taz2024full: Analysing German Newspapers for Gender Bias and Discrimination across Decades

Anonymous ACL submission

Abstract

Open-access text corpora are crucial for advancing research in natural language processing (NLP) and computational social science (CSS). Despite the growing availability of datasets, resources for languages other than English, such as German, remain scarce. This limits large-scale studies on linguistic, cultural, and societal trends and hinders research of complex issues like gender bias and discrimination. To address this gap, we present `taz2024full`, to our knowledge, the largest publicly available dataset of German newspaper articles to date. Comprising over 1.8 million articles from the German newspaper "taz" spanning 1980 to 2024. Unfortunately, including other sources in the corpus was impossible, as no other German newspaper provided free access to their data or allowed the publication of such a dataset. While access could have been obtained through paid licensing, this would not have guaranteed full data availability, and legal restrictions would have prohibited the release of the corpus for public use. As a result, "taz" remains the sole source for this dataset.

To demonstrate the potential of the corpus for bias and discrimination research, we analyse how references to different genders have evolved over more than four decades of reporting. Our findings reveal a persistent imbalance, with men consistently appearing more frequently in articles and receiving more textual space. However, we also observe a gradual shift towards a more balanced representation of genders in recent years. By adapting and scaling an existing pipeline for detecting gender bias and discrimination in news media, we provide researchers with a structured approach to studying actor representation, sentiment, and linguistic framing in German journalistic texts.

The `taz2024full` corpus and its accompanying pipeline support a wide range of research applications, from studying language evolution to investigating media bias and discrimination. By making this resource publicly available and demonstrating its application, we aim to facilitate interdisciplinary research, foster inclusivity in language technologies, and contribute to a more informed selection of training data for NLP models.

1 Introduction

Publicly available text corpora are invaluable resources for advancing research in natural language processing (NLP) and computational social science (CSS). However, most of these resources are frequently limited to English, restricting language-specific studies in other languages, such as German. The scarcity of open-access German newspaper datasets has hindered research on German news media's linguistic, cultural, and societal aspects. Specifically, the research on gender discrimination and bias detection suffers from a lack of resources as large collections of preferably diverse texts are necessary to analyse these phenomena *at scale*.

Given that the term "*gender*" is not always used uniformly, we define its usage in the scope of this paper: We refer to "*gender*" as a social construct, which is non-binary, flexible, shaped by individuals and by those perceiving them, rather than a biological given (West and Zimmerman, 1987; Devinney et al., 2022). However, due to methodological and linguistic constraints, we are limited to working within a binary gender spectrum in this work. We use pronouns to assume the gender of actors. As we are working on German data, we are limited to the commonly used pronouns in this language. Since German lacks widely used non-binary pronouns, non-binary genders could not be included in our analysis. We further employ the definition of "*biases*" as "all notions and beliefs a person has towards another person or group of persons" (Mateo and Williams, 2020). These biases can manifest in (written) discrimination, wherein a person or group is intentionally mistreated based on specific characteristics (Reisigl, 2017).

Contribution: In this work, we present a twofold contribution to the research area of gender bias and gender discrimination in written language and potentially also beyond:

1. We introduce the corpus "`taz2024full`", a comprehensive dataset that can serve as a valuable resource for analysing German newspaper articles published between 1980 and 2024, a period spanning more than four decades. Beyond our use case (see below), this allows analysing various other phenomena in the German language over time.

2. Furthermore, we demonstrate the potential of the corpus for bias and discrimination research by analysing how references to different genders have developed over more than 40 years of newspaper reporting, examining both their frequency and how they are discussed.

2 Related Work

Most publicly available news corpora focus on English newspapers, providing this type of diverse NLP research resource only for a single language. The "*Chronicling America*" dataset, provided by the Library of Congress, offers access to historical newspapers and digitised pages (spanning from 1690 to the present day) (Library of Congress). The "*BBC News Summary Dataset*" includes 2,225 documents from 2004 to 2005, covering five topical areas (Gupta et al., 2022). The "News Category Dataset"[1] contains around 210,000 headlines and descriptions from `https://www.huffpost.com`, spanning 2012 to 2022 (Misra and Grover, 2021). The "*News Articles Dataset*" comprises articles from 2015 to 2017, scraped from `https://www.thenews.com.pk/`, focusing on business and sports (Mahmood, 2017). Larger corpora such as RealNews, constructed from CommonCrawl dumps, offer 120GB of deduplicated news articles from 2016 to 2019, targeting large-scale training and evaluation of LLMs (Zellers et al., 2019). Additionally, the "*20 Newsgroups*" dataset[2], widely used in NLP, contains approximately 20,000 documents grouped into 20 categories, though it is neither based on newspaper content nor does it specifically look at developments over time (Lang, 1995).

