Measuring Sexism in US Elections: A Comparative Analysis of 𝕏 Discourse from 2020 to 2024

Anna Fuchs[♣] **Elisa Noltenius**[♣] **Caroline Weinzierl**[♣]
Bolei Ma[♣,♡] **Anna-Carolina Haensch**[♣,♡,♢]

[♣]LMU Munich [♡]Munich Center for Machine Learning
[♢]University of Maryland, College Park

{anna.fuchs,elisa.noltenius,caroline.weinzierl}@campus.lmu.de, {bolei.ma,c.haensch}@lmu.de

Abstract

Sexism continues to influence political campaigns, affecting public perceptions of candidates in a variety of ways. This paper examines sexist content on the social media platform 𝕏 during the 2020 and 2024 US election campaigns, focusing on both male and female candidates. Two approaches, single-step and two-step categorization, were employed to classify tweets into different sexism categories. By comparing these approaches against a human-annotated subsample, we found that the single-step approach outperformed the two-step approach. Our analysis further reveals that sexist content increased over time, particularly between the 2020 and 2024 elections, indicating that female candidates face a greater volume of sexist tweets compared to their male counterparts. Compared to human annotations, GPT-4 struggled with detecting sexism, reaching an accuracy of about 51%. Given both the low agreement among the human annotators and the obtained accuracy of the model, our study emphasizes the challenges in detecting complex social phenomena such as sexism.

Disclaimer: This paper contains content that can be offensive or upsetting.

1 Introduction

Sexism is defined as prejudice, stereotyping, or discrimination based on sex, typically against women (Oxford English Dictionary, 2023). Despite progress toward gender equality, it remains prevalent in many areas of society, from workplaces and education to the media, shaping perceptions and limiting opportunities for women. One area where sexism is particularly prominent is politics, where women are underrepresented and often unfairly judged compared to their male counterparts (Fox and Lawless, 2004; Lovenduski, 2014). Female candidates often face scrutiny over their leadership and competence abilities simply because of their gender. The media further intensifies these biases by focusing on their looks and personal lives instead of their political views. As social media plays an important role in shaping voters' opinions, it reinforces existing gender biases and gender-based criticism, particularly affecting female politicians (Tromble and Koole, 2020).

Detecting sexism and understanding the intentions behind it are essential steps in overcoming deeply embedded gender norms and biases, especially in contexts where women seek leadership positions, such as presidential candidacy in politics. However, sexist comments do not always exhibit obvious negative emotions (Becker and Wright, 2011). Sexism can be subtle, often unnoticed, making it challenging to identify since it is embedded in cultural and societal norms (Swim and Cohen, 1997). Therefore, it becomes crucial to investigate these implicit forms of sexism and their impact on individuals and society.

The 2020 and 2024 US election cycles present a unique opportunity for researchers to examine whether gender continues to influence the chances of presidential candidates. The 2020 presidential election featured Joe Biden and Donald Trump as primary candidates. In contrast, the 2024 election was remarkable for a candidate switch during the campaign, as Joe Biden announced his resignation on 21 July, with Kamala Harris subsequently launching her campaign (CNN Politics, 2024).

Our paper analyzes 𝕏 (formerly Twitter) data using tweets sampled from three time frames over two election periods (2020 and 2024). These periods represent two different candidate scenarios: male vs. male and female vs. male. The selected tweets were chosen based on election-specific keywords, from the data source publicized by previous research (Balasubramanian et al., 2024).

We use GPT 4.0 (OpenAI, 2024) to categorize these tweets, setting up two different approaches. The first approach directly classifies tweets into non-sexist and more granular sexist categories. The

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 130–147

second approach involves a two-step process: initially identifying tweets as either non-sexist or sexist, followed by categorizing the sexist tweets into finer-grained categories. A consistent set of prompts is applied to compare the two approaches, while a small subsample dataset with manually annotated data is used as a reference for evaluating GPT's annotation capabilities.

Using this set-up, we address the following research questions (RQs):

- **RQ1:** How do two-step and single-step GPT-4-based categorization approaches compare for identifying and classifying sexist tweets?
- **RQ2:** Have sexist patterns and categories of sexist content in US election-related discourses on 𝕏 changed over the three election time frames?

2 Literature Review

This review offers an overview of sexism in political discourse, discussing approaches for classifying sexist content, with a focus on methods that use generative AI, and prompting techniques for the automated detection of sexist language.

Sexism in Politics. Literature on sexism in politics often focuses on the gender-biased representation of female politicians and the undermining of women's leadership roles, highlighting how such biases influence public opinion and election outcomes. Systematic marginalization and societal structures within political institutions contribute to underrepresentation and limited political participation of women (Lovenduski, 2014). Despite similar qualifications, women express less political ambition due to lack of encouragement to run for office and a lower self-perception of qualifications (Fox and Lawless, 2004). The 2016 US presidential election between Hillary Clinton and Donald Trump served as a crucial case for studying gender dynamics in politics. Research shows that sexism played a substantial role in Hillary Clinton's defeat, as women candidates face challenges and unequal evaluations compared to their male counterparts (Knuckey, 2019).

Sexism also shaped voter favorability: men showed a much stronger preference for Trump than women, while attitudes toward Clinton were similar between genders (Glick, 2019; Ratliff et al., 2019). Political sexism, defined as the belief that men are better suited emotionally for politics than women, strongly predicted support for Trump, especially among white voters (Bracic et al., 2019). Hostile sexism, defined as having negative views towards individuals who defy traditional gender stereotypes, emerged as a key factor benefiting Trump's candidacy, while benevolent sexism, which is positive in tone but yet connotes inferiority to men, increased support for Clinton without affecting Trump (Glick and Fiske, 2001; Ratliff et al., 2019). From a broader point of view, Falk (2010) examines nine female political campaigns, uncovering how media portrayals often frame female candidates as unviable or incompetent. Analyzing political sexism in social media, particularly using 𝕏 data, has already been addressed by Tromble and Koole (2020). This study reports no clear differences in the tone of messages directed at male and female politicians across three countries, including the US.

Sexism Classification. Lots of research on sexism in social media has focused primarily on detecting misogyny and hateful language directed at women (Guest et al., 2021; Pamungkas et al., 2020). However, sexism often operates in more nuanced ways. To capture its complexity, researchers have developed various classification frameworks. A common method is a two-step approach: first identifying sexist tweets, then categorizing them into more granular categories (Jiang et al., 2022; Plaza et al., 2023). These finer-grained categories can be defined from multiple perspectives. According to ambivalent sexism theory (Glick and Fiske, 2001), sexism can be hostile (overtly negative) or benevolent (seemingly positive but reinforcing inferiority). Other studies classify the degree to which sexism manifests itself - blatant, subtle, or covert (Swim et al., 2004). Studies also explored multi-label classification, with varying granularity. For instance, Rodríguez-Sánchez et al. (2021) define five categories, while Parikh et al. (2019) define 23. Some approaches incorporate cultural perspectives (Jiang et al., 2022), focus on specific forms of harassment (Sharifirad et al., 2018), or distinguish sexism by target (individual or generic) (Jiang et al., 2022) and intention (Plaza et al., 2023). These different classification frameworks highlight the complexity of sexism and the need for approaches to successfully recognize its various forms.

