely involve multiple nested clauses or ambiguous pronouns.

## 1.5 Overview of Chapters

**WRITE AFTER IMPLEMENTATION PART**

# 2 Theoretical Background and Related Work

This section outlines key findings of related work on gender bias in MT, with a focus on the English-German (EN-DE) language pair to build the theoretical knowledge base. The research aims are to (1) define the core concept of gender bias in MT, (2) establish the relevance of the topic, (3) identify the research gap, and (4) justify technical design choices.

## 2.1 Definitions

This section explains the key terms and concepts needed to understand gender bias in English-to-German MT. It defines important ideas like natural language processing (NLP), MT, and gender bias. These concepts provide the background necessary to follow the thesis.

### 2.1.1 Natural Language Processing and Machine Translation

**NLP** refers to the development of machine systems that can process and generate human language. The goal is to mimic and understand it as fluently as possible (Smacchia et al. 2024; Ullmann 2022). Common applications are chatbots, translation tools, speech recognition, and image captioning.

**MT** is a direct application of NLP. It is used to automatically translate text from one language to another (Lin and Chien 2009). MT systems have gone through several stages of development; earlier approaches like rule-based and statistical MT used manually defined grammar rules or pattern matching from large translation corpora (Chakravarthi et al. 2021). For example:

> "The girl reads a book" → "Das Mädchen liest ein Buch"
>
> Rules: "girl" → "Mädchen", "reads" → "lesen", "book" → "Buch"

These systems often struggled with full sentences and complex expressions because they fail to capture context and phrase-level meaning. "She gave him a hand" might be translated literally, missing its idiomatic meaning.

Most modern systems, including Google Translate and DeepL, use **neural machine translation (NMT)** (Wu et al. 2016; DeepL 2021). These systems are trained on large sets

of translated texts. They learn to represent the meaning of whole sentences as mathematical structures and generate more fluent and accurate translations. Unlike earlier systems, they aim to consider the full context of a sentence, which helps reduce errors and improves the handling of ambiguous or idiomatic language.

### 2.1.2 Bias in Machine Translation Systems

Similarly to how humans are shaped by their environemnt, MT models learn from data they are trained on. Existing biases are thus reflected and reinforced in the final models, creating "machine bias" (Stanczak and Augenstein 2021; Smacchia et al. 2024). Shah et al. (2020), as described by Ullmann (2022), differentiates between four origins of biases affecting NLP systems:

- **Selection Bias:** Happens when the training data does not reflect the context in which the model is used (e.g., using Wikipedia data for detecting harmful language on Twitter).
- **Label Bias:** Occurs when annotations in the dataset are incorrect or skewed. This can be influenced by the annotators' own biases or lack of awareness of diverse linguistic expressions.
- **Model Overamplification:** During training, models can exaggerate patterns found in the data. If a dataset predominantly associates cooking with women, the assumption can be reinforced that cooking is an activity exclusive to women.
- **Semantic Bias:** Stems from associative relationships within the data, where certain words or phrases are frequently co-occurring with specific genders (e.g., "he" with "doctor").

Ullmann (2022) notes that the scale of training data (e.g., 175 billion parametres for GPT-3) makes it practically impossible to review all of it, allowing misinformation or offensive content to be reproduced by the system. The author also points out that platforms like Wikipedia and Reddit are male-dominated and often contain harmful or false content.

### 2.1.3 Manifestiations of Gender Bias

This section draws from the main studies analyzing gender bias in EN-DE MT (Ullmann 2022; Rescigno and Monti 2023; Lardelli et al. 2024; Kappl 2025). Since existing research

does not clearly define the different manifestations, the findings are grouped here into three main categories.

### Defaulting to Masculine Forms

In both singular and plural contexts, the *generic masculine* refers to the default use of the masculine grammatical gender. For example, the sentence "Die Studenten sind im Hörsaal" (translation: "The students are in the lecture hall") uses the masculine plural form to refer to a group of students regardless of their gender.

It is commonly used in spoken German (Lardelli et al. 2024; Schmitz 2022), although research has consistently shown that the generic masculine creates a male bias in mental representations, leading readers or listeners to think more of male than female examples (Sczesny et al. 2016). In MT, the generic masculine can lead to inaccurate or unfair representations of gender in translated text. Rescigno and Monti (2023) observed a predominance of masculine forms in translation outputs (approximately 90% in Google Translate and 85–88% in DeepL for EN-IT and EN-DE), even when the original sentences contained relatively few masculine references. This shows that the bias is not minor but occurs quite heavily in those systems.

### Reinforcement of Stereotypes

Stereotypes and gender roles stem from historical and cultural perceptions of men's and women's societal roles, many of which are obsolete but still influential. For example, when men and women often take on different roles at work and at home, it shapes how people think about their personalities and qualities. Correspondence bias can emerge, where people infer attributes from observable behaviours (Godsil et al. 2016). These associations can then be reinforced by popular media, such as TV and advertisements (Godsil et al. 2016), just as much as it can be influenced by MT tools.

A common manifestation of this are **stereotypical job associations**. This can be seen in cases where models assign he/him pronouns to roles like doctors and pilots, and she/her pronouns to roles like nurses and flight attendants (Shrestha and Das 2022), with an even stronger tendency in male-domniated fields such as STEM (Prates et al. 2019). In addition, NLP models have also been shown to **link certain adjectives and traits to genders**. Traits like "masterful," "assertive," and "competitive" are often associated with men, while "friendly," "unselfish," and "emotionally expressive" are more commonly linked to women (Godsil et al. 2016).

#### Neglecting Contextual Information

Coreference resolution refers to the process of using contextual information to determine the correct gender in translation (Stanczak and Augenstein 2021). In MT, this means identifying links between words like pronouns and the nouns they refer to. While human translators use both linguistic cues (such as pronouns and grammar) and real-world knowledge to correctly assign gender (Rescigno and Monti 2023), MT systems often fail to do so reliably, especially when gender information appears earlier in the text or across sentence boundaries (Cho et al. 2019; Stanovsky et al. 2019). For example, if a biography introduces a person with a female name at the beginning, but later refers to that person only by name, translation systems may lose the link and default to masculine forms for the remaining text.

Rescigno and Monti (2023) found that including previous sentences improved coreference resolution and reduced masculine defaults, though some systems benefited more than others. However, the use of context also introduced occasional new errors. Additionally, Savoldi et al. (2024) highlighted that correcting biased translations toward feminine forms required significantly more time and edits than masculine ones, revealing a notable cost disparity.

Similarly, Lardelli et al. (2024) showed that even with natural passages from Wikipedia and Europarl, systems still largely defaulted to masculine forms. Feminine and inclusive translations remained rare, while gender-neutral alternatives appeared mainly when the noun itself suggested them.

