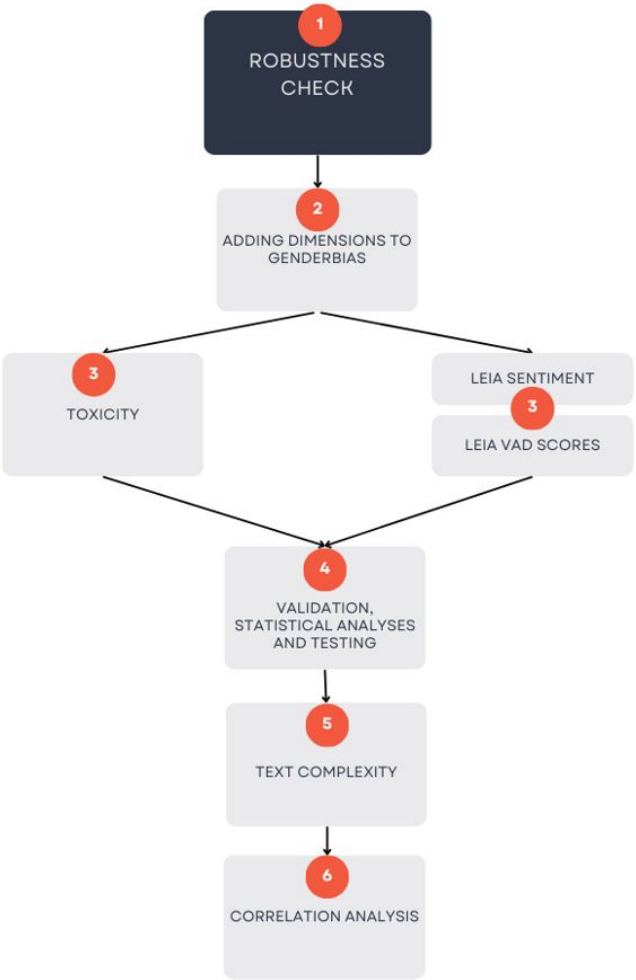
Sentimental Gender Bias: Insights from Reddit Posts

Final Presentation

Andri Rutschmann, Peer Saleth and Jördis Strack

Project Overview

1. Robustness check of Marjanovic et al. (2022)
2. We add Toxicity and Sentiment-VAD scores
3. Implement it!
4. Validate and compare to Marjanovic et al. (2022)
5. Enter Text Complexity
6. Correlation Analysis Sentimental Gender Bias vs. Text Complexity



Theory

Sentimental Gender Bias

Can we observe significant differences in sentiment for posts about female and male US politicians?

What are sentiments?

1. Sentiments as discrete classes
2. **Sentiments as groups in a multi-dimensional space**

Valence, Arousal, Dominance

Sentiments as combination of three dimensions regarding the sentiments'

1. **Valence**: How 'good' or 'bad'? 😄 vs. 😢
2. **Arousal**: How energizing? 🐝 vs. 🦥
3. **Dominance**: How controlling? 👸 vs. 🫠

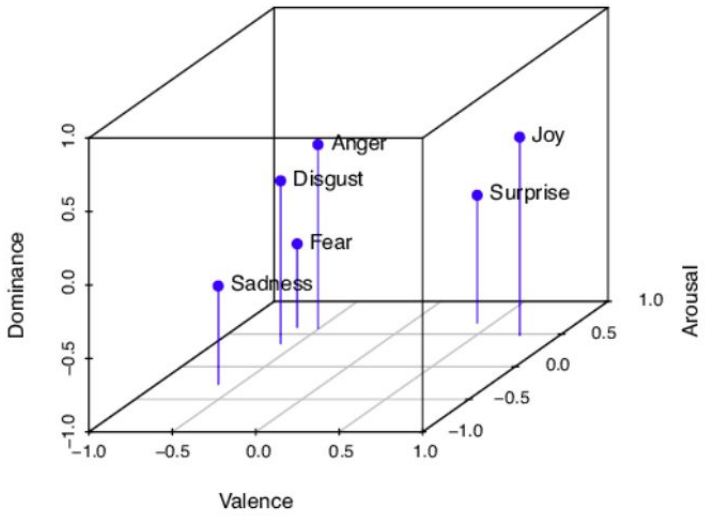


Figure from Buechel & Hahn (2017) showing Ekman's six basic emotions

Dimensions of Gender Bias

1. Hostile Sexism

- **Dominative Paternalism**: The belief that women should be controlled by men.
- **Competitive Gender Differentiation**: Negative stereotypes that assert men are superior to women.
- **Heterosexual Hostility**: Viewing women merely as sexual objects and fearing they use sexual attraction to gain power over men.

These attitudes reinforce traditional gender roles and maintain patriarchal social structures by portraying women as inferior and dependent on men (Glick & Fiske, 1996).

 measure with proxy toxicity

Dimensions of Gender Bias

Toxicity is defined as rude, disrespectful, or unreasonable comments that are likely to make someone leave a discussion. This includes offensive, abusive, and hateful comments (Pavlopoulos et al., 2020).

Proxying Hostile Sexism through Toxicity

- Hostile sexism often manifests in toxic behaviors and language directed towards women.
- Toxic comments are a direct expression of negative attitudes and beliefs consistent with hostile sexism.
- Contents of toxic social media posts are often motivated by stereotypical gender biases (Cheng et al., 2022)

Challenges and Considerations:

- Context Sensitivity: Accurate detection requires understanding the context in which comments are made, as toxicity and sexism can be context-dependent (Pavlopoulos et al., 2020).
- Data Annotation: Effective models need large datasets annotated with context-aware labels

Dimensions of Gender Bias

2. Benevolent Sexism

- **Protective Paternalism**: The idea that men should protect and provide for women due to their perceived weakness.
- **Complementary Gender Differentiation**: Idealizing women for traits like purity and nurturing, which are seen as complementary to male characteristics.
- **Intimate Heterosexuality**: Romanticizing women as necessary for a man's fulfillment, thus perpetuating dependency.

These attitudes appear kind and affectionate but ultimately reinforce women's subordinate status by valuing them only within traditional roles (Glick & Fiske, 1996).

 hard to measure, sentiment VAD scores?

Sentimental Gender Bias and Text Complexity

Syntactic and semantic text complexity:

- **Campo (2011):** emotional content is reflected in the semantic representations of abstract words
- **Garcia, D., Garas, A. & Schweitzer (2012):** length of words and emotional content determine frequency of use → positive bias in many languages
- **Morzhov (2021):** toxic texts display fewer unique words and ‘less inventive’ use of language
- **Kayam (2018):** political speeches vary in total number of words, sentences and syntactic complexity → especially Trump’s political speeches score low on readability indices

Sentimental Gender Bias and Text Complexity

How should the influence of sentiment on text complexity be measured?

- **Ofek (2023):** text sentiment relies on context → position & order of words matter for sentiment analysis
- **González-Bailón and Paltoglou (2015**): dictionary size alone might produce poor sentiment scores due to irrelevant vocabulary
- **Haselmayer and Jenny (2017):** dictionaries are often domain specific and do not generalize well to texts from new domains

A new approach is required to assess the influence of sentiment on text complexity!

Hypotheses

Hypotheses

Research Question: How are sentimental gender bias and text complexity related?

- **H1:** Sentimentally gender biased (VAD) Reddit posts display lower text complexity.
- **H2:** Reddit posts that are perceived as toxic display lower text complexity.
- **H3:** Sentimental gender bias operationalized through toxicity and sentimental gender bias operationalized through VAD correlate positively.