Large Language Models and Prompt Design. Generative AI is emerging as a powerful tool for annotation and is being extensively researched as a substitute for human-annotated data, due to the human annotation challenges associated with the lat-

ter (Kern et al., 2023). Studies comparing human-annotated data with annotations using the ChatGPT show promising results in detecting hateful, offensive, and toxic (HOT) language (Li et al., 2024), with high accuracy. However, some highlight the persisting presence of additional bias in LLM annotations, given different contextual variations (Das et al., 2024; Okpala and Cheng, 2025). Huang et al. (2023) emphasize that hate speech detection is subjective and context-dependent, yet ChatGPT performs well even with identifying implicit hate speech. Maximizing the performance of an LLM relies on using qualitative prompts. Few-shot prompting, where the model is asked to perform a task with a few examples, generally performs better than zero-shot prompting, where no examples are given (Brown et al., 2020). Prompt engineering strategies, such as those introduced by White et al. (2025), offer adaptable structures for better results. Despite advancements, the use of LLMs as an annotation tool for sexism-related data remains sparsely researched. Given the widespread and evolving nature of sexism on social media, particularly in political discourse, further research is essential.

3 Research Design and Methodology

In this section, we provide an overview of the data used for this analysis, how we define the sexism categories, and the methodology we apply to answer our research questions.

3.1 Data

The data for this analysis consists of three distinct periods from US presidential election cycles:

- Biden vs. Trump 2020 (12 - 20 July 2020)
- Biden vs. Trump 2024 (12 - 20 July 2024)
- Harris vs. Trump 2024 (22 - 30 July 2024)

We will refer to these time frames as **BT2020**, **BT2024**, and **HT2024**, respectively, throughout the remainder of this paper.

The time frames BT2020 and BT2024 allow for a year-on-year comparison, providing insights into shifts in sexism in political discourse over time. The HT2024 time frame additionally allows for an analysis of sexism across different candidate scenarios, as it includes not only an election with two male candidates (Biden vs. Trump) but also a race featuring a female vs. male candidates (Harris vs. Trump). Including the two male vs. male candidacies aims to provide a clearer understanding of whether sexist content has increased over time alone while keeping the candidates constant.

To extract tweets for these three different time periods, we made use of two public GitHub repositories from previous research capturing discourse on 𝕏 related to the US presidential elections (Chen et al., 2022; Balasubramanian et al., 2024).

Filtering for Relevant Tweets. Political discourse on social media covers a range of topics. To limit the tweets to more relevance regarding sexism, we filtered the tweets for specific keywords. The keywords used were: she, her, woman, women, men, man, female, girl, girls, lady, feminism, feminist, gender, sex, sexism, and sexist. This allows us to pre-filter the tweets for relevancy.

Data Retrieval for BT2020. Chen et al. (2022) provide a publicly available repository containing tweets from January 2020 to June 2021. These tweets were extracted using 227 different keywords and account references. The repository consists of several `.txt` files, organized by year, month, date, and hour, with each `.txt` file containing multiple tweet IDs. The `.txt` files covering our selected time frame, BT2020, were merged together, to then randomly select a sample of tweet IDs. To retrieve the actual tweet content corresponding to the IDs, access to the 𝕏 API is required. For this analysis, we used the Basic version of the 𝕏 API v2 (X Developer Platform, 2025) and extracted tweet texts and the creation date using the `tweepy` package in Python (Roesslein, 2020). Our access period for the Basic version spanned from January 20 to February 23, 2025. The sample size for this time frame was set to 15,000 since the Basic 𝕏 API version allows retrieval of up to 15,000 tweets per month. Of the sampled tweet IDs for which requests were sent via the 𝕏 API, we ultimately obtained 6,316 tweets for this analysis. Several factors contributed to this reduction in available data.

First, we restricted our data set to English-language tweets, meaning that any non-English tweets were automatically excluded. Additionally, a noteworthy number of tweet IDs belong to already deleted tweets, making it impossible to retrieve their content. Furthermore, the retrieved tweets included both original tweets and retweets. Due to a limitation of the 𝕏 API and the `tweepy` package, the full text of the retweets cannot be retrieved. Instead, only a truncated version is available, making such data unsuitable for this analysis. Since the 𝕏 API registers each request - regardless of whether

the tweet text is available, deleted, truncated, or not in English - this leads to a considerably lower number of tweets retrieved than originally anticipated. The implications of these limitations are discussed further in section 7. The tweets were categorized into two groups, according to the keywords mentioned in the previous paragraph. All tweets containing a keyword (172) were used and a sample was chosen from tweets not containing the keywords, resulting in 431 tweets. The reason for the difference in sampled tweets arises from the piecewise approach to the OpenAI limit (see subsection 3.3).

Data Retrieval for BT2024 and HT2024. For the two election time frames in 2024 (BT2024 and HT2024) the data used for this analysis was previously extracted by Balasubramanian et al. (2024) using 44 different keywords. The corresponding public GitHub repository contained tweets from May until July 2024 and provided multiple `.csv.gz` files consisting of tweets related to the US election and information such as the tweet ID, text, url, date, number of retweets, view count, etc. For our analysis, we kept the tweet ID, the text, and the date of the tweet. After selecting the tweets that correspond to our two time frames, that is, July 12-20 and July 22-30, 2024, the tweets were filtered into two groups, according to previously mentioned keywords with relevance to sexism (see subsection 3.1), one group with tweets containing the keywords and the other group without containing them. For the BT2024 time frame, 3,000 tweets were randomly sampled per group, resulting in 6,000 tweets together. For the HT2024 time frame, 1,000-2,000 tweets were sampled per group, resulting in 3,000 tweets together. As for BT2020, the number in final categorized tweets per group differ slightly due to the piecewise approach to the OpenAI limit.

The final data used for the analysis consisted of 8,870 tweets, whose statistics are shown in Table 1.

	BT2020	BT2024	HT2024
Total Number	431	5,630	2,809
With keywords	172	2,788	930
Without keywords	259	2,842	1,879

Table 1: Statistics of the final dataset.

3.2 Sexism Categories

We classify sexism into distinct categories using definitions similar to those of other studies (Glick, 2019; Jiang et al., 2022; Rodríguez-Sánchez et al., 2021; Sharifirad et al., 2018; Swim et al., 2004).

The sexism categories were defined as follows:

- **Sexist:** Tweets that discriminate, demean, or reinforce stereotypes based on gender, including offensive language, objectification, slurs, or preserving harmful gender roles. Tweets that discuss the topic of sexism but not in a way that is offensive towards people of certain genders.
- **Non-Sexist:** Tweets unrelated to gender bias, respectful or inclusive in tone, and free of gender-based stereotypes or discrimination.

