Detecting Gender Discrimination on Actor Level Using Linguistic Discourse Analysis

Stefanie Urchs[1], Veronika Thurner[1], Matthias Aßenmacher[2,3], Christian Heumann[2], Stephanie Thiemichen[1],

[1]Faculty for Computer Science and Mathematics,
Hochschule München University of Applied Sciences, [2]Department of Statistics, LMU Munich,
[3]Munich Center for Machine Learning (MCML), LMU Munich,

Correspondence: stefanie.urchs@hm.edu

Abstract

With the usage of tremendous amounts of text data for training powerful large language models such as ChatGPT, the issue of analysing and securing data quality has become more pressing than ever. Any biases, stereotypes and discriminatory patterns that exist in the training data can be reproduced, reinforced or broadly disseminated by the models in production. Therefore, it is crucial to carefully select and monitor the text data that is used as input to train the model. Due to the vast amount of training data, this process needs to be (at least partially) automated. In this work, we introduce a novel approach for automatically detecting gender discrimination in text data on the actor level based on linguistic discourse analysis. Specifically, we combine existing information extraction (IE) techniques to partly automate the qualitative research done in linguistic discourse analysis. We focus on two important steps: Identifying the respective person-named-entity (an actor) and all forms it is referred to (*Nomination*), and detecting the characteristics it is ascribed (*Predication*). As a proof of concept, we integrate these two steps into a pipeline for automated text analysis. The separate building blocks of the pipeline could be flexibly adapted, extended, and scaled for bigger datasets to accommodate a wide range of usage scenarios and specific ML tasks or help social scientists with analysis tasks. We showcase and evaluate our approach on several real and simulated exemplary texts.

1 Introduction

Ethical considerations as, e.g., formulated in the UNESCO's Recommendations on the Ethics of Artificial Intelligence, as well as emerging legislation such as the EU AI Act, require that any AI system adheres to fundamental values such as "the inviolable and inherent dignity of every human" (UNESCO, 2022). Specifically, this demand also holds true for systems based on large language models (LLMs). This implies that systems based on LLMs must carefully ensure that they do *not* reproduce, reinforce or broadly disseminate any existing biases, stereotypes or other discriminatory patterns, as this would violate the inherent human dignity.

However, LLMs are trained on existing data. If this input data is pervaded by stereotypes, biases and discrimination (as is often the case), the resulting model will reflect these discriminatory patterns. Thus, if developers need to ensure that an LLM-based system adheres to the ethical standards mentioned above, they can take one of two approaches: filter the LLM's output downstream to ensure that it is free from discrimination – or purge the input data from any discriminatory patterns, to ensure that the LLM itself will be free from discrimination in the first place.

Research on downstream gender bias mitigation in word embeddings by Gonen and Goldberg (2019) shows that downstream mitigation only hides bias and does not remove it. Thus, the effective alternative is to address bias upstream by selecting unbiased training data.

As the training corpora for LLMs need to be very extensive, it is impossible to ensure their quality manually. Therefore, technical means need to be developed that automatically detect discrimination in vast amounts of natural language texts.

What we read and see in media shapes our reality (Lippmann, 1929). If we are surrounded by bias and discrimination, we are likely to include these in our reality and act on them. That explains why media, notably text, plays an important role in the striving for equality for all genders. By detecting bias and especially discrimination against particular genders, it is possible to be wary of these texts and not distribute them. This is particularly important when choosing training data for natural language processing (NLP) tasks.

The term gender has at least three different notations: the linguistic gender, sex, and the social gender. The linguistic or grammatical gender can

August 16, 2024 ©2024 Association for Computational Linguistics

be defined as follows: "*[...] grammatical gender in the narrow sense, which involves a more or less explicit correlation between nominal classes and biological gender (sex).*" (Janhunen, 2000). For example, in German, nouns could be female, male, or neutral. The sex, however, refers to a "biological" notion of gender that is "*binary, immutable and physiological*" (Keyes, 2018). This notion is flawed because intersex humans do exist, as well as trans-persons, thus refuting the binary and immutable part of this notion. For our work, we use the third notion, the social gender. This notion defines gender as a social construct represented by a person's intentional and unintentional actions to represent their gender and the reception of these actions. Therefore, the social gender is non-binary, flexible, and constructed by the person themselves and the persons perceiving them (West and Zimmerman, 1987; Devinney et al., 2022). We use the terms woman for persons who can be read as female-identifying, men for persons who can be read as male-identifying, and non-binary for persons who do not adhere to the before mentioned.