### 2.1.4 What Gender Bias Means in Machine Translation

A clear definition of gender bias has not yet been established (Stanczak and Augenstein 2021). Determining which features in text indicate bias is difficult, and the characteristics of non-biased text are often unclear. This makes it challenging to hold users accountable for gender bias, detect all harmful signals, and develop standard evaluation benchmarks (Barclay and Sami 2024; Shrestha and Das 2022; Stanczak and Augenstein 2021).

Since there is no clear definition, this work defines gender bias based on the three manifestations described above: **Defaulting to Masculine Forms, Reinforcement of Stereotypes and Neglecting Contextual Information.** Any text that exhibits one or more of these forms will be considered gender biased.

## 2.2 Related Works

### 2.2.1 Literature Search Process

For the literature review I combined incremental and conceptual literature review methods, where each source led to the identification of the next. Based on this progression, I identified key concepts and used them to organize and interpret the literature, aligning with a conceptual approach. The structure followed the qualitative Information Systems framework by Schryen (2015) and was further informed by Shrestha and Das (2022) and Savoldi et al. (2025a), who both conducted systematic reviews on gender bias in ML and MT respectively.

#### Search Sources and Tools

Sources were primarily searched on Google Scholar and Perplexity, which served as an additional search engine. Prompts and outputs from Perplexity have been saved and are included in the appendix. To organize and manage the collected sources, Zotero was used throughout the process.

#### Literature Review Framing

To answer the four research aims, I have defined the key concepts in Table 2.1. Key search terms consisted of *gender bias*, *machine translation*, *AI*, *machine learning*, *German*, *stereotypes*, and *detection*, which were combined with *AND/OR*. The focus was on literature published between 2019 and 2025 to maintain relevance and currency, while foundational and definitional works from earlier periods were selectively included. The initial search for the term *gender bias in machine translation* returned over 18,000 results. Through my iterative selection process, this was narrowed down to 34 core sources.

#### Citation Tracking

Backward citation searching involved reviewing references cited by selected papers, prioritizing frequently cited and foundational works relevant to gender bias in MT. Forward citation searching used Google Scholar's "cited by" function to identify newer research citing those key papers. Filtering with specific terms (e.g., *German* and *machine translation*) was applied during forward search to maintain focus. Beyond these systematic methods, I also included supplementary sources when needed while writing. These consist of contextual references, statistics, or secondary citations that support specific points but were not part of the core conceptual or methodological framework. Supplementary sources were defined as

| Key Concept | Description |
|---|---|
| Foundations of Gender Bias in Natural Language Processing | Traces early research that identified gender bias in language. Focuses on foundational studies that showed why the issue matters and how later work builds on these findings. |
| Sources and Manifestations of Bias | Explains how stereotypes shape language and persist over time. Describes how societal bias enters training data, model design, and system feedback. Shows how bias appears in machine translation and everyday language. |
| Linguistic Challenges in English-German Translations | Explores key grammatical differences between English and German that affect translation. Focuses on how the lack of gender in English and its presence in German can lead to biased outputs. |
| Mitigation Strategies and Current Limitations | Reviews how current research tries to reduce gender bias in NLP. Highlights what these methods can and cannot do. Helps identify where a classification-based approach could fill gaps and improve bias detection in translations. |

Table 2.1: Key concepts relevant to this thesis

materials identified outside the systematic search, such as papers found through backward citations or targeted queries for statistics and news, which provided support for subordinate arguments without being central to the study's theoretical or analytical structure.

### Selection Criteria and Screening Process

Titles and abstracts were manually screened to select relevant studies. **Inclusion criteria** required sources to specifically address gender bias in MT, provide examples or discussions of gender-related errors, or explain the significance of gender bias in this context. Sources also had to be available in full text without access restrictions. **Exclusion criteria** filtered out studies focusing on general NLP bias without a direct link to MT, non-gender biases, and highly technical papers lacking contribution to the general understanding of gender bias or that did not provide additional knowledge beyond what was already found in previously published papers. Full texts were reviewed after initial screening to confirm relevance and extract insights. Redundant sources not providing new perspectives aligned with the thesis goals were excluded.

### 2.2.2 Linguistic Challenges in English-German Translation

Although both English and German originate from the Indo-European language family (Baldi 2008), they have different characteristcs. English does not assign grammatical gender to nouns. The article "the" is used universally, independent of what it refers to. On the contrary, German assigns one of three grammatical gendered articles to nouns: "der" (m), "die" (f) and "das" (n). The form or ending of a noun may also change depending on its grammatical gender. While English has a few gendered word pairs, such as "actor" (m) and "actress" (f), gender distinctions in German apply broadly across the entire noun system. "Der Student" refers to a male student, whereas "die Studentin" refers to a female student. Note that grammatical gender has no connection to societal or biological gender. It is a rule of the language rather than a reflection of identity. For example, the German word Mädchen (girl) is grammatically neuter and takes the article "das". This is not because the referent lacks gender, but because the suffix "-chen" automatically assigns neuter gender. Grammatical gender in German follows structural rules, even when they contradict real-world gender associations.

### 2.2.3 German Gender-Fair Language

Gender-fair language (GFL) refers to the use of language that treats all genders equally and aims to reduce stereotyping and discrimination (Sczesny et al. 2016). Three common approaches to plural mentionings in German are:

- **Gender-neutral rewording:** This uses neutral terms instead of gendered nouns, e.g., *die Studierenden lernen.* A challenge for this version is that neutral alternatives do not exist for every noun and cannot be consistently applied (Lardelli et al. 2024).
- **Gender-inclusive characters:** This combines masculine, feminine and non-binary forms by using a character like *, :, or _, e.g., *die Student*innen lernen.* This method is consistent but may interrupt reading flow and lacks standardization (Lardelli et al. 2024).
- **Pair form:** This names both gender forms, e.g., *die Studentinnen und Studenten lernen.* It is currently the most used GFL form in German (Waldendorf 2024), briefly surpassing the star and colon characters as seen in Figure 2.1.

These examples apply when the gender of the subjects is ambiguous. But when gender is known, especially in singular mentions, the generic masculine should be avoided. However,