[1] `https://huggingface.co/datasets/heegyu/news-category-dataset`

[2] `https://huggingface.co/datasets/google-research-datasets/newsgroup`

Regarding news datasets specifically for the German language, the "*One Million Posts Corpus*" is one of the most prominent examples. It is derived from online discussions on the Austrian newspaper DER STANDARD's website. It contains user posts from 2015-06-01 to 2016-05-31, with 11,773 labelled and 1,000,000 unlabelled entries, providing valuable insights into user-generated content (Schabus et al., 2017). This is, however, notably different from our resource as we do not focus on user-generated content, but on the articles themselves. In addition to this, several linguistic newspaper corpora exist, offering access to German news data. These include DWDS (Berlin-Brandenburgischen Akademie der Wissenschaften), the TüPP-D/Z corpus (Seminar für Sprachwissenschaft), the Mannheim German Reference Corpus (DeReKo) (Kupietz and Keibel, 2009), Leipzig Wortschatz (Universität Leipzig et al.), TIGER (Sabine et al., 2004), and others (Schiller et al., 1999; Nolda et al., 2021). However, these resources are often limited to keyword searches and return only sentence-level results, with most of them being unavailable for public use.

These datasets highlight the diversity of resources available for English and German news research but also reveal limitations, such as restricted access and narrow use cases. This underscores the need for openly available, comprehensive datasets like `taz2024full` to support language-specific studies and address gaps in research on German news media.

The detection of bias and discrimination in NLP has received significant attention in recent years. Blodgett et al. (2020), Sun et al. (2019), and Shah et al. (2020) provide broad overviews of existing approaches, highlighting the diverse methodologies employed to identify, mitigate, and evaluate biases in textual data. These works discuss various strategies, from detecting stereotypical associations in embeddings to evaluating fairness in predictive systems. Their surveys cover theoretical frameworks and practical implementations, offering valuable insights into the state of the field.

Despite the breadth of these reviews, none of the techniques discussed are directly comparable to the approach proposed by Urchs et al. (2024), which we adopt for evaluating our dataset. Unlike many conventional bias detection methods, which focus on linguistic patterns, embeddings, or statistical measures, the method of Urchs et al. (2024)

combines information extraction and linguistic discourse analysis to identify markers of bias and discrimination at the actor level. This actor-focused approach enables a more nuanced examination of how individuals and groups are represented in text, making it particularly suited for analysing bias in news media.

3 Dataset

We introduce the `taz2024full` newspaper corpus, a German newspaper corpus containing 1,834,370 publicly available articles published between 1980 and 2024 in the German newspaper "taz". This extensive dataset provides a valuable resource for linguistic, cultural, and societal research. It enables analyses across more than four decades of journalistic content. To the best of our knowledge, this is a unique collection of articles from a single news source over such an extended period, offering insights beyond specific use cases and opening avenues for long-term, diachronic studies (cf. section 4).

The dataset covers a wide range of historical contexts, including events of global significance, such as 9/11, the financial crisis in Europe, and the COVID-19 pandemic, all of which have shaped discourse on an international scale. Additionally, it encompasses events with particular relevance to Germany, such as the reunion, the 2015 migrant crisis, and several political changes, providing researchers with a lens to explore national and regional impacts on public discourse. The corpus's temporal span allows for the analysis of how language, societal attitudes, and journalistic practices have evolved in response to these events.

For our corpus, we exclusively used "taz" as a data source because no other German newspaper granted us free access to their archives. Licensing fees would have been required, and even then, access to the full dataset would not have been guaranteed. Furthermore, publishing the corpus would have been prohibited due to legal restrictions on data redistribution.

