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

List of Tables

1 Introduction

Machine Translation (MT) helps millions of people communicate across languages, in daily life and in areas like healthcare, law, and business (**?**). Services like Google Translate handle over 200 million users every day (**??**). It is a fast-growing market. A report by **?** 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 (**???**).

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 (**???**). 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 (**?**). 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 (**??**).

Language also shapes thought. Research shows that readers often interpret masculine forms as male-specific, even if they are supposed to be generic (**?**). 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 (**???**). Common methods include counting gendered forms in outputs and comparing them to demographic baselines or human expectations (**???**). 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, **?** 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 (**?**). These outputs

often reflect social stereotypes rather than objective translations, yet current systems offer no mechanism to detect or signal when such bias occurs (**?**). 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 (**?**). 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 (**?**). 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 **?** 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 (**?**). 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.

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 **?** and was further informed by **?** and **?**, who both conducted systematic reviews on gender bias in ML and MT respectively.

2.1 Literature Search Process

2.1.1 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.

2.1.2 Literature Review Framing

To answer the four research aims, I have defined the key concepts in **??**. 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.

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

2.1.3 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 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.

2.1.4 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 Understanding Gender Bias in English-to-German Machine Translation

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.

2.2.1 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 (**??**). 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 (**?**). 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 (**?**). 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)** (**??**). 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.

2.2.2 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" (**??**). **?**, as described by **?**, 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").

? 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.

2.2.3 What Gender Bias Means in Machine Translation

A clear definition of gender bias has not yet been established (**?**). 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 (**???**).

Since there is no clear definition, this work defines gender bias based on specific manifestations described in the following subsection. Any text that exhibits one or more of these forms will be considered gender biased.

2.2.4 Manifestiations of Gender Bias

This section draws from the main studies analyzing gender bias in EN-DE MT (**????**). 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 (**??**), 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 (**?**). In MT, the generic masculine can lead to inaccurate or unfair representations of gender in translated text. **?** 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 (**?**). These associations can then be reinforced by popular media, such as TV and advertisements (**?**), 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 (**?**), with an even stronger tendency in male-domniated fields such as STEM (**?**). 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 (**?**).

Neglecting Contextual Information

Coreference resolution refers to the process of using contextual information to determine the correct gender in translation (**?**). 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 (**?**), MT systems often fail to do so reliably, especially when gender information appears earlier in the text or across sentence boundaries (**??**). 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.

? 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, **?** highlighted that correcting biased translations toward feminine forms required significantly more time and edits than masculine ones, revealing a notable cost disparity.

Similarly, **?** 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.

2.2.5 Linguistic Challenges in English-German Translation

Although both English and German originate from the Indo-European language family (**?**), 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.

2.2.6 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 (**?**). 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 (**?**).
- **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 (**?**).
- **Pair form:** This names both gender forms, e.g., *die Studentinnen und Studenten lernen*. It is currently the most used GFL form in German (**?**), briefly surpassing the star and colon characters as seen in **??**.

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

2.3 Societal Relevance and Impact of Gender Bias in Machine Translation

This section outlines why gender bias is a subject of research in the first place and where it connects to broader social and ethical questions. It first looks at early studies that brought