Bias against a particular gender entails discriminating against this gender. While bias contains all notions and beliefs towards a person/group (Mateo and Williams, 2020), (social) discrimination is a more intentional act: an offender treats someone or a group of people differently in a negative way, based on a specific feature of this person/-group (Reisigl, 2017). Textual discrimination is a special kind of (social) discrimination because the offender is not always apparent.

Linguistics and sociology have studied discrimination for over eighty years, mainly focusing on racism in the early research (Myrdal et al., 1944; Razran, 1950; Allport et al., 1954). During this period, different definitions of discrimination were defined, leading to different approaches for detecting it. One of these approaches is linguistic discourse analysis (LingDA), which inspects discourse to identify discriminating tendencies by combining research from sociology and linguistics (Bendel Larcher, 2015). Computational linguistics integrates LingDA and computer science into computational discourse analysis. So far, this discipline concentrates on the quantitative parts of LingDA, mostly focusing on coherence and cohesion (Dascalu, 2014). We concentrate on the qualitative parts of LingDA and partly automate the discrimination detection within the text.

2 Problem Formulation and Goals

Existing approaches for automatic discrimination detection often focus on identifying drastic wording, which is relatively easy to detect by simple comparison with a database of discriminatory terms. However, in many cases, textual discrimination manifests more subtly, requiring a more semantic approach to detect it.

To achieve our goal of automatically identifying discrimination and biases in text, we seek to enhance computational discourse analysis (CompDA) by integrating two fundamental, qualitative strategies from linguistic discourse analysis for detecting gender discrimination on the actor level: Identifying the respective person named entity (an actor) and all forms in which it is referred to (*Nomination)*, and then detecting the traits, characteristics, qualities, and features that are ascribed to this actor (*Predication*). By focusing on actors, we aim to reveal even subtle gender-specific discrimination. Furthermore, we can analyse the text's meaning on a deeper level.

To automatically process large amounts of input text data, we implement a pipeline for automated text analysis that integrates nomination and predication by using IE techniques (cf. Figure 1). Specifically, as a first step, we identify nominations by extracting the actors and detecting their pronouns. Second, we extract the predication of these actors and finally use the extracted information to analyse the whole text for discrimination. By ensuring a modular structure built from exchangeable components, we aim to make our pipeline flexibly adaptable, accommodating a wide range of usage scenarios and specific ML tasks. For example, the pipeline should be able to scale from single texts to a whole corpus, process different languages, and focus on different criteria, thus reflecting cultural differences.

Finally, we evaluate our approach and implementation by analysing several sample texts, two real-world examples, and three generated texts, and discuss the discrimination markers identified in these samples.

3 Background

This work combines qualitative research on LingDA with IE, thus enhancing quantitative CompDA methods for detecting gender discrimination in text. Discrimination is a form of bias. We define discrimination and its relation to bias.

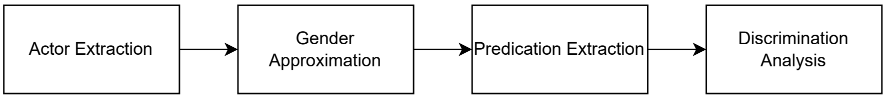


Figure 1: Visualisation of the flexible and language agnostic pipeline introduced in this work.

3.1 Linguistic Discourse Analysis

In LingDA, discourse is defined as a collection of text about a topic relevant to society (Bendel Larcher, 2015). This contrasts with computational linguistics, which defines discourse as "any multi-sentence text (Grishman, 1986). The focus of LingDA is the so-called actor. Actors are the entities in a text that perform some action. Actors can be individuals, groups, institutions, or organisations (Spitzmüller and Warnke, 2011).

Discourse is normally analysed on the corpus level as an extension of text linguistics that analyses single texts (Niehr, 2014). For our work, we concentrate on the level of single texts, especially on written text, potentially extending the approach to a whole corpus in future work. In this work, we disregard multimodal media, conversations, and pictures in general to scope our research. When analysing texts, Bendel Larcher (2015) points out that the nomination and predication is one of six aspects that should be considered. Nomination comprises how and what an actor in a text is named (Knobloch, 1996). The predication of the actor is what the text conveys about traits, characteristics, qualities, and features ascribed to this actor (Kamlah and Lorenzen, 1996).