We have chosen not to provide a predefined train-test split for the dataset. Data splitting strategies may vary significantly depending on research objectives and use cases. For instance, studies focusing on the impact of specific historical events may require custom temporal splits, while others analysing long-term trends might need broader, cross-temporal divisions. Allowing users to define their splits ensures maximum flexibility and adaptability for diverse research applications.

By offering a corpus that captures the breadth and depth of "taz" reporting across decades, we aim to provide a foundation for a wide range of studies, from exploring shifts in public discourse to examining linguistic phenomena and identifying patterns of bias and discrimination. This adaptability makes the `taz2024full` corpus a critical resource for researchers in NLP, CSS, and related fields.

To our knowledge, this corpus is the largest German newspaper corpus available. Other publicly accessible German newspaper corpora do not provide full access to their data; instead, they typically allow only keyword-based searches or sentence-level queries through linguistic databases, making large-scale analysis impossible.

3.1 Data Source

The "taz" (*die Tageszeitung*, `https://taz.de/`) is a German daily-occurring, left-leaning newspaper based in Berlin. First published on September 22, 1978, it transitioned to a daily publication schedule on April 17, 1979 (taz, 2018). Known for its progressive editorial stance, "taz" has built a reputation as a prominent voice in the German media landscape, with 13,800 subscribers and an additional 39,000 paying for digital content (taz, 2024). The newspaper plans to cease daily printing on October 17, 2025, transitioning to an online-only format while retaining a weekly Saturday print issue. This transition marks a shift reflective of broader trends in the news industry.

The "taz" newspaper is known for its diverse and comprehensive reporting, covering a wide array of topics. Its editorial structure is organised into various sections, including breaking news, politics, society, culture, and sports. Additionally, "taz" has dedicated segments such as "Öko," which covers economics, ecology, labour, consumption, transport, science, and the network economy. There are also regional sections focusing on Berlin and northern Germany, as well as "Wahrheit," which features unconventional content like satire, commentary, and creative formats. Unfortunately, the metadata we crawled did not include explicit labels indicating the section in which an article was published. As a result, this information could not be incorporated into the dataset. Nonetheless, the corpus reflects the full spectrum of "taz" journalism, encompassing a rich variety of topics and perspec-

tives.

3.2 Data Collection

The `taz2024full` corpus was created by crawling publicly available content from the "taz" website (`https://taz.de/`) between August 2024 and November 2024. Permission has been granted to use and release this dataset for academic research, though its use for commercial purposes is strictly prohibited. The dataset only contains articles with more than three tokens (measured with SoMaJo tokeniser (Proisl and Uhrig, 2016)), thus excluding articles that only contain text fragments. The `taz2024full` corpus will be publicly available on Zenodo for non-commercial, academic purposes[3].

The articles are stored in JSON format. Figure 1 provides an example of the full JSON structure, illustrating how the metadata and article components are organised in the dataset. The metadata was extracted directly from the HTML of the crawled articles, and no modifications were made to the entries. The JSON format contains the following fields:

- **`"published_on"`**: The publishing date, stored as a string in the format `YYYY-MM-DDThh:mm:ss+01:00`.
- **`"contains_actors"`**: A boolean indicating whether person entities were detected in the article.
- **`"crawled_on"`**: The crawling date, saved as a string in the format `YYYY-MM-DD hh:mm`.
- **`"language"`**: The language of the article, always set to `"de"`.
- **`"type"`**: Mostly set to `"article"`.
- **`"author"`**: The person who wrote the article.
- **`"keywords"`**: Keywords relevant to the article, could be used for topic recognition.
- **`"token_count"`**: The number of tokens in the article.

The article-data consists of three parts, though not all are always available. However, every entry contains at least one `"text"` component:

- **`"title"`**: The headline of the article.
- **`"teaser"`**: A short description of the content or a short introduction.
- **`"text"`**: The main body of the article.

```
{
    ID: {
        "metadata": {
            "published_on": string,
            "contains_actors": boolean,
            "crawled_on": string,
            "language": "de",
            "type": string,
            "author": string,
            "keywords": string,
            "token_count": int
        },
        "text": {
            "title": string,
            "teaser": string,
            "text": string
        }
    }
}
```

Figure 1: Structure of the elements in the corpus, including all available metadata collected alongside the raw texts.