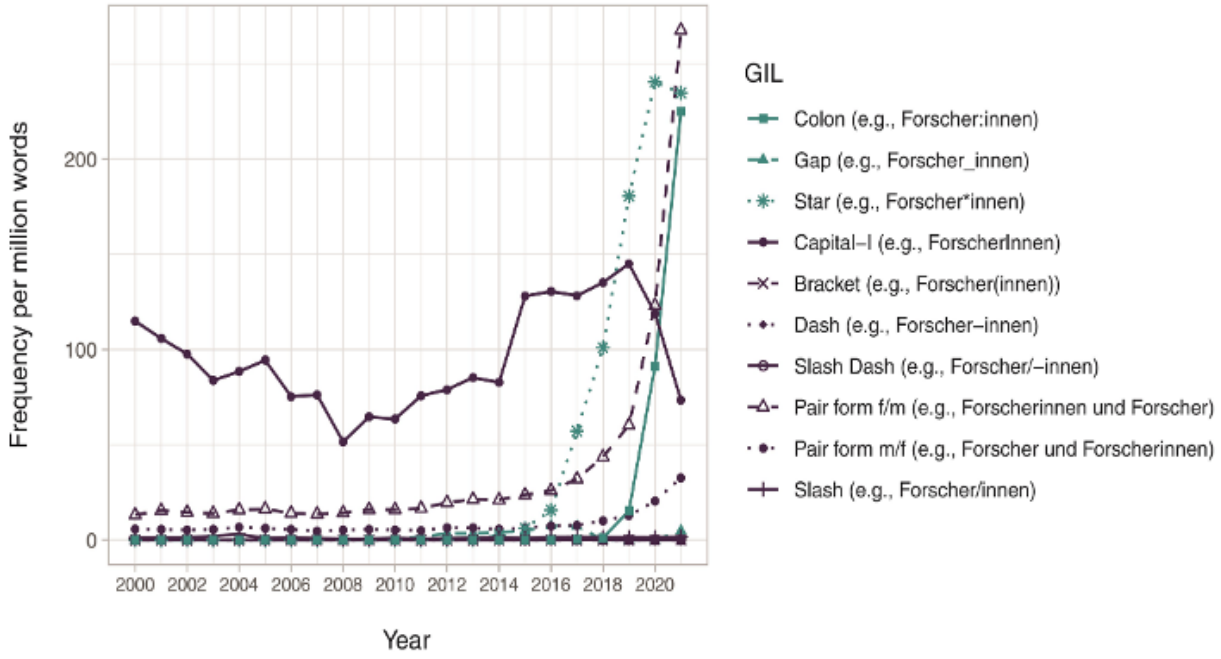

Figure 2.1: Frequency of different types of gender-inclusive language. Source: **?** p. 367.

attention to gender patterns in language technologies and raised awareness of their social impact. Understanding these origins helps explain why it continues to be relevant today.

2.3.1 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 **?** in 2014. Since then, there has been a general increase in research papers focusing on this topic, especially between 2019 and 2023 (**?**).

? 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 **?** to change their translation policy: Google Translate began showing both feminine and masculine forms for ambiguous inputs (**?**) (see **??**).

Building on this, **?** 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 **?**. Together, these studies helped spark the ongoing research interest in gender bias in MT.

2.3.2 Why it matters

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

It also contributes to the invisibility of women in male-dominated professions (**?**). Studies show that biased language in machine-generated text, such as children's stories or job ads, can **influence how young people view themselves** (**??**). It may shape their interests, hobbies, and career choices. This is especially visible in STEM fields (**?**), 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 (**?**). 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 (**?**). Using inclusive language can therefore offer both social and competitive benefits for companies.

Furthermore, a study by **?** 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 (**??**). Consequently this effort gap also translates into **higher economic costs**, suggesting a measurable **quality-of-service disadvantage that disproportionately affects women**. **?** 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.