When detecting nomination, the following aspects could be considered (Bendel Larcher, 2015):

1. **Proper Names**: Are actors referred to with their full name, surname, or just the first name?
2. **Generic Names**: When actors are not referred to by their proper names but with generic terms. Reisigl (2017) lists the following categories of problematic generic names: Negatively annotated general descriptions, ethnonyms, metaphorical slurs, animalistic metaphors, proper names used for a general description, and referring to an actor by their relation to someone else.
3. **Pronouns**: Pronouns can distance oneself from others (we vs. them), which is the basis for treating someone differently. Furthermore, using the wrong pronouns for someone (misgendering) is a clear aggression. Using the "generic masculine" in gendered languages like German can be considered problematic. Women and non-binary people are not directly addressed but are "included" in the word's meaning. Therefore, women and non-binary people are not represented by the language.
4. **Deagentification**: The actor of the text is not named. The text only generally describes what is happening without giving credit to the person.

The predication detection analyses the text for characteristics, features, and qualities attributed to an actor. These can convey stereotypes and biases that can be extracted by looking at the following grammatical indicators (Reisigl, 2017; Bendel Larcher, 2015):

- **Attributes**: e.g. skinny, bright
- **Prepositional Attributes**: e.g. the professor living in Munich
- **Collocations** e.g. working mom
- **Relative Clauses** e.g. the tennis player who has a nice dress

For this work, we focus on indicators of discrimination based on the actor's gender.

3.2 Computational Discourse Analysis

CompDA focuses on the analysis of cohesion and coherence. Cohesion describes how sentences are grammatically and lexically linked together to reflect the status of an actor through discourse. Typical methods include topics, coreferencing, and lexical and semantic word relatedness from ontologies. CompDA differentiates between referential cohesion (how often words, concepts, and phrases are repeated or related through the text) and causal cohesion (explicit use of connectives) (Dascalu, 2014). Coherence addresses the "*continuity of senses*" (De Beaugrande and Dressler, 1981) throughout the text. In other words, coherence conveys to the reader that the text is semantically connected. Dascalu (2014) distinguishes informational level coherence (causal relations between utterances, lexical chains, and centring theory) and intentional level coherence (tracing of the changes in the mental state of the discourse participants during the discourse).

Our approach combines cohesion and coherence by analysing the text using methods used in cohesion analysis to track actors (and their states) throughout the text.

3.3 Bias and Discrimination

Text can contain a lot of problematic properties regarding gender. The most problematic ones are biases and discrimination. However, also insults, defamation or misinformation should be avoided.

Mateo and Williams (2020) define bias as follows: "*Biases are preconceived notions based on beliefs, attitudes, and/or stereotypes about people pertaining to certain social categories that can be implicit or explicit.*". They continue that discrimination is the manifestation of biases through behaviour and actions. Reisigl (2017) has a clearer definition of discrimination: "*[...] social discrimination occurs when someone disadvantages or favours (i.e., treats unequally) a particular group or members of that group through a linguistic or other act or process, in comparison to someone else and on the basis of a particular distinguishing characteristic (such as an alleged 'race' or 'sexual orientation').*" leading to the following five parts of discrimination:

1. Offender
2. Victim (beneficiary in case of 'positive discrimination')
3. Disadvantaging (or favouring) act, process
4. Comparison group that is treated differently
5. Distinguishing feature on which the disadvantaging or favouring is grounded

Discrimination in written text is a manifestation of social discrimination. We consider discrimination as the manifestation of biases. Therefore, we consider the author of the text as the *offender*, and the *victim* is an actor of the text. The *feature that distinguishes* the victim from its *comparison group* is their gender. To scope our work, we only explore gender discrimination, even though we are aware that other kinds of discrimination, especially the intersection of different kinds of discrimination, exist and should not be part of NLP training data or other text. We extract the *disadvantaging act/process* from the text by quantifying differences between genders using LingDA and IE.

In manual LingDA researchers focus on the context of a text: was it released for a specific group of people from a specific kind of people? In the proper context, some kind of language that is offensive outside a group is acceptable if it is uttered by one person of a group towards another person of this group if it has an in-group context. Furthermore, some texts are seen as products of their time and represent the social norms of these times. However, when training NLP models, the context of a text is lost. The models learn equally on all text data. Therefore, we always have to assume an out-group context and the current social norms when evaluating textual data for training purposes.

Not removing discrimination and biases from training data leads to representational harms: gender stereotypes are spread in generated texts and, therefore, hardened in readers' minds. This harms all genders. Furthermore, not representing non-binary individuals in text generated by large language models (LLM) decreases their visibility. However, non-binary individuals are a part of our world and should be visible in LLM-generated texts. A text corpus not containing non-binary representation can not be considered balanced.