Data

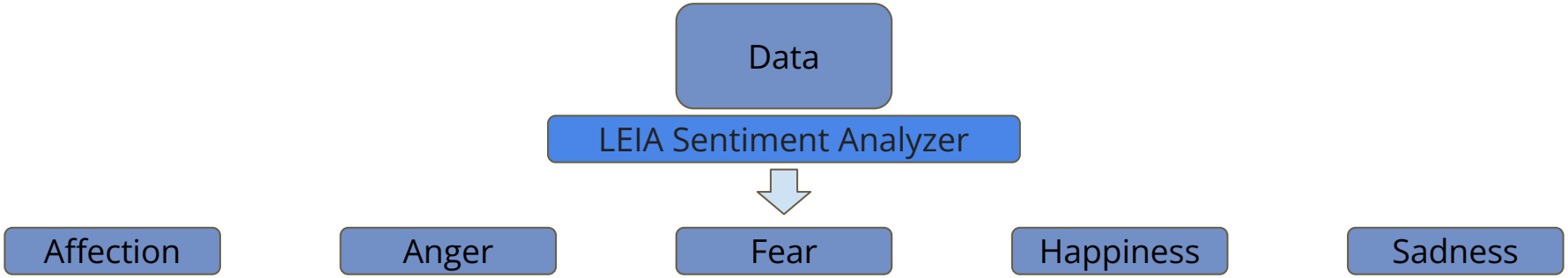
Data

Quantifying gender biases towards politicians on Reddit - Data Set (Marjanovic et al., 2022)

- Original data contains over 10 Mio Reddit posts
- Random sample due to computational resources (~137k posts)
- Data was gathered to investigate gender bias on Reddit
- Key variables include:
    - body, sex, nrc_valence, nrc_arousal, nrc_dominance

Methods - LEIA VAD

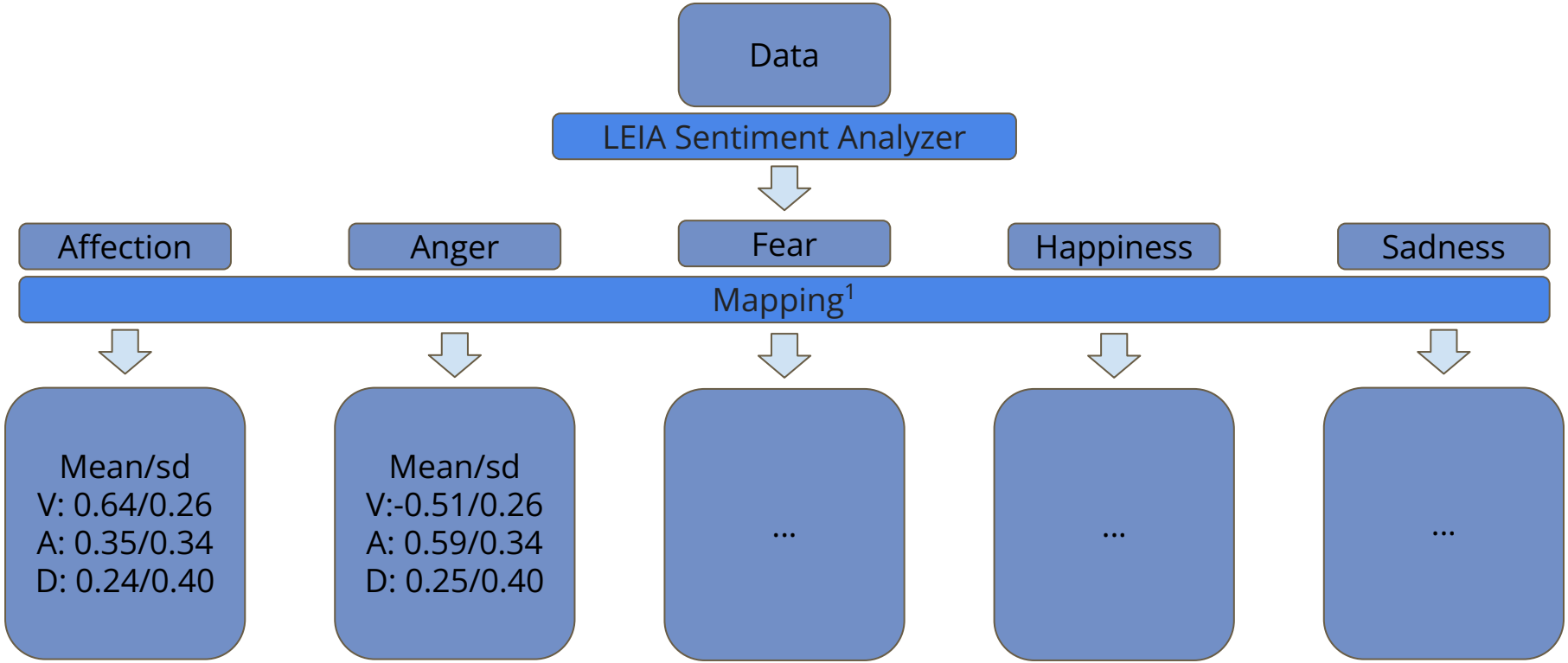
LEIA VAD Overview



LEIA VAD Mapping

- Validated against human labeled posts:
  - N = 100
  - F1 = 0.59
  - Precision = 0.67
  - Recall = 0.61

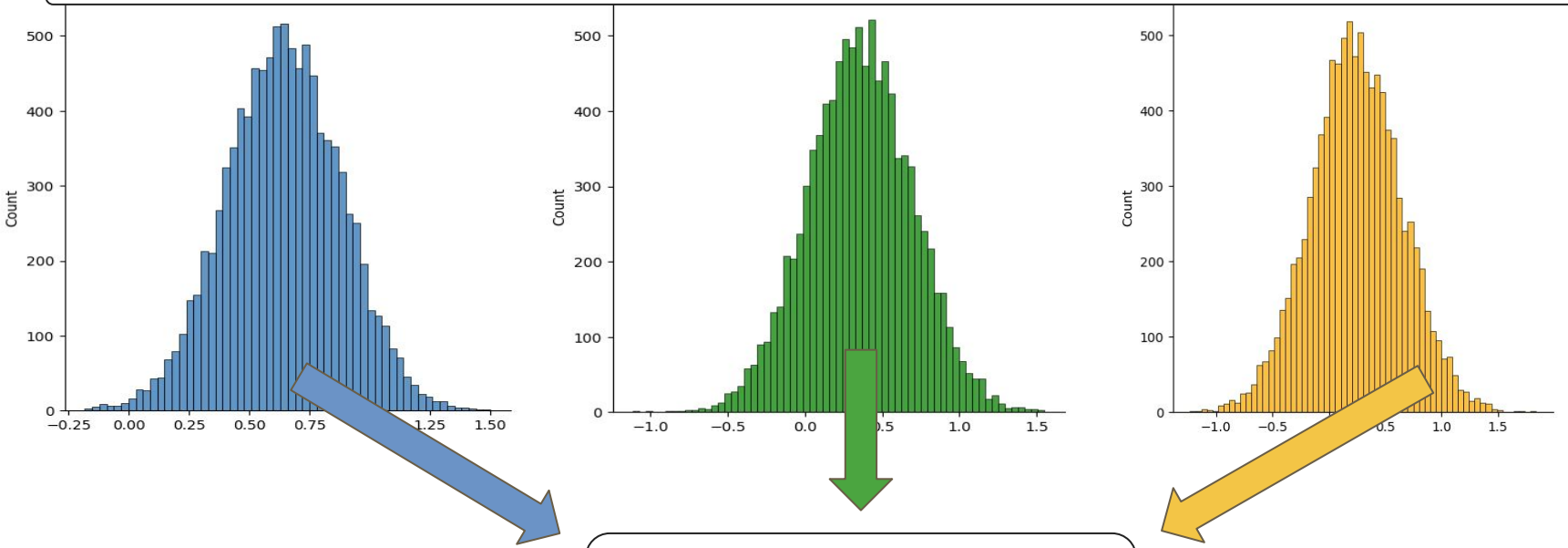
LEIA VAD Overview



1: The VAD mapping uses Russell & Mehrabian's over all N annotators averaged findings (N ~ 30)