3.3 Dataset Statistics

The `taz2024full` corpus consists of 1,834,026 newspaper articles published in the German newspaper taz between 1980 and 2024 (cf. Figure 2). From 1980 onwards, there was a steady increase in the number of published articles, reaching a peak of 73,002 in 2004. An unexpected dip occurred in 1991, though the reason for this anomaly is unclear. After 1993, publication numbers stabilised again before declining after the peak in 2004. This decline may be linked to changes in the publishing strategy, potentially involving an increase in paid content from 2007 onwards. Only publicly available content is used for this corpus, thus leading to declining article numbers from 2007 onwards.

The corpus contains 6,944,197 unique tokens, as determined using the SoMaJo tokeniser (Proisl and Uhrig, 2016). Additionally, 83% of all articles mention specific individuals, allowing for an analysis of gender bias and discrimination in how different actors are represented in the article. Table 1 provides an overview of token, sentence, and article lengths. The maximum token and sentence lengths suggest that the tokeniser did not always ideally segment the articles. However, an average of four characters per token and 17 tokens per sentence appears reasonable.

[3] put URL here for publication

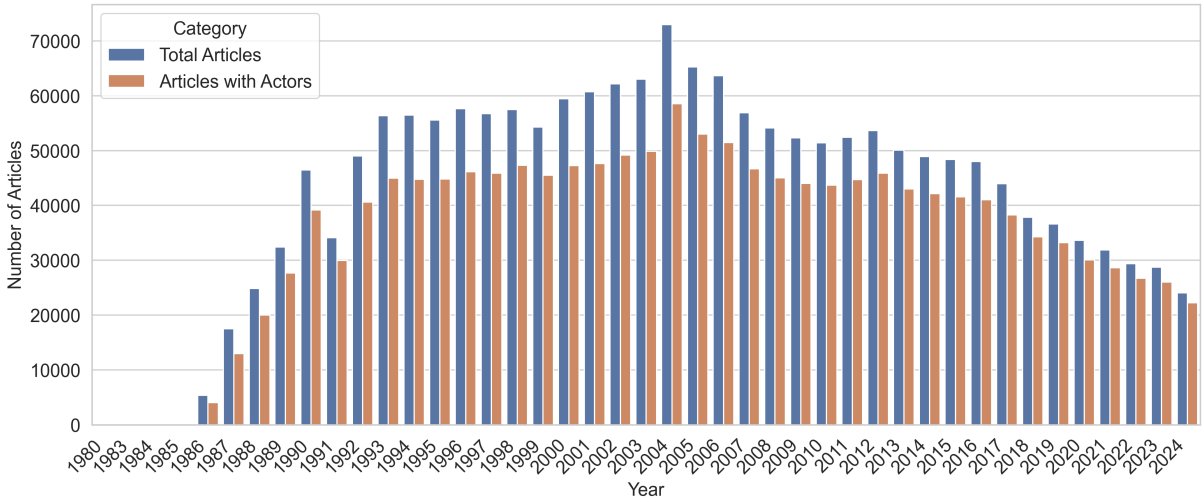

Figure 2: Number of newspaper articles and articles containing actors per year from 1980 to 2024. Years before 1986 include four articles or fewer.

	min	max	mean	median
token length	1	2,238	5.15	4
sentence length	1	18,872	20.07	17
article length (token)	3	26,855	396.89	276
article length (sentences)	1	1,027	19.77	13

Table 1: Statistics for the `taz2024full` corpus.

3.4 Dataset Applications

The `taz2024full` dataset offers numerous applications for researchers in NLP, CSS, and related fields. Its extensive temporal span supports investigations into the evolution of the German language, shifts in public discourse, and changes in thematic emphasis. For instance, it is possible to explore how language has adapted over decades, whether specific topics follow seasonal patterns, or how societal events have influenced reporting.

The dataset is particularly suited for examining patterns of bias and discrimination, mainly because it focuses on actors in texts. By leveraging this resource, insights can be obtained into representation trends and linguistic framing over time.

By providing this comprehensive dataset, we aim to support a wide range of research efforts, from linguistic studies to bias detection, thereby contributing to a deeper understanding of German media discourse.

4 Experiments

In our work, we build upon the pipeline introduced by Urchs et al. (2024) for detecting biases and discrimination on actor level. Rather than single English newspaper texts (as in Urchs et al. (2024)), we apply and scale an extended version of the pipeline to the German `taz2024full` corpus and compare discrimination markers in newspaper texts through more than four decades.

