Sentimental Gender Bias: Insights from Reddit Posts

Andri Rutschmann[1], Peer Saleth[1], Jördis Strack[1]

[1]Konstanz University

[andri.rutschmann, peer.saleth, joerdis.strack]@uni-konstanz.de
https://github.com/joerdisstrack/gender_bias_CSS_RP

Abstract

This study explores the relationship between sentimental gender bias and text complexity in online discussions about U.S. politicians on Reddit. Building on the work of Marjanovic et al. (2022), we investigate whether posts exhibiting gender bias—quantified through valence, arousal, and dominance (VAD) values and toxicity measures—correlate with lower text complexity. Utilizing a sample of approximately 137,000 Reddit posts, we apply the LEIA sentiment analyzer and map the resulting sentiments onto the VAD dimensions. Using Perspective API we assess toxicity. We then examine correlations between these metrics and syntactic (Gunning Fog Index) and semantic (Type Token Ratio) text complexity measures. Our findings indicate no substantial evidence that gender-biased or toxic posts are associated with lower text complexity, challenging assumptions that emotional or biased language correlates with simpler textual structures. Despite low correlation coefficients across all hypotheses, this research contributes to understanding how online gender bias manifests in language complexity.

Introduction

Gender bias on social media has been an upcoming topic for research, revealing how these platforms often reinforce existing gender inequalities. Social media algorithms prioritize content based on engagement, which can promote misogynistic and sexist content. This occurs because such content often generates high levels of interaction, including both support and outrage, which the algorithms interpret as valuable engagement (Gerrard et al., 2022). Consequently, harmful stereotypes and biases are not only maintained but also normalized within online communities, creating hostile environments for women and other marginalized groups. Social Media platforms often lack sufficient mechanisms to moderate gendered harassment effectively, leading to environments where women are more likely to experience hostile behavior online. This can have significant impacts, including self-censorship and reduced participation in public discourse by women (ODI, 2023).

Gender bias on social media is not only present in general discourse, but is also particularly evident in political discussions. Previous work by Haraldsson and Wängnerud (2019) and Barker and Jurasz (2019) indicates that female politicians are more likely to face gender based disadvantages and gendered attacks, including comments that focus on their appearance, personal life, or gender roles, rather than their policies or political actions. This form of bias is not only demoralizing, but also reinforces traditional gender stereotypes, which can undermine the public's perception of female politicians and their legitimacy as leaders. A meta-study by Haraldsson and Wängnerud (2019) found that media sexism significantly reduces the political ambition of women on a global scale. Specifically, their research demonstrated that in countries where media sexism is more prevalent – measured by the lower proportion of women portrayed as news subjects or experts – there is a corresponding decrease in the share of women who become political candidates. This suggests that biased media portrayals create a discouraging environment that creates barriers for women to participate in politics. A study by Barker and Jurasz (2019) found that online misogyny and violence against women in politics create significant barriers to women's equal participation in public and political life. Their research highlighted how these online abuses, often driven by gender discrimination, prevent women from freely expressing their opinions and participating in public discourse. The study underscores the need for more robust legal and policy responses to address the growing issue of online misogyny.

Our study aims to contribute to research on gender bias in political discussions about and directed at (US) politicians. We build on previous work by Marjanovic et al.'s "Quantifying Gender Biases Towards Politicians on Reddit" (2022). This paper offers a comprehensive analysis of how gender bias manifests itself in political online discussions on Reddit. Utilizing an extensive dataset comprising 10 million posts from 2018 to 2020, Marjanovic et al. aimed to investigate inequalities in how male and female politicians are discussed on the platform. They found no significant coverage bias, as both male and female politicians received similar levels of public attention. However, female politicians were more often discussed alongside male politicians, indicating a tendency for cross-gender mentions. Additionally, a nominal bias was observed: male politicians were more frequently referred to by their surnames (a sign of respect), while female politicians were often addressed by their full or first names, suggesting a subtle gender bias in the way they are addressed. Evidence for sentimental gender bias was found as well.
For the purpose of our analysis we focus on sentimental gender bias since we believe that this is the only dimension we can meaningfully analyze using VAD scores and toxicity.
Marjanovic et al. (2022) operationalized sentimental gender bias as the mean difference in valence, arousal, and dominance between those posts directed at women and those directed at men. Valence describes the amount of pleasant sentiment a post contains. Love, generally speaking, is a rather pleasant sentiment and thus scores higher on valence than anger or hate. Arousal informs us about how energizing the sentiment(s) contained in a post are. Excitement, for example, is associated with higher arousal than boredom or depression. Dominance reflects to what degree controlling language is used in a post. Anger or hate, for example, score higher on the dominance dimension than fear.
Valence, arousal, and dominance scores for each individual post were obtained using the NRC VAD Lexicon (Mohammad, 2018) to assess the individual VAD scores of each word and subsequently averaging the values for each of the three dimensions over the whole post.
Using t-tests of means Marjanovic et al. (2022) show that posts directed at men generally score higher on the valence and dominance dimensions. Predicted effect sizes for these phenomena are extremely small, though.

In the following we will delve into the theoretical background of this project, present our methodological approaches and conclude with the presentation and discussion of the results of our analysis, naming limiting factors and proposing avenues for future research.

Theory and Background

In this section we give a detailed account of our theoretical assumptions and background. We first present the concepts of hostile and benevolent sexism and explain how we believe them to be connected to sentiment and toxicity. We then introduce the concept of text complexity and theoretically connect it to sentimental gender bias. Finally, we present our theoretically motivated research question and hypotheses.

Sexism

Sexism is often conceptualized in two distinct dimensions that contribute to broader gender bias: hostile sexism and benevolent sexism. Hostile sexism involves openly negative attitudes toward women, characterized by beliefs that women seek to control men, are inferior, or are untrustworthy. This form of sexism justifies male dominance, power, and the exclusion of women from positions of authority, reinforcing traditional gender roles through derogatory stereotypes (Glick & Fiske, 1996). In contrast, benevolent sexism consists of seemingly positive attitudes that idealize women in traditional roles, portraying them as pure, nurturing, and in need of male protection. While these views may appear favorable, they contribute to gender bias by confining women to subordinate roles deemed "naturally" suitable for them. Both dimensions work together to maintain patriarchal social structures, with one using overt hostility and the other employing subtle idealization to perpetuate traditional gender roles (Glick & Fiske, 1996).

Sentimental Gender Bias and Sexism

Valence, arousal, and dominance are used to encode different sentimental dimensions. As explained, they provide information about a sentiment's pleasant, energizing, and controlling nature. Since hostile and benevolent sexism are clearly associated with different attitudes and choice of language (Glick & Fiske, 1996) we assume to be able to observe sentimental gender bias by investigating differences in VAD scores across posts directed at women and men.
We do not want to imply that sexism can be directly detected via VAD scores but we venture that general tendencies of sentiment towards gender can be unraveled by analyzing VAD

scores. Since hostile sexism is generally associated with negative and dominant emotions we assume it to show itself in low valence and high dominance language. Benevolent sexism on the other hand might manifest itself in high valence language since it is mostly associated with pleasant sentiments or emotions.

Toxicity and Hostile Sexism

We use targeted toxicity as a measurement for hostile sexism due to the direct relationship between toxic behaviors and the negative attitudes that characterize hostile sexism. Hostile sexism often manifests in language and behaviors that are openly derogatory and hostile toward women, reflecting beliefs that women are inferior, manipulative, or untrustworthy. Toxic comments, which are defined as rude, disrespectful, or unreasonable, often embody these very attitudes and can therefore be a strong indicator of hostile sexism. Research shows that toxic language is frequently motivated by stereotypical gender biases, making it a suitable proxy for measuring hostile sexism (Cheng et al., 2022). The use of tools like Perspective API[1], which is designed to detect and quantify toxicity in online comments, allows for the identification and analysis of these attitudes on a large scale. This approach is particularly useful in examining social media environments, where hostile sexism may be expressed more freely and anonymously. However, not every toxic text directed at a woman is necessarily grounded in sexist motivation. Toxicity can arise from various sources of frustration or anger that may not be directly related to gender. We average the toxicity scores in text targeted at women, so we can identify a general toxic tonus toward women. Additionally, text which mentions female politicians is not automatically targeted towards them. The aggregated measure reveals a more broader pattern of hostility that may not be evident from individual comments, but nonetheless contributes to gender bias. The context sensitivity of toxicity and hostile sexism is a major task for natural language processing. Approaches using Large Scale Machine Learning models try to ensure that the analysis accounts for the specific contextualization of the comments like irony or negations, which is crucial because both sexism and toxicity can be highly context-dependent (Pavlopoulos et al., 2020).

[1] https://perspectiveapi.com/

Text Complexity and Gender Bias

A growing body of literature provides evidence of a potential link between sentiment and text complexity on both a syntactic and semantic level. Kousta et al. (2011) have found that the choice of abstract words in personal expression is affected by one's emotional state and words to be associated with different sentiments. Similarly, the length of words and their emotional content appear to determine their frequency of use (Piantadosi et al., 2011). Positive words tend to be used more often during conversation, reflecting a positive bias in word use that appears to be independent of language (Garcia et al., 2012). This positive bias can be translated into a higher number of tokens for positive words than for those with a negative connotation (Rozin et al., 2010). Yet, the impact of the latter might be asymmetrically larger with negative words being more differentiated in their meaning and resulting in stronger emotional impact upon their evaluation by readers (Peeters & Czapinski, 1990). Rozin et al. (2010) translate this observation as positive words appearing in larger numbers and thus with a larger number of tokens, while negative words are expressed using more unique expressions and thus a larger number of types.