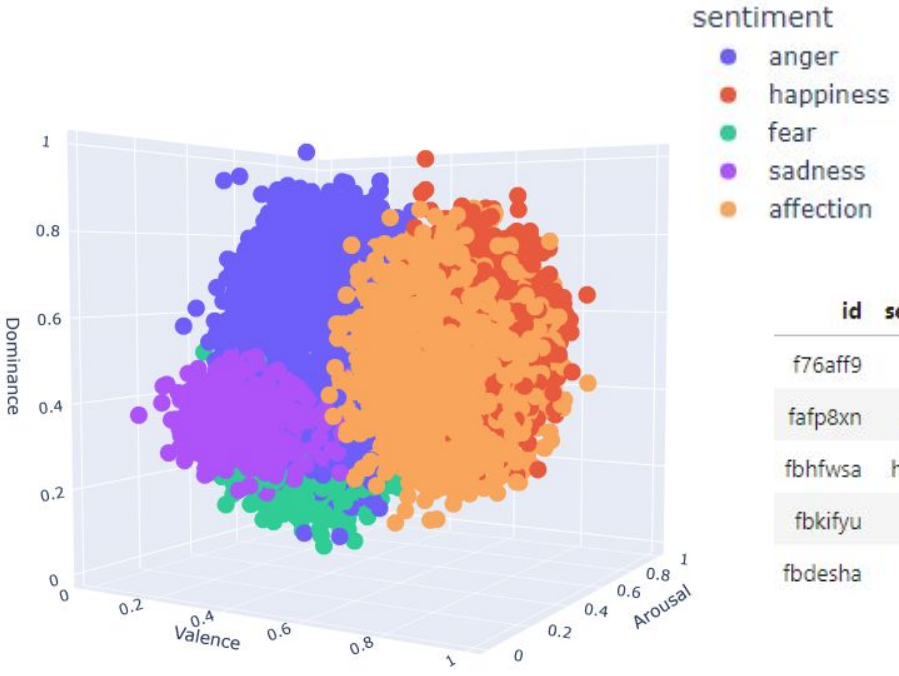
LEIA VAD Mapping

VAD Normal Distributions for ‘Affection’



V: 0.71; A: 0.46; D: 0.36

LEIA VAD Mapping



id	sentiment	Valence	Arousal	Dominance	nd_valence	nd_arousal	nd_dominance
f76aff9	fear	0.179638	0.132858	0.157276	0.235063	0.516702	0.348295
fafp8xn	affection	0.267769	0.130692	0.208923	0.605192	0.374887	0.462084
fbhfwsa	happiness	0.222412	0.155235	0.173824	0.721993	0.413671	0.610274
fbkifyu	fear	0.162178	0.124044	0.141444	0.328853	0.487382	0.233346
fbdesha	anger	0.151745	0.134800	0.174145	0.347222	0.552486	0.412906

LEIA VAD Analysis

- VAD values not normally distributed
  - Shapiro Wilk Test statistics ~0.80 - 1**
  - Monte Carlo Test p-value approximation
- Mann-Whitney U Test (female vs. male directed posts)

	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]

* $p < 0.05$; ** $p < 0.001$

Methods - Toxicity

Perspective API

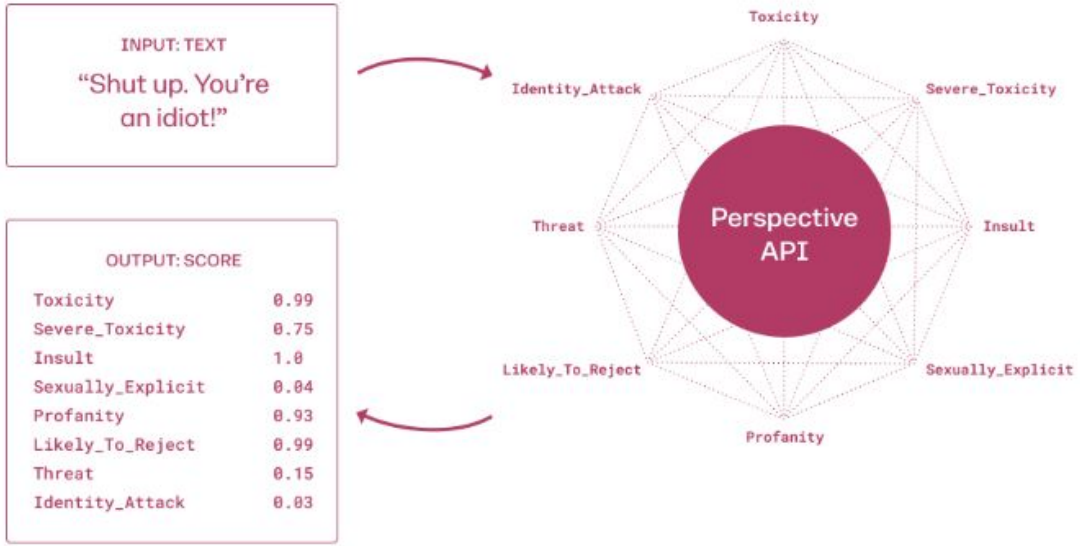
‘When people discuss male and female politicians, do they express equal levels of **toxicity** in words chosen?’

We utilize **perspective API** to operationalize toxicity:

Models (Transformer CNN, RNN, LSTM) trained on millions of comments from a variety of sources, including comments from online forums such as Wikipedia and The New York Times, across a range of languages (annotated)

- Performance on Jigsaw Multilingual Toxic Comments Challenge AUC-ROC [0.87 ; 0.94] (Lees et al., 2022)
- Performance on TweetEval dataset macro F1: [0.53; 0.57] (Lees et al., 2022)
- Performance on CivilComments-WILDS dataset Avg. Accuracy: [0.89 ; 0.94] (Lees et al., 2022)
- Performance on English-only HatemojiCheck Accuracy: 90.8%, F1: [0.89 ; 0.93] (Lees et al., 2022)

Perspective API



- probability score between 0 and 1
- higher score indicates a greater likelihood that a reader would perceive the comment as containing the given attribute
- less effective in finding nuanced Misogynoir (Kwarteng et al. 2022)
- might change as model gets updated

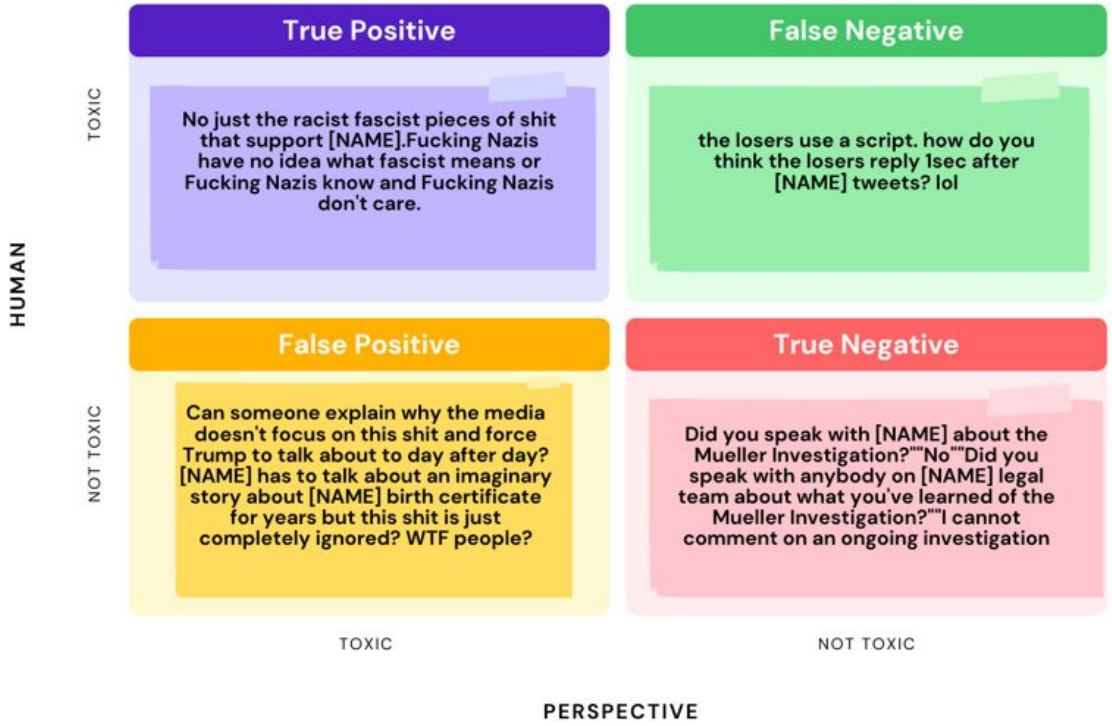
https://developers.perspectiveapi.com/s/about-the-api?language=en_US

