Research Projects in Computational Studies of Social Phenomena

— Midway Presentation —

Contents

Theory: Toxicity and Sentimental Gender Bias

Sentimental gender bias influences people through traditional gender stereotypes and personal feelings, favoring one gender over the other (Marjanovic et al. 2022)

- **Glick & Fiske (1996):** Benevolent and Hostile Sexism
- **Pavlopoulos et al. (2020):** Toxicity as umbrella term for ‘offensive’, ‘abusive’, and ‘hateful’ language
  - Implies connection between toxicity and (hostile) sexism
- **Cheng et al. (2022):** Contents of toxic social media posts are often motivated by stereotypical gender biases
- **Caliskan et al. (2022):**
  - Male-associated words → dominance & female-associated words → valence

Theory: Sentimental Gender Bias and Text Complexity

"Machine learning algorithms trained in natural language processing tasks have exhibited various forms of systemic racial and gender biases" (Sharma et al., 2020, p. 1)

Sentimental gender bias is found in NLP tasks such as:

- learned word embeddings (Bolukbasi et al., 2016; Brunet et al., 2019)
- natural language inference (He et al., 2019a)
- hate speech detection (Park et al., 2018)

Toxicity and text complexity:

- **Kayam (2018):** Analysis of Trump's speeches revealed them to be less complex and easier to read then other politicians
- **Morzhov (2021):** Toxic texts display fewer unique words and 'less inventive' use of language

Research Question & Hypotheses

- Research Question
  - How do sentimental gender bias and text complexity appear (together) in Reddit posts?
- Hypotheses
  - H1: Sentimentally gender biased (VAD) Reddit posts about or directed at female US politicians display lower text complexity.
  - H2: Toxic Reddit posts about or directed at female US politicians display lower text complexity.
  - H3: Sentimental gender bias operationalized through toxicity and sentimental gender bias operationalized through VAD correlate positively.

Data

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

- Original data contains over 10 Mio Reddit posts
- Sampled due to computational resources (~137k posts)
    - apply random sample directly for paper replication
    - post-stratification of sample to adjust for imbalanced group sizes to test hypotheses
- Data was gathered to investigate gender bias on Reddit
- Key variables include:
    - content, named_entity, sex, valence, arousal, dominance

Methods: Sentimental Gender Bias

'When people discuss male and female politicians, do they express equal sentiment and power levels in the words chosen?'

- Sentimental gender bias is operationalized through VAD: Valence, Arousal, Dominance
- VAD NCR lexicon: 20.000 entries with scores for valence, arousal and dominance
  → compute average score per post
    - Authors: Version 2018 → Our Replication: Version 2022
- Test for significance of variables using t-tests of means
  → evaluate strength of effects using Cohen's D

Methods: Valence, Arousal and Dominance

Validate Marjanovic et al. (2022) following Atmaja & Akagi (2019)[1]

- 1. Analyze Reddit posts using a combination of CoreNLP (Manning et al., 2014) and Valence, Arousal, Dominance (VAD) Lexicon derived from Affective norms for English words (ANEW) (Bradley & Lang, 1999)
- 2. Rerun the analysis performed by Marjanovic et al. using the ANEW VAD scores
    - t-tests for VAD differences between genders
    - Cohen’s Delta analysis to gauge effect size of sentimental gender bias

1. GitHub: bagustris/text-vad

Methods: Text Complexity

Text complexity is evaluated following Kayam (2018):

- 1. The percentage of complex words
- 2. The average number of words per sentence (sentence length)
- 3. The average number of syllables per word
- 4. The average number of characters per word (word length)
  → Validate sample

Reliability and Validity:

- Gunning Fog Index → how many years of education are required?
- SMOG
- Flesch-Kincaid Readability Score → scaled between 0 (very easy to read) and 100 (very difficult to read)

Methods: Toxicity

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

Toxicity: "A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."

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)

Methods: Toxicity

COMMENT

You're a real idiot, you know that.

☐ This comment is not in English or is not human-readable.

Rate the toxicity of this comment.

Very toxic: A comment that is very hateful, aggressive, disrespectful, or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.

Toxic: A comment that is rude, disrespectful, unreasonable, or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.

- ○ Very toxic
- ○ Toxic
- ○ Maybe, not sure
- ○ Not Toxic

Does this comment contain obscene or profane language?

Profanity/obscenity: Swear words, curse words, or other obscene or profane language.

- ○ Yes
- ○ Maybe, not sure
- ○ No

Does this comment contain identity-based negativity?

Identity-based negativity: A negative, discriminatory, stereotype, or hateful comment against a group of people based on criteria including (but not limited to) race or ethnicity, religion, gender, nationality or citizenship, disability, age, or sexual orientation.

- ○ Yes
- ○ Maybe, not sure
- ○ No

Does this comment contain insulting language?

Insults: Inflammatory, insulting, or negative language towards a person or a group of people. Such comments are not necessarily identity specific.