Given the choice of words based on the emotional state of a text's writer and the information they want to convey (Kousta et al., 2011), it is crucial to investigate how sentiment might be reflected in a text's semantics and syntax. Because text design is far from arbitrary: During the age of online communication text sentiment is a vital tool to convey a specific political ideology and is used intentionally to convince a large audience (Mohammad et al., 2015). Tweets that are part of incumbent parties' election campaigns have been found to express more positive sentiment, thus higher valence, than tweets by party members who are not in office (Crabtree et al., 2020).

Gender stereotypes influence political communication similarly, with women choosing positive words that express emotions such as joy (Scott & McDonald, 2022), representing higher valence. An experiment by Stapleton and Dawkins (2022) found participants who observed an angry and emotionally charged discussion between politicians of opposing parties to adopt feelings of anger and report a greater likelihood of voting. These findings are supported by further evidence of the potential to mobilize voters through emotionally charged texts has been presented by Jones et al. (2013). This suggests that different pieces of texts might be capable of expressing higher levels of arousal

and dominance, which are then adopted by potential voters.
An analysis of semantic and syntactic complexity of US politicians' speeches revealed Trump to score notably lower across readability indices and his speeches to consist of shorter sentences and a lower TTR, which classified them as less complex than speeches of his opponents (Kayam, 2018). Interestingly, social media posts that were classified as toxic in a study presented by Morzhov (2021) also displayed a lower TTR than 'clean' texts. This contradicts previous findings that associated texts with a larger vocabulary and higher TTR to contain more words with negative connotations (Rozin et al., 2010), and requires further investigation.

Combining these theoretical constructs, we want to research the question: How are sentimental gender bias and text complexity related? In order to investigate the uncertain link between text complexity and sentiment we propose the following hypothesis:

H1: Sentimentally gender biased (VAD) Reddit posts display lower text complexity.

Depending on the hostile or benevolent nature of sentimental gender bias we expect different results. In case of a general tendency towards hostile treatment of women we expect the number of swear words to increase and overall text complexity to decrease. Since toxic content is often characterized by its use of inappropriate language we therefore motivate the following hypothesis:

H2: Reddit posts that are perceived as toxic display lower text complexity.

Finally, we aim to connect our two operationalizations of sentimental gender bias by investigating the interplay between VAD scores and toxicity by testing the hypothesis:

H3: Sentimental gender bias operationalized through toxicity and sentimental gender bias operationalized through VAD correlate positively.

Methods

In this section we present our methodological approaches. First, we explain how we derived VAD scores from categorical sentiment labels, collected toxicity ratings from the Perspective API and then used these new variables to reinvestigate the sentimental gender bias detected by Marjanovic et al. (2022). Finally, we outline our operationalization of text complexity.

Sentiment-derived VAD Scores

The study of emotions or sentiments can be divided into two main approaches. The first one defines sentiments as discrete classes (see Ekman, 2008). The other approach understands sentiments as clusters or groups in a multidimensional space. Russell and Mehrabian (1977) define this space to be spanned by the variables valence, arousal, and dominance.
Due to the importance of sentiment analysis many data sets have been put together employing either one of these general approaches to sentiment classification. Seeing the potential in unifying this non-uniform data treasure, where sentiments are sometimes classified via explicit labels and at other times as a tuple of VAD values, efforts to unify this data have been made. Buechel and Hahn (2017) propose the mapping between discrete sentiment labels and VAD dimensions. Using this idea they are able to achieve near human performance when mapping VAD values to discrete labels with an automated approach.
To test the robustness of Marjanovic et al.'s (2022) findings regarding sentimental gender bias we thus propose to employ the LEIA sentiment analyzer (Aroyehun et al. 2023) to determine the sentiment of individual posts, subsequently map these sentiments to the VAD dimensions following a mapping by Russell and Mehrabian (1977), and finally rerun Marjanovic et al.'s statistical analysis using the sentiment-derived VAD values.

The Linguistic Embeddings for the Identification of Affect (LEIA) sentiment analyzer developed by Aroyehun et al. in 2023 is a transformers based model that incorporates word masking during pre-training to enhance the learning of emotion words. It builds on BertTweet (Nguyen et al., 2020) which was trained using English social media posts from Twitter. LEIA's performance both on in-domain as well as out-of-domain data sets, such as the enISEAR (Troiano et al., 2019) or SemEval (Mohammad et al., 2018), clearly outperforms baseline approaches using the NRC emotion lexicon (Mohammad & Turney, 2010). For the enISEAR data set LEIA achieves F1 scores of 0.6 to 0.8 depending on the individual emotions classified with it achieving the best performance on *happiness* and performing worst when classifying *sadness*. The other sentiments LEIA is able to classify being *anger*, *affection*, and *fear*.
Due to this rather impressive performance in sentiment classification and low computational

requirements we chose LEIA in order to analyze the sentiments of the posts in our data set. A manual validation of LEIA's performance on a sample of 100 randomly sampled and human-annotated posts from the data yielded a weighted F1 score of 0.59. The worst per class performance was achieved for *sadness* with an F1 of 0. The validation set contained 5 posts labeled as *sadness* but LEIA did not classify a single one of them likewise. Best performance was achieved for *anger* which was the most prevalent class in the validation sample with an F1 of 0.78 and a total of 62 posts belonging to that class. For the other classes F1 scores ranged from 0.25 to 0.4. A detailed summary of these results can be found in Table 1.

After classifying the sentiment of each post in the data set we used a sentiment label to VAD representation mapping by Russell and Mehrabian (1977) to determine VAD scores for each individual post. Russell and Mehrabian tasked a group of annotators to assign VAD scores to numerous sentiments with affection, anger, fear, happiness, and sadness among them. For each of these sentiments we thus have the average rating for valence, arousal, and dominance of roughly 30 annotators per sentiment as well as the standard deviation. The exact VAD values for all five sentiments we inspect here are shown in Table 2. Due to the fact that the annotators' rating of valence, arousal, or dominance for an individual sentiment always differed – specific sentiments do not feel exactly the same to everybody after all – we decided to not just map the individual sentiments to their respective mean VAD values but rather draw VAD values from normal distributions created by the means and standard deviations of valence, arousal, and dominance of each sentiment. So instead of each post classified as *anger* receiving the same VAD values we get different VAD values for each of these posts drawn from the normal distributions of valence, arousal, and dominance as assigned to the sentiment *anger* by the annotators. Figure 13 shows a three dimensional representation of the sentiment-derived VAD values with each dot representing a single post from the data set.

Finally, we used the sentiment-derived VAD scores assigned to each post to test for differences between posts directed at female and male US politicians.
Marjanovic et al. (2022) used t-tests of mean for this purpose but visual and statistical investigation of the data revealed that VAD values for our subsample of the original data were not normally distributed. The results of respective Shapiro-Wilk tests can be found in Table 3. Given non-normally distributed variables we used the non-parametric Mann-Whitney U test (Bonferroni-corrected to account for multiple testing).
Using the original VAD values assigned by Marjanovic et al. (2022) we found posts directed at women to score significantly lower on valence, arousal, and dominance compared to posts directed at men with $p < 0.001$. Rerunning the same analysis with our sentiment-derived VAD values found posts directed at women to score higher on valence as opposed to those directed at men with $p < 0.05$. This possibly hints at the presence of benevolent sexism as posts directed at women are more likely to be associated with more pleasant sentiments. The effects for arousal and dominance were not statistically significant at $p < 0.05$. An overview of the statistical analysis along with bootstrapped confidence intervals of group means can be found in Table 4.

Toxicity

Toxicity, in the context of online communication, is typically defined as rude, disrespectful, or unreasonable comments that are likely to make someone leave a discussion. This definition encompasses offensive, abusive, and hateful comments and is central to understanding and measuring hostile sexism in digital spaces (Pavlopoulos et al., 2020). The Perspective API, a widely used tool for detecting toxicity, operationalizes this concept by providing a probability score that indicates the likelihood of a comment being perceived as toxic. The higher the score, the greater the likelihood that a comment will be recognized as containing toxic elements.
The Perspective API uses advanced models, including Transformer, CNNs, RNNs, and LSTMs, trained on millions of comments from diverse sources like Wikipedia and The New York Times. These models have shown robust performance across multiple datasets, such as the Jigsaw Multilingual Toxic Comments Challenge and the CivilComments-WILDS dataset, with AUC-ROC scores ranging between 0.87 and 0.94, demonstrating high accuracy in identifying toxic content (Lees et al., 2022). Despite this, there are inherent challenges in using such tools, particularly when it comes to detecting more nuanced forms of toxicity, such as Misogynoir – an intersectional form of sexism and racism targeted at Black women. The Perspective API has been noted to be less effective in identifying these nuanced cases, which require a deeper understanding of context and intersectionality (Kwarteng et al., 2022).
Scores from the Perspective API are typically interpreted on a scale from 0 to 1, with higher scores indicating a greater likelihood of

perceived toxicity. However, the use of this black-box API comes with several limitations. For instance, the models may not always account for the context in which comments are made, which can lead to misinterpretation of the content. Additionally, as the models are updated over time, the scores might change, leading to inconsistencies in long-term studies. Furthermore, relying on a black-box system means that how the model processes and interprets data is not fully transparent, which can pose challenges for researchers who need to understand the specific biases and limitations of the tool (Pavlopoulos et al., 2020).