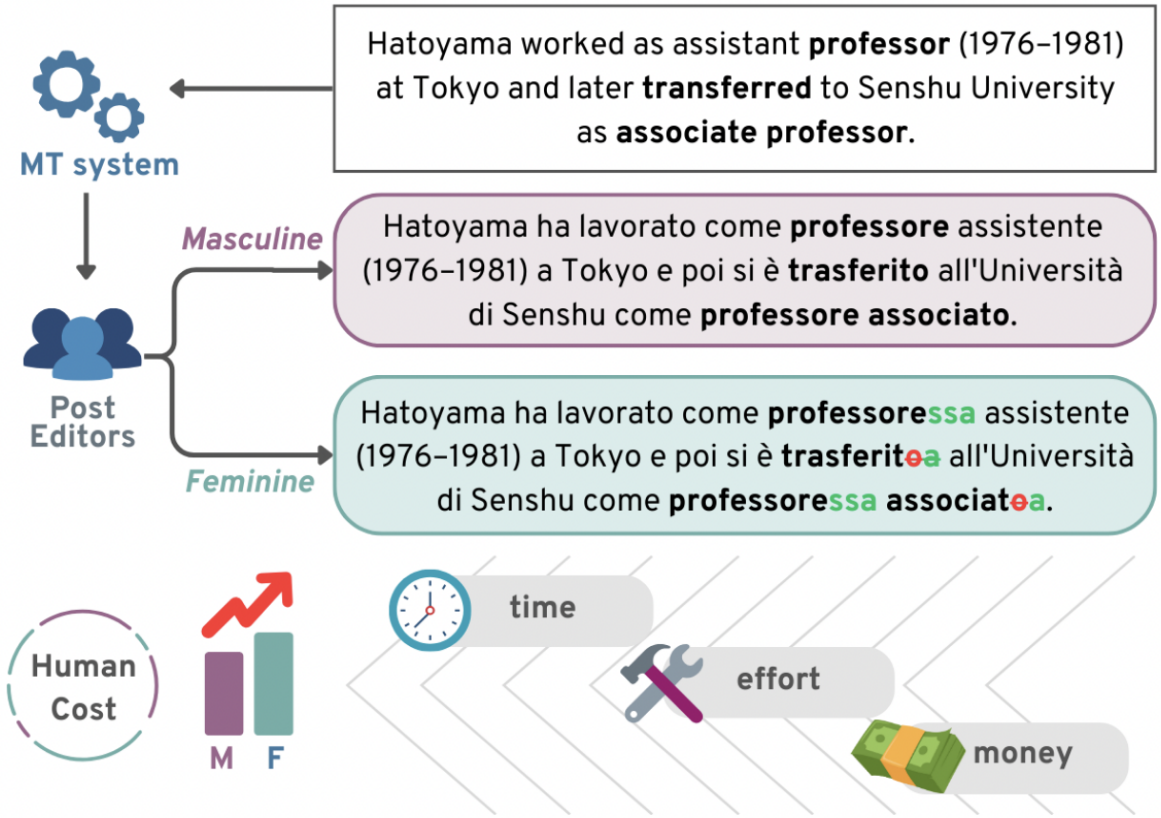

Figure 2.2: Study design. Post-editing of an MT output into both feminine and masculine gender. Source: **?** p. 18048

2.4 Research Gaps

2.5 Approach and Justification of the Technical Setup

2.5.1 Binary Classification in NLP

Binary Classification means categorizing whether an element belongs into a distinct class. 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" (**?**). The ML algorithms use information from past examples to create a model or find key rules for making correct decisions. **This thesis attempts to label a translation as "potentially gender biased" or "neutral"**.

As explained in **??**, since there is no clear definition of what counts as gender biased, there is also no clear definition of what is unbiased. I have set rules for what I consider biased, but that does not mean everything else is automatically unbiased. For this reason, I use the term "neutral" instead.

2.5.2 Transformer Architecture

Transformers are a neural network architecture that incorporate a **self-attention mechanism** (**?**). This mechanism lets the model capture relationships between all elements simultaneously, meaning it looks at all the words in a sentence at once to better grasp the general meaning. They're commonly used for NLP tasks. For example, in the sentence "The dog chased the ball because it was fast," the model learns that "it" refers to "the dog" by attending to the relevant parts of the sentence.

The transformer architecture follows the general **encoder-decoder** framework. **?? ?** gives an overview of its components.

The encoder on the left hand side is responsible for processing the input sentence and converting it into a sequence of vector representations that capture the meaning the input. The input text (bottom left side) is split into tokens (like words or subwords) are first mapped to vectors—numerical representations that carry semantic meaning. Since the model has no inherent sense of word order, positional encodings are added to these vectors. These encodings provide information about the position of each token in the sequence, allowing the model to understand word order.

The encoder consists of a stack of multiple identical layers. Each layer takes the output of the previous one and refines it further, allowing the model to gradually build up more complex and abstract representations of the input.

Each layer contains two key sub-layers. The first is a multi-head self-attention mechanism. This mechanism allows each word to focus on other words in the same sentence and weigh their importance. The term "multi-head" means this process is done in parallel several times, with each head learning to capture different kinds of relationships.

The second sub-layer is a position-wise feed-forward network, which is a simple neural network applied to each token vector separately. It helps transform and refine the information at each position.

To improve training and preserve important information, the model uses residual connections. These are shortcut paths that skip each sub-layer by adding its input directly to its output. After this addition, the result is normalized, a step referred to as Add Norm. This structure helps the model learn better and prevents certain training issues.

At the top of the encoder stack, the final output is a sequence of context-rich vector representations. These are passed to the decoder, which uses them to generate the output.

The decoder (right side of Figure 1) generates the output sentence one token at a time. It also begins with input embeddings and positional encodings. However, the input here is the target sequence—the output sentence the model is trying to generate. These tokens are shifted to the right, meaning that when predicting a word, the model only sees the words that came before it. This ensures that predictions do not rely on future tokens, which would break the natural flow of generation.