For the finer-grained categories, the following were chosen:

- **Covert and Subtle Sexism:** Tweets that show unequal treatment that is not overtly hostile but reinforces systemic inequality. Masking sexism as a positive sentiment, depicting women as incompetent or unsuited for specific roles.
- **Discrediting:** Tweets that undermine women's competence, achievements, or worth without meaningful critique, often dismissing them outright or marginalize women from decision-making and public discussions.
- **Objectification and Sexual Harassment:** Tweets that reduce women to their physical appearance, treating them as objects of desire rather than individuals with agency or intellect. Tweets that use sexualized language to intimidate women in the political sphere.
- **Remarks - Awareness and Advocacy:** Remarks or information highlighting sexism or advocating for gender equality in a way that is not offensive or derogatory. These kind of tweets often aim to expose, discuss, or address sexism constructively.
- **Stereotyping:** Tweets that enforce traditional gender roles or suggest that women should occupy lower social, economic, or political statuses due to traditional or ideological beliefs.

These finer-grained categories were chosen because they capture types of sexism that are particularly relevant within the context of political discourse. The category *Remarks - Awareness and advocacy* was specifically included to analyze whether informative discussions about sexism increase over time and whether a female presidential candidate leads to more public discussions, awareness, and potentially more positive narratives about sexism in politics.

For a more detailed overview of the categories, including the complete definitions and corresponding examples, refer to the prompts in Appendix A.

3.3 Methods

Classification Approaches. To classify tweets into defined sexism categories, we compare two classification approaches using GPT-4. The first approach follows a **single-step categorization**, in which GPT-4 directly categorizes each tweet as either *Non-Sexist* or into one of the finer-grained categories. The second approach consists of a **two-step categorization process**: First, GPT-4 classifies the tweets as either *Sexist* or *Non-Sexist*; all tweets that were classified as *Sexist* are further categorized into finer-grained categories.

To compare the two-step and single-step GPT-4-based classification approaches, we begin by addressing RQ1. As metrics for overall comparison of the prompting approaches for RQ1, we used accuracy and Cohen's Kappa index; for category-wise comparison, we used recall and precision.

Tweet examples illustrating the alignment and difference between single-step and two-step categorization are provided in the Appendix B.

Human Annotation. For the comparison of the two classification approaches, a subsample data set of 300 tweets was selected and manually annotated. To ensure that the annotated tweets represent different cases, the selected 300 tweets are composed as follows: 25% of the tweets were labeled as *Sexist* but classified into different finer-grained categories by both approaches. 25% of the tweets were labeled as *Sexist* by one approach but *Non-Sexist* by the other approach. 40% of the tweets were classified as the same finer-grained *Sexist* category by both approaches. 10% of the tweets were labeled *Non-Sexist* by both approaches. This selection guarantees the representation of cases where the two approaches differed or aligned. The 300 tweets were then manually annotated by three annotators. First, two annotators independently annotated all 300 tweets. For these two annotators, the agreement on the 300 selected tweets, which included both the *Sexist* category (subdivided into the five fine-grained categories) and the *Non-Sexist* category, resulted in a Cohen's Kappa score of 0.394. This score is generally considered minimal agreement (McHugh, 2012). Because of this low agreement, a third annotator reviewed the annotations. If both of the first two annotators assigned the same category to a tweet, that category was retained. In cases where their categorization differed, the third annotator reviewed the tweet and either chose the more appropriate category or accepted both if either categories were deemed valid. This serves as the final human annotation, used for the analysis. The purpose of this annotation was to determine which of the GPT-4-based approaches better aligned with human judgment, used as the ground truth in this analysis.

4 Results

In this section, the results obtained from the annotated tweets are presented. First, the prediction quality of the different categorization approaches is evaluated by comparing them to the human annotations (**RQ1**). Then, the change in the frequency of the sexism categories over time is analyzed (**RQ2**).

In total, 8,870 tweets were annotated by both single-step and two-step categorization: 430 tweets for the time frame BT2020, 5,630 tweets for BT2024, and 2,809 tweets for HT2024.

4.1 Comparison of Single- and Two-Step Categorization

When comparing the human annotation with GPT-4 categorization, a tweet is counted as correctly annotated by GPT-4 if the given category corresponds to one of the final human-annotated categories. In the following results, we refer to the human annotation as the ground truth.

Metric	**Single-Step**	**Two-Step**
Accuracy	0.510	0.503
Confidence Interval	[0.452, 0.568]	[0.445, 0.561]
Cohen's Kappa	0.416	0.380

Table 2: Classification metrics for single- and two-step categorization, taking human-annotated data as the ground truth. The square brackets show the confidence interval: [lower bound, upper bound].

In Table 2, the classification metrics chosen to compare the categorization approaches are depicted. The GPT-4 predictions for the single-step categorization have an accuracy of 51.0%, which means that 51.0% of the tweets were assigned to the correct category. Two-step categorization attained a similar accuracy of 50.3%. The accuracy confidence intervals for both approaches overlap, meaning there is no statistical significant difference between the two approaches. Cohen's Kappa lies at 0.416 for the single-step process, which is considered weak agreement, and at 0.380 for the two-step process, which is considered minimal agreement (McHugh, 2012).

To get a better impression of how well the two approaches categorize the tweets, it is useful to

additionally look at classification metrics per finer-grained sexism category. Figure 1 depicts two confusion matrices, one for each categorization method, showing the agreement (in %) between the GPT-4 categorization and the human annotations. Darker fields indicate higher percentages and, therefore, higher agreement, while lighter fields represent lower agreement.

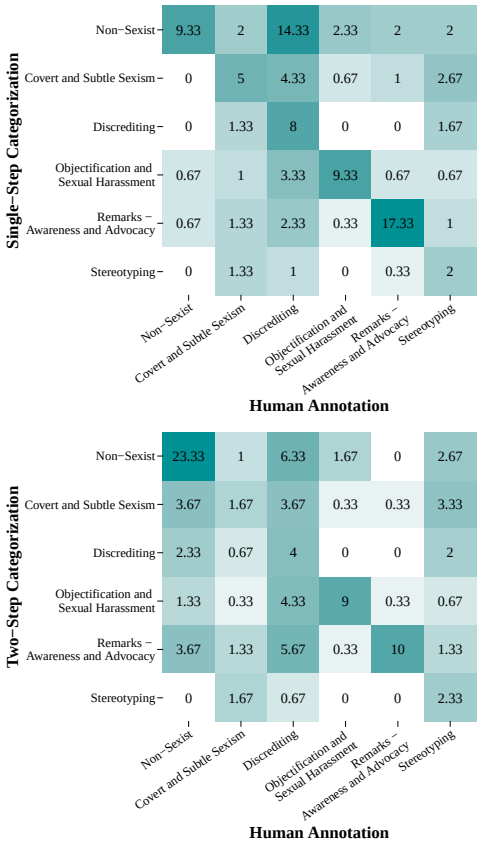


Figure 1: Confusion matrix of agreement between single-step (top) and two-step (bottom) categorization and human annotation

In the confusion matrices for single-step categorization (top) and two-step categorization (bottom), the off-diagonal elements for the single-step approach are slightly lighter, indicating fewer misclassifications. Additionally, the diagonal values for the single-step classification are mostly higher than those of their corresponding cells in the two-step matrix, suggesting that the single-step categorization achieves greater overall agreement with human annotations.