The full code is published on GitHub[4]. See also appendix A for a sample report of 2023.

4.1 Data

We analyse the entire corpus for discrimination markers to demonstrate its potential for bias and discrimination research.

The dataset encompasses all topics covered by "taz" without filtering, ensuring a comprehensive representation of the newspaper's reporting across politics, culture, economy, society, and more.

The chosen time-frame enables a meaningful comparison, capturing how taz's reporting on actors has evolved over 44 years. This period includes significant societal, cultural, and journalistic developments, which are likely to influence how gender bias and discrimination manifest in the texts.

4.2 Method

Urchs et al. (2024) automate linguistic discourse

[4] https://anonymous.4open.science/r/corpus_pipeline-1468

analysis by applying information extraction techniques to analyse English newspaper articles for discriminatory content.This approach focuses on two key aspects of linguistic discourse analysis: nomination, which examines how actors in a text are named, and predication, which analyses how they are described.

Their pipeline is designed to identify actors within the text by leveraging named entity recognition (NER) to detect person entities. The authors utilise spaCy[5] to extract all named entities from the text, retaining only those classified as persons. Additionally, generic terms referring to individuals, such as "mother," "father," "woman," and "man", are extracted. Entities that share the same name or name components (e. g. if only a first name is present or only a surname) are grouped together under a single actor. To further refine the analysis, coreferee[6] is employed for coreference resolution, linking pronouns to the corresponding actors. This approach enables text analysis based on distinct actors rather than relying solely on pronoun frequency.

The pipeline subsequently assumes the gender of the actors based on the extracted pronouns, categorising them as `woman`, `man`, `non-binary`, or `undefined`. If more than 70% of the pronouns associated with an actor are either feminine or masculine, the actor is categorised as `woman` or `man`, respectively. Actors are labelled `undefined` if no pronouns are present or the thresholds for other categories are not met.

Urchs et al. (2024) extract sentences mentioning the actors to analyse their predication. Therefore each sentence that either contains the name of the actor or a pronoun linked to the actor is extracted. This data is used to identify the following markers of discrimination across the gender categories:

- Number of actors in text per gender category
- Count of mentions per gender category / individual
- Sentiment towards gender category / individual[7]
- Count of feminine-coded words and masculine-coded words in the predication of each gender category/individual[8]
- Abusive words in predications (not used in this work)

These markers are then compiled into a discrimination report and visually represented for each text.

In this work, we adapt the existing pipeline to analyse German newspaper articles, expanding its scope from processing individual articles to handling thousands at a time. Our adaptations include the following:

- **Gender Assumption:** Instead of assuming the gender of actors, we determine the primary pronoun used to refer to them. Additionally, we scanned the corpus for German neo-pronouns[9] and found that only five texts contained such pronouns. Consequently, our analysis focuses on the pronouns she/her and he/him. For the results section, we categorise actors primarily referred to with she/her pronouns as women and those with he/him pronouns as men.

- **Pronoun Driven Analysis:** We include only actors for whom co-reference with pronouns could be established. All other actors are not part of the analysis.

- **Generic Masculine/German Gender-Neutral Language:** We introduce a marker to determine whether an article employs the generic masculine and one for the German gender-neutral language.

- **Pairwise Mutual Information (PMI):** Additionally, we identify the top 10 adjectives with the highest PMI in each actor's predication. PMI (cf. equation 1) measures the probability of two words x and y co-occurring by chance or meaningfully. A higher PMI indicates a more meaningful relation between these words (Jurafsky and Martin, 2000). In our pipeline, we calculate the PMI for each

[5] https://spacy.io/

[6] https://spacy.io/universe/project/coreferee

[7] The spaCy pipeline component spacytextblob (https://spacy.io/universe/project/spacy-textblob) is used for sentiment analysis. In our German version we use the pre-traind model "oliverguhr/german-sentiment-bert", which is trained on German newspaper articles, from huggingface (https://huggingface.co/transformers/v3.0.2/main_classes/pipelines.html)

[8] We utilise the lists of feminine- and masculine-coded words provided on the Gender Decoder website https://www.msl.mgt.tum.de/rm/third-party-funded-projects/projekt-fuehrmint/gender-decoder/wortlisten/, which are derived from the work of Gaucher et al. (Gaucher et al., 2011). Their research demonstrates that women are more frequently associated with communal traits, while men are linked to agency-related terms.