3.4 Information Extraction

IE locates predefined information in natural language text. According to Grishman (2015), the following steps are performed during IE (not necessarily in the order mentioned):

1. **Named Entity Recognition**: extraction of entities with proper names (persons, organisations, places, or suchlike)
2. **Syntactic Analysis**: extraction of syntactic information from sentences and tokenisation
3. **Coreference Resolution**: combining several mentions of an entity into one (e. g. a text mentions Dr. Ruth Harriet Bleier, further mentions may take the form of "Dr. Bleier", "Ms. Bleier", "R. H. Bleier", "R. B." or "she") (we also add generic names to form the full nomination of an actor)
4. **Semantic Analysis**: extracting relations between entities and mapping of sentences containing an entity to this entity (predication of an actor)
5. **Resolution of Cross-Document Coreferences**: coreferencing an entity through several documents (We are not exploring this step in this work.)

4 Methodical Approach

Our analysis pipeline can be subdivided into four consecutive steps that build on each other (cf. Figure 1): The first task is to extract the actors, followed by a gender approximation for each actor. In these steps, we save the nomination of each actor in our knowledge base. The third step expands the knowledge base with the predication of each actor detected in step one. As the fourth and final step of the pipeline, we analyse the extracted information for potential discrimination.

4.1 Nomination

The nomination process starts with the tokenisation of the text. No further preprocessing is applied to retain the full semantic meaning of the text. Subsequently, the dependency trees are parsed for each sentence. Therefore, each token is annotated with its relation to its semantic neighbours and its part of speech. All tokens that are proper nouns are analysed using named entity recognition (NER). Person entities are the actors of the text. As actors are mentioned more than once in a text, it is essential to coreference all mentions of the same actor. Coreferencing combines all references of one actor (this can be done in one text or the whole corpus). Therefore, the full name of an actor is matched to its name parts (e.g. first name, last name, last name, and abbreviations of first name), pronouns, and titles. In less formal settings, actors are referred to by generic names. These are not detected as proper nouns during NER. Therefore, generic names must be detected in an additional step and coreferenced with actors. We use a list of commonly used generic names to detect the generic names. All coreferenced entities and pronouns are the nomination of the actor. These are saved into a knowledge base using the same key for later use.

Every actor in the knowledge base is assigned one of the following gendered entries: woman, man, non-binary, unknown. The gendered entry is assigned by pronouns in the actor nomination.

4.2 Predication

The predication analyses what is ascribed to an actor. Ideally, the predication should only contain text that describes an actor. If a sentence contains more than one actor, this sentence should be split and matched accordingly. Furthermore, if an actor describes another actor, the sentence should only match the described actor and not the active one. For our proof of concept implementation, we simplify the sentence-matching process and assign a sentence to an actor if the actor is contained in this sentence. The predication is also stored in the knowledge base.

4.3 Discrimination Detection

We analyse the nomination for common derogatory terms for each entry in the knowledge base. To scope the research, we only use lists of derogatory terms referring to women, men, and transgender people[1]. For all predication sentences, the sentiment of the sentence is computed. Furthermore, the predication is analysed for feminine-coded words and masculine-coded words[2]. The authors show that women are associated with communal traits and men with more agency-related terms. Overusing gender-coded language can embed stereotypes. Using the computed information, we compile a discrimination report. For detailed report components, see Section 5.3.

5 Implementation and Validation

As mentioned in Section 4, we start by collecting the nomination of actors and subsequently enhance our knowledge base with the predication of the actors. The content of the knowledge base is subsequently analysed for discrimination and biases[3]. The code for our pipeline can be found on GitHub[4].

5.1 Nomination

SpaCy can perform tokenisation, dependency parsing, part of speech tagging, and named entity recognition out of the box. The named entity recognition can detect all actors in the text. When manually evaluating the results of our pipeline in the sample texts, we found that one actor's name was not classified as a person. Still, the error was not severe enough to justify changing libraries. We use the person entities as seed for the nomination.