Validation

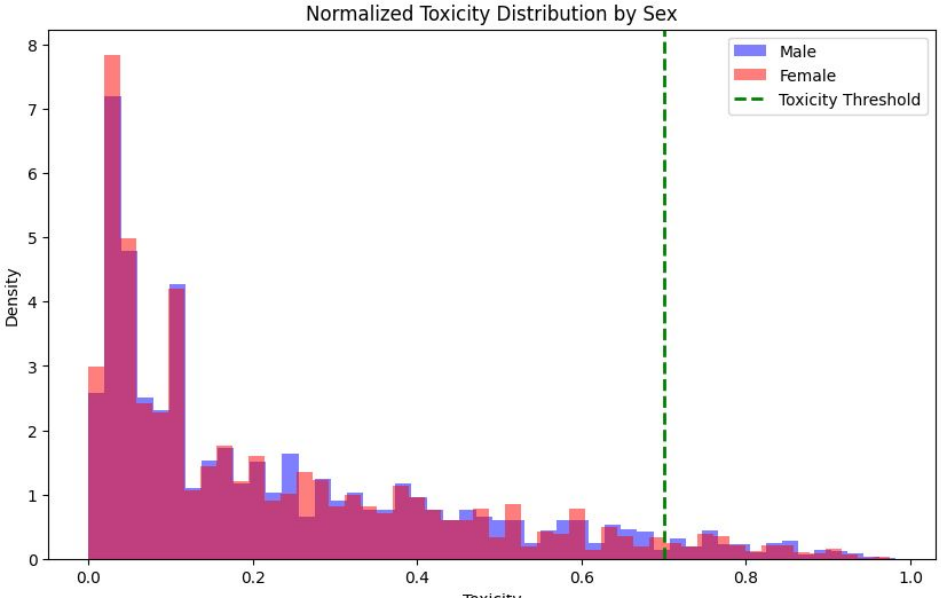
Sample randomly 25 toxic and 25 non toxic posts (threshold = 0.7)

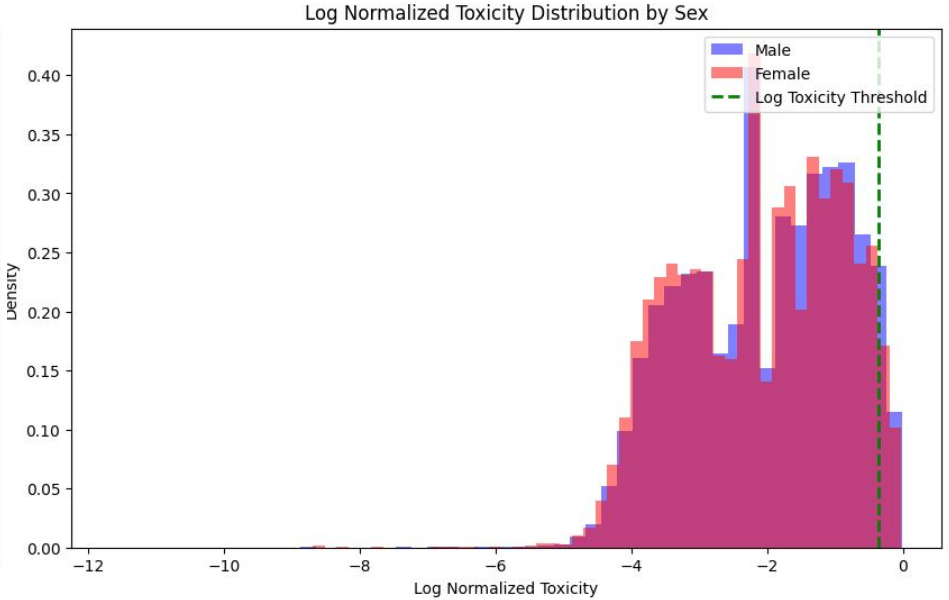
- Annotate them by myself
- Accuracy: 0.88
- F1 Score: 0.875
- Precision: 0.913
- Recall: 0.840

Validation


True Positive
No just the racist fascist pieces of shit that support [NAME].Fucking Nazis have no idea what fascist means or Fucking Nazis know and Fucking Nazis don't care.
False Negative
the losers use a script. how do you think the losers reply 1sec after [NAME] tweets? lol
False Positive
Can someone explain why the media doesn't focus on this shit and force Trump to talk about to day after day? [NAME] has to talk about an imaginary story about [NAME] birth certificate for years but this shit is just completely ignored? WTF people?
True Negative
Did you speak with [NAME] about the Mueller Investigation?""No""Did you speak with anybody on [NAME] legal team about what you've learned of the Mueller Investigation?""I cannot comment on an ongoing investigation
HUMAN
TOXIC
NOT TOXIC
TOXIC
NOT TOXIC
PERSPECTIVE

Toxicity Scores





Toxicity Analysis

Normality Test (Shapiro-Wilk):

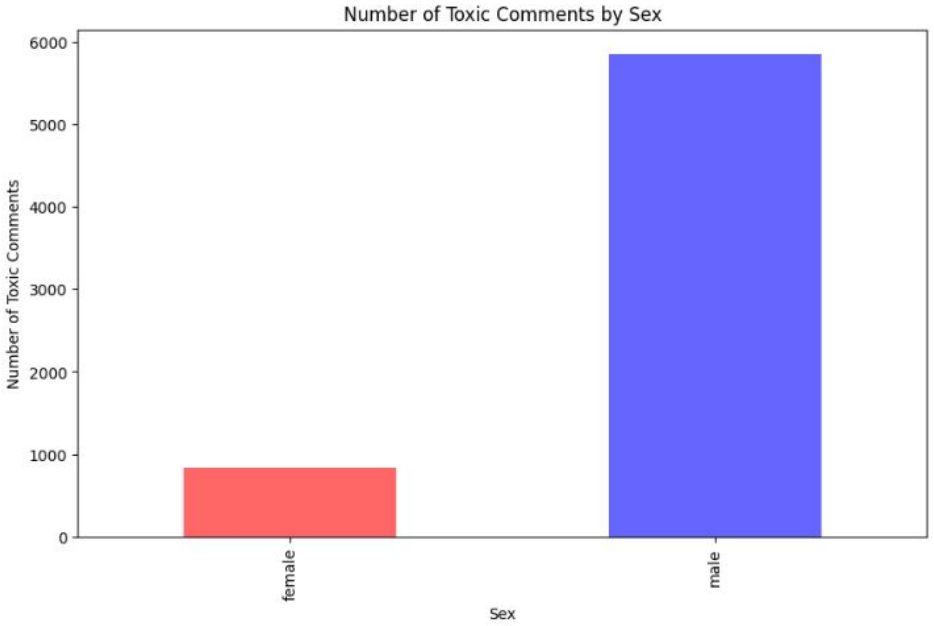
- **Statistic:** 0.964
- **P-value:** 0.000
- **Monte Carlo Test p-value approximation:** 0.0002
- **Interpretation:** Data is not normally distributed.