We retrieved the Perspective API scores for the whole sample on July 20, 2024 at 1:38 PM. The scores were validated on a sample of 25 toxic and 25 non-toxic comments, with a threshold of 0.7 for classifying a text as toxic. This sample was annotated by a human, which then was used as ground-truth for a performance metric. It is noteworthy to underscore that the specific threshold value plays a pivotal role in this analysis; however, the outlined observation generally remains consistent across varying thresholds. A crucial point of clarification pertains to the nature of the toxicity metric, which reflects not the extent of toxicity, but rather the predictive probability assigned by the model to a given text being toxic. While one might infer that comments rated as more toxic possess a heightened likelihood of being toxic, such a direct link cannot be definitively established. The validation of the Perspective API's toxicity detection model showed strong performance metrics, with an accuracy of 0.88, indicating that the model correctly classified 88% of the comments. The F1 Score, which balances precision and recall, was 0.875, reflecting a well-rounded performance in detecting true positives while minimizing false positives and negatives. The model's precision was 0.913, meaning that 91.3% of the comments it identified as toxic were indeed toxic. The Matrix in Figure 1 underscores the drawbacks of using toxicity as a measurement for hostile sexism. The true positive case is clearly toxic, but not grounded on gender nor targeted at gender. This shows two problems: First, texts with swear words are classified as toxic even if they don't attack a person, this is shown by the example post in the False Positive case. Second, toxic text classified and targeted at a person doesn't infer that this toxicity is grounded on gender stereotypes, like shown in the false negative example post. Also, the toxicity in the false negative example is not targeted towards the detected entity. Therefore, we look at the accumulated toxicity to measure a more general tonus towards women, which can indirectly reflect toxicity based on gender stereotypes.

In figure 2 the distribution of toxicity scores for comments targeting males and females are shown. We can see that the scores tend to be power law distributed, indicating a high share of non-toxic posts and few toxic ones. The toxicity threshold is visualized by a line. There are no visual differences in toxic posts between male and female targeted posts, even when moving the threshold.

To further test for differences, we applied several statistical tests. First we log transformed the data and tested whether the log toxicity scores are normal distributed. We found a Shapiro-Wilk Statistic of 0.964 ($p < 0.001$). Due to python limitations of the scipy[2] Shapiro-Wilk implementation, p values with an $N > 5000$ might not be accurate. Therefore, we implemented a Monte Carlo Test p-value approximation ($p < 0.001$) for robustness. Data was not found to be normally distributed. Therefore, we applied a non-parametric test, to test for differences in distribution of toxicity scores between posts directed at men and women. The Mann-Whitney U Test Statistics of 1,105,051,493.0 ($p < 0.001$) found a significant difference between these distributions. Although this difference, the effect size was minimal. The analysis revealed that the mean toxicity score for comments directed at males was 0.23, with a 95% confidence interval of [0.23, 0.23]. For females, the mean toxicity score was slightly lower at 0.22, with a 95% CI of [0.22, 0.22]. The mean difference in toxicity scores between posts directed at women and men (females - males) was -0.01, with a 95% CI of [-0.01, -0.01]. The confidence intervals were bootstrapped using 1,000 iterations, ensuring robust estimates of the mean and the difference between the groups. These findings contradict the assumption of accumulated hostile sexism towards women being present in this dataset.

We found similar patterns for the total amount of toxic comments. Figure 3 shows the aggregated amount of toxic comments towards women and men. Figure 4 shows the ratio of toxic comments per total comments. These Figures also visualize how unbalanced the dataset is, with comments towards men being way more present. When accounting for this by looking at the ratio, the amount of toxic comments is also slightly higher for comments targeted at men.

To get some deeper insights we expanded the analysis and visualized toxicity scores for different subreddits. Figure 5 shows mean toxicity scores for comments targeted at males

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

and females for each subreddit present in the dataset. There are multiple insights such as that the subreddit *Feminisms* only contains comments targeted at males. For the subreddits *Socialism* and *USPolitics* the average toxicity towards females is higher but with high uncertainty. For the subreddit *Mensright* the mean toxicity towards men is higher.

The analysis revealed that toxicity scores for comments targeting both males and females are nearly identical, with only a minimal difference. Despite a statistically significant difference in distributions, the effect size is small. This suggests that there is no substantial difference in the level of toxicity directed toward males versus females in this dataset, contradicting the assumption of the presence of hostile sexism towards women. Additionally, when considering the total amount of toxic comments, the ratio slightly favors higher toxicity towards men, particularly when adjusted for the dataset's imbalance. Subreddit-specific analysis showed variability, but with high uncertainty, indicating that any observed differences are not consistent or significant across the board.

Text Complexity

Frantz et al. (2015) define text complexity as the level of challenge and sophistication required to read and understand a piece of text. It includes both syntactic and semantic components, which can be measured using the Gunning Fog index to assess who is targeted by whom using grammar (Gunning, 1952, 1969), and the type token ratio to capture the range of a text's vocabulary and its emotive content (Guiraud, 1954).

The Gunning Fog index (Gunning, 1952, 1969), is a grade-level index that is especially suited for the English language (Rodríguez Timaná et al., 2020). Based on the average number of words per sentence and the share of multisyllabic words against all words, it returns a score typically between 0 and 20 that expresses the years of education that would be required to understand a piece of text. A comparison of all posts in figure 6 reveals that subreddits like *Republican*, *Feminisms* and *Teenagers* exhibit higher Gunning Fog scores, suggesting that posts in these communities tend to have more complex sentence structures and may require a higher reading level to comprehend. Conversely, subreddits such as *Q434706* display much lower scores, indicating simpler text that is easier to read.

The focus on complex words is of special interest in this case, given the link between sentiment, word length and relayed information (Garcia et al., 2012; Piantadosi et al., 2011). The TTR is computed by dividing the number of unique words against the total number of words. So a text with a higher TTR utilizes a larger vocabulary and vice versa. Following the findings of Rozin et al. (2010), this would indicate a text containing a larger number of negative words over a relatively smaller number of positive ones. Figure 7 shows that subreddits such as *Socialism* and *USPolitics* demonstrate higher TTR, indicating the use of a more diverse vocabulary in these communities. In contrast, subreddits like *Republican* and *TwoXChromosomes* show lower TTR, suggesting a more repetitive or limited use of vocabulary.

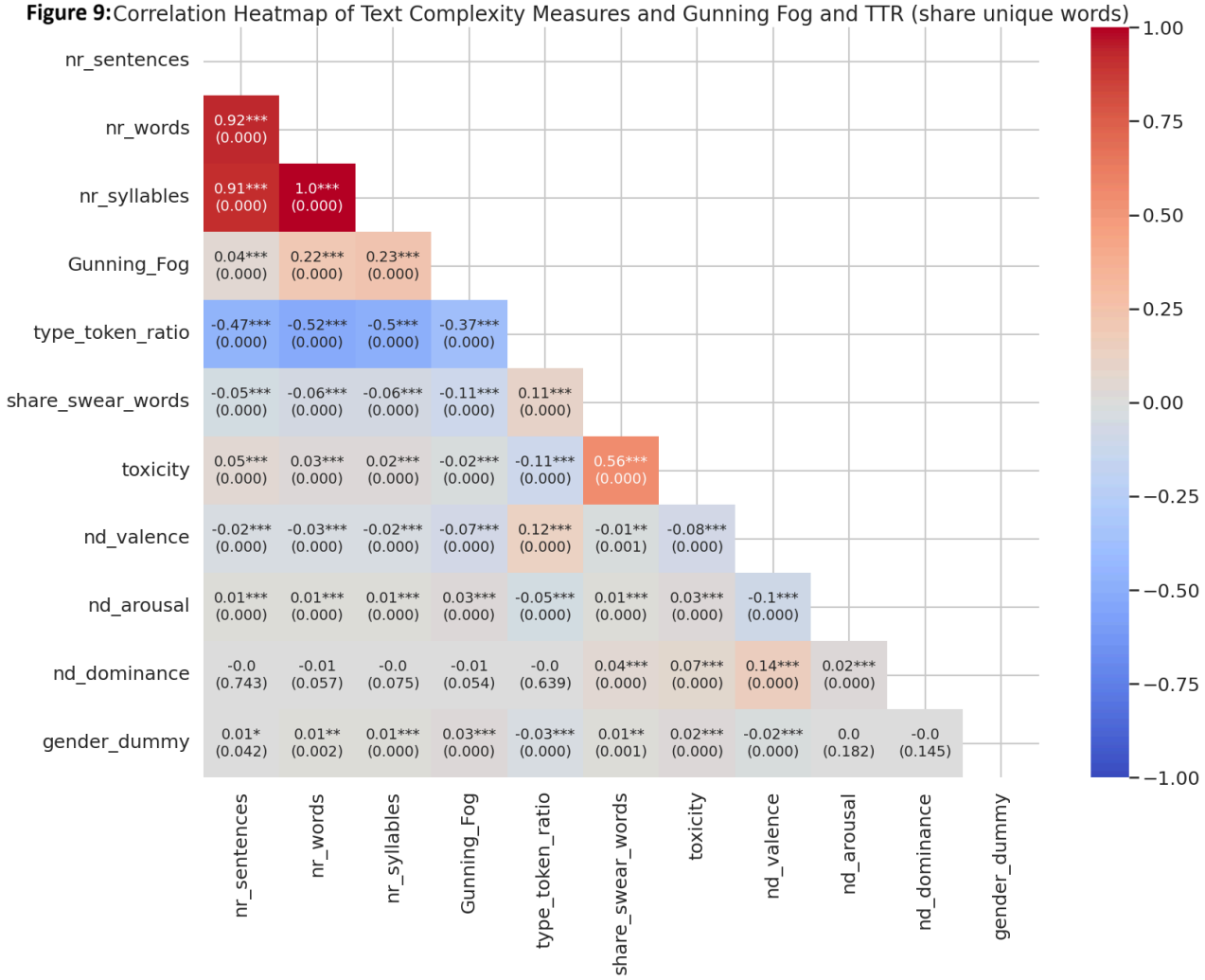
Additionally, the use of swear words was detected and presented in figure 8 using the better_profanity[3] Python library, which allows for detection and filtering of offensive language within text. The profanity check was performed to identify profane words in relation to toxicity and their occurrence over subreddits: Subreddits such as *MensRights*, *Australia*, and *Teenagers* show a higher mean share of swear words, indicating that discussions in these communities may contain more offensive or strong language. Others like *Q434706* and *TwoXChromosomes*, display a lower incidence of swear words, suggesting a more moderate or formal tone in the language used.