To further compare how the approaches performed, we looked at the precision and recall for each finer-grained category (see Table 3).

For single-step categorization, recall is highest for *Remarks - Awareness and Advocacy* (0.754), *Discrediting* (0.727), and *Objectification and Sexual Harassment* (0.596). The remaining categories have a recall below 0.5. For the two-step categorization, the highest recall is for *Non-Sexist* (0.667), which in the single-step categorization has the lowest recall. The categories *Objectification and Sexual Harassment* and *Stereotyping* achieve a similar recall in the single-step and two-step categorization. However, all other categories have a recall below 0.5 in the two-step categorization, which is lower than for the single-step categorization.

The precision in single-step categorization is highest for *Non-Sexist* (0.875), followed by *Remarks - Awareness and Advocacy* (0.812) and *Objectification and Sexual Harassment* (0.737). The other categories have a precision below 0.5. In two-step categorization, precision is higher for *Remarks – Awareness and Advocacy* (0.938) but lower for *Non-Sexist* (0.680) compared to single-step categorization. The precision for *Objectification ans Sexual Harassment* remains similar for both approaches. The other categories have a precision below 0.3. Table 3 also confirms the results seen in Table 2, where we assessed overall performance of the two approaches: better values are achieved for single-step categorization compared to two-step categorization.

Overall, single-step categorization outperformed two-step in most categories, as both recall and precision are higher. The accuracy is similar for both approaches, but Cohen's Kappa is higher for single-step categorization. However, despite the higher Kappa for single-step categorization, it remains quite low, indicating only minimal agreement with human annotations. Consequently, these results should be interpreted with caution.

In the remainder of this section, where the category distribution is analyzed over time, and to answer **RQ2**, only the results for the single-step categorization are reported. All corresponding analyses for two-step categorization can be found in Appendix A.

4.2 Relative Frequencies of Categories

In Table 4, the relative frequencies of sexism categories are presented for different time frames, determined by single-step categorization. Since we have different numbers of tweets per time frame, the relative frequencies are assessed instead of the absolute. Table 4 shows that the relative frequency is highest for the *Non-Sexist* category across all three time frames: 90.72%, 85.22%, and 58.70% for BT2020, BT2024, and HT2024, respectively. When compar-

	Single-Step		Two-Step	
	Recall	Precision	Recall	Precision
Non-Sexist	0.292	**0.875**	**0.667**	**0.680**
Covert and Subtle Sexism	0.366	0.417	0.128	0.250
Discrediting	**0.727**	0.240	0.444	0.162
Objectification and Sexual Harassment	**0.596**	**0.737**	**0.562**	**0.794**
Remarks - Awareness and Advocacy	**0.754**	**0.812**	0.448	**0.938**
Stereotyping	0.429	0.200	0.500	0.189

Table 3: Classification metrics per sexism category for single- and two-step categorization.

	BT2020	BT2024	HT2024
Non-Sexist	**90.72**	**85.22**	**58.70**
Covert and Subtle Sexism	0.70	1.55	3.06
Discrediting	5.80	9.01	**27.91**
Objectification and Sexual Harassment	1.16	1.17	1.32
Remarks - Awareness and Advocacy	1.62	2.42	8.22
Stereotyping	0.00	0.64	0.78

Table 4: Relative frequency of sexism categories according to single-step categorization by time frame

ing BT2020 with BT2024 - the two election time frames where we had male vs. male candidates - single-step categorization suggests that *Non-Sexist* tweets decreased slightly (-5.50%). Whereas, when the election periods where two males were candidates, BT2020 and BT2024, are compared to the time frame HT2024 (female vs. male), we can see that *Non-Sexist* tweets became increasingly less prevalent (-32.02% and -26.52%, respectively).

The relative frequency of sexist tweets additionally increases when comparing male vs. male with female vs. male election periods, especially for the sexism categories *Covert and Subtle Sexism*, *Discrediting*, and *Remarks - Awareness and Advocacy*. The category *Discrediting* has the highest relative frequency (27.91%) for the election period HT2024 compared to the other categories and the election periods BT2020 and BT2024. In Appendix C, Table 6 the additive and multiplicative changes between the three time frames are displayed.

When looking at the multiplicative change in relative frequency, we can observe the following. Comparing BT2020 with BT2024, single-step categorization suggests that sexist tweets became increasingly prevalent. *Covert and Subtle Sexism* had the largest relative increase, more than doubling in prevalence. *Discrediting* and *Remarks - Awareness and Advocacy* each increased by about 50%.

Comparing BT2020 to HT2024, these three categories (*Covert and Subtle Sexism*, *Discrediting*, and *Remarks - Awareness and Advocacy*) showed an even greater increase – up to 5 times as much. Meanwhile, *Non-Sexist* tweets decreased by 35%. The category *Objectification and Sexual Harassment* exhibited the least change in time frames. In particular, in the first time frame, no tweets were classified as *Stereotyping*, though it is essential to consider that fewer tweets were classified in this period, which may have affected the results.

These results indicate that sexist tweets seem to have slightly increased between 2020 and 2024 and increase even more when a female is running for presidency. These results will be discussed in more detail in section 6.

It is important to keep in mind that these interpretations are based on single-step categorization, which, as shown earlier in this section, has only limited reliability. In Appendix C, Table 5 show the category distribution according to two-step categorization and its changes over time. However, it is crucial to note that the two-step approach performed comparatively poorly, making its distribution and observed changes over time less reliable.

In Figure 2, the distribution of sexist categories over time is shown for the year 2024. The figure reveals that shifts in the distribution occurred suddenly rather than gradually, particularly when the presidential candidates changed and Harris replaced Biden. When looking at the shift for each sexism category, *Discrediting* and *Remarks - Awareness and Advocacy* have the steepest increase. This also reflects the results seen in Table 4.

In Appendix C in Figure 3 the same figures can be seen for two-step categorization. Also, in Appendix C, Figure 4 the distribution for all categories (*Sexist* and *Non-Sexist*), according to single-step and two-step categorization, can be seen for each of the three time frames.

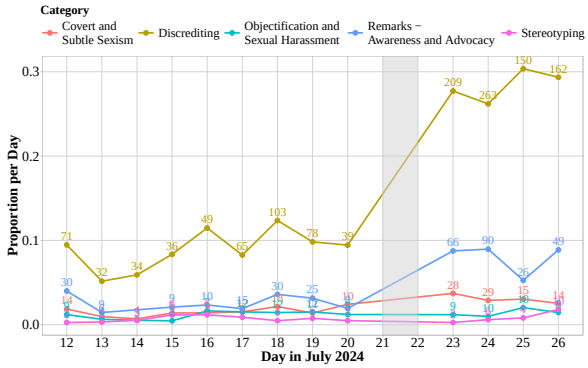


Figure 2: Distribution of sexist categories according to single-step categorization over time (gray area separating the two time frames BT2024 and HT2024)

5 Discussion

The results for **RQ1** show that single-step categorization outperforms the two-step approach. The findings indicate a higher Cohen's Kappa, as well as better precision and recall for most sexist categories. However, the single-step categorization approach achieved weak agreement with human annotations, indicating that GPT-4 struggles to categorize finer instances of sexism. Although GPT-annotation has demonstrated promising results in the identification of hateful, offensive, and toxic language on social media (Li et al., 2024), these results contradict our findings. This could be due to the complexity of sexist language, making it harder for GPT-4 to detect, or the differences in methodology and the limitations presented in section 7. The minimal inter-annotator agreement additionally indicated that sexist content is complex and can be perceived differently by individuals. This shows that the challenge lies not only in the limitations of GPT-4 but also in the subjective nature of sexism classification itself. In contrast to Plaza et al. (2023) and Jiang et al. (2022), where classifying content as sexist or non-sexist before classifying it into finer categories proved more effective, our findings suggest the opposite.