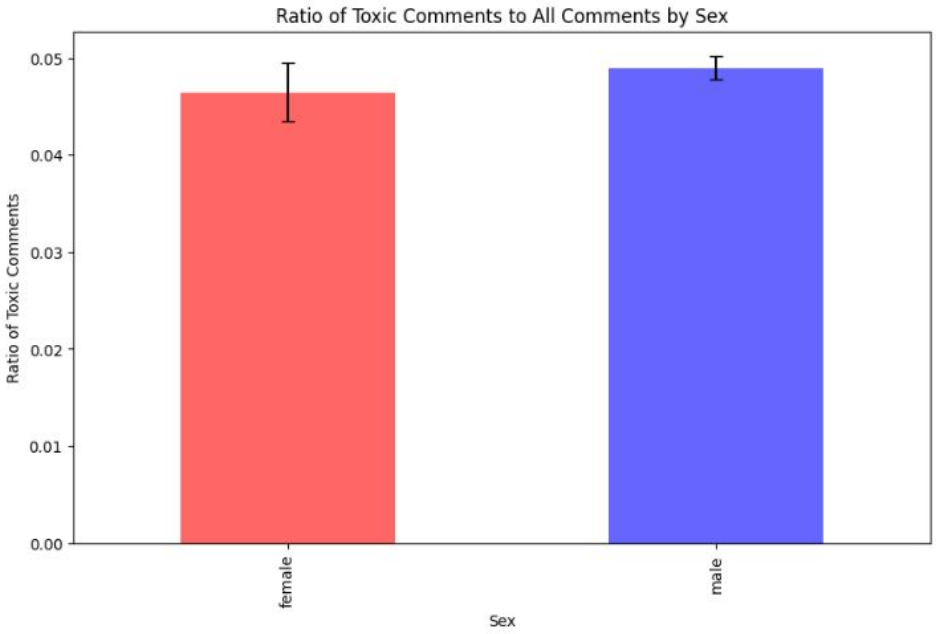
Mann-Whitney U Test:

- **U-statistic:** 1,105,051,493.0
- **P-value:** 1.37e-11
- **Interpretation:** Significant difference in distributions: can imply a difference in central tendencies

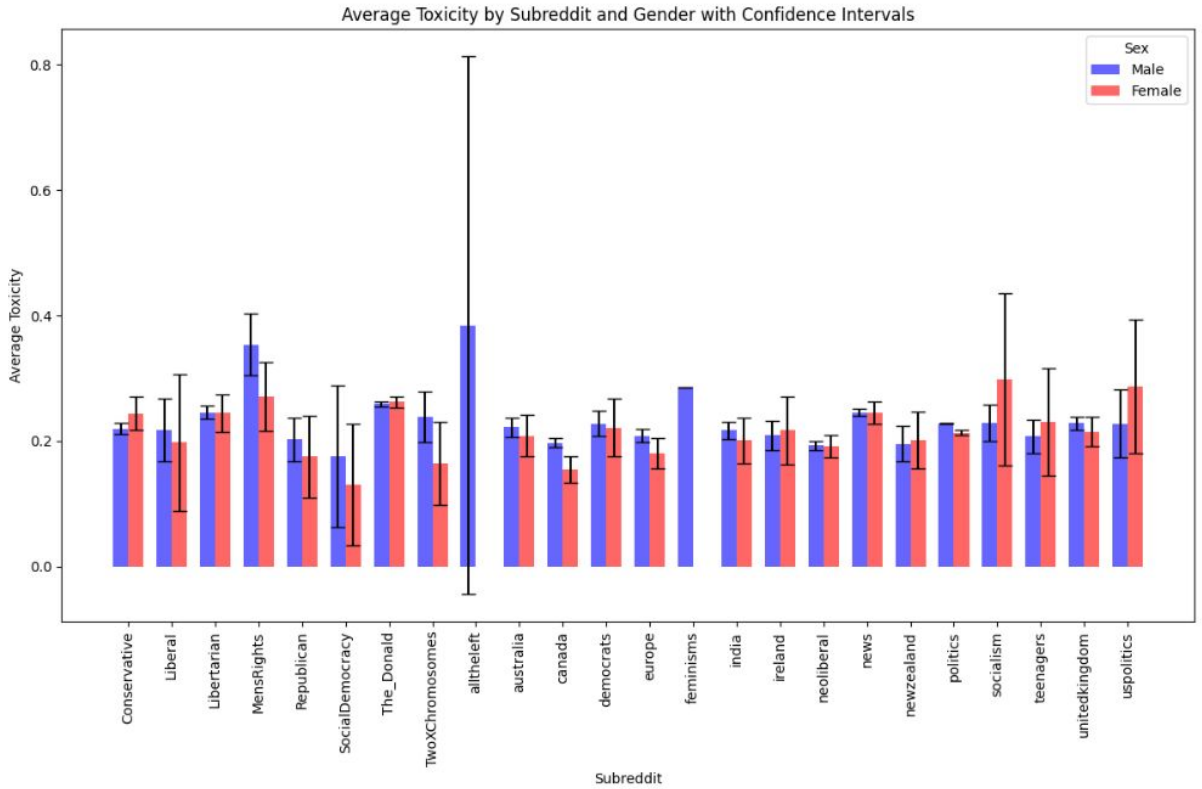
- **Mean toxicity score for males: 0.23 [0.23, 0.23] 95CI**
- **Mean toxicity score for females: 0.22 [0.22, 0.22] 95CI**
- **Mean difference in toxicity scores (female - male): -0.01 [-0.01, -0.01] 95CI**

Toxicity Analysis





Toxicity Analysis



Methods - Text Complexity

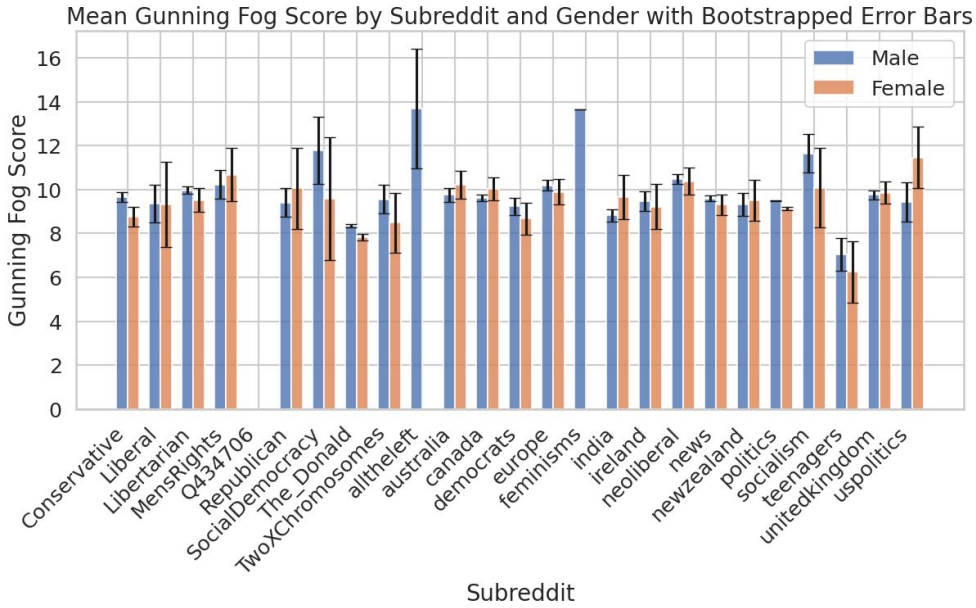
Syntactic and Semantic Text Complexity

- **Syntactic Complexity:**
  - **Gunning-Fog Index:** $0.4 \cdot \left(\left[\frac{\#\ words}{\#\ sentences}\right] + 100 \cdot \left[\frac{\#complex\ words}{\#\ words}\right]\right)$
    returns a score between 0 and 20 that resembles years of education that are required to understand a text
    → the longer the sentences & more complicated words, the higher the score
    → are emotional texts less readable?

- **Semantic Complexity:**
  - **Type Token Ratio:** The share of unique words against the total count of words
  - a greater variety of words used indicates a richer vocabulary
    → do emotional texts display a smaller vocabulary?