Results and Discussion

To test for evidence regarding the potential link between sentimental gender bias and text complexity, we performed a correlation analysis, see figure 9. Results for the first hypothesis of sentimentally gender-biased Reddit posts displaying lower text complexity vary. Posts that were labeled as more positive with higher valence correlate negatively with the Gunning Fog index (-0.07, $p < 0.001$) and positively with TTR (0.12, $p < 0.001$).
This implies more positive Reddit posts to be of lower syntactic and higher semantic text complexity at the same time, making them easier to understand while presenting a larger vocabulary. Yet, findings for arousal contradict this pattern: Posts that are more stimulating correlate positively with Gunning Fog (0.03, $p < 0.001$) and negatively with TTR (-0.05, $p < 0.001$). This indicates posts to be more syntactically challenging, but also the use of a smaller vocabulary. Correlations including dominance are not significant.

[3] https://pypi.org/project/profanity-check/

Figure 9: Correlation Heatmap of Text Complexity Measures and Gunning Fog and TTR (share unique words)



The low Gunning Fog scores for texts with higher valence support findings of Peeters and Czapinski (1990), who stated negative words to be more informative, and thus might make a text consisting of positive words and high valence easier to read. The positive bias present in the English language (Garcia et al., 2012) might explain the unexpected positive correlation between TTR and valence. The reverse pattern for arousal might be explained by people's urge to respond to content they engage with online in a manner that is less deliberate, leading to potentially convoluted and self-repeating texts, which is supported by significant positive correlations for the number of sentences and words.

Second, we hypothesize that Reddit posts, which are perceived as toxic, display lower scores for Gunning Fog and TTR. We find significant, negative correlations that affirm this expectation: The higher the probability of a Reddit post being classified as toxic, the lower the Gunning Fog score (−0.02, $p < 0.001$) and the TTR (−0.11, $p < 0.001$). While the reported correlations are weak, they are consistently negative and thus yield supporting evidence for the second hypothesis. These results go hand in hand with findings of toxic social media posts utilizing a smaller vocabulary and being less syntactically creative (Morzhov, 2021). Nevertheless, while the correlations are significant, they are small in magnitude. Reddit Posts that are perceived as toxic are sparse and as Figure 10 indicates, there is no clear pattern visible that would point towards a link between Gunning Fog score, TTR and toxicity. This observation holds for a more detailed inspection of text complexity and toxicity across all LEIA sentiments, which can be seen in Figure 11.

Finally, we expected a positive correlation between toxicity and the sentiment-derived VAD scores, as two different operationalisations of

sentimental gender bias. As with H1, findings are significant yet inconsistent: Posts that display higher valence scores again correlate negatively with toxicity (−0.08, $p < 0.001$), whereas the dimensions arousal (0.03, $p < 0.001$) and dominance (0.07, $p < 0.001$) correlate positively with toxicity. While this is only partial support for H3, it corresponds to the definitions of the VAD dimensions, as a post that is perceived as toxic likely evokes negative sentiments and an urge to act.

urge to react. The negative correlation between valence and toxicity might be explained by Perspective API's definition of toxicity, which does not include toxic positivity and would thus likely not classify posts as toxic which include numerous positive words.

To control for gender as a potential confounder, we included a dummy variable that took on the values 1 for women and 2 for men. It was significantly positively correlated with all variables but the sentiment-derived arousal and dominance scores (which were not significant) and TTR with a correlation of −0.03 ($p < 0.001$), indicating that posts targeting men used fewer unique words on average.
Notably, the correlation between the gender dummy and toxicity of 0.02 ($p < 0.001$) as well as a negative correlation with sentiment-derived valence of 0.02 ($p < 0.001$) correspond to the toxicity analysis: Being male correlates positively with the probability of a text being perceived as toxic and negatively with valence. This highlights the slight difference in toxic post distribution across genders of the toxicity analysis. It further corroborates findings of women being often targeted using benevolent sexism, whereas men experience more dominant toxic attacks (Glick & Fiske, 1996).
Similarly, being male correlates positively with the number of sentences, words, syllables and swear words of posts with a constant value of 0.01 and varying levels of significance. Albeit small, these correlations coincide with a larger positive correlation with the Gunning Fog Index of 0.03 ($p < 0.001$), which further evokes an interpretation of posts targeting men to be longer, more dominant and verbally explicit.
Yet, the results depend on the mapping between LEIA's predicted sentiments and VAD scores. Figure 12 reveals that 78,9 % of Reddit posts are classified as angry by LEIA, causing them to take on high scores for arousal and dominance. This imbalance in the distribution of LEIA's predicted sentiments likely affected the analysis.

The correlation between toxicity and the share of swear words lies at 0.56 ($p < 0.001$), and the higher the share of swear words, the shorter the post as indicated by the negative significant correlations with the number of sentences, words and syllables. Further, a higher share of swear words appears to reduce the Gunning Fog scores given the correlation of −0.11 ($p < 0.001$) and increases the TTR by 0.11 ($p < 0.001$). These findings potentially support evidence presented by Stapleton and Dawkins (2022) as well as Jones et al. (2013).

Limitations

The preprocessing of the data, where names of targeted politicians were redacted, poses a limitation to our analysis. While this procedure was necessary to ensure privacy, it may have distorted the analysis by removing important contextual information, such as the order of words and grammatical structures that help identify the target of a post and are important for sentiment analysis. Additionally, the distribution of toxic comments in the dataset was highly skewed, with very few toxic comments compared to a majority of non-toxic comments. This imbalance makes it challenging to analyze toxicity effectively. Furthermore, the non-random selection of Reddit comments into the original dataset, with a higher proportion of posts from men, resulted in a sample that is not fully representative, which potentially limits the generalizability of the findings.

By looking at a sample of data from one platform, we don't control for specific platform effects. For instance, Reddit is not fully representative of broader political discussions across social media, and its users may exhibit unique characteristics. Additionally, content moderation and data preprocessing on Reddit can significantly influence the sample, potentially providing only limited insights when attempting to measure constructs like gender bias in political discourse online.
Moreover, using toxicity as a proxy for hostile sexism introduces measurement errors. While hostile sexism often includes toxic content, not all toxic content contains hostile sexism or is grounded on gender stereotypes. The same applies to the use of VAD values. Macro-level differences of VAD scores certainly provide evidence for some sort of gender bias. The same does not hold for the micro-level, though. High valence is no sure sign of benevolent sexism the same way high dominance is of hostile sexism. For this we simply lack contextual information, especially information regarding what entity (or not) VAD scores refer to.

Conclusion

In conclusion, the analysis provides evidence for nuanced relationships between sentiment, text complexity, and gender bias in Reddit posts. Positive sentiment is associated with simpler syntax but richer vocabulary, whereas more stimulating content tends to be more complex yet less diverse lexically. Posts that express more controlling sentiments did not significantly correlate with either semantic or syntactic complexity. Our results thus partially support the first hypothesis, even though correlations are generally weak.
Posts perceived as toxic are characterized by lower semantic and syntactic text complexity, supporting the notion that toxic language often employs simpler and less creative expressions. This is prominent in posts targeting men, which correlate higher with toxicity and more explicit language, whereas posts targeting women tend to be shorter and of more positive sentiment. The results thus support the second hypothesis.
Finally, there is inconclusive evidence for the third hypothesis: Significant yet inconsistent correlations between toxicity and the sentiment-derived VAD scores suggest that posts with higher valence are less likely to be toxic, whereas those with higher arousal and dominance are more likely to be perceived as toxic.
Overall, the findings highlight the dynamics of sentiment, complexity, and gender bias in online communication. To advance our understanding of social and political communication online, future research is invited to provide more nuanced perspectives on gender bias, including sentiment and toxicity, in order to capture the role sentiment plays in shaping textual communication.

Contributions

Peer Saleth: Introduction, Theory: Sexism, Toxicity and Hostile Sexism, Methods: Toxicity, Limitations
Jördis Strack: Theory: Text Complexity and Gender Bias, Methods: Text Complexity, Results and Discussion, Limitations, Conclusion
Andri Rutschmann: Introduction, Theory: Sentimental Gender Bias and Sexism, Methods: Sentiment-derived VAD scores, Limitations

Ethical Considerations

When working with online data one must take special care to protect individuals' privacy. While this type of data is publicly available for most anyone to retrieve we still have to consider that individual actors have not actively consented to be part of our research. This is especially important when dealing with sensitive content, such as toxicity or bias.
Regarding this sensitive content, especially its annotation, we must also take special care to inform and educate annotators while providing them with psychological support, if needed.
Finally, we must clarify that this study, by no means, assumes gender or sex to be binary. We only investigate differences between female and male post targets because of the scarcity of non-binary and intersex persons in our data. We hope that future research is able to remedy this shortcoming.