In the first step, we extract all compounds of an actor's name; the head element of the compound is used as a key in a dictionary of actors. In a text

[1] derogatory terms were collected from the following websites (accessed on 2024-05-08): https://en.wikipedia.org/wiki/Category:Pejorative_terms_for_women, https://en.wikipedia.org/wiki/Category:Pejorative_terms_for_men, https://genderkit.org.uk/slurs/, https://en.wiktionary.org/wiki/Category:English_swear_words

[2] We use the lists of feminine/masculine coded words as found on the gender decoder website https://gender-decoder.katmatfield.com/about, which is based on work from Gaucher et al. (Gaucher et al., 2011)

[3] We use Python (version 3.9.18) and the NLP library SpaCy (Honnibal and Montani, 2017) in version 3.7.2, in combination with the en_core_web_lg model, for our experiments. Furthermore, we use the packages coreferee (version 1.4.1) and spacytextblob (version 4.0.0).

[4] https://github.com/Ognatai/nomination_predication

about Bill Clinton, the key Bill Clinton contains the values Bill Clinton, Clinton, President, and unexpectedly trail. We can also extract titles; for example, the key Kirsten Gillibrand contains the values Sen. and Kirsten Gillibrand. This implementation combines all actors with the same first or last names into one nomination.

In the second step, keys that are part of the value of another key are merged into the other key. Thus, all nomination keys are full names (if the actor is mentioned with their last name; otherwise, the key is a first name), and first names and last names are assumed to be unambiguous. These nominations are extended by a list of generic names found in the text and not coreferenced to other actors.

We determine the pronouns and, therefore, approximate the gender of the actors by using coreferee. This package references pronouns to actors. Unfortunately, coreferee has problems identifying gender-neutral/non-binary pronouns. In two of three test texts, it cannot detect the non-binary actors. Due to the lack of better-performing packages, we use coreferee nonetheless. Actors are assigned woman or man if the majority (at least 70%) of used pronouns refer to one of these gendered entries (we use a majority of at least five pronouns to be able to react to software problems stemming from the matching algorithm of coreferee). A non-binary entry is only assigned if gender-neutral/non-binary pronouns are used consistently. Otherwise, the gender is listed as unknown.

The last step of the nomination detection is to combine all information into a knowledge base stored as a pandas (pandas development team, 2023) data frame.

5.2 Predication

In the predication phase, the knowledge base is extended by all sentences that mention the corresponding actor. Each token object contains information about its position in the text. Therefore, we generate a text span with the size of the token and obtain the sentence that includes the text span of the token. Duplicates within one actor are removed. If a sentence contains more than one actor, this sentence is matched to all contained actors.

5.3 Discrimination Detection

For the discrimination detection, we extend the knowledge base by the sentiment of each predication sentence and the gender-coded words used in the predication. We use the package spacytextblob[5], which builds upon the textblob[6] library, to assign a value between -1 (very negative sentiment) and 1 (very positive sentiment) to each sentence. The sentiment analysis utilises a naive Bayes classifier trained on movie reviews. To detect gender-coded words, we use a list of feminine-coded and masculine-coded word stems by Gaucher et al. (2011) and test if these stems occur in the predication. We create a discrimination report for a text, building on the information of the knowledge base we created for this text. The report contains the following information:

- count of woman, man, non-binary, and undefined actors overall and per actor
- count of woman, man, non-binary, and undefined actor mentions overall and per actor
- sentiment towards woman, man, non-binary, and undefined actors overall and per actor
- count of feminine-coded words and masculine-coded words in the actor predication of woman, man, non-binary, and undefined actors overall and per actor
- abusive words used for woman, man, non-binary, and undefined actors and overall

5.4 Validation

Most NLP tasks like hate speech detection or sentiment analysis tend to utilise short utterances, like tweets or social media posts, for training purposes. In contrast, our approach aims to analyse longer texts like news articles or blog posts that describe one or more persons.

For testing our pipeline, we generate three texts with ChatGPT (OpenAI, 2023) that contain several actors, with at least one respectively using feminine, masculine, or gender-neutral/non-binary pronouns. All these actors have a full name and interact with each other. The content of all three generated texts is rather generic and not biased. We generated these texts mainly to test the pipeline on non-binary actors, but we do not further discuss the results of these texts because of their generic nature[7]. Instead, we collected texts about Bill and Hillary Clinton from Fox News[8].