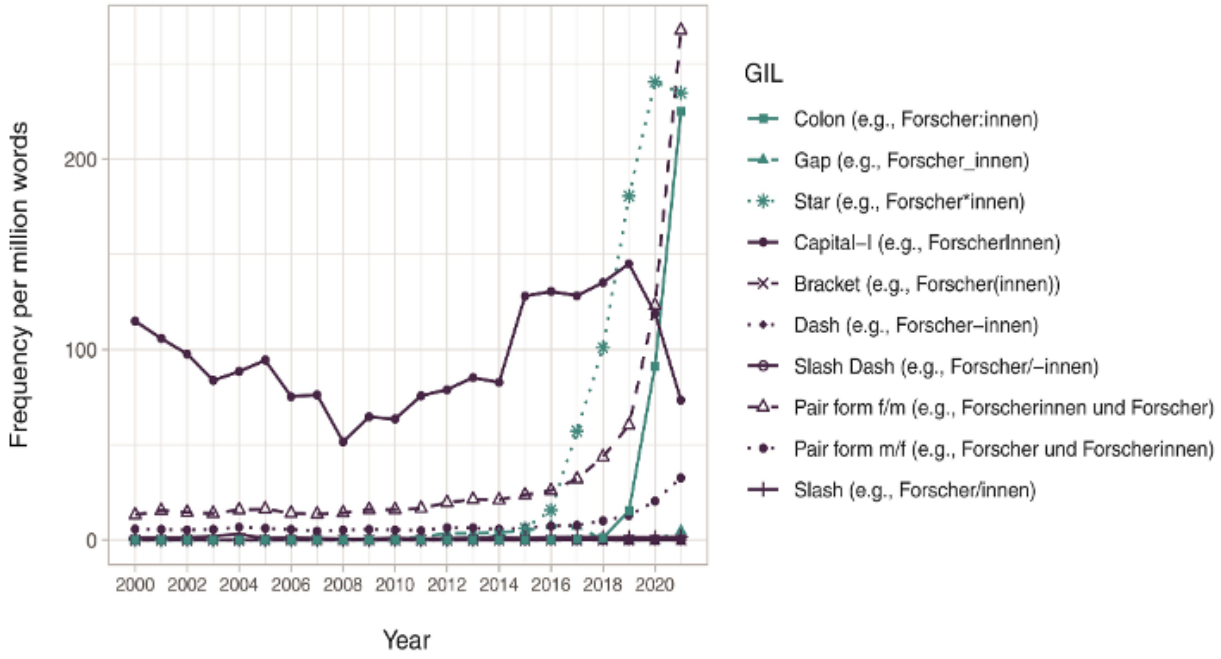


Figure 2.1: Frequency of different types of gender-inclusive language. Source: Waldendorf (2024) p. 367.

in the same way as gender bias has no clear definition, there is **no agreed standard for GFL** (Lardelli et al. 2024; Savoldi et al. 2025a). "Fairness" therefore heavily depends on personal views, culture, and context, which raises ethical questions about debiasing systems.

### 2.2.4 Foundational studies

The existence of gender bias in MT is well-documented. First mentions of this issue date back to over a decade ago, having been recognized by a paper by Schiebinger in 2014. Since then, there has been a general increase in research papers focusing on this topic, especially between 2019 and 2023 (Savoldi et al. 2025a).

**Prates et al. (2019)** conducted a large-scale study using Google Translate to translate sentences like "[Gender-neutral pronoun] is an engineer" from twelve gender-neutral languages into English. The results showed a strong bias toward male pronouns, especially in STEM occupations. This could not be explained by real-world labor statistics, pointing instead to imbalances in the system's training data. The study received wide media attention, leading Google to change their translation policy: Google Translate began showing both feminine and masculine forms for ambiguous inputs (Google 2018) (see Figure 3).

Building on this, **Stanovsky et al. (2019)** created WinoMT, a benchmark for evaluating gender bias in English-to-multilingual translations. It focused on occupations in contexts designed to challenge stereotypes. The study found that systems were more accurate for stereotypical gender roles but struggled in non-stereotypical cases, confirming the trends observed by Prates et al.. Together, these studies helped spark the ongoing research interest in gender bias in MT.

### 2.2.5 Why it matters

Gender bias in MT can lead to **representational harm**, meaning biased or reductive portrayals of a particular gender continue to spread (Stanczak and Augenstein 2021).

It also contributes to the invisibility of women in male-dominated professions (Kappl 2025). Studies show that biased language in machine-generated text, such as children's stories or job ads, can **influence how young people view themselves** (Soundararajan and Delany 2024; Kappl 2025). It may shape their interests, hobbies, and career choices. This is especially visible in STEM fields (Prates et al. 2019), where stereotypes are more persistent. When job descriptions or mock interviews use gender-exclusive pronouns, women report feeling less belonging, lower motivation, and weaker identification with the role (Godsil et al. 2016). Many self-select out of applying, shrinking the female talent pool and **reinforcing gender gaps in the workforce**.

Research also shows that using GFL like "she and he" or "one" can improve how women respond to job ads. It reduces stereotype threat and helps them engage more positively with opportunities (Godsil et al. 2016). Using inclusive language can therefore offer both social and competitive benefits for companies.

Furthermore, a study by Savoldi et al. (2024) employed behavioral metrics such as time to edit and the number of edits, measured through human-targeted error rate, to quantify the effort required. The results showed that post-editing feminine translations required nearly twice as much time and four times the number of editing operations compared to masculine counterparts (Figure 2.2). Consequently this effort gap also translates into **higher economic costs**, suggesting a measurable **quality-of-service disadvantage that disproportionately affects women**. Savoldi et al. concluded that current automatic bias metrics do not sufficiently capture these human-centered disparities, emphasizing the need for evaluation methods that reflect real user experience.