References

Alodat, A. M., Al-Qora'n, L. F., & Abu Hamoud, M. (2023). Social Media Platforms and Political Participation: A Study of Jordanian Youth Engagement. *Social Sciences*, 12 (7), 402. https://doi.org/10.3390/socsci12070402

Aroyehun, S., Malik, L., Metzler, H., Haimerl, N., Natale, A., & Garcia, D. (2023). LEIA: Linguistic Embeddings for the Identification of Affect. *EPJ Data Science*, 12 (1), 52. https://doi.org/10.1140/epjds/s13688-023-00427-0

Barker, K., & Jurasz, O. (2019). Online Misogyny: A Challenge for Digital Feminism? *Journal of International Affairs*, 72(2), 95-114. https://www.jstor.org/stable/10.2307/26760834

Boulianne, S., Koc-Michalska, K., & Vedel, T. (2021). Gender and Online Politics: Digital Media as Friend and Foe in Times of Change. *Social Science Computer Review*, 39 (2), 175–180. https://doi.org/10.1177/0894439319865511

Buechel, S. & Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In Lapata, M., Blunsom, P., & Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 578–585). Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2092

Crabtree, C., Golder, M., Gschwend, T., & Indriason, I. H. (2020). It Is Not Only What You Say, It Is Also How You Say It: The Strategic

Use of Campaign Sentiment. *The Journal of Politics*, 82 (3), 1044–1060. https://doi.org/10.1086/707613

Diepeveen, S. (2022, March). Break the bias to challenge gender norms on social media. Retrieved August 20, 2024, from https://odi.org/en/insights/break-the-bias-to-challenge-gender-norms-on-social-media/

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6 (3-4), 169–200. https://doi.org/10.1080/02699939208411068

Frantz, R. S., Starr, L. E., & Bailey, A. L. (2015). Syntactic Complexity as an Aspect of Text Complexity. *Educational Researcher*, 44 (7), 387–393. https://doi.org/10.3102/0013189X15603980

Garcia, D., Garas, A., & Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science*, 1 (1), 3. https://doi.org/10.1140/epjds3

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7), 1266-1286. https://doi.org/10.1177/1461444820912540

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70* (3), 491–512. https://doi.org/10.24963/ijcai.2023/689

Guiraud, P. (1954). Les caractères statistiques du vocabulaire : Essai de méthodologie. Presses universitaires de France. https://doi.org/10.1093/fs/IX.1.87

Gunning, R. (1952). The technique of clear writing. New York: McGraw-Hill International Book Co.

Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, 6 (2), 3–13. https://doi.org/10.1177/002194366900600202

Haraldsson, A., & Wängnerud, L. (2019). The effect of media sexism on women's political ambition: Evidence from a worldwide study. *Feminist Media Studies, 19*(4), 525-541. https://doi.org/10.1080/14680777.2018.1468797

Jones, P. E., Hoffman, L. H., & Young, D. G. (2013). Online emotional appeals and political participation: The effect of candidate affect on mass behavior. *New Media & Society*, 15 (7), 1132–1150. https://doi.org/10.1177/1461444812466717

Kayam, O. (2018). The Readability and Simplicity of Donald Trump's Language. *Political Studies Review*, 16 (1), 73–88. https://doi.org/10.1177/1478929917706844

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140 (1), 14–34. https://doi.org/10.1037/a0021446

Marjanovic, S., Stánczak, K., & Augenstein, I. (2022). Quantifying gender biases towards politicians on Reddit. *PLOS ONE*, 17 (10), e0274317. https://doi.org/10.1371/journal.pone.0274317

Mohammad, S. (2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In Gurevych, I. & Miyao, Y. (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 174–184). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1017

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 1–17. https://doi.org/10.18653/v1/S18-1001

Mohammad, S. & Turney, P. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In Inkpen, D. & Strapparava, C. (Eds.), *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26–34). Association for Computational Linguistics. https://aclanthology.org/W10-0204

Mohammad, S., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51 (4), 480–499. https://doi.org/10.1016/j.ipm.2014.09.003

Morzhov, S. V. (2021). Modern Approaches to Detecting and Classifying Toxic Comments Using Neural Networks. *Automatic Control and Computer Sciences*, 55 (7), 607–616. https://doi.org/10.3103/S0146411621070117

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In Liu, Q. &

Schlangen, D. (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Peeters, G., & Czapinski, J. (1990). Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects. *European Review of Social Psychology*, 1 (1), 33–60. https://doi.org/10.1080/14792779108401856

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. In *Proceedings of the National Academy of Sciences*, 108 (9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Rodríguez Timaná, L. C., Saavedra Lozano, D. F., & Castillo García, J. F. (2020). Software to determine the readability of written documents by implementing a variation of the Gunning Fog Index using the Google linguistic corpus. In Botto-Tobar, M., Zambrano Vizuete, M., Torres-Carrión, P., Montes León, S., Pizarro Vásquez, G., & Durakovic, B. (Eds.), *Applied technologies* (Vol. 1193, pp. 409–420). Springer International Publishing. https://doi.org/10.1007/978-3-030-42517-3_31

Rozin, P., Berman, L., & Royzman, E. (2010). Biases in use of positive and negative words across twenty natural languages. *Cognition & Emotion*, 24 (3), 536–548. https://doi.org/10.1080/02699930902793462

Russell, J. A. & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11 (3), 273–294. https://doi.org/10.1016/0092-6566(77)90037-X

Scott, Z. A., & McDonald, J. (2022). Tell Us How You Feel: Emotional Appeals for Votes in Presidential Primaries. *American Politics Research*, 50 (5), 609–622. https://doi.org/10.1177/1532673X221106432

Singh, G. V., Ghosh, S., & Ekbal, A. (2023). Promoting Gender Equality through Gender-biased Language Analysis in Social Media. *IJCAI*, 6210–6218. https://doi.org/10.24963/ijcai.2023/689

Song, Y., Wang, X., & Li, G. (2024). Can social media combat gender inequalities in academia? Measuring the prevalence of the Matilda effect in communication. *Journal of Computer Mediated Communication*, 29 (1), zmad050. https://doi.org/10.1093/jcmc/zmad050

Stapleton, C. E., & Dawkins, R. (2022). Catching My Anger: How Political Elites Create Angrier Citizens. *Political Research Quarterly*, 75 (3), 754–765. https://doi.org/10.1177/1065912921102

Troiano, E., Padó, S., & Klinger, R. (2019). Crowdsourcing and validating event-focused emotion corpora for German and English. In Korhonen, A., Traum, D., & Màrquez, L. (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4005–4011). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1391