- ○ Yes
- ○ Maybe, not sure
- ○ No

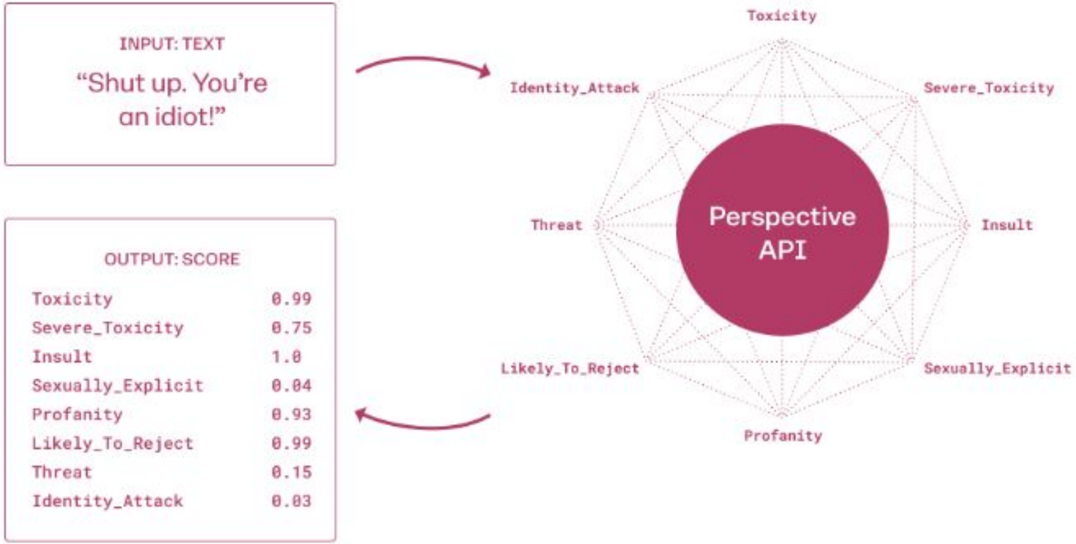
Does this comment contain threatening language?

Threatening: Language that is threatening or encouraging violence or harm, including self-harm.

- ○ Yes
- ○ Maybe, not sure
- ○ No

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

Methods: Toxicity

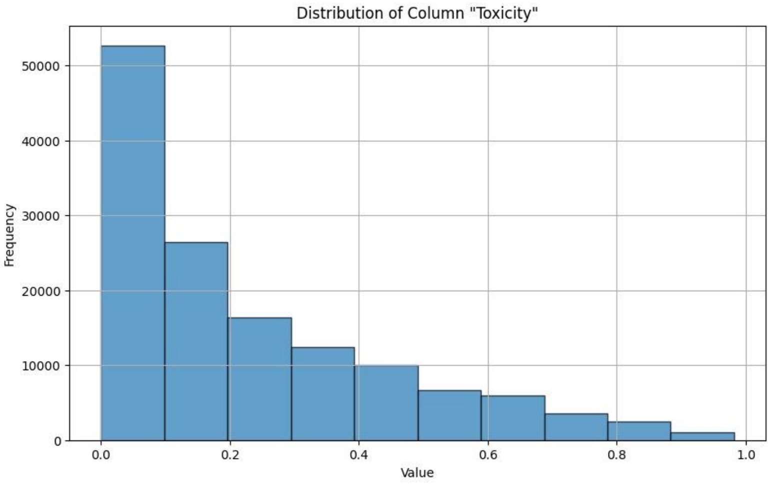


https://developers.perspectiveapi.com/s/about-the-api?language=en_US

- 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

Methods: Toxicity

Check if operationalization are normally distributed:



- Validate sample
- Test for significance of variables using log transformation and t-tests or Mann-Whitney U Test
- Bootstrapping (differences in means)
- Cohen's Delta analysis to gauge effect size of toxicity gender bias (t-test)

Methods: Correlation Analysis

- **Correlation Coefficients:** Compute correlation coefficients to examine the relationship between gender bias (measured via VAD/toxicity) and readability scores.
- **Spearman Correlation:** Use this if the data does not meet the assumptions of Pearson correlation (normal distributed and linear relation).
- **Visualization:** Scatter plots to show correlation between variables
- **Bootstrap Correlations:** Get Confidence Intervals to see how stable the results are

Open Questions

- Validation:
  - Bimodal NRC VAD values vs. normal distributed ANEW VAD values
  - NRC Dict from 2022 or 2018 (used in Marjanovic et al., 2022)
  - Should we use bootstrapped results for paper replication?
- Sample:
  - Resample?
- Statistical tests:
  - Enough?!

References

- Atmaja, B. T. & Akagi, M. 2019. Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text. https://doi.org/10.31227/osf.io/fhu29
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf
- Bradley, M. M. & Lang, P. J. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Vol. 30, no. 1. Technical report C-1, the center for research in psychophysiology, University of Florida.
- Brunet, M.- E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the Origins of Bias in Word Embeddings. In K. Chaudhuri & R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning (pp. 803–811, Vol. 97). PMLR. https://proceedings.mlr.press/v97/brunet19a.html
- Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society,156–170. https://doi.org/10.1145/3514094.3534162
- 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
- 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/10.1037/0022-3514.70.3.491

References II

- He, Q., Wang, H., & Zhang, Y. (2020). Enhancing Generalization in Natural Language Inference by Syntax. Findings of the Association for Computational Linguistics: EMNLP 2020, 4973–4978. https://doi.org/10.18653/v1/2020.findings-emnlp.447
- 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
- Kwarteng, Joseph, et al. 2022 "Misogynoir: challenges in detecting intersectional hate." Social Network Analysis and Mining 12.1: 166.
- Lees, A. et al. 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.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.
- Marjanovic, S., Stańczak, K. & Augenstein, I. 2022. Quantifying gender biases towards politicians on Reddit. 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
- Pavlopoulos J., Sorensen J., Dixon L., Thain N., & Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter?. arXiv. https://doi.org/10.48550/arXiv.2006.00998
- Sharma, S., Dey, M., & Sinha, K. (2021). Evaluating Gender Bias in Natural Language Inference [Version Number: 1]. https://doi.org/10.48550/ARXIV.2105.05541