Like the encoder, the decoder is made up of six identical layers. Each decoder layer has three sub-layers. The first is a masked multi-head self-attention mechanism. It works like the encoder's self-attention but includes a mask to block information from future positions. This ensures that each word is predicted based only on earlier words in the sequence.

The second sub-layer is the encoder-decoder attention mechanism. Here, the decoder uses the encoder's output to find relevant parts of the input sentence. This is done using a query-key-value structure: the decoder's output so far is used as queries, while the encoder's output provides the keys and values. This allows the decoder to focus on the most important parts of the input when generating each word.

The third sub-layer is again a position-wise feed-forward network, identical in function to the one used in the encoder.

After passing through all six decoder layers, the final output is processed by a linear layer and a softmax function. This converts the decoder's output into a probability distribution over the vocabulary. The word with the highest probability is selected as the next token in the output sequence.

This combination of attention, layered processing, and position-aware inputs allows the

Transformer model to understand context, maintain order, and generate accurate outputs without relying on recurrence or convolution.

2.5.3 Pre-trained Language Model: BERT

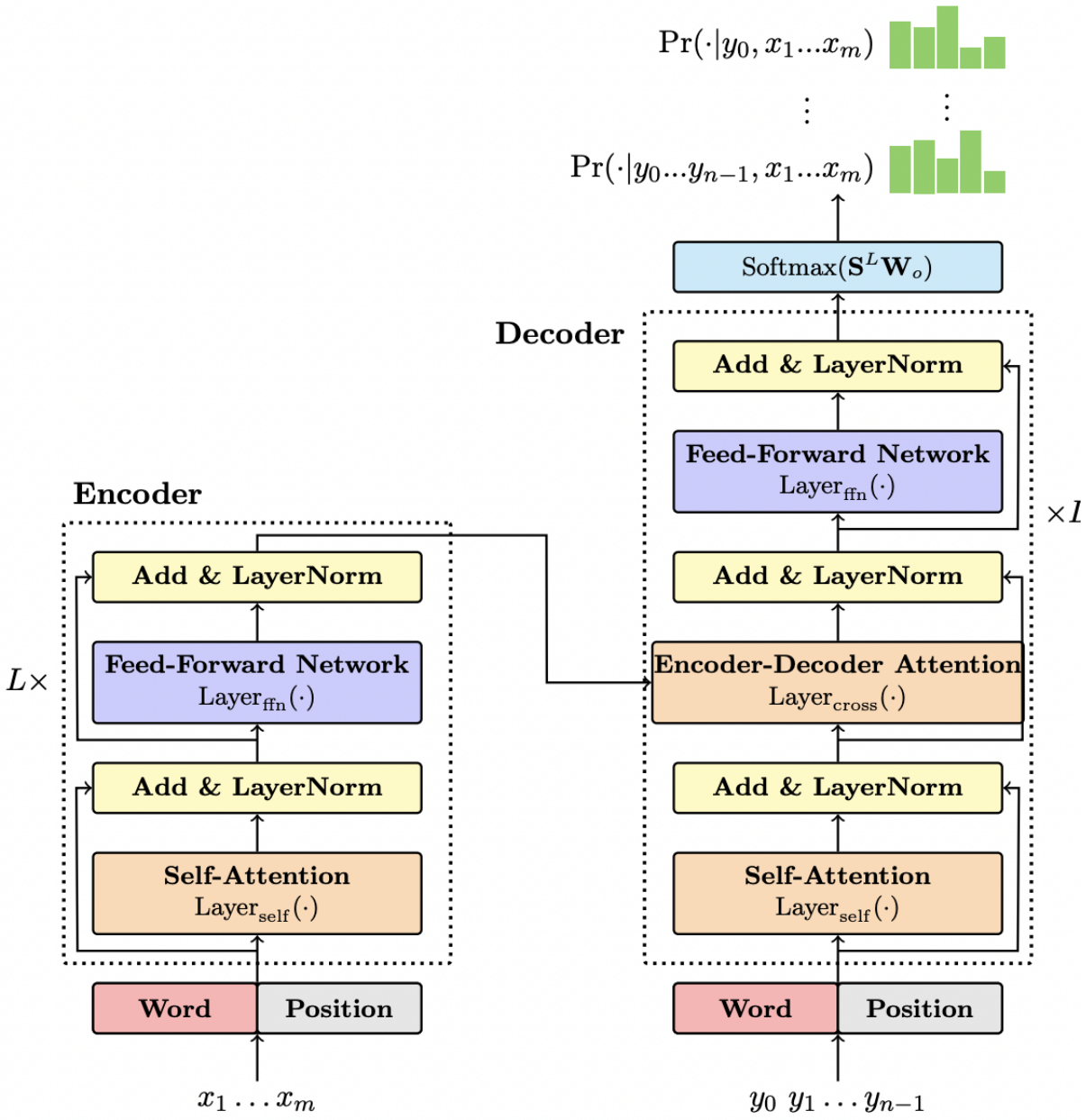

Figure 2.3: Architecture of the Transformer model, composed of an encoder and a decoder. Source: **?** p. 6.

3 Conceptual Framework

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.

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.

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

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. (2025). A decade of gender bias in machine translation. *Patterns*, page 101257.

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.

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.

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

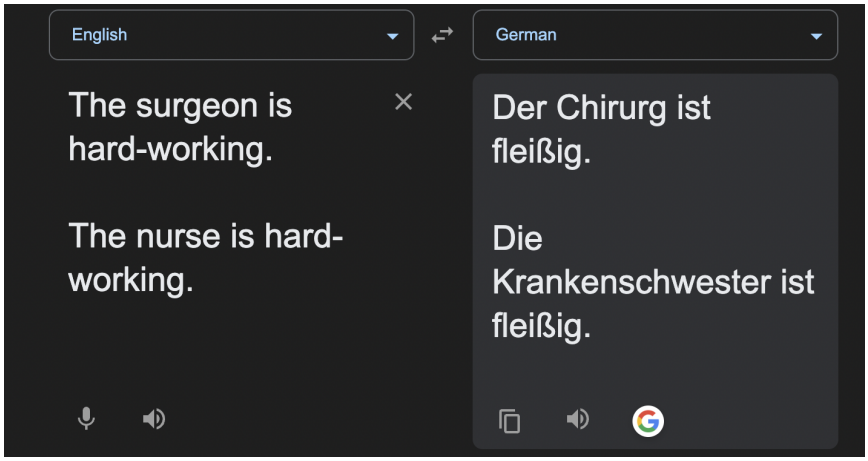

Figure 1: Google Translate assigns stereotypical genders to occupational roles.

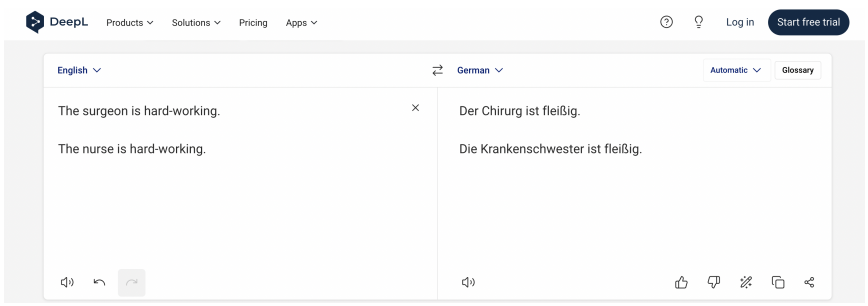

Figure 2: DeepL shows a similar bias in the same sentence, highlighting consistent patterns across MT tools.

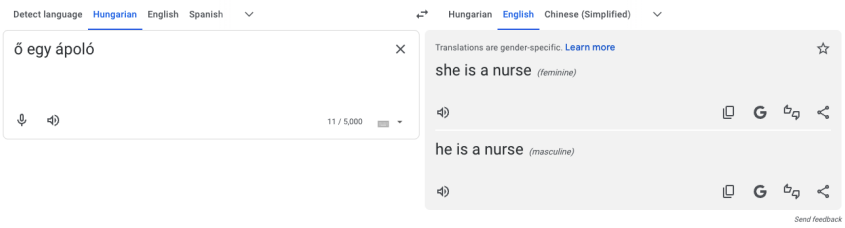

Figure 3: Gender-specific translation by Google Translate for ambiguous pronouns.

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 10, 2025

.....

(Unterschrift des Verfassers)