Appendix

Figure 1: Toxicity Confusion Matrix with examples

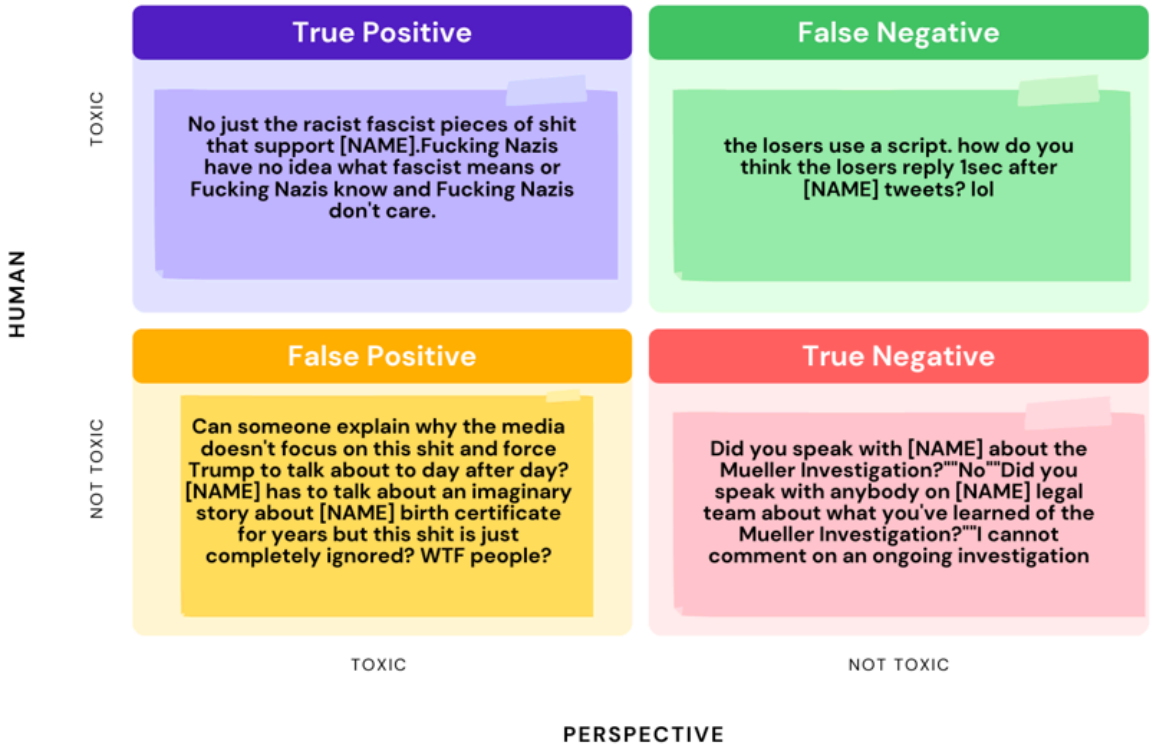


Figure 2: Perspective API Toxicity Distribution by sex

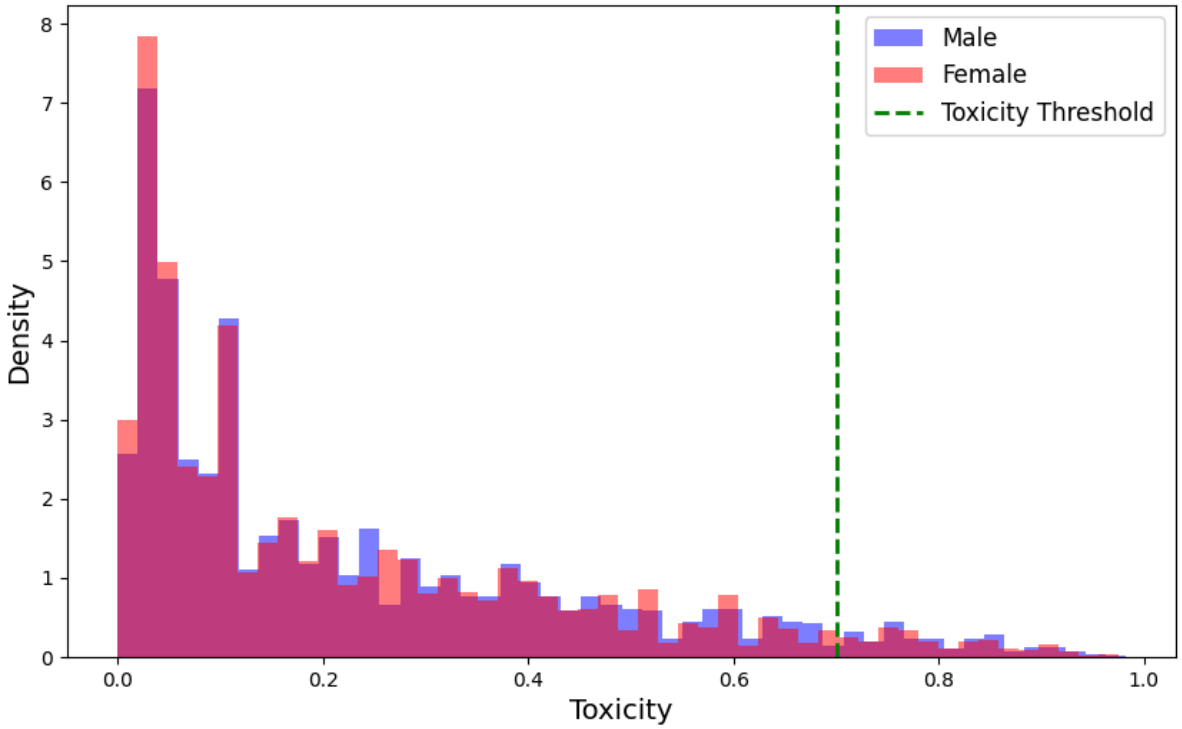


Figure 3: Number of Toxic Comments by Sex

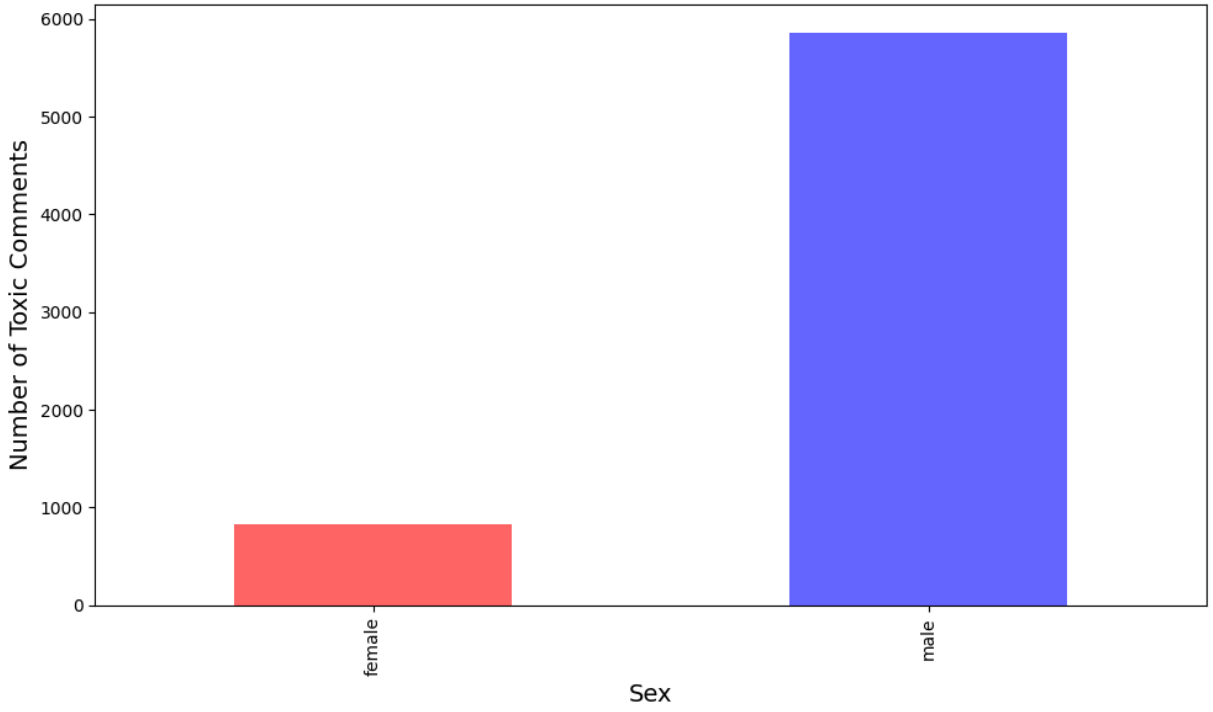


Figure 4: Ratio of Toxic Comments by Sex with 95% CI

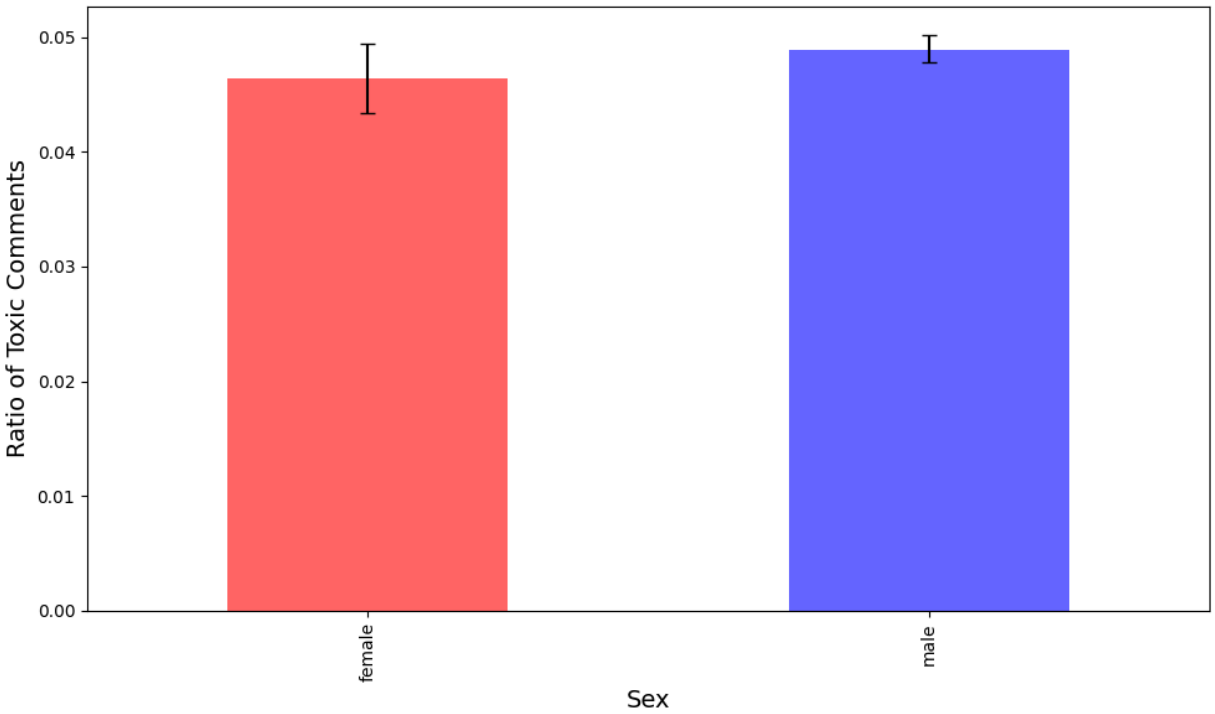


Figure 5: Mean Toxicity per Subreddit and sex with 95 % CI

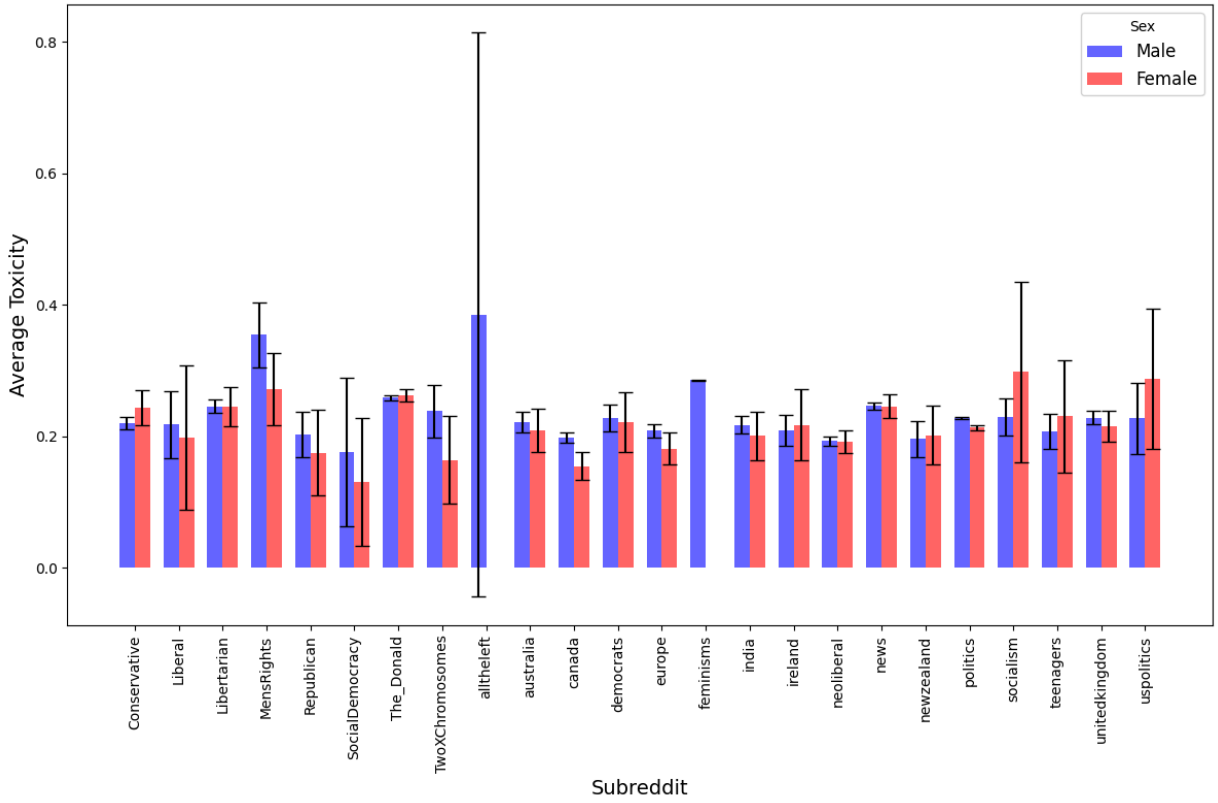


Figure 6: Distribution of Gunning Fog scores over all sampled subreddits

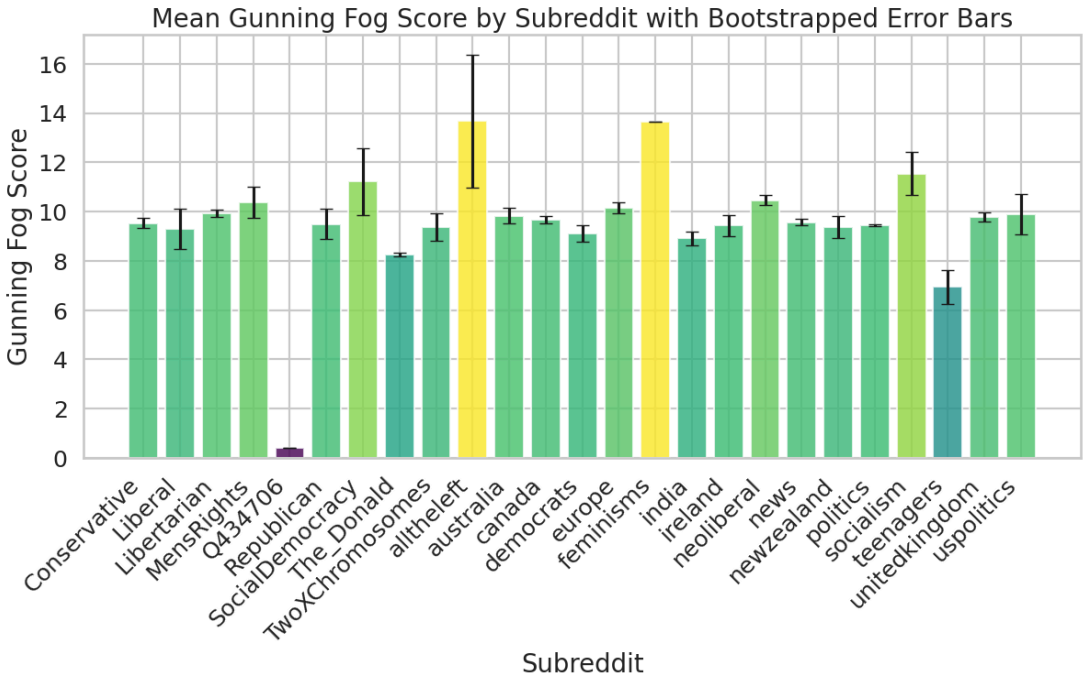


Figure 7: Share of Type Token Ratio over all sampled subreddits

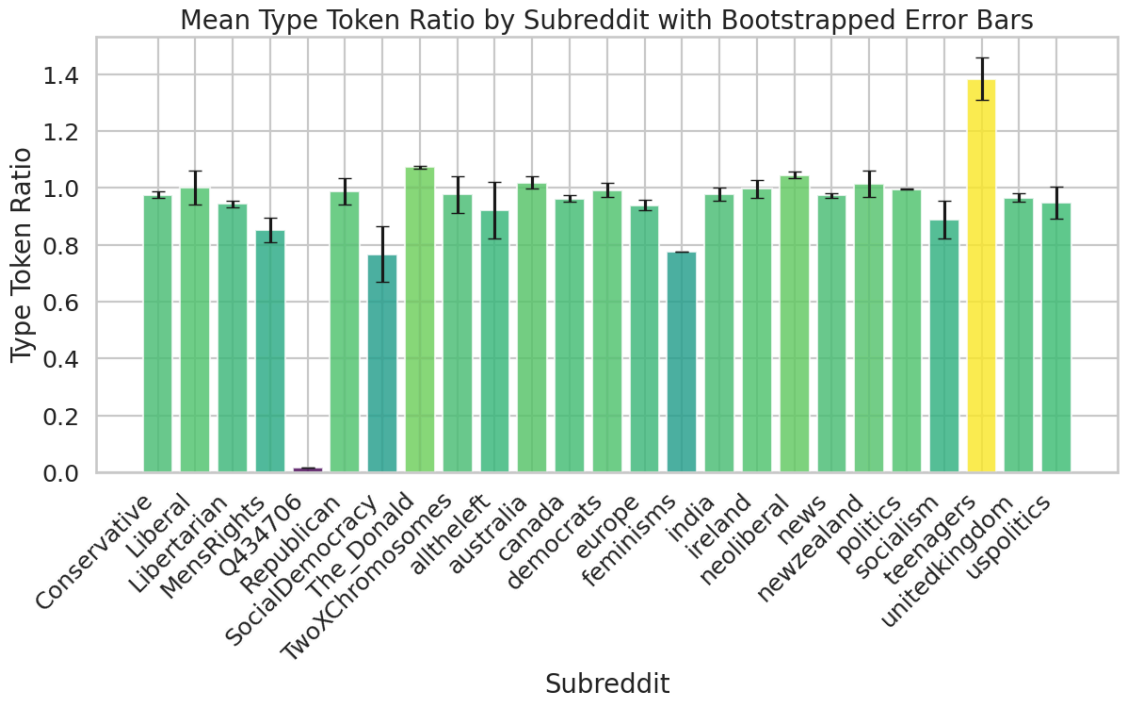


Figure 8: Share of swear words over all sampled subreddits

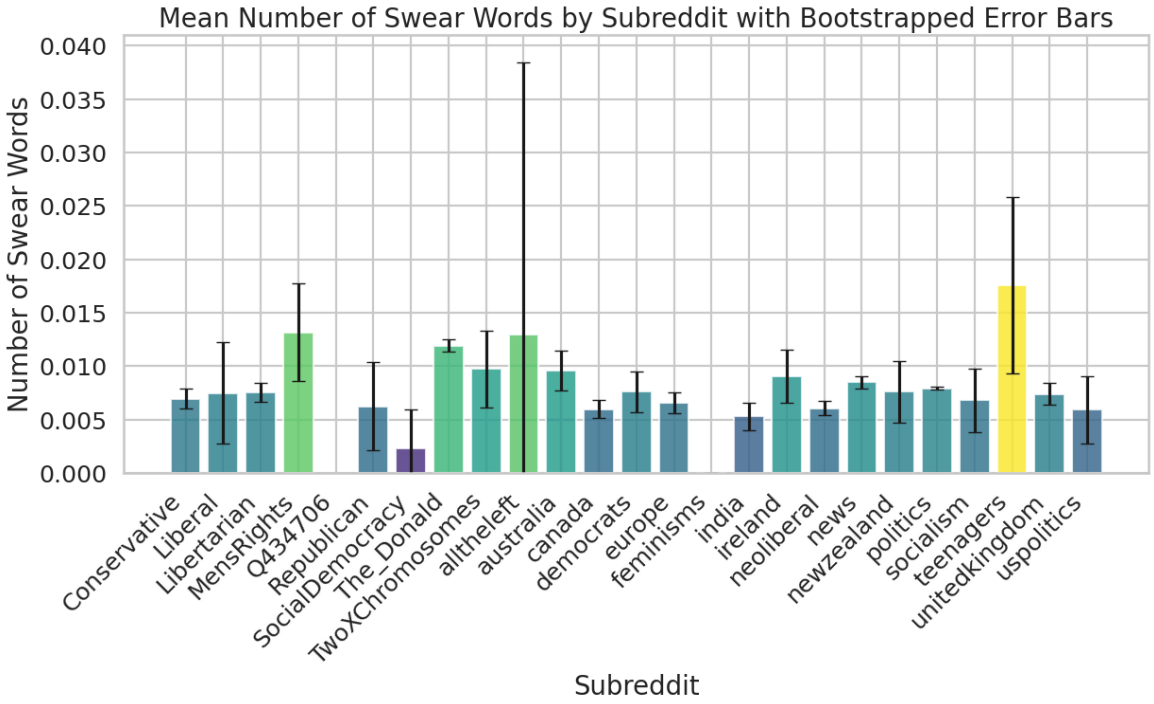


Figure 10: Toxicity distribution across Gunning Fog and TTR