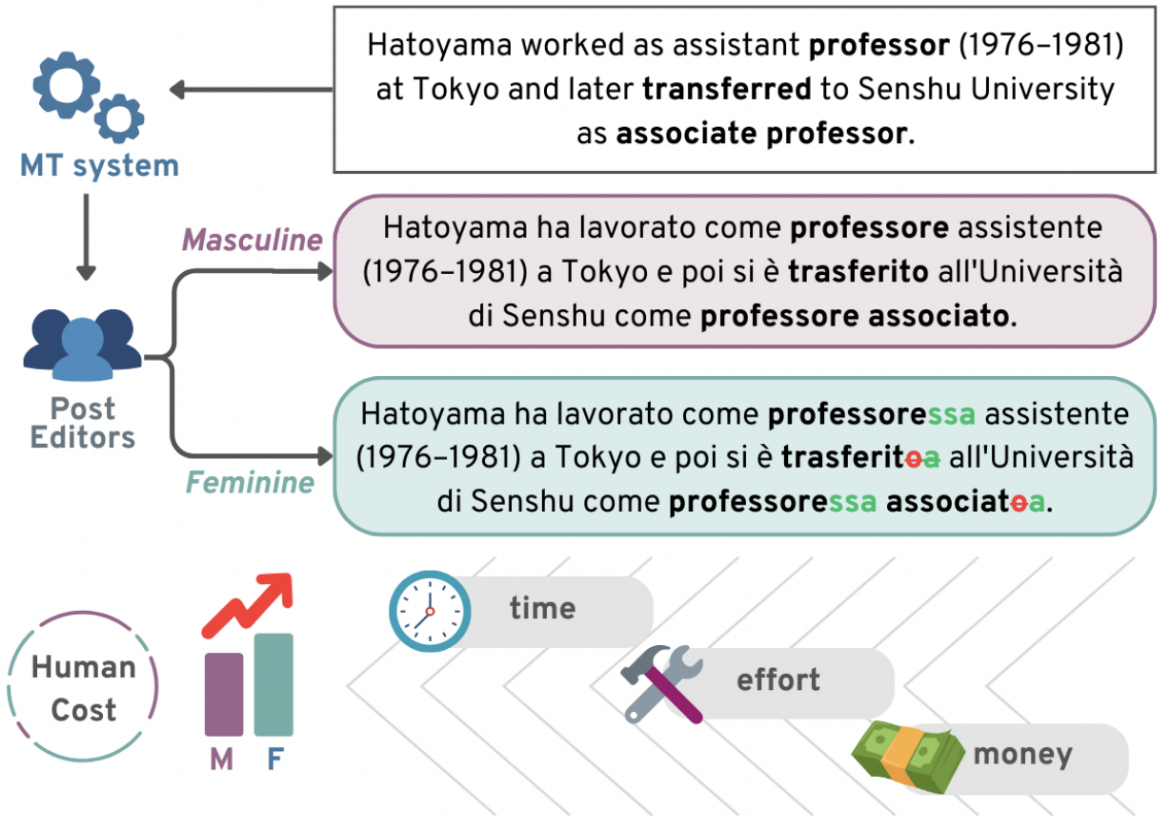


Figure 2.2: Study design. Post-editing of an MT output into both feminine and masculine gender. Source: Savoldi et al. (2024) p. 18048

## 2.3 Research Gaps

## 2.4 Approach and Justification of the Technical Setup

### 2.4.1 Binary Classification in NLP

Binary classification means sorting items into two clear groups. It is the most common task in ML and is frequently found in every day life, such as automatically flitering e-mails as "spam" or "not spam" (Quemy 2019). The ML algorithms use information from past examples to create a model or find key rules for making correct decisions. **This thesis tries to label a translation as either "potentially gender biased" or "neutral".**.

As explained in subsection 2.1.4, gender bias is difficult to define precisely, and the same applies to what counts as unbiased. I set specific rules to label something as biased, but this does not imply that everything else is unbiased. Therefore, I use the term "neutral" instead of "unbiased."

While it is possible to extend the classification beyond two categories, such as distinguishing types of bias or including labels like "gender-fair", this would require much more data and training. Given the practical aim of this work, which is to help users quickly identify whether their text might contain gender bias, the model focuses on a simple binary decision.

### 2.4.2 Transformer Architecture

Since BERT is based on the transformer architecture and is used for classification in this thesis, this section provides a brief overview of its structure. The transformer *transforms* an input sequence into an output sequence, such as in translation. To effectively achieve that, they deploy a self-attention mechanism (Phuong and Hutter 2022).

#### Self-attention mechanism

The self-attention mechanism allows the model to weigh the significance of all input elements simultaneously (Xiao and Zhu 2023), unlike traditional methods like Recurrent Neural Networks (RNNs), which process data sequentially. As a result, self-attention captures global dependencies and contextual relationships more accurately, creating "context-aware" representations, which is crucial for detecting subtle gender biases that depend on context within translations.

**Encoder-Decoder Framework**

The transformer architecture consists of two main components: the encoder and the decoder. The **encoder**'s job is to read the input sentence and turn it into a series of vectors the model can understand. Each vector is a list of numbers representing the meaning and structure of each word (Xiao and Zhu 2023). The encoder works as follows (see Figure 2.3):

1. It receives **input embeddings**, which represent the words, and **positional encodings**, which tell the model the order of the words.

2. The data then passes through several identical layers. Each layer has two main parts:
   a. **Multi-head self-attention** runs several attention processes in parallel. Each attention head focuses on different details to help the model understand the sentence better.
   b. A **Feed-forward network** processes each word vector separately, refining the information like a small filter.
   c. **Add & Layer Norm** combines a shortcut connection (Add) and normalization (Layer Norm). The Add step passes the original input forward to keep useful information. Layer Norm adjusts the output values to a stable range so the model can learn more reliably.

3. Each layer builds on the output of the previous one, helping the model form more complex and abstract ideas about the input sentence.

4. Finally, the encoder outputs a sequence of *hidden states*. These are continuous vector representations for each input token. They encode contextual information from the entire sentence. For example, in the sentence "The cat sat on the mat," the vector for "cat" reflects its relationship to words like "sat" and "mat."

The **decoder** generates the output sentence one word at a time by using the information from the encode (Xiao and Zhu 2023). However, since BERT uses only an **encoder-only architecture**, the decoder is not relevant for this work and is therefore excluded from the discussion.

### 2.4.3 Language model: BERT

*This entire subsection summarizes the input representation methodology from Devlin et al. (2019).*

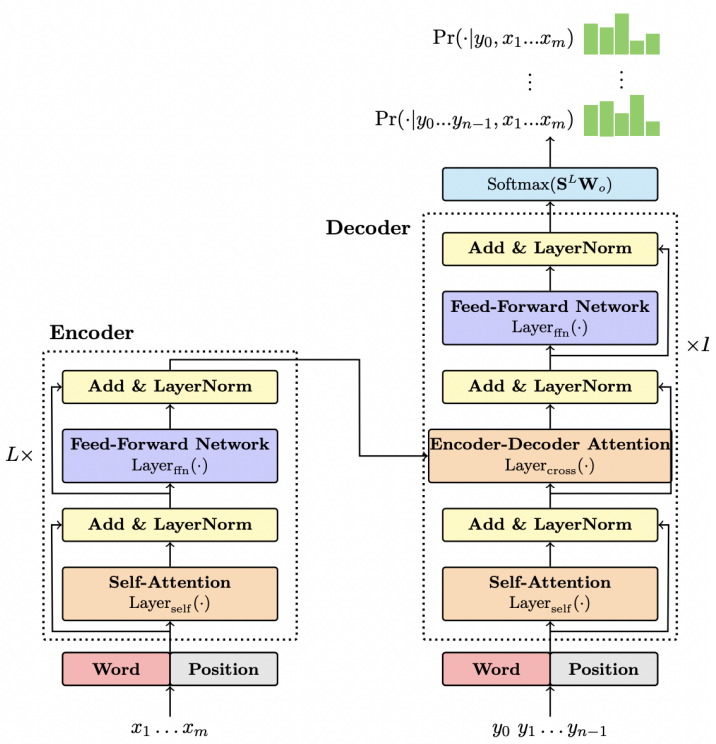


Figure 2.3: Transformer encoder-decoder architecture. The encoder (left) processes input tokens $x_1, \ldots, x_m$ through: (1) a self-attention layer for contextual relationships, (2) a feed-forward network for feature transformation, and (3) residual connections with layer normalization. The decoder (right) generates outputs by attending to both the encoder's representations and its previous outputs ($y_0$ to $y_{n-1}$), producing the next-token probability distribution. Figure and description adapted from Xiao and Zhu (2023), p. 6.

BERT is a language model that stands for "Bidirectional Encoder Representations from Transformers" and was introduced by Google in 2018 (Devlin et al. 2019). After pre-training, BERT can be adapted to many NLP tasks by adding a simple output layer and fine-tuning, without needing major changes to its design.