[9] We used the neo-pronouns listed at https://gleichstellung.tu-dortmund.de/projekte/klargestellt/neo-pronomen

actor and each adjective in their predication, excluding stop-words. Therefore, we can identify the most influential adjectives for each actor.

$$\text{PMI}(x, y) = log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

- **Aggregated Report:** We introduce a human-readable report to facilitate easy corpus analysis.

We deliberately chose not to use large language models (LLMs) for the analysis, despite known challenges with pronoun detection and co-referencing. While LLMs could potentially enhance performance, we opted against them due to concerns about the biases they may introduce and the lack of transparency in their outputs. Instead, we relied on well-established, interpretable methods, acknowledging their limitations. This decision prioritises ethical considerations and ensures greater methodological control.

We conduct a yearly analysis of the whole corpus, examining changes and trends over the 44-year span of our corpus.

4.3 Results

The analysis of the corpus over the years provides valuable insights into how taz has written about women and men across the decades. Figure 3 illustrates the proportion of actors whose gender could be co-referenced with pronouns, distinguishing between women and men (Woman Actors/-Man Actors). Additionally, the figure presents the frequency with which these actors are mentioned within texts (Woman Mentions/Man Mentions)[10]. Before 1990, the data is too sparse to allow definitive interpretations. However, from the 1990s onward, a clear trend emerges: "taz" reported significantly more on men than on women. This imbalance is present across all decades, with man actors not only more frequently included in articles but also mentioned more often. While a shift towards greater inclusion of women actors becomes apparent from the 2010s onwards, this pattern of men actors being both more often the subject of reporting and more frequently referenced persists. Even in recent years, where gender representation appears almost balanced in terms of actor inclusion, men continue to receive more textual space, indicating a continued dominance in media visibility.

Figure 3: Comparison of the number of actors per article based on detected genders and the frequency of their mentions within each article.

Beyond the quantity of mentions, the sentiment towards men and women in "taz" articles is also revealing. Figure 4 illustrates the sentiment associated with women and men actors over time. Sentiment values range from -1 (highly negative) to +1 (highly positive), with 0 representing neutrality. The data shows that "taz" articles generally lean towards a neutral but slightly negative sentiment. More strikingly, across the entire 44-year period, sentiment towards women actors is consistently slightly more negative than sentiment towards men actors. While the differences are not extreme, this persistent pattern suggests that women in "taz" articles are, on average, portrayed in a slightly more negative light than their male counterparts.

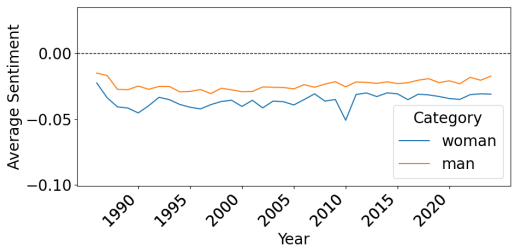

Figure 4: Sentiment towards the detected genders through the years.

Looking at language use more closely, an analysis of adjectives with the highest PMI association with women and men actors reveals that these descriptors remain relatively stable over time and do not exhibit strong gender differentiation. Additionally, a targeted analysis of female-coded and

[10]We differentiate between the number of actors in a text and the number of mentions to account for cases where, for instance, a single woman actor is referenced multiple times, whereas multiple male actors might be mentioned only once. Without this distinction, a text featuring one woman mentioned ten times and another featuring ten men mentioned once each would appear equivalent in terms of gender representation, despite their differing narrative emphases.

male-coded words shows that these were used only rarely. This suggests that "taz" makes little use of explicitly gender-coded language. However, our analysis found no evidence that "taz" systematically uses German gender-neutral language. While a few instances of gender-neutral forms were manually observed, these were rare exceptions rather than a common practice. Despite ongoing discussions about inclusive language in German, "taz" does not appear to have adopted gender-neutral writing conventions in its standard editorial style.