The results for **RQ2** show a shift in sexist content across the three election periods. The relative frequency of sexist tweets increased from 2020 to 2024, with a particularly higher rise during the Harris vs. Trump 2024 election. These findings align with the previous research, showing that female candidates face increasing sexism in political discourse online (Bracic et al., 2019; Knuckey, 2019).

As seen in section 4, *Covert and Subtle Sexism* became increasingly more prevalent from 2020 to 2024, suggesting that sexist comments are becoming less explicit and more complex over time. The frequency of tweets that exhibited *Remarks - Awareness and Advocacy* regarding sexism particularly increased between the two male vs. male elections and HT2024. This indicates that a female candidate contributes to more discussions surrounding information about sexism or advocating for gender equality. The category *Discrediting* also had one of the highest relative frequency changes between elections with male vs. male candidates and female vs. male candidates. This increase in *Discrediting* tweets observed during HT2024 aligns with research by Falk (2010), which found that female politicians are often portrayed as less competent or natural compared to male candidates. However, the findings contradict Tromble and Koole (2020), who found no clear differences in the tone of messages directed at female and male politicians. This discrepancy could be explained due to the increasing polarization of US politics in recent years, especially on the platform 𝕏.

6 Conclusion

This research aimed to investigate sexist content in political discourse on social media during the 2020 and 2024 US election campaigns, comparing different time frames and candidate gender scenarios. Two approaches were used to detect sexism, and GPT-4's role as a data annotation tool was evaluated. For **RQ1**, the results showed that the single-step categorization outperformed the two-step approach, but both had limited reliability and low agreement with human annotations. This highlights GPT-4's limitations in sexism detection and the need for improved classification methods for social phenomena such as sexism. For **RQ2**, sexist discourse increased between 2020 and 2024, with a notable rise when Kamala Harris was a presidential candidate. These findings suggest female candidates continue to face gender-based discrimination in political discussions. At the same time, the challenges of detecting sexism are reflected both in the low human inter-annotator agreement and the model's accuracy. This underlines the need for further research on capturing complex social phenomena such as sexism in computational research and emphasizes the importance of refining LLM-based sexism detection to support research on gender bias.

7 Limitations

This additional section points out several key limitations, which could potentially pave the way for future research.

Data Retrieval and API Constraints. A major limitation in the BT2020 timeframe is the availability and retrieval of tweets. Since data retrieval relied on tweet IDs from an existing dataset (Chen et al., 2022), many tweets were no longer accessible at the time of retrieval. Tweets that had been deleted by users or removed by the platform could not be retrieved, yet they still counted as requests due to the 𝕏 API's limitations. As controversial or highly offensive tweets may be more likely to be deleted, this introduces a potential bias. The BT2020 timeframe could underrepresent more extreme and offensive types of sexism, as tweets that provoked backlash or violated platform policies could have been removed. Additionally, the 𝕏 API does not provide full-text access to retweets. Since retweets were included in the tweet ID dataset from Chen et al. (2022), when retrieving them, we obtained a truncated text, making them unsuitable for this analysis. Since the API does not allow pre-filtering based on whether a tweet is an original post or a retweet, extensive computational time was spent obtaining tweets that were not usable. The Basic 𝕏 API version also limits the number of queries to 15 requests per 15 minutes, resulting in a long data collection period. Future research could explore alternative data retrieval methods, such as higher-level API access or pre-filtered data sets such as the data set for 2024 (Chen et al., 2022), to minimize data loss and computational time. Other research ideas for the future could expand the analysis beyond 𝕏. With the increasing role of platforms like TikTok, future research could use the TikTok API - which allows for quick keyword-based data collection without high computational time or major limitations - to reproduce this analysis. This would also enable researchers to examine sexist discourse across multiple social media platforms, providing a more comprehensive picture of sexism in online political discourse.

Annotation Bias. The annotation procedure potentially introduces a source of bias due to the limited number of annotators and their sociodemographic diversity. All three annotators in this study are white females with a shared social and cultural background, potentially influencing the perception of sexist content. More diverse annotators, including individuals of different genders, ethnicities, and political perspectives, could provide a broader, less biased understanding of how to define sexist language in political debate. Additionally, the very low level of agreement between the first two annotators indicates that classifying sexism into fine-grained categories is a challenging and subjective task, even among individuals with similar backgrounds. As a result. the reported accuracy scores of GPT-4 should be interpreted with caution. Future research could focus on extending the annotation process in order to improve the classification reliability and strengthen the results of the research questions.

Keyword Discrepancies Between Time Frames. The 2020 data set was created using 227 keywords and account references (Chen et al., 2022), while the 2024 data sets are based only on 44 keywords (Balasubramanian et al., 2024). When comparing these, we found that only 10 keywords were identical in both data sets. Although some differences in the keywords are obvious, e.g., election-specific keywords such as "Trump2020" or "Harris2024," the overall difference in keyword quantity may have influenced the comparability of sexist content between the three time frames. A potential extension of this paper could be to reproduce the analysis by first generating a new list of keywords and extracting new tweets for each election time frame. This approach would address the issue of keyword discrepancies and also resolve the challenge of retrieving previously deleted tweets, as described in the Data Retrieval and API Constraints paragraph.

Platform Evolution. An important limitation when comparing 2020 and 2024 is the change in the social media platform 𝕏. Following Elon Musk's acquisition of Twitter in October 2022, there were significant shifts in content moderation policies (Conger and Hirsch, 2022). While some previously suspended right-leaning accounts were restored, many left-leaning users left the platform (Barrie, 2023). As a result, the user base between 2020 and 2024 changed, which may have influenced the types of content shared and the tone of political discourse. This implies that the results should be interpreted within the context of 𝕏 specifically, rather than as representative of the general population in the US. Future research could address this limitation by incorporating data from other platforms (e.g., TikTok or Reddit) or modeling changes in the

platform's user base over time. Despite this limitation, the results still indicate a notable increase in sexist content from the BT2024 time frame to the HT2024.

Candidate-Specific Factors. Sexist language is rarely isolated from other forms of marginalization. For instance, Kamala Harris is biracial, a stepmother, and a female candidate in a male-dominated office. This study centers on sexism and does not take other factors such as race, religion, or family structures into account. Consequently, some tweets labeled as sexist may be intersectional, while other tweets motivated by sexism but amplified by race or parental status could be under-captured. A better picture of online hostility might come from extending the taxonomy to include overlapping categories to control for other candidate-specific factors.