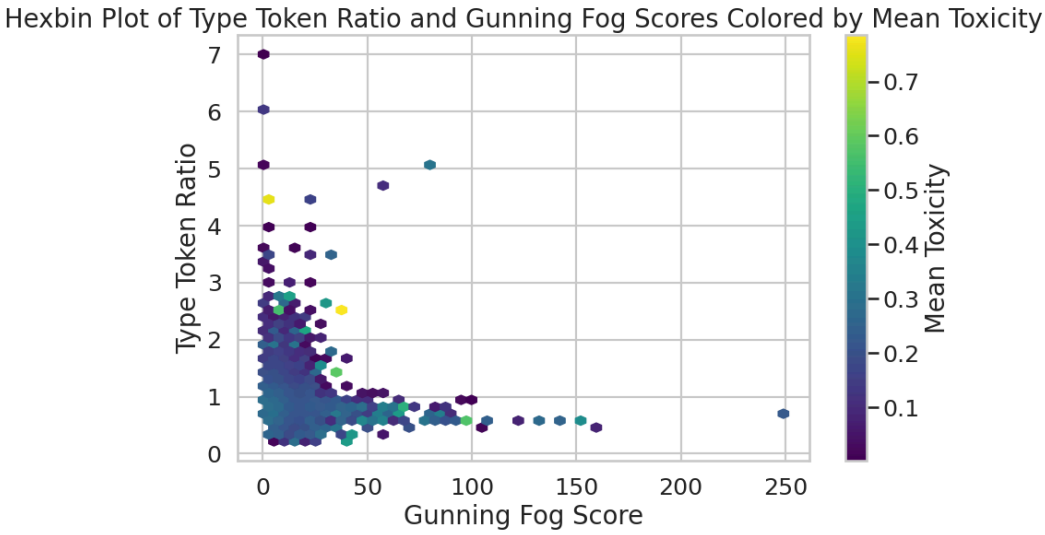


Figure 11: Toxicity distribution across Gunning Fog, TTR and LEIA sentiment

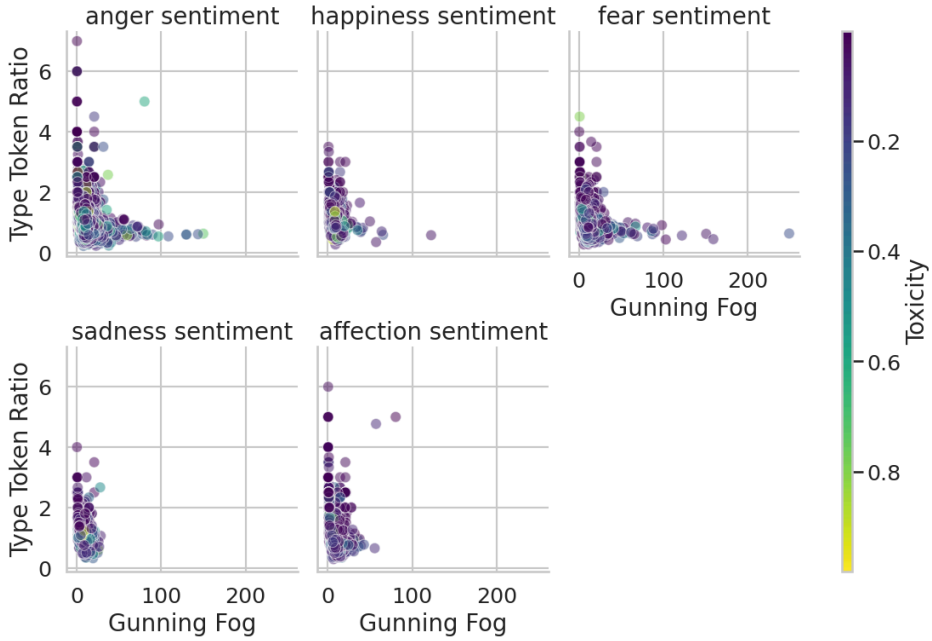


Figure 12: Distribution of LEIA's predicted sentiments

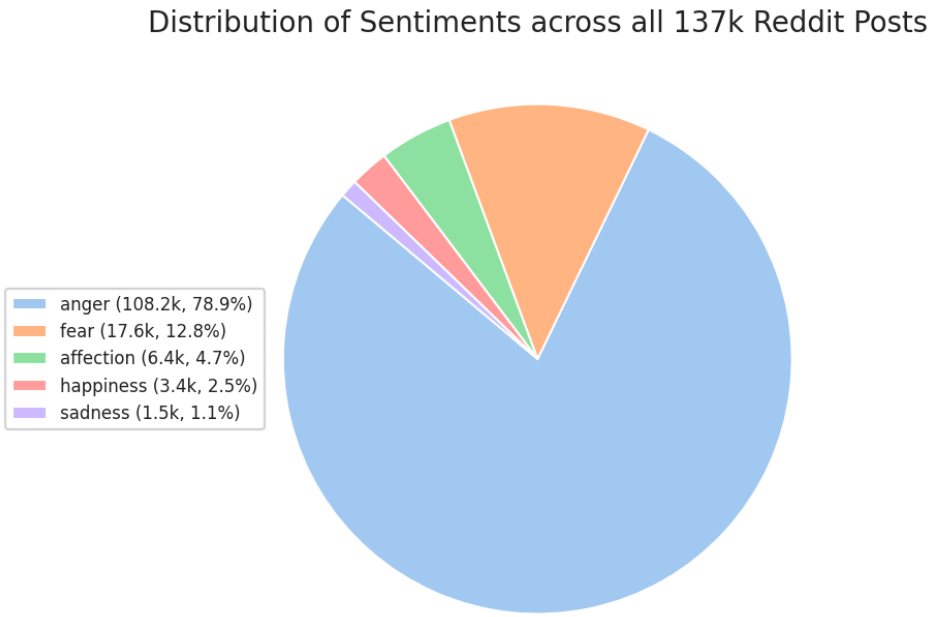


Figure 13: 3 Dimensional Representation of Posts' Sentiment-derived VAD scores

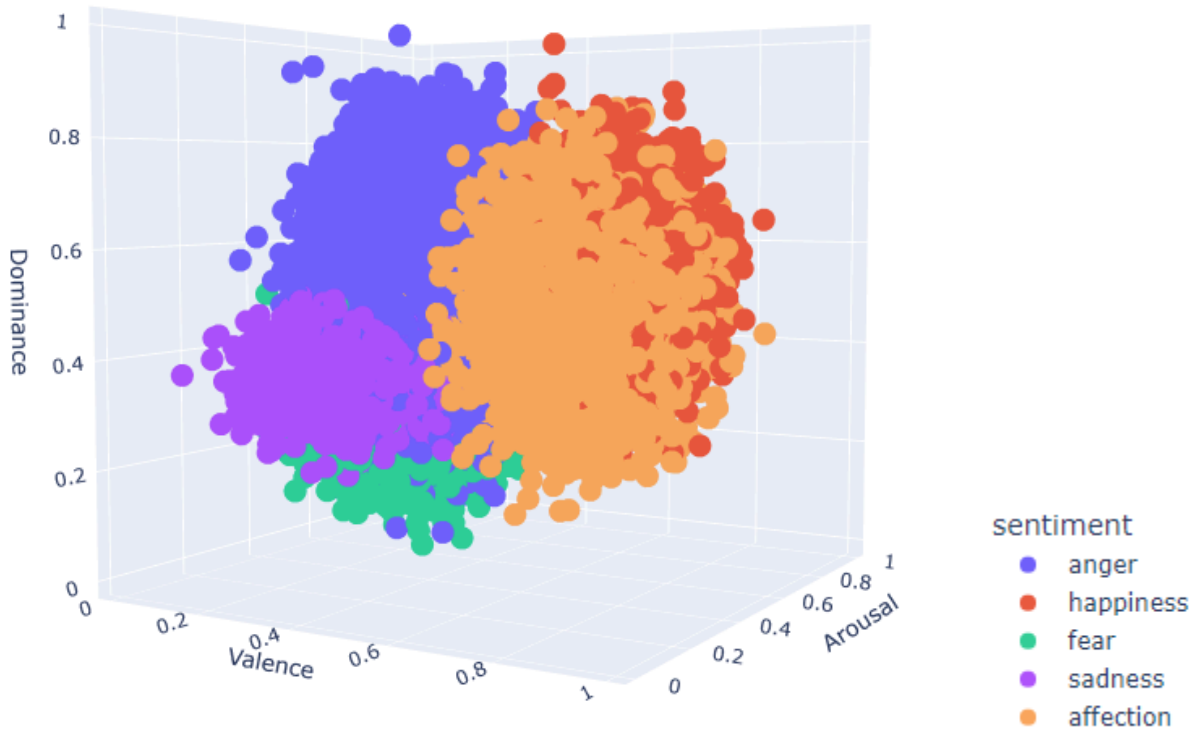


Table 1: LEIA Validation Metrics (N = 100)

	Precision	Recall	F1 - score	N
Affection	0.38	0.20	0.26	15
Anger	0.73	0.84	0.78	62
Fear	0.14	1.00	0.25	2
Happiness	1.00	0.25	0.40	16
Sadness	0.00	0.00	0.00	5
Weighted Avg	0.67	0.61	0.59	100

Table 2: Sentiment Annotator Mean and SD

	Valence (Mean/SD)	Arousal (Mean/SD)	Dominance (Mean/SD)
Affection	0.64 / 0.26	0.35 / 0.34	0.24 / 0.40
Anger	-0.51 / 0.20	0.59 / 0.33	0.25 / 0.39
Fear	-0.64 / 0.20	0.60 / 0.32	-0.43 / 0.30
Happiness	0.81 / 0.21	0.51 / 0.26	0.46 / 0.38