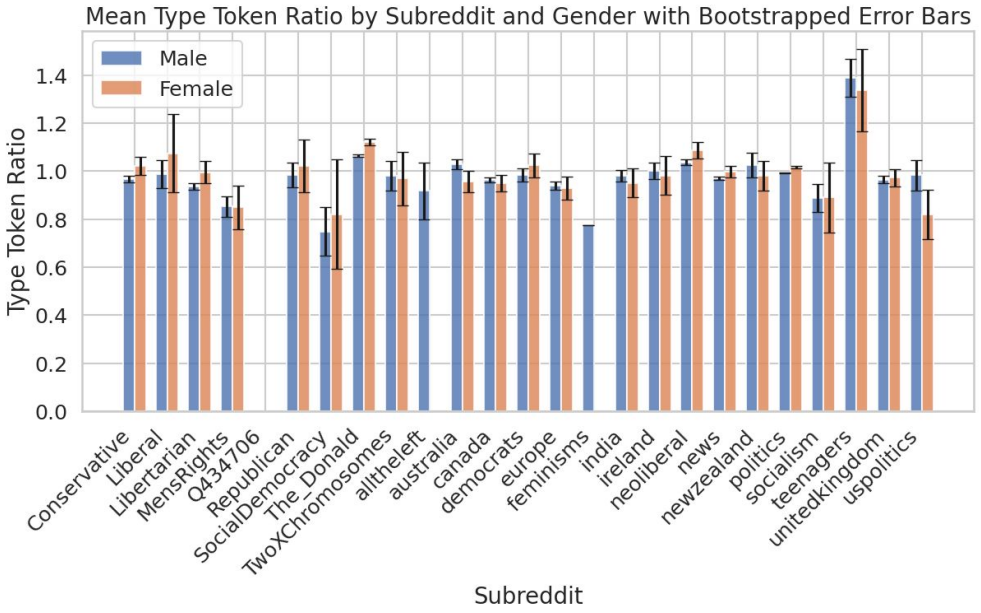
Gunning-Fog

- most subreddits display similar scores
- subreddits with high Gunning-Fog scores are:
  - ‘SocialDemocracy’
  - ‘alltheleft’
  - ‘feminisms’
    → only when men are targeted
- the subreddit 'teenagers' has the lowest average Gunning Fog score across men and women



Type Token Ratio

- No clear pattern apparent
- the subreddits ‘alltheleft’ and ‘feminisms’ do not target women in their sampled posts
- the subreddit ‘teenagers’ displays the highest TTR
- standard errors are smaller



Results

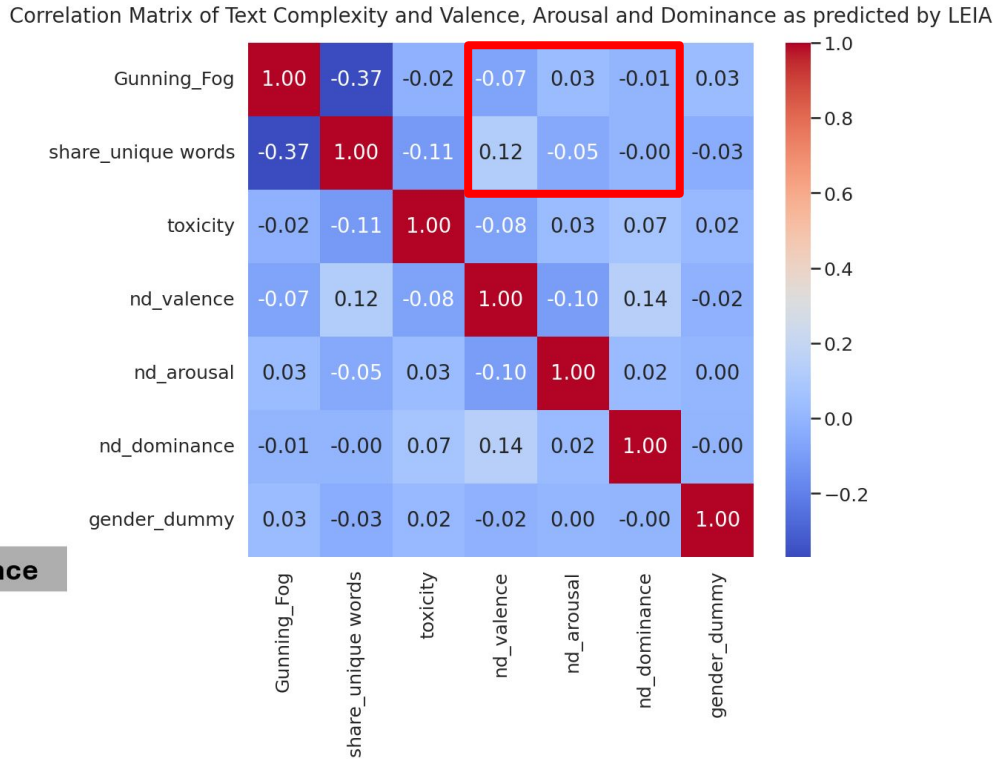
Correlation Analysis: H1

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

→ correlations between text complexity and LEIA VAD are low

→ interestingly, a higher TTR correlates negatively with Gunning-Fog (-0.37)