Contextual Differences Between Election Periods. Finally, when analyzing the results, the political context surrounding the 2020 and 2024 elections must be considered. The 2020 election period occurred during the COVID-19 pandemic. Although the 2020 election was dominated by online discussions and political discourse surrounding the pandemic, the 2024 election took place after the pandemic, which could lead to different discussion topics and a greater focus on other time-relevant topics. The presence or absence of major external events may have changed the way sexism manifested in online political discourse, making direct comparisons between time frames more complex. To account for this limitation, the tweets for both time frames were already filtered by specific keywords that could potentially be linked to sexism, as described in subsection 3.3. One approach to further extend this analysis could focus on longitudinal tracking of sexist discourse beyond election cycles. Instead of focusing on a short 9-day election period, future research could analyze sexism in political discourse during a broader time period. This could help better understand whether the increase or decrease in sexist content in political discussions is temporary and event-driven or whether it indicates a broader societal trend.

References

Ashwin Balasubramanian, Vito Zou, Hitesh Narayana, Christina You, Luca Luceri, and Emilio Ferrara. 2024. A public dataset tracking social media discourse about the 2024 u.s. presidential election on twitter/x. *Preprint*, arXiv:2411.00376.

Christopher Barrie. 2023. Did the Musk takeover boost contentious actors on Twitter? *Harvard Kennedy School (HKS) Misinformation Review*, 4(4).

Julia C Becker and Stephen C Wright. 2011. Yet another dark side of chivalry: Benevolent sexism undermines and hostile sexism motivates collective action for social change. *Journal of personality and social psychology*, 101(1):62.

Ana Bracic, Mackenzie Israel-Trummel, and Allyson F Shortle. 2019. Is sexism for white people? gender stereotypes, race, and the 2016 presidential election. *Political Behavior*, 41(2):281–307.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Emily Chen, Ashok Dev, and Emilio Ferrara. 2022. #election2020: the first public twitter dataset on the 2020 us presidential election. *Journal of Computational Social Science*, 5:1–18.

CNN Politics. 2024. Biden will not seek reelection; endorses harris. Accessed: 2025-03-15.

Kate Conger and Lauren Hirsch. 2022. Elon Musk Completes $44 Billion Deal to Own Twitter. *The New York Times*.

Amit Das, Zheng Zhang, Najib Hasan, Souvika Sarkar, Fatemeh Jamshidi, Tathagata Bhattacharya, Mostafa Rahgouy, Nilanjana Raychawdhary, Dongji Feng, Vinija Jain, Aman Chadha, Mary Sandage, Laura-marie Pope, Gerry Dozier, and Cheryl Seals. 2024. Investigating annotator bias in large language models for hate speech detection. In *Neurips Safe Generative AI Workshop 2024*.

Erika Falk. 2010. *Women for president: Media bias in nine campaigns*. University of Illinois Press.

Richard L Fox and Jennifer L Lawless. 2004. Entering the arena? gender and the decision to run for office. *American journal of political science*, 48(2):264–280.

Peter Glick. 2019. Gender, sexism, and the election: did sexism help trump more than it hurt clinton? *Politics, Groups, and Identities*, 7(3):713–723.

Peter Glick and Susan T Fiske. 2001. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Jonathan Knuckey. 2019. "i just don't think she has a presidential look": Sexism and vote choice in the 2016 election. *Social Science Quarterly*, 100(1):342–358.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Trans. Web*, 18(2).

Joni Lovenduski. 2014. The institutionalisation of sexism in politics. *Political Insight*, 5(2):16–19.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Ebuka Okpala and Long Cheng. 2025. Large language model annotation bias in hate speech detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1389–1418.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Oxford English Dictionary. 2023. sexism, n.². Accessed: 2025-03-15.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information processing & management*, 57(6):102360.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of exist 2023: sexism identification in social networks. In *European Conference on Information Retrieval*, pages 593–599. Springer.

Kate A. Ratliff, Liz Redford, John Conway, and Colin Tucker Smith. 2019. Engendering support: Hostile sexism predicts voting for donald trump over hillary clinton in the 2016 u.s. presidential election. *Group Processes & Intergroup Relations*, 22(4):578–593.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Joshua Roesslein. 2020. Tweepy: Twitter for python! *GitHub*.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 107–114, Brussels, Belgium. Association for Computational Linguistics.

Janet K Swim and Laurie L Cohen. 1997. Overt, covert, and subtle sexism: A comparison between the attitudes toward women and modern sexism scales. *Psychology of women quarterly*, 21(1):103–118.

Janet K Swim, Robyn Mallett, and Charles Stangor. 2004. Understanding subtle sexism: Detection and use of sexist language. *Sex roles*, 51:117–128.

Rebekah Tromble and Karin Koole. 2020. She belongs in the kitchen, not in congress? political engagement and sexism on twitter. *Journal of Applied Journalism & Media Studies*, 9(2):191–214.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2025. A prompt pattern catalog to enhance prompt engineering with chatgpt. In *Proceedings of the 30th Conference on Pattern Languages of Programs*, PLoP '23, USA. The Hillside Group.

X Developer Platform. 2025. X API v2. Accessed: 2025-03-15.

A GPT-4 Prompts

A.1 Single-Step Approach

Single-Step Approach

I want you to perform a data annotation task. In your output, I want you to categorize a tweet into one of the following categories, considering the respective enclosed definition and three examples:
A tweet belongs to the category **Non-Sexist** if it is neutral or unrelated to sexism and it does not contain any discriminatory or gendered undertones. Examples for this category would be:
- "She gave a great speech on climate policy!"
- "We need more candidates like her!"
- "He would be better president."

A tweet belongs to the category **Remarks - Awareness and Advocacy** if it includes a remark or information highlighting sexism or advocating for gender equality in a way that is not offensive or derogatory. These comments often aim to expose, discuss, or address sexism constructively. Examples:
- "We need more women in leadership roles—diversity strengthens democracy."
- "It's disheartening to see how women candidates are treated differently from men in this election."
- "She's breaking barriers and paving the way for future generations—proud of her resilience in the face of sexism."

A tweet belongs to the category **Discrediting** if it attempts to undermine women's competence, achievements, or worth without engaging in meaningful critique, often dismissing them outright (explicit negative sentiment targeting women, often demeaning their character, intentions, or abilities), or if it is an attempt to preserve male control or marginalize women from decision-making and public discussions. Examples:
- "What does she know about politics?"
- "She's only there to meet diversity quotas, not because she's actually qualified."
- "Politics is no place for a woman—leave it to the men."

A tweet belongs to the category **Stereotyping** if it enforces traditional gender roles or suggests that women should occupy lower social, economic, or political statuses due to traditional or ideological beliefs. Examples:
- "She's just another angry feminist, probably can't even cook."
- "Her place is to support men, not lead."
- "She's just overreacting, like all women do."