Sadness	-0.63 / 0.23	-0.27 / 0.34	-0.33 / 0.22

Note: The Values drawn from the normal distributions created with these values were rescaled to [0, 1] in order to correspond to the VAD values introduced by Marjanovic et al. (2022).

Table 3: VAD Shapiro Wilk Test Results

	NRC-VAD	LEIA-VAD
Valence	0.821*	0.802*
Arousal	0.836*	0.997*
Dominance	0.827*	0.998*

* : $p < 0.001$ (p-values were approximated using Monte Carlo tests due to a SciPy limitation mentioned above)

Table 4: Female directed vs. Male directed Posts
(Bonferroni-corrected Mann-Whitney U)

	Valence	Arousal	Dominance
NRC-VAD	-3.88**	-7.17**	-6.96**
BS CIs (Mean)	[0.312, 0.318]; [0.312, 0.318]	[0.241, 0.246]; [0.251, 0.252]	[0.290, 0.296]; [0.301, 0.303]
LEIA-VAD	3.21*	-1.18	-1.46
BS CIs (Mean)	[0.337, 0.341]; [0.333, 0.334]	[0.531, 0.533]; [0.5325, 0.534]	[0.487, 0.490]; [0.486, 0.488]

** : $p < 0.001$
* : $p < 0.05$