5 Conclusion and Future Work

In this work, we introduced `taz2024full`, a comprehensive German newspaper corpus spanning over four decades. This dataset represents a significant resource for linguistic, cultural, and societal research, particularly in the areas of gender bias and discrimination in media. By leveraging a structured approach to analysing gender representation through actor mentions and predications, we provided insights into how "taz" has reported on women and men over time. Our findings highlight a persistent imbalance in gender representation, with men not only appearing more frequently as actors but also receiving greater textual space.

Furthermore, we demonstrated the adaptability of an existing bias detection pipeline, originally designed for English texts, to large-scale German-language data. Our extensions included modifications tailored to the German linguistic landscape, such as pronoun-based gender identification and an analysis of the generic masculine. These enhancements offer a more refined approach to studying gender-related language use in German news media.

Looking ahead, several avenues for future research emerge. One priority is updating the corpus with new data beyond 2024 to enable ongoing diachronic analysis. Additionally, incorporating topic modelling could provide deeper insights into the contextual framing of gender representation. Since the current implementation would model topics at the sentence level, a necessary improvement would involve incorporating broader textual context to enhance topic coherence around actor predications.

Another promising direction is argument mining, which could refine our understanding of the implicit and explicit biases embedded in journalistic discourse. By identifying argument structures and rhetorical strategies, we could further uncover how gender bias manifests in media narratives.

Ultimately, we hope that `taz2024full` serves as a valuable resource for researchers in NLP, computational social science, and related fields, facilitating future studies on bias, representation, and media discourse in the German language.

The `taz2024full` corpus and its language-agnostic pipeline provide a foundation for analysing bias and discrimination in German news media. Future work will focus on expanding the corpus with newer data to ensure it remains relevant and reflective of contemporary discourse. Enhancements to the pipeline include training sentiment analysis models tailored to German newspapers, improving pronoun coreference resolution for German texts, and incorporating argument mining for discrimination detection. Additionally, we aim to extend the pipeline to support multimodal analysis by integrating text with non-textual data such as images and videos, enabling a comprehensive understanding of media content.

Use of AI

The authors are not native English speakers; therefore, ChatGPT and Grammarly were used to assist with writing English in this work.

Limitations

The `taz2024full` corpus has several limitations to consider when interpreting research results. Firstly, the dataset reflects the views of a Berlin-based, left-leaning publication and does not represent the entire spectrum of German discourse. This inherent bias limits its applicability for studying nationwide or ideologically diverse perspectives.

Additionally, bias and discrimination detection within corpora is inherently subjective, as no universally accepted gold standard exists. Different users may have varying values and interpretations of bias or discrimination, complicating evaluating such tasks. Therefore, we are not able to decide if a corpus is discriminatory or biased, as we do not know the use cases for each corpus. Additionally, our understanding of discrimination or bias might differ from the users understanding. Thus, we limit the output to a discrimination report, allowing each user to determine whether any adjustments to the corpus are necessary or to compare different corpora based on the calculated metrics.

Further challenges arise when applying the

language-agnostic, flexible pipeline to detect gender discrimination and bias in texts. Co-reference resolution, particularly on German data, remains problematic due to the lower accuracy of current models for this language. This can affect the precision of gender-related analyses. Furthermore, the lack of directly comparable works or benchmarks complicates evaluating the pipeline's performance.

Ethical Considerations

The `taz2024full` corpus is intended exclusively for academic research purposes, and exploiting it for commercial use would harm the publisher, taz Verlags und Vertriebs GmbH. To prevent such misuse, we strongly emphasise the importance of adhering to the dataset's intended purpose: fostering academic exploration and understanding.

Defining discrimination can be addressed abstractly, but implementing the concrete pipeline requires concrete values and definitions. Since we do not know the pipeline users' specific use cases and value systems, we opted for a flexible analysis of discrimination markers, highlighting potentially problematic content, thus refraining from definite judgments. We intentionally leave the final interpretation to human users. This allows them to apply their understanding of what constitutes discrimination in their specific use cases.

We acknowledge that our pipeline could be misused to curate datasets with specific biases or intentionally exclude particular genders. We strongly discourage using this system to manipulate datasets in ways that reinforce or amplify discrimination and biases. Instead, we aim to promote fairness and inclusivity by providing insights that help users curate discrimination and bias-free data sets.

By maintaining human oversight in the evaluation process, we aim to balance automated analysis with ethical responsibility, ensuring that the system supports diverse needs while promoting fairness in language technologies.