	Valence	Arousal	Dominance
Gunning-Fog	-0.07	0.03	-0.01
TTR	0.12	-0.05	-0.00



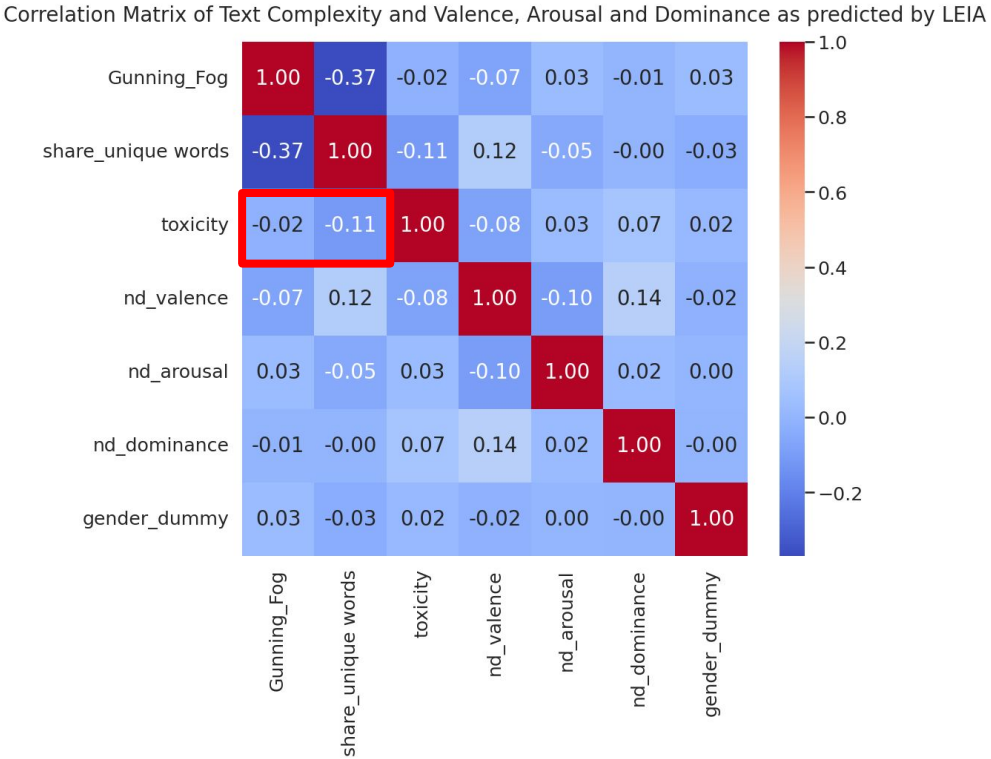
Correlation Analysis: H2

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

→ Gunning-Fog: -0.02

→ TTR: -0.11

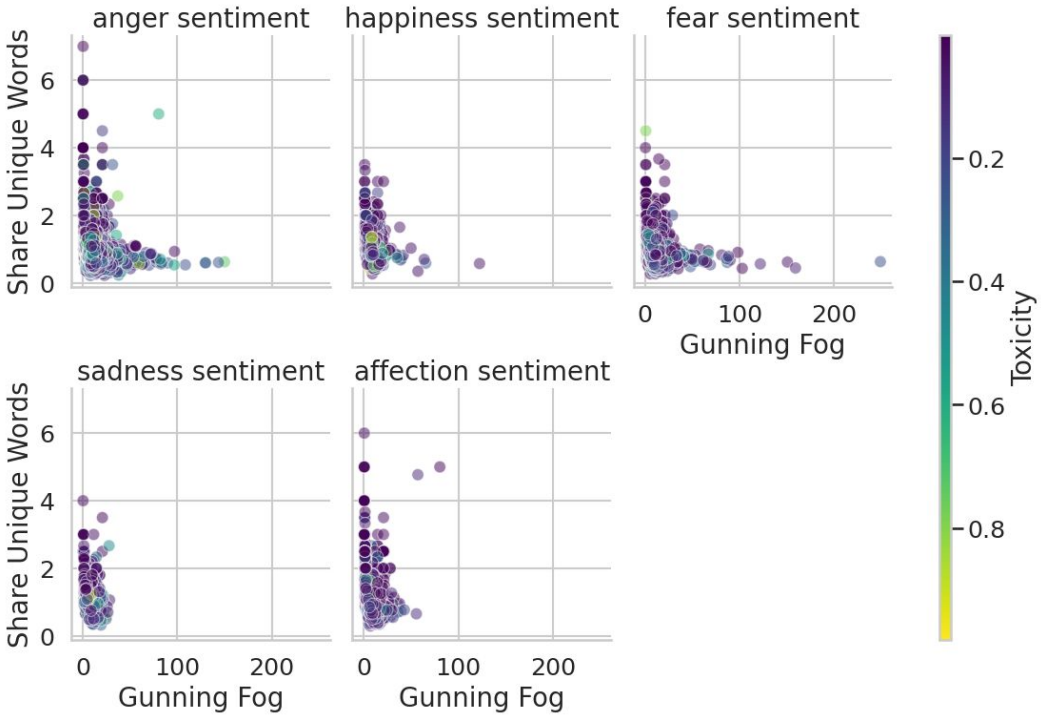
correlations are low but negative



Correlation Analysis: H2

- The sentiments ‘anger’, ‘fear’ and ‘affection’ have both higher TTR values and more outliers for Gunning-Fog

- Overall, the probability at which posts are perceived to be toxic is low across all five sentiments and does not seem to affect either text complexity measure



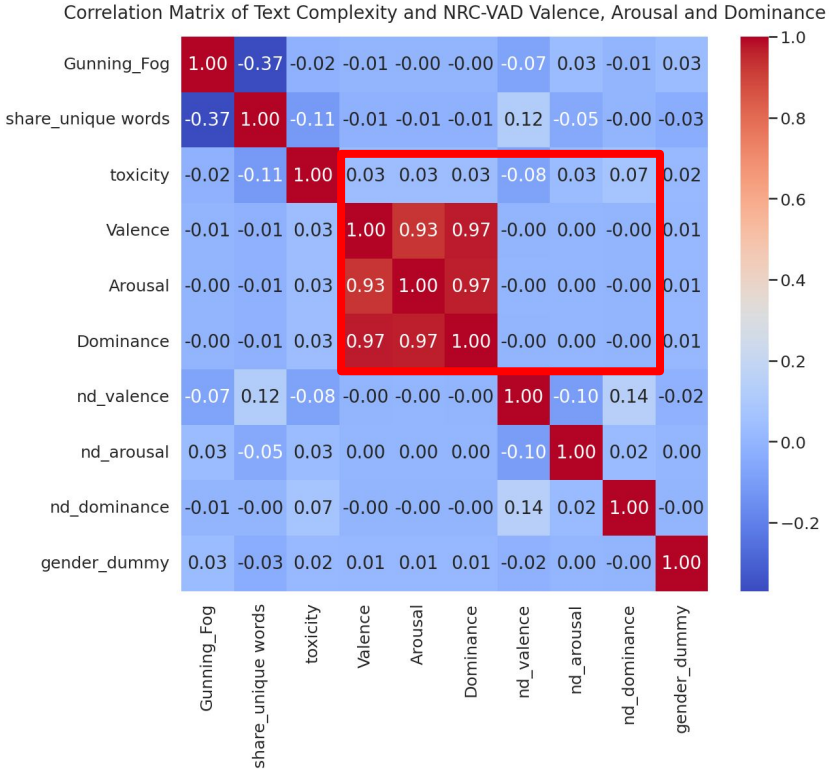
Correlation Analysis: H3

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

→ LEIA VAD scores do NOT correlate with NRC-VAD, potentially due to mapping

→ correlations across NRC-VAD are constant: 0.03

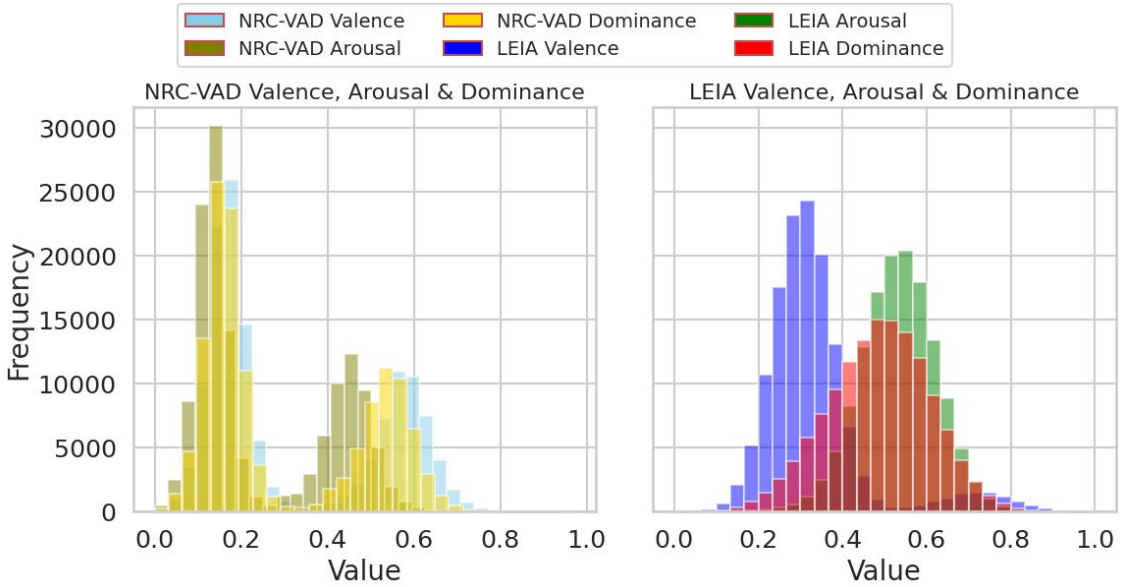
→ correlations across LEIA VAD differ slightly, but are low



Correlation Analysis: H3

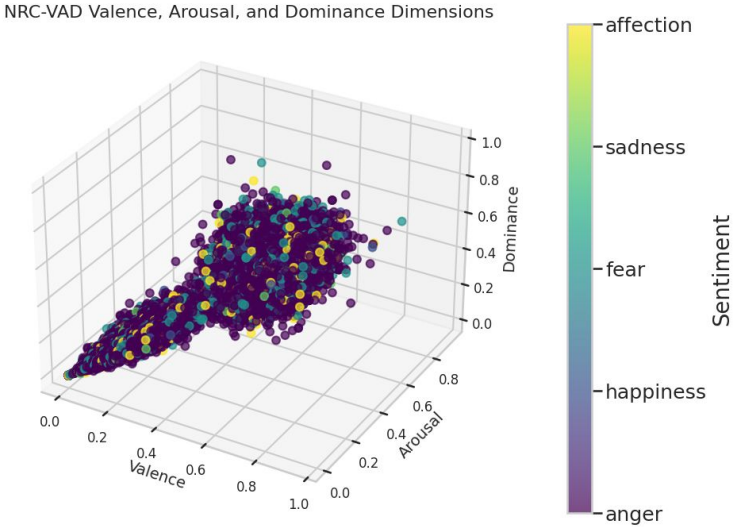
The SD-adjusted LEIA sentiment mapping does not correlate with the NRC-VAD scores but the shift of distributions across the x-axis might explain the differences in correlations