The Hillary Clinton text describes Hillary Clin-

[5] https://spacy.io/universe/project/spacy-textblob

[6] https://textblob.readthedocs.io/en/dev/

[7] All text are available on GitHub: https://github.com/Ognatai/nomination_predication

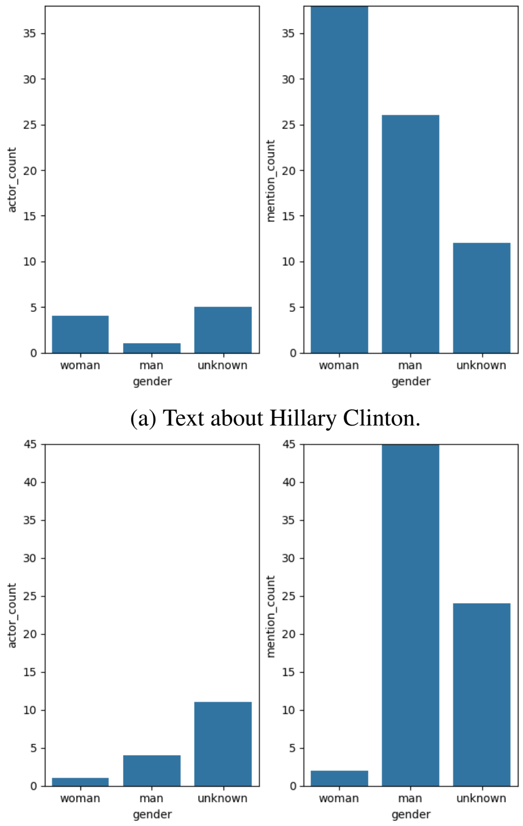
[8] https://www.foxnews.com/

ton's controversial statement that Trump followers should be 'deprogrammed' and reactions to this statement. The Bill Clinton text details how Bill Clinton "*reemerges as Democrat surrogate after being silenced by #MeToo movement*".[9]

We use our pipeline on these texts and compare the results by manually checking the corresponding texts for the correctness of the results.

The pipeline can detect all actors contained in the texts. Only the texts generated with ChatGPT contain non-binary actors. When analysing these texts, we found that `coreferee` has problems matching gender-neutral/non-binary pronouns to actors. Non-binary actors are detected in only one of three texts. Otherwise, our pipeline can mainly match the correct pronouns to the corresponding actor. We encounter problems in the text about Hillary Clinton. Here, `coreferee` has problems matching a pronoun from a partial sentence to one of the three actors mentioned before.

To count the mentions of each actor, we count all entries in the nomination and pronoun columns of the knowledge base. This leads to a minor problem since titles are not part of the name token and are counted as additional mentions. In our test data, this behaviour leads to one to two additional mentions per actor. In a future version of the pipeline, this behaviour will be fixed. Figure 2a and Figure 2b shows how many actors of a specific gender are part of the text and how often actors of a specific gender are mentioned throughout the text. Both texts do not contain non-binary actors. Interestingly, in the text about Hillary Clinton (Figure 2a, we detect four women (mentioned 38 times) and one man (mentioned 26 times). However, of the 38 women mentioned, Hillary Clinton is mentioned 26 times. Therefore, Donald Trump, the only recognised man, is mentioned as often in a text about Hillary Clinton as Hillary Clinton herself. However, the text describes how Hillary Clinton criticises Donald Trump's followers; therefore, many mentions make sense. In the text about Bill Clinton (Figure 2b, we detect four men, which are mentioned 45 times; 35 are mentions of Bill Clinton.



(a) Text about Hillary Clinton.

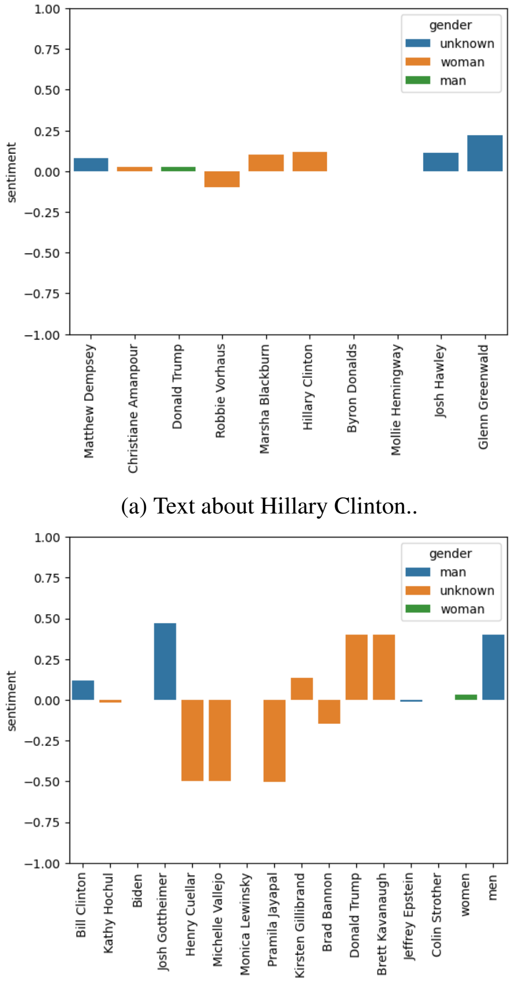
(b) Text about Bill Clinton.