### Model architecture

As explained in subsection 2.4.2, BERT uses an **encoder-only architecture** (see Figure 2.4). It processes and understands input thoroughly but does not generate new text. That makes BERT **well suited for binary classification tasks**, since it can analyze each word's meaning and decide accurately between two categories.

It uses a multi-layer bidirectional Transformer encoder. Multi-layer means it stacks 12 encoder layers that refine the information step by step. Bidirectional means the model reads

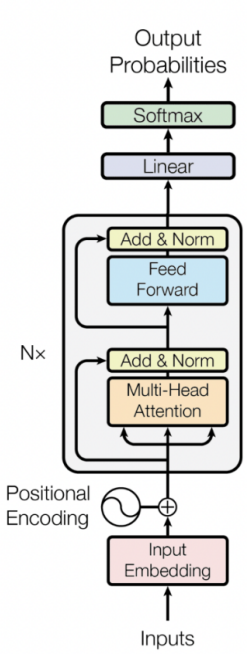


Figure 2.4: BERT's encoder-only architecture. Figure by Smith (2024).

both the words before and after a given word at the same time. This is a key improvement over earlier models, which could only read text in one direction, usually from left to right.

### Input and Output Representations

BERT processes an input by splitting whole words or subword units into *tokens*. This allows the model to handle rare and compound words. For example:

`unbelievable` → `un`, `##believable`

Special tokens are reserved tokens added to input sequences to indicate boundaries or roles, helping the model distinguish parts of the text and process it correctly.

- `[CLS]` (classification) marks the start of the sequence,
- `[SEP]` separates sentence pairs.

In this work, each input combines an English source sentence and its German translation as:

`[CLS] english sentence [SEP] german translation [SEP]`

After processing, BERT outputs a hidden vector for each token. The hidden vector for the `[CLS]` token represents the entire sequence and is used for tasks like classification. For example, for the input sequence:

`[CLS] the nurse is kind [SEP] die krankenschwester ist nett [SEP]`

BERT creates a vector for each token:

$$\text{Vector}([CLS]), \text{Vector}(\text{the}), \text{Vector}(\text{nurse}), \ldots, \text{Vector}([SEP])$$

The `[CLS]` vector aggregates the meaning of both sentences and feeds into the classifier for bias detection.

### Pre-training and Fine-tuning

BERT was pre-trained on a large corpus of unlabeled text (Devlin et al. 2019). Since this data contains no manual labels, the model learns to understand language by using patterns in the data to learn token relationships. The result is a task-agnostic base model. In Figure 2.4, this base model includes all components below the purple linear layer.[1]

**Fine-tuning adjusts this base model for a specific task**, in this case, detecting gender bias in translations. Since this thesis focuses on the fine-tuning process, the following explanation provides more technical background for this part of the model.

A task-specific **classification head**, comprising a linear layer followed by a softmax function, is added on top of BERT's output. The **linear layer** applies a learned transformation to the final hidden state vector of the `[CLS]` token.

$$z = Wx + b$$

Here, $x$ is the `[CLS]` embedding, $W$ is the weight matrix, and $b$ is the bias vector. Both $W$ and $b$ are parameters learned during training to help map BERT's output to the task labels. This changes BERT's output into two numbers (logits), one for each class: biased or neutral. Then, **softmax** turns these numbers into probabilities:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

[1] The explanation in this section is based on a prompt-response pair generated using the Perplexity.ai search engine. The original prompt and result are included as a PDF in the appendix.

Each logit $z_i$ is exponentiated to ensure positivity. The result is then normalized by dividing by the sum of all exponentials, producing a probability distribution over the classes. $K$ is the number of possible classes. The class with the highest probability is selected as the model's prediction.

**Example: Predicting Gender Bias with Softmax**

When the model processes the sentence pair "the nurse is kind" / "die krankenschwester ist nett," it creates a `[CLS]` embedding summarizing the input. This embedding captures features such as word choice, syntactic structure, and semantic associations.

The linear layer then transforms this embedding into two logits, for example, $[2.0, 1.0]$. The first logit corresponds to the "biased" class, and the second to "neutral". These values are not manually assigned but result from learned weights in the linear layer. During training, the model adjusts these weights to associate certain input patterns with higher scores for one class. In this case, the presence of a gendered profession ("nurse") and a feminine translation ("krankenschwester") may lead the model to assign a higher logit to the "biased" class, based on patterns seen in the training data.

To convert these logits into probabilities, the softmax function is applied:

$$\text{softmax}([2.0, 1.0]) = \left[\frac{e^2}{e^2 + e^1}, \frac{e^1}{e^2 + e^1}\right] \approx \left[\frac{7.39}{7.39 + 2.72}, \frac{2.72}{7.39 + 2.72}\right] = [0.73, 0.27]$$

Here, $K = 2$ because there are two classes. The model assigns a 73% probability to the "biased" class and 27% to "neutral". Since the "biased" probability is higher, the model predicts the translation likely contains gender bias.

### 2.4.4 Multilingual BERT

There are multiple variants of the original BERT model. Even the standard version was released in two sizes: BERT-Base and BERT-Large, which differ in the number of layers, attention heads, and overall model capacity (Devlin et al. 2019). Since then, many other versions have been developed. Most of them modify either BERT's pre-training objectives or the underlying Transformer architecture (Libovický et al. 2019).

For this thesis, I use multilingual BERT **(mBERT)** (Devlin et al. 2019). mBERT uses the same configuration as BERT-Base, but it is pretrained on Wikipedia data from 104 languages, including both English and German. There is no explicit indication of the

input language, nor is there a training objective that aligns languages bilingually; instead, multilingual capabilities emerge naturally from training on a large multilingual text corpus (Pires et al. 2019).

Monolingual models like German BERT do not support English input. Larger multilingual models, such as XLM-RoBERTa, require more computational resources and training time, which was not feasible here. mBERT offers a good balance between language coverage, model size, and training efficiency, making it a practical choice detecting gender bias in EN-DE translations.

### 2.4.5 Interactive Demo

The fine-tuned model is intended to be presented through an interactive demonstration. Since the focus lies on showcasing the model's functionality rather than creating a fully developed application, Streamlit was chosen. Streamlit allows for quick and easy development of lightweight user interfaces in Python, providing a simple setup and effective performance.

For live translation, an open-source tool supporting EN-DE pairs was required. Opus-MT (Tiedemann and Thottingal 2020) meets these criteria and integrates smoothly into the demonstration. While state-of-the-art translators like Google Translate or DeepL would have been preferred for their quality, they do not meet the requirements for this setup. Therefore, a separate tab for manual translation input was added, allowing users to paste translations directly and bypass this limitation.

# 3 Methodology

This chapter explains the overall approach and structure of the project. It covers how data is handled, how the model is built and trained, and how the demo application is designed.