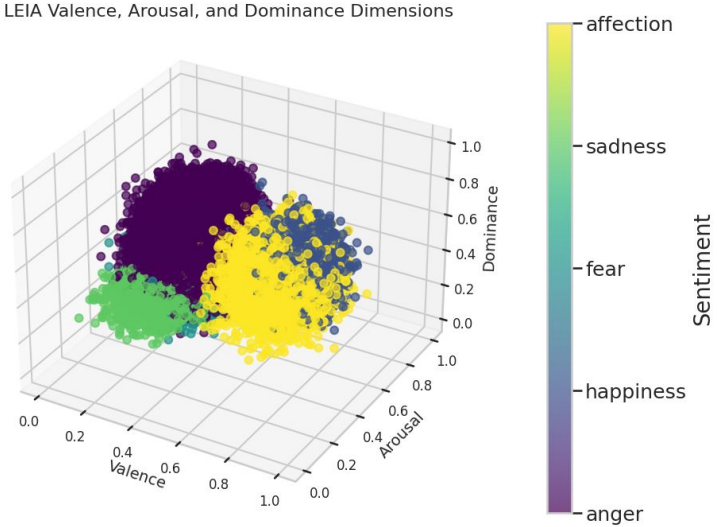
0.03	0.03	0.03	-0.08	0.03	0.07
1.00	0.93	0.97	-0.00	0.00	-0.00
0.93	1.00	0.97	-0.00	0.00	-0.00
0.97	0.97	1.00	-0.00	0.00	-0.00



Correlation Analysis: H3

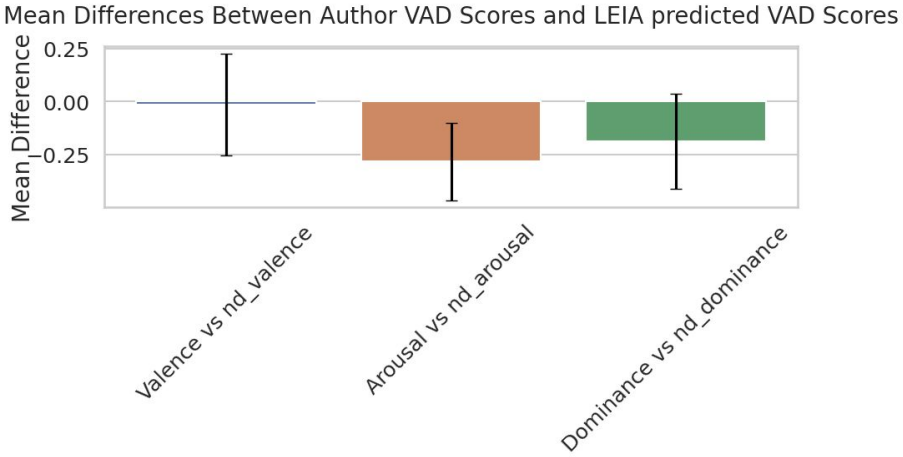
The SD-adjusted LEIA sentiment mapping does not correlate with the NRC-VAD scores

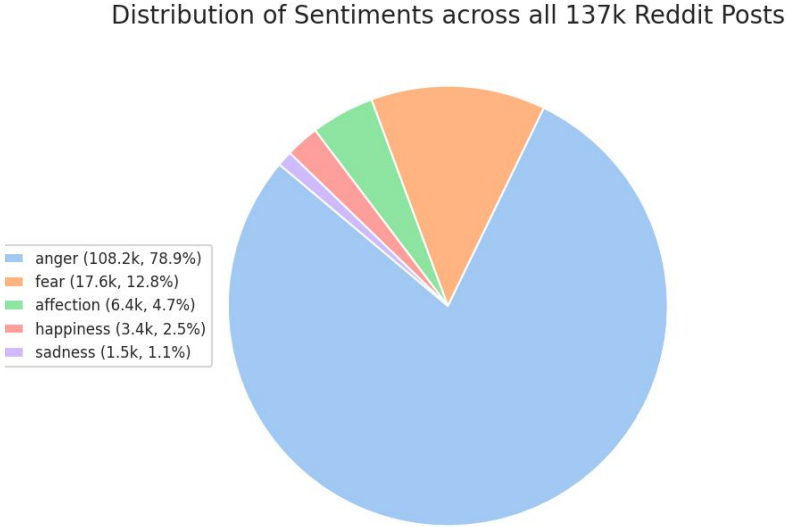




Correlation Analysis: H3

So... Which one is realistic? And is LEIA just angry?





Conclusion

H1: We do NOT find sufficient evidence to say that gender biased Reddit posts display lower text complexity:

> Correlations are low and mixed in direction across both semantic and syntactic text complexity

H2: We do NOT find sufficient evidence to say that Reddit posts that are perceived as toxic display lower text complexity:

> The probability at which posts are perceived to be toxic is low across all five sentiments and does not seem to affect either text complexity measure

H3: We do NOT find sufficient evidence to say that Sentimental gender bias operationalized through toxicity and sentimental gender bias operationalized through VAD correlate positively:

> Correlations are mixed in direction and low, the LEIA VAD scores noticeably do not correlate with each other at all.

Ethics

Ethics

- **Annotating Toxic comments**
  - was okay for us :D
  - cultural background influenced annotation of texts (white - european biased?!, very weird!)
- **Anonymized Data**
  - names are redacted
  - consideration of ‘Gender-Groups’ that are large enough so that no individual can be traced
- **Reproducibility**:
  - all code and figures are reported on GitHub
  - data with all new variables is stored XXXX

References

- **Cheng, L., Mosallanezhad, A., Silva, Y. N., Hall, D. L., & Liu, H. (2022)**. Bias Mitigation for Toxicity Detection via Sequential Decisions. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1750–1760. https://doi.org/10.1145/3477495.3531945
- **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
- **Glick, P., & Fiske, S. T. (1996)**. The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. 70 (3), 491–512.
- **González-Bailón, S., & Paltoglou, G. (2015).** Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. The ANNALS of the American Academy of Political and Social Science, 659(1), 95–107. https://doi.org/10.1177/0002716215569192
- **Haselmayer, M., & Jenny, M. (2017)**. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. Quality & Quantity, 51(6), 2623–2646. https://doi.org/10.1007/s11135-016-0412-4
- **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

References II

- **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
- **Kwarteng, J., Perfumi, S. C., Farrell, T., Third, A., & Fernandez, M. (2022).** Misogynoir: Challenges in detecting intersectional hate. Social Network Analysis and Mining, 12(1), 166. https://doi.org/10.1007/s13278-022-00993-7
- **Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022)**. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3197–3207. https://doi.org/10.1145/3534678.3539147
- **Marjanovic, S., Stańczak, K., & Augenstein, I. (2022).** Quantifying gender biases towards politicians on Reddit (N. J. Shook, Ed.). PLOS ONE, 17(10), e0274317. https://doi.org/10.1371/journal.pone.0274317
- **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

References III

- **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
- **Ofek, N. (2023)**. Sentiment Analysis for Social Text. In L. Rokach, O. Maimon, & E. Shmueli (Eds.), Machine Learning for Data Science Handbook (pp. 801–831). Springer International Publishing. https://doi.org/10.1007/978-3-031-24628-9_35
- **Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020).** Toxicity Detection: Does Context Really Matter? [Version Number: 1]. https://doi.org/10.48550/ARXIV.2006.00998
- **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
- **Sven, B., & Hahn, U. (2016).** Emotion Analysis as a Regression Problem — Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation, 1114–1122. https://doi.org/doi:10.3233/978-1-61499-672-9-1114