A tweet belongs to the category **Objectification and Sexual Harassment** if it reduces women to their physical appearance, treats them as objects of desire rather than individuals with agency or intellect, or uses sexualized language or harassment to intimidate or silence women in the political sphere. Examples:
- "Politics is for people with brains, not just pretty faces."
- "She needs to get laid, maybe she'll calm down."
- "She's asking for it, dressing like that in public."

A tweet belongs to the category **Covert and Subtle Sexism** if it shows unequal treatment that is not overtly hostile but perpetuates systemic inequality through normalized attitudes and actions, or if it masks sexism as positive sentiment, often portraying women as needing protection, incapable of independence, or inherently suited for specific roles. Examples:
- "She should smile more if she wants to connect with people."
- "It's surprising how well she handles tough questions for a woman."
- "We need women in politics to bring a softer touch."

Your answer should only contain the name of the given category. Do not provide any other outputs or any explanation for your output.

A.2 Two-Step Approach

Two-Step Approach: First Step Prompt

I want you to perform a data annotation task. In your output, I want you to categorize a tweet into one of the following categories, considering the respective enclosed definition and three examples:
A tweet belongs to the category **Sexist** if it discriminates, demeans, or reinforces stereotype based on gender. This includes offensive language, objectification, gender-based slurs, or preserving harmful/negative gender roles. A tweet also belongs to this category if it discusses the topic of sexism, gender discrimination, or stereotypes but not in a way that is offensive towards people of certain gender. Examples for this category would be would be
- "It is insulting to women to have the obey-clause remain in the marriage service",
- "Girls shouldn't be allowed to be commentators for football games",
- "who asked you? Stupid bitch"

A tweet belongs to the category **Non-Sexist** if it is not related to sexism and do not contain any form of gender-based bias, discrimination, or stereotyping. The tweet is neutral, respectful, or positively inclusive in tone and content regarding gender. Examples for this category would be
- "We need more women in leadership roles—diversity strengthens democracy.",
- "It's disheartening to see how women candidates are treated differently from men in this election.",
- "She's breaking barriers and paving the way for future generations—proud of her resilience in the face of sexism."

Your answer should only contain the name of the given category. Do not provide any other outputs or any explanation for your output.

Two-Step Approach: Second Step Prompt

I want you to perform a data annotation task. In your output, I want you to categorize a tweet into one of the following categories, considering the respective enclosed definition and three examples:
A tweet belongs to the category **Remarks - Awareness and Advocacy** if it is/includes a remark or information highlighting sexism or advocating for gender equality in a way that is not offensive or derogatory. These kind of comments often aim to expose, discuss, or address sexism constructively. Examples for this category would be
- "We need more women in leadership roles—diversity strengthens democracy.",
- "It's disheartening to see how women candidates are treated differently from men in this election.",
- "She's breaking barriers and paving the way for future generations—proud of her resilience in the face of sexism.".

A tweet belongs to the category **Discrediting** if it is/includes an attempt to undermine women's competence, achievements, or worth without engaging in meaningful critique, often dismissing them outright (explicit negative sentiment targeting women, often demeaning their character, intentions, or abilities), or if is an attempt to preserve male control or marginalize women from decision-making and public discussions. Examples for this category would be
- "What does she know about politics?",
- "She's only there to meet diversity quotas, not because she's actually qualified.",
- "Politics is no place for a woman-leave it to the men.".

A tweet belongs to the category **Stereotyping** if it is enforcing traditional gender roles or suggesting that women should occupy lower social, economic, or political statuses due to traditional or ideological beliefs. Examples for this category would be
- "She's just another angry feminist, probably can't even cook.",
- "Her place is to support men, not lead.",
- "She's just overreacting, like all women do.".

A tweet belongs to the category **Objectification and Sexual Harassment** if it is reducing women to their physical appearance, treating them as objects of desire rather than individuals with agency or intellect, or if it using sexualized language or harassment to intimidate or silence women in the political sphere. Examples for this category would be
- "Politics is for people with brains, not just pretty faces.",
- "She needs to get laid, maybe she'll calm down.",
- "She's asking for it, dressing like that in public.".

A tweet belongs to the category **Covert and Subtle Sexism** if it shows unequal treatment that is not overtly hostile but perpetuates systemic inequality through normalized attitudes and actions, or if it is masking sexism as positive sentiment, often portraying women as needing protection, incapable of independence, or inherently suited for specific roles. Examples for category would be
- "She should smile more if she wants to connect with people.",
- "It's surprising how well she handles tough questions for a woman.",
- "We need women in politics to bring a softer touch.".

Your answer should only contain the name of the given category. Do not provide any other outputs or any explanation for your output.

B Alignment and Differences between Single-Step and Two-Step Approach

Alignment between Approaches

Tweet
The first female president should be honorable and should not have slept her way to the top.
Categorization for both Single-Step and Two-Step Approach
Discrediting

Difference between Approaches

Tweet
She has no skills. So maga it is
Categorization for Single-Step Approach
Discrediting
Categorization for Two-Step Approach
Non-Sexist

C Additional Results

	BT2020	BT2024	HT2024
Non-Sexist	98.144	95.702	88.501
Covert and Subtle Sexism	0.232	0.320	0.570
Discrediting	0.696	2.007	5.732
Objectification and Sexual Harassment	0.000	0.675	0.783
Remarks - Awareness and Advocacy	0.000	0.711	3.667
Stereotyping	0.928	0.586	0.748

Table 5: Relative frequency of sexism categories according to two-step categorization by time frame

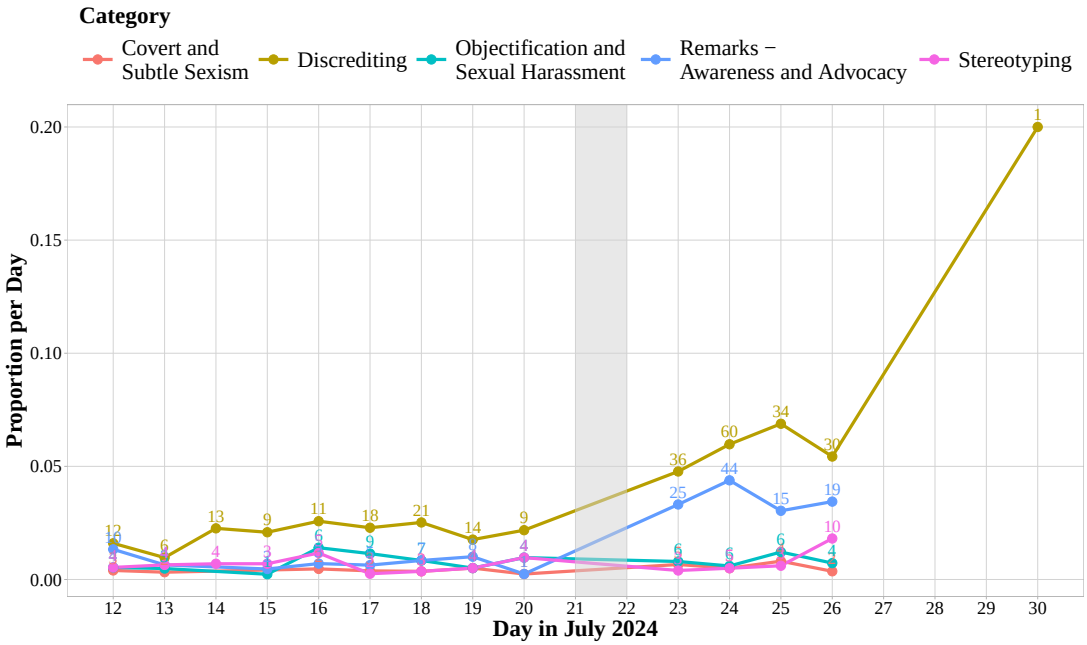


Figure 3: Distribution of sexist categories according to two-step prompting over time in 2024 (gray area separating the time frames BT2024 and HT2024)