## 3.1 Goal of the project

The goal is to build a gender bias detection model for real-world MT scenarios. This includes cases like translating everyday sentences or job descriptions. The focus is to flag bias at the sentence level, so users do not have to find the specific sentences causing bias themselves.

Thus, the model processes each sentence independently. If multiple sentences are inputted, bias is evaluated for each one separately. Context across sentences is not considered, as it does not reflect the intended use case. This approach is also reflected in the design of the training data, where each sentence pair is treated as a standalone instance.

## 3.2 Workflow

The project begins by selecting and combining datasets from previous work (see Figure 3.1). The model building phase then follows, as shown in the purple boxes. It starts with cleaning and preparing the data, followed by extracting features for training. A pre-trained German BERT model is then fine-tuned for the classification task. Its performance is measured using standard evaluation metrics. In the final step, the trained model is integrated into a demo application for user interaction and testing.

## 3.3 Dataset Handling

Since there was no ready-to-use dataset for this task and no prior work that built a similar model, I first had to define: **(1)** the number of samples required, and **(2)** the desired structure and content of my dataset.

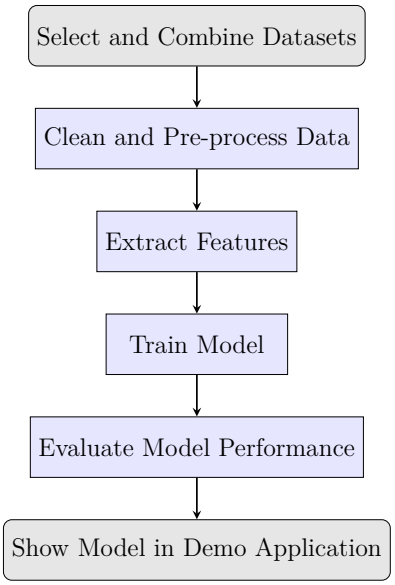


Figure 3.1: Workflow of the project.

**(1)** For a binary classification task of detecting gender bias using `mBERT`, general guidelines suggest between 100 and 5000 labeled samples for fine-tuning (Pecher et al. 2024). However, the complex nature of gender bias often requires a larger dataset for robust detection since the number of samples depends mainly on the task type. Multi-class tasks need fewer samples (around 100), while binary tasks can require up to 6,000 or more. **I will therefore aim to create a dataset with 5000-6000 samples.**

**(2)** Ideally, I wanted to make use of past EN-DE datasets to minimize manual labour. My options were `mGeNTE en-de` (Savoldi et al. 2025b), `Building Bridges Dictionary` (Lardelli et al. 2024), and `Translated Wikipedia Biographies` (Stella et al. 2021).

While analyzing the `Translated Wikipedia Biographies` dataset, I found issues that prevented automatic reuse. For example, the `perceivedGender` column sometimes contained subject names instead of expected labels like Male, Female, or Neutral, which would require manual review. Moreover, the dataset only provided neutral labels (0) since the phrases were correctly gendered. Because my other two datasets are already balanced and include enough neutral examples, I decided to exclude this dataset.

The `mGeNTE` dataset was suitable for automatic reuse. It includes columns `SET-G`, `REF-G`, and `REF-N`. The `SET-G` column indicates if the source sentence has a gendered subject. For each entry with a gendered source (`SET-G`), I add two rows: one with `REF-G` as the correctly gendered (neutral) label, and one with `REF-N` as a neutral translation that loses the original gender, making it biased. For source sentences without gendered subjects (`SET-N`), I apply

| Dataset | Description | Content |
|---|---|---|
| `mGeNTE en-de` (Savoldi et al. 2025b) | Multilingual dataset to assess gender bias in MT. | ~1,500 gender-ambiguous and gendered English sentences with gender-neutral and gendered German translations. |
| `Building Bridges Dictionary` (Lardelli et al. 2024) | Bilingual dictionary designed to support gender-fair EN-DE translation. | ~230 German gender-neutral and gender-inclusive singular and plural sentences with English equivalents. |
| `Translated Wikipedia Biographies` (Stella et al. 2021) | Automatically translated Wikipedia biographies for evaluating gender bias. | ~1,500 translated biography sentences with English source text and gender-accurate translated German equivalent. |

Table 3.1: Overview of available EN-DE datasets based on past works.

the same approach but reversed: one row with the neutral source and a correctly gendered translation as biased. This method creates two rows for each original entry, resulting in about 3,000 balanced instances of biased and neutral sentences.

`Building Bridges Dictionary` did not contain full sentences but rather a gender-fair dictionary of nouns. While this made it a valuable resource for studying gender-fair language, I needed full sentences for my task. To address this, I used prompt engineering with Google Gemini 2.5 Flash to synthetically expand the dataset. The prompt used for generation is included in the appendix. The generated sentences can be found in the code files. I used the nouns from the original dataset to create multiple grammatically correct sentence variations, covering singular, plural, gender-neutral, and gender-inclusive forms. This process resulted in 2,718 sentences.

### 3.3.1 Final Dataset Composition

The final dataset used for training is a combination of the transformed `mGeNTE` dataset and the synthetically extended `Building Bridges Dictionary` dataset. Each source contributes 2,600 sentence pairs, split evenly into 1,300 biased and 1,300 neutral examples, resulting in a balanced dataset of 5,200 entries. The dataset is shuffled before training to prevent

source-specific bias. Each entry contains:

- an English source sentence (`english`),
- its German translation (`german`),
- and a binary label (`label`), where `1` denotes gender bias and `0` denotes neutrality.

A sample of the dataset is shown below:

| English | German | Label |
|---|---|---|
| The laypeople are intelligent. | Die Laiinnen sind intelligent. | 1 |
| The ceramist is responsible. | Der Keramiker ist verantwortlich. | 1 |
| The forest keepers are responsible. | die Forstwart*innen sind verantwortlich. | 0 |
| I am sorry that the Commissioner responsible for agriculture is not here and that she does not have the courage to face us. | Ich bedauere, dass die für die Landwirtschaft zuständige Kommissarin nicht anwesend ist und nicht den Mut hat, uns gegenüberzutreten. | 0 |
| Mrs Brok, I have no problem whatsoever with this motion. | Ich habe überhaupt kein Problem mit diesem Antrag des Kollegiumsmitglieds Brok. | 1 |

Table 3.2: Sample entries from the final training dataset.

## 3.4 Data Pre-processing

## 3.5 Feature extraction

## 3.6 Model Selection and Training

### 3.6.1 Hyperparameters

## 3.7 Evaluation

## 3.8 Demo Application Design

# 4 Implementation

## 4.1 Project Structure

## 4.2 Environment Setup

## 4.3 Core components and their interaction

## 4.4 Demo Result

# Bibliography


Baldi, P. (2008). English as an Indo-European Language. In Momma, H. and Matto, M., editors, *A Companion to the History of the English Language*, pages 127–141. Wiley, 1 edition.