References

Berlin-Brandenburgischen Akademie der Wissenschaften. Der deutsche wortschatz von 1600 bis heute. online.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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.

Anushka Gupta, Diksha Chugh, Anjum, and Rahul Katarya. 2022. Automated news summarization using transformers. In *Sustainable Advanced Computing*, pages 249–259, Singapore. Springer Singapore.

Daniel Jurafsky and James H. Martin. 2000. *Speech and language processing*. Prentice-Hall.

Marc Kupietz and Holger Keibel. 2009. The mannheim german reference corpus (dereko) as a basis for empirical linguistic research. *Working papers in corpus-based linguistics and language education*, 3:53–59.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Library of Congress. Newspaper datasets and api access. online.

Asad Mahmood. 2017. News articles. online.

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.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

Andreas Nolda, Adrien Barbaresi, and Alexander Geyken. 2021. *Das ZDL-Regionalkorpus: Ein Korpus für die lexikografische Beschreibung der diatopischen Variation im Standarddeutschen*, pages 317–322. De Gruyter, Berlin, Boston.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

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

Brants Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2:597–620.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Universität Tübingen Seminar für Sprachwissenschaft. Das korpus tüpp-d/z. online.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 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, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

taz. 2018. taz-geschäftsführer als zeitzeuge: Es begann in einem kalten ladenlokal | taz.de. online.

taz. 2024. Die seitenwende | taz.de. online.

Universität Leipzig, Sächsischen Akademie der Wissenschaften zu Leipzig, and Instituts für Angewandte Informatik. Willkommen beim wortschatz-portal. online.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2024. Detecting gender discrimination on actor level using linguistic discourse analysis. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 140–149, Bangkok, Thailand. Association for Computational Linguistics.

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

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A Corpus Report 2023

```
Aggregated Report for 2023
==========================================================

Total Texts: 26357
Total Actors: 17161
Pronoun Distribution: {'he_him': 9088, 'she_her': 8073}
Total Mentions: 109634
Mentions by Pronoun: {'he_him': 60008, 'she_her': 49626}

Mean Metrics:
  total_actors: 1.67
  total_mentions: 10.66
  total_feminine_coded_words: 0.27
  total_masculine_coded_words: 0.11
  contains_majority_gender_neutral: 0.00
  generic_masculine: 0.31
  pronoun_distribution_she_her: 0.79
  pronoun_distribution_he_him: 0.88
  mentions_pronoun_distribution_she_her: 4.83
  mentions_pronoun_distribution_he_him: 5.84
  feminine_coded_words_pronoun_distribution_she_her: 0.13
  feminine_coded_words_pronoun_distribution_he_him: 0.14
  masculine_coded_words_pronoun_distribution_she_her: 0.05
  masculine_coded_words_pronoun_distribution_he_him: 0.06
  average_sentiment_all: -0.03
  sentiment_by_pronoun_she_her: -0.03
  sentiment_by_pronoun_he_him: -0.02

Median Metrics:
  total_actors: 1.00
  total_mentions: 5.00
  total_feminine_coded_words: 0.00
  total_masculine_coded_words: 0.00
  contains_majority_gender_neutral: 0.00
  generic_masculine: 0.00
  pronoun_distribution_she_her: 1.00
  pronoun_distribution_he_him: 1.00
  mentions_pronoun_distribution_she_her: 2.00
  mentions_pronoun_distribution_he_him: 2.00
  feminine_coded_words_pronoun_distribution_she_her: 0.00
  feminine_coded_words_pronoun_distribution_he_him: 0.00
  masculine_coded_words_pronoun_distribution_she_her: 0.00
  masculine_coded_words_pronoun_distribution_he_him: 0.00
  average_sentiment_all: 0.00
  sentiment_by_pronoun_she_her: 0.00
  sentiment_by_pronoun_he_him: 0.00

Top PMI Adjectives Table:
All                  she/her              he/him
--------------------------------------------------------------------------------
letzten              junge                letzten
deutschen            letzten              russischen
junge                deutschen            deutschen
russischen           jungen               politische
berliner             deutsche             berliner
deutsche             berliner             politischen
politische           russischen           ukrainischen
politischen          nächsten             ukrainische
jungen               alten                russische
nächsten             alte                 deutsche
```

Listing 1: Full discrimination report of 2023.