Figure 2: Comparison of how often actors of a certain gender occur in the text and how often actors of a certain gender are mentioned. Both texts do not contain non-binary actors.

The sentiment analysis we use in our pipeline encounters problems when used for news articles. Figure 3b shows a moderately negative sentiment for `Henry Cuellar` and `Michelle Vallejo` which refers to the sentence "*During the trip, Clinton will rally with Rep. Henry Cuellar and Democratic candidate Michelle Vallejo – each of whom is locked in a difficult contest with Republicans.*" The sentence has a very neutral tone. In contrast, the model detects almost no negative sentiments in the text about Hillary Clinton (see Figure 3a. However, the predication of Hillary Clinton contains the following sentences: *"Sen. Marsha Blackburn, R-Tenn., posted to X, "Hillary Clinton wants Trump supporters to be formally reeducated., Independent journalist Glenn Greenwald shredded Clinton over the comments, saying, "As she gets increasingly bitter about her 2016 defeat – even when you think there's no way she can – Hillary Clinton is more and more the liberal id: she just spews what liberals really think and feel but know not to say., Clinton's 'deprogramming' hopes for Trump supporters a long shot in the era of political silos Clinton has had sharp words for Trump supporters over the years, once calling them 'deplorables'.*" The sentence contains a negative sentiment towards Hillary Clinton, but `spacyblob` cannot detect those neg-

[9] All text are available on GitHub: `https://github.com/Ognatai/nomination_predication`

ative sentiments. These examples showcase that the language used in news articles is too different from that used in movie reviews (which are one of the standard sources of training data for sentiment analysis approaches). Therefore, it is impossible to use a model trained on movie reviews for every domain; in future work, a domain-specific sentiment model will be utilised.



(a) Text about Hillary Clinton..

(b) Text about Bill Clinton.

Figure 3: Visualisation about the sentiments towards certain actors. Both texts do not contain non-binary actors.

In all texts, gender-coded words are rarely used. Both “real-world” texts contain a few feminine-coded words (Bill Clinton: 1, Hillary Clinton: 6) but no masculine-coded ones. Nevertheless, these could be an interesting feature if used for the whole corpus. We have a very explicit list of abusive words, but none are used in our sample texts. This list should be exchanged with domain-specific hate speech detection.

6 Discussion

Our method shows promising first results, even on our limited test data.

6.1 Strengths

Our pipeline can detect how different actors in a text are described. By approximating the gender of the actors, we can analyse if the text differentiates between genders and discriminates against a particular gender. Texts with very negative sentiments towards certain genders could then be excluded from model training, for instance. Our pipeline differentiates from other discrimination detection methods by focusing on actors and not the text as a whole. Therefore, it is possible to detect more subtle discrimination. Our pipeline is modular and, therefore, flexible. Single modules can be exchanged for domain-specific modules, and the pipeline can be extended anytime. Other discrimination detection approaches like hate speech detection or word lists can be included. The flexibility of the pipeline offers the possibility of even changing the languages of the texts analysed. Our proof of concept verifies the assumption that we can partly automate the qualitative parts of linguistic discourse analysis. Our discrimination report helps, for example, social scientists to decide if a text may contain discrimination or biases. This pipeline will be scaled to the corpus level to fully analyse the discourse within the corpus.

6.2 Limitations

Our proof-of-concept pipeline is tailored to detect actors in text. We cannot analyse the text if the text does not describe specific actors but a general situation. We combine actors with the same first and/or last name into one and do not coreference generic nominations to already detected actors. The predication should only consider text parts that attribute something to an actor. Currently, we use all sentences that contain the actor. If a sentence contains more than one actor, we match this sentence to all actors instead of doing an in-depth analysis of which parts of the sentence could belong to which actor. This also affects the sentiment analysis. A sentence containing an actor is not always a sentence containing a sentiment towards this actor. Another source of limitations is the general-purpose models we use in our pipeline. These are not tailored to the domain of news articles, leading to a sub-optimal performance. These general-purpose models also have problems in detecting gender-neutral/non-binary pronouns.