Barclay, P. J. and Sami, A. (2024). Investigating Markers and Drivers of Gender Bias in Machine Translations.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*.

Chakravarthi, B. R., Rani, P., Arcan, M., and McCrae, J. P. (2021). A Survey of Orthographic Information in Machine Translation. *SN Computer Science*, 2(4):330.

Cho, W. I., Kim, J. W., Kim, S. M., and Kim, N. S. (2019). On Measuring Gender Bias in Translation of Gender-neutral Pronouns.

DeepL (2021). How does DeepL work? https://www.deepl.com/en/blog/how-does-deepl-work.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Godsil, R. D., Tropp, L. R., Goff, P. A., Powell, J. A., and MacFarlane, J. (2016). The Effects of Gender Roles, Implicit Bias, and Stereotype Threat on the Lives of Women and Girls. *THE SCIENCE OF EQUALITY*, 2(Perception Institute).

Google (2018). Reducing gender bias in Google Translate. https://blog.google/products/translate/reducing-gender-bias-google-translate/.

Kappl, M. (2025). Are All Spanish Doctors Male? Evaluating Gender Bias in German Machine Translation.

Lardelli, M., Attanasio, G., and Lauscher, A. (2024). Building Bridges: A Dataset for Evaluating Gender-Fair Machine Translation into German. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7542–7550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Libovický, J., Rosa, R., and Fraser, A. (2019). How Language-Neutral is Multilingual BERT?

Lin, G. H.-c. and Chien, P. S. C. (2009). Machine Translation for Academic Purposes. *Proceedings of the International Conference on TESOL and Translation 2009*, pages pp.133–148.

Pecher, B., Srba, I., and Bielikova, M. (2024). Comparing Specialised Small and General Large Language Models on Text Classification: 100 Labelled Samples to Achieve Break-Even Performance.

Phuong, M. and Hutter, M. (2022). Formal Algorithms for Transformers.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is Multilingual BERT?

Prates, M. O. R., Avelar, P. H. C., and Lamb, L. (2019). Assessing Gender Bias in Machine Translation – A Case Study with Google Translate.

Quemy, A. (2019). Binary Classification in Unstructured Space With Hypergraph Case-Based Reasoning.

Rescigno, A. A. and Monti, J. (2023). Gender Bias in Machine Translation: A statistical evaluation of Google Translate and DeepL for English, Italian and German. In *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*, pages 1–11, UNIOR NLP Research Group, University of Naples "L'Orientale", Naples, Italy. INCOMA Ltd., Shoumen, Bulgaria.

Savoldi, B., Bastings, J., Bentivogli, L., and Vanmassenhove, E. (2025a). A decade of gender bias in machine translation. *Patterns*, page 101257.

Savoldi, B., Cupin, E., Thind, M., Lauscher, A., Piergentili, A., Negri, M., and Bentivogli, L. (2025b). mGeNTE: A Multilingual Resource for Gender-Neutral Language and Translation.

Savoldi, B., Papi, S., Negri, M., Guerberof-Arenas, A., and Bentivogli, L. (2024). What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Schiebinger, L. (2014). Scientific research must take gender into account. *Nature*, 507(7490):9–9.

Schmitz, D. (2022). In German, all professors are male.

Schryen, G. (2015). Writing Qualitative IS Literature Reviews—Guidelines for Synthesis, Interpretation, and Guidance of Research. *Communications of the Association for Information Systems*, 37.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., and Lockhart, J. W. (2020). Diagnosing Gender Bias in Image Recognition Systems. *Socius*, 6:2378023120967171.

Sczesny, S., Formanowicz, M., and Moser, F. (2016). Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7.

Shah, D., Schwartz, H. A., and Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264.

Shrestha, S. and Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in Artificial Intelligence*, 5:976838.

SkyQuest (2025). Machine Translation (MT) Market Size, Growth & Trends Report | 2032. https://www.skyquestt.com/report/machine-translation-market.

Smacchia, M., Za, S., and Arenas, A. (2024). Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools. In Braccini, A. M., Ricciardi, F., and Virili, F., editors, *Digital (Eco) Systems and Societal Challenges*, volume 72, pages 355–373. Springer Nature Switzerland, Cham.

Smith, B. (2024). A Complete Guide to BERT with Code.

Soundararajan, S. and Delany, S. J. (2024). Investigating Gender Bias in Large Language Models Through Text Generation. *Association for Computational Linguistics*, Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024):410–424.

Stanczak, K. and Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation.

Stella, R., Austermann, A., Johnson, M., Linch, M., Niu, M., Pushkarna, M., Shah, A., and Webster, K. (2021). A Dataset for Studying Gender Bias in Translation. https://research.google/blog/a-dataset-for-studying-gender-bias-in-translation/.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – Building open translation services for the World. *European Association for Machine Translation*, Proceedings of the 22nd Annual Conference of the European Association for Machine Translation:479–480.

Ullmann, S. (2022). Gender Bias in Machine Translation Systems. In Hanemaayer, A., editor, *Artificial Intelligence and Its Discontents*, pages 123–144. Springer International Publishing, Cham.

United Nations (2023). Achieve Gender Equality And Empower All Women and Girls. https://sdgs.un.org/goals/goal5.

Waldendorf, A. (2024). Words of change: The increase of gender-inclusive language in German media. *European Sociological Review*, 40(2):357–374.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Xiao, T. and Zhu, J. (2023). Introduction to Transformers: An NLP Perspective.

# Appendix

## 1 Prompt and Output for Pre-training/Fine-tuning Explanation

perplexity

## Linear and Softmax Layers in BERT Fine-tuning: Formulas and Functions

In BERT fine-tuning, the **linear and softmax layers** form the classification head that processes the final hidden representations from the transformer to produce predictions. Here's a comprehensive explanation of their mathematical formulas and functions:

### Linear Layer (Classification Head)

The linear layer is a fully connected layer that transforms the BERT output representation into class scores.

**Formula:**
$ z = W \cdot h + b $

Where:

- $ z $ is the output logits (raw prediction scores)
- $ W $ is the weight matrix with dimensions $ num_classes \times hidden_size $
- $ h $ is the hidden representation from BERT (typically the [CLS] token representation)
- $ b $ is the bias vector with dimensions $ num_classes $

**What it does:**

- **Transforms representations**: Converts the high-dimensional BERT output (usually 768 dimensions for BERT-base) into class-specific scores[1] [2]
- **Learns task-specific mappings**: The weight matrix $ W $ learns which features in the BERT representation are most important for each class
- **Provides linear transformation**: Each output neuron computes a weighted sum of all input features plus a bias term

### Softmax Layer (Activation Function)

The softmax function converts the raw logits from the linear layer into probability distributions.