Change from BT2020 to BT2024

	BT2020	BT2024	BT2020 to BT2024 (additive)	BT2020 to BT2024 (multiplicative)
Non-Sexist	90.72	85.22	-5.50	0.94
Covert and Subtle Sexism	0.70	1.55	+0.85	2.22
Discrediting	5.80	9.01	+3.21	1.55
Objectification and Sexual Harassment	1.16	2.42	+0.01	1.01
Remarks - Awareness and Advocacy	1.62	2.42	+0.80	1.49
Stereotyping	0.00	0.64	+0.64	Inf

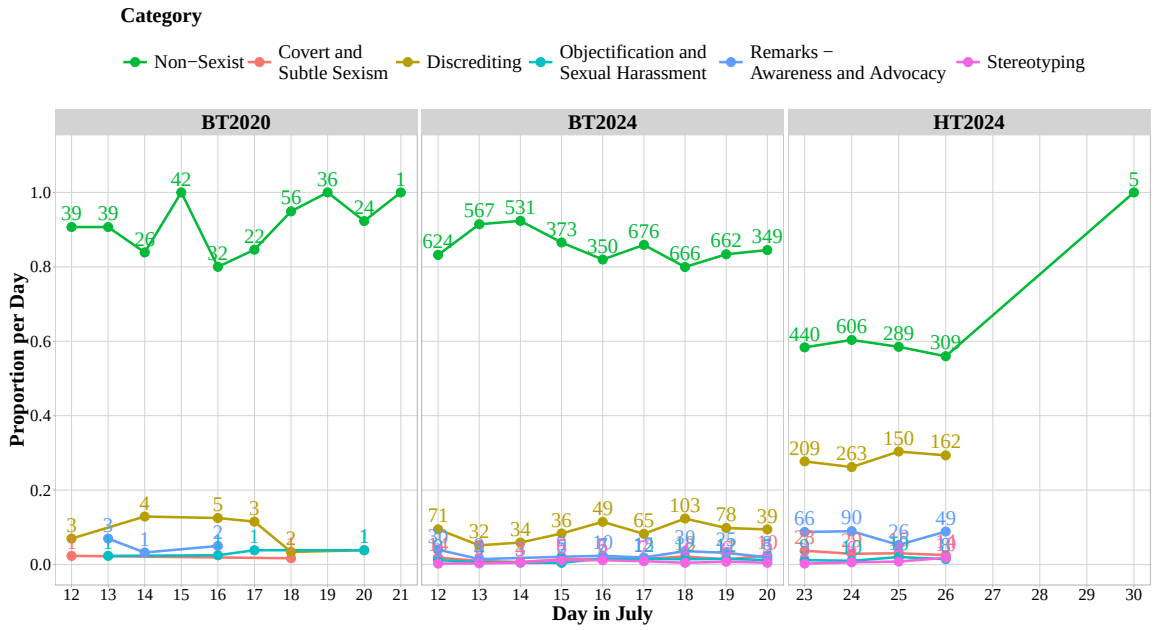
Change from BT2020 to HT2024

	BT2020	HT2024	BT2020 to HT2024 (additive)	BT2020 to HT2024 (multiplicative)
Non-Sexist	90.72	58.70	-32.02	0.68
Covert and Subtle Sexism	0.70	3.06	+2.36	4.40
Discrediting	5.80	27.91	+22.11	4.81
Objectification and Sexual Harassment	1.16	1.32	+0.16	1.14
Remarks - Awareness and Advocacy	1.62	8.22	+6.60	5.06
Stereotyping	0.00	0.78	+0.78	Inf

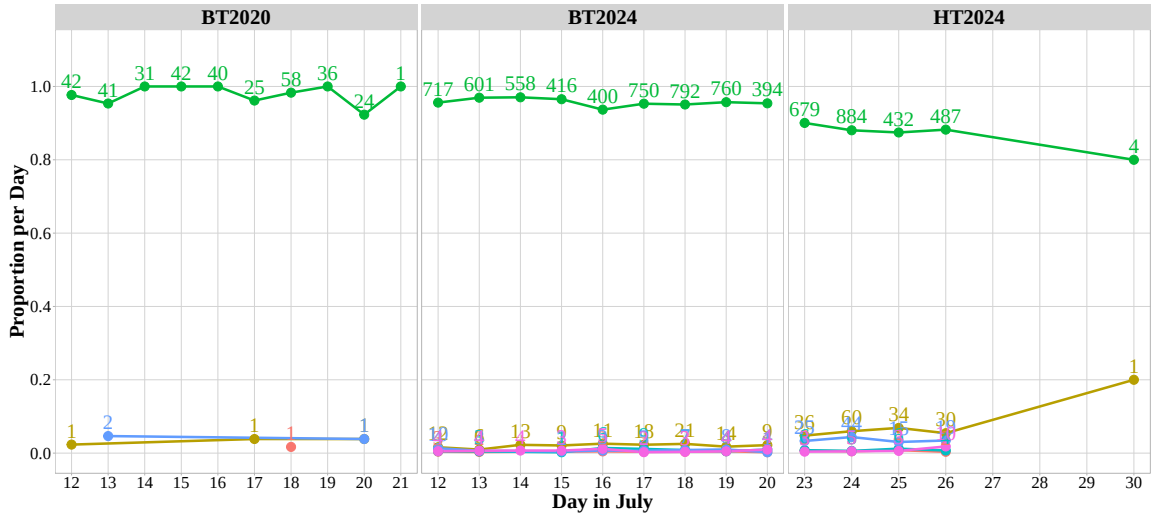
Change from BT2024 to HT2024

	BT2024	HT2024	BT2024 to HT2024 (additive)	BT2024 to HT2024 (multiplicative)
Non-Sexist	85.22	58.70	-26.52	0.69
Covert and Subtle Sexism	1.55	3.06	+1.51	1.98
Discrediting	9.01	27.91	+18.90	3.10
Objectification and Sexual Harassment	1.17	1.32	+0.15	1.12
Remarks - Awareness and Advocacy	2.42	8.22	+5.80	3.40
Stereotyping	0.64	0.78	+0.14	1.22

Table 6: Change in relative frequency of sexism categories according to single-step categorization by time frame



(a) Single-Step Categorization



(b) Two-Step Categorization

Figure 4: Distribution of all categories over time according to single-step (a) and two-step (b) categorization