7 Conclusion and Future Work

In this work, we build a flexible pipeline to analyse newspaper articles and blog posts about people. We use linguistic methods to detect how actors are described within a text. In contrast to common discrimination detection methods, we do not treat the whole text as one object. By focusing on actors and the gender of the actors, we can do more nuanced text analyses that can detect subtle discrimination on a gender basis. First, limited tests on newspaper articles show that we can detect how actors are treated differently, depending on their gender. The first proof-of-concept pipeline implementation has some limitations that will be addressed in future work.

Other future work includes using the pipeline in different languages, such as German. Furthermore, instead of analysing one text at a time, we will scale the input to several documents, analysing complete corpora. We will also experiment with different pipeline components, for example, exchanging the simplistic abusive language detection with a sophisticated hate-speech detection or coreferencing detected actors with real-world actors to detect their pronouns. As today's discourse is not only written, analysis of multi-modal data might also be an interesting endeavour.

Ethical Consideration Statement

Defining discrimination for LLM training data means defining the value system for internationally used systems, but we do not share one common international value system. We can all agree on international human rights. However, an LLM also generates texts containing opinions about religion, race, gender, and sexual orientation. There are currently no common international values regarding these topics. As computer scientists, we define the values and opinions that our systems should convey. However, we are only able to adhere to our value system. Therefore, it is essential to work in diverse teams. The author team enriches their perspective by discussing our research with researchers from fields outside of computer science and from different cultural backgrounds. Our team consists of white Western European researchers. Three of us identify as women, representing the feminine and masculine gender spectrum but not the non-binary. Nevertheless, our group's diversity helps analyse gender-specific discrimination. Our understanding of discrimination stems from the system of beliefs and values based on Western European culture.

Acknowledgements

This work was written by an author team working in different projects. Stefanie Urchs' project "Prof:inSicht" is promoted with funds from the Federal Ministry of Education and Research under the reference number 01FP21054. Matthias Aßenmacher is funded with funds from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581. Responsibility for the contents of this publication lies with the authors.

References

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. *The nature of prejudice*. Addison-wesley Reading, MA.

Sylvia Bendel Larcher. 2015. *Linguistische Diskursanalyse: Ein Lehr-und Arbeitsbuch*. Narr Francke Attempto Verlag.

Mihai Dascalu. 2014. *Computational Discourse Analysis*, page 53–77. Springer International Publishing.

Robert-Alain De Beaugrande and Wolfgang U Dressler. 1981. *Introduction to text linguistics*, volume 1. longman London.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Ralph Grishman. 1986. *Computational linguistics: an introduction*. Cambridge University Press.

Ralph Grishman. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Juha Janhunen. 2000. Grammatical gender from east to west. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 124:689–708.

Wilhelm Kamlah and Paul Lorenzen. 1996. *Die Elementare Prädikation*, pages 23–44. J.B. Metzler, Stuttgart.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Clemens Knobloch. 1996. *Nomination: Anatomie eines Begriffes*, pages 21–53. VS Verlag für Sozialwissenschaften, Wiesbaden.

Walter Lippmann. 1929. *Public Opinion: By Walter Lippmann*. Macmillan Company.

Camila M Mateo and David R Williams. 2020. More than words: a vision to address bias and reduce discrimination in the health professions learning environment. *Academic medicine*, 95(12S):S169–S177.

Gunnar Myrdal et al. 1944. *An American dilemma; the Negro problem and modern democracy.(2 vols.)*. Harper.

Thomas Niehr. 2014. *Einführung in die linguistische Diskursanalyse*. WBG (Wissenschaftliche Buchgesellschaft).

OpenAI. 2023. ChatGPT(November 06 version).

The pandas development team. 2023. pandas-dev/pandas: Pandas.

Gregory Razran. 1950. Ethnic dislikes and stereotypes: a laboratory study. *The Journal of Abnormal and Social Psychology*, 45(1):7.

Martin Reisigl. 2017. *Sprachwissenschaftliche Diskriminierungsforschung*, pages 81–100. Springer Fachmedien Wiesbaden, Wiesbaden.

Jürgen Spitzmüller and Ingo Warnke. 2011. *Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse*. Walter de Gruyter.

UNESCO. 2022. Recommendation on the ethics of artificial intelligence. PDF Document. Page 18.

Candace West and Don H. Zimmerman. 1987. Doing gender. *Gender & Society*, 1(2):125–151.