**Formula:**
$ softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} $

Where:

- $ z_i $ is the logit for class $ i $
- $ K $ is the total number of classes
- $ e $ is the exponential function

**What it does:**

- **Normalizes outputs**: Ensures all class probabilities sum to 1[3] [4]
- **Amplifies differences**: Uses exponential function to emphasize higher-scoring classes
- **Enables probabilistic interpretation**: Converts raw scores into meaningful probabilities for decision-making

### Combined Process in BERT Fine-tuning

The complete classification pipeline follows this sequence:

1. **BERT Processing**: Input text → BERT layers → [CLS] token representation $ h $
2. **Linear Transformation**: $ z = W \cdot h + b $
3. **Softmax Activation**: $ p = softmax(z) $
4. **Prediction**: $ \hat{y} = \arg\max(p) $

### Training Process

During fine-tuning, these layers are optimized using:

- **Cross-entropy loss**: $ L = -\sum_{i=1}^{K} y_i \log(p_i) $
- **Backpropagation**: Gradients flow back through softmax → linear → BERT layers
- **Parameter updates**: Both the linear layer weights and BERT parameters are updated

### Practical Implementation

In practice, the linear and softmax layers work together as follows[2] [5]:

```
# Conceptual implementation
linear_output = torch.matmul(bert_output, weight_matrix) + bias
probabilities = torch.softmax(linear_output, dim=-1)
```

### Key Functions

**Linear Layer Functions:**

- Feature extraction and dimensionality reduction
- Task-specific weight learning
- Bias adjustment for class imbalances

**Softmax Layer Functions:**

- Probability normalization
- Confidence scoring
- Multi-class decision making

### Variations and Enhancements

Recent research has explored alternatives to the standard linear+softmax approach:

- **Multiple linear layers**: Adding depth to the classification head[2]
- **Attention mechanisms**: Using attention to weight different parts of the representation[4]
- **Alternative activation functions**: Exploring functions beyond softmax for specific tasks[6]

The linear and softmax layers, while conceptually simple, are crucial for translating BERT's rich contextual representations into task-specific predictions. Their mathematical simplicity belies their importance in the overall fine-tuning process, serving as the bridge between pre-trained language understanding and downstream task performance.

⁂


1. https://www.semanticscholar.org/paper/b964afe5b755022f1f1e6915d23df9a7f65c911c
2. https://ieeexplore.ieee.org/document/10742347/
3. https://www.mdpi.com/1424-8220/23/3/1481
4. https://journals.sagepub.com/doi/10.3233/KES-230066
5. https://www.mdpi.com/2078-2489/14/8/467
6. https://arxiv.org/abs/2408.08803

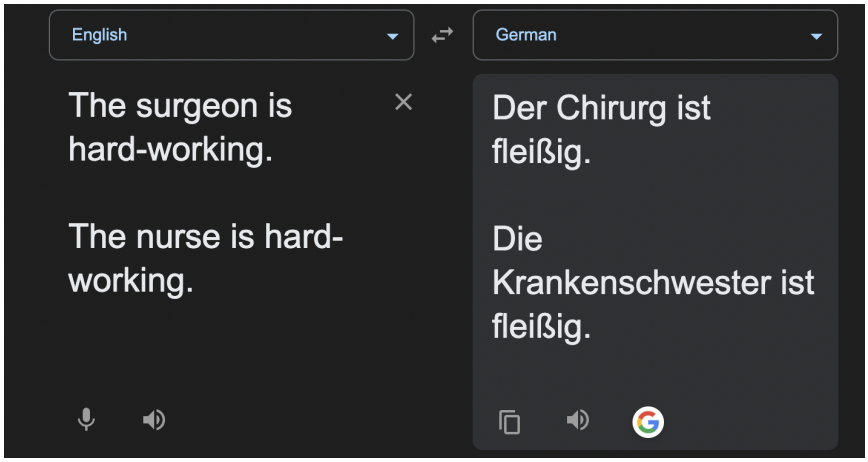


Figure 1: Google Translate assigns stereotypical genders to occupational roles.

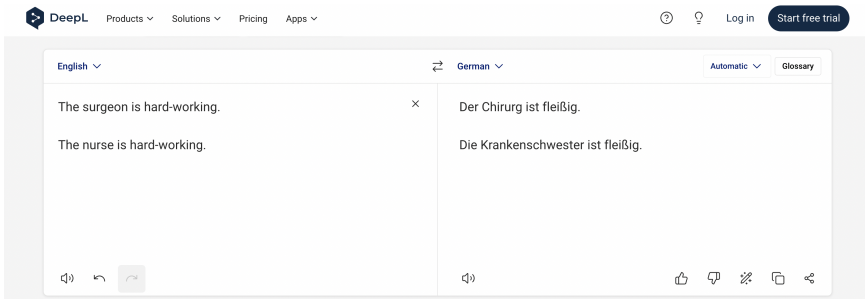


Figure 2: DeepL shows a similar bias in the same sentence, highlighting consistent patterns across MT tools.

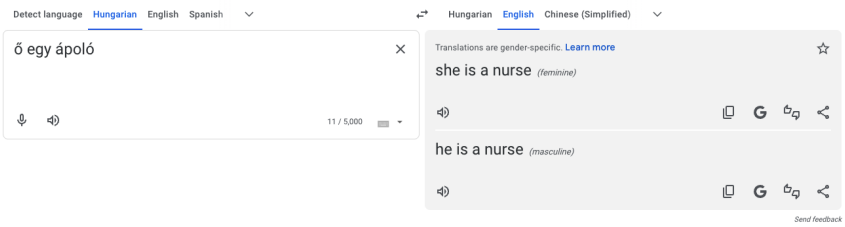


Figure 3: Gender-specific translation by Google Translate for ambiguous pronouns.

1. Hiermit versichere ich,

   - dass ich die von mir vorgelegte Arbeit selbständig abgefasst habe,
   - dass ich keine weiteren Hilfsmittel verwendet habe als diejenigen, die im Vorfeld explizit zugelassen und von mir angegeben wurden,
   - dass ich die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen und KI-basierte Tools) entnommen sind, unter Angabe der Quelle kenntlich gemacht habe und
   - dass ich die vorliegende Arbeit noch nicht für andere Prüfungen eingereicht habe.

2. Mir ist bewusst,

   - dass ich diese Prüfung nicht bestanden habe, wenn ich die mir bekannte Frist für die Einreichung meiner schriftlichen Arbeit versäume,
   - dass ich im Falle eines Täuschungsversuchs diese Prüfung nicht bestanden habe,
   - dass ich im Falle eines schwerwiegenden Täuschungsversuchs ggf. die Gesamtprüfung endgültig nicht bestanden habe und in diesem Studiengang nicht mehr weiter studieren darf und
   - dass ich, sofern ich zur Erstellung dieser Arbeit KI-basierter Tools verwendet habe, die Verantwortung für eventuell durch die KI generierte fehlerhafte oder verzerrte (bias) Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate trage.

Berlin, den July 15, 2025

.....

*(Unterschrift des Verfassers)*