Visualization of Social Stereotypes and Biases in Toxic Language Detection Models

Xuhui Zhou*
xuhui_zhou@gatech.edu
Georgia Institute of Technology

Chengyu Bao†
cbao37@gatech.edu
Georgia Institute of Technology

Zhili Luo‡
zluo79@gatech.edu
Georgia Institute of Technology

Renjun Lu§
rlu73@gatech.edu
Georgia Institute of Technology

1 INTRODUCTION

We will introduce our toxic language detection model which aims to detect whether a given statement implies stereotypes and social bias and provide detailed explanations for the implied harms. Using the SBF-GPT based classification model [15], we will focus our analyses on three existing English Twitter data sets annotated for toxic or abusive language from Founta et al. [5], Waseem and Hovy [17] and Davidson et al. [3].Then we will visualize the output via a user-interactive web app.

2 LITERATURE SURVEY

Social stereotypes are prevalent in language detection models. Nadeem et al. [11] built the first benchmark to evaluate social biases embedded in large-scale language models on a test dataset they built. Their synthetic dataset, however, provided little insight about how the stereotype is manifested in our daily language. On the other hand, we will test our models using three existing English Twitter datasets which collect the daily language in online content. Sap et al. [15] also introduce a new corpus containing 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups. While the dataset was large, annotations of social stereotypes lacked sufficient quality control. Blodgett et al. [1] examined Sap et al.'s dataset and found only 13% of the datasets were not affected by the pitfalls they pre-coded. In our work, the design of the benchmark will be elusive in terms of what is really being measured, which leads to the idea of better data exploration and interaction.

In our study, we will consider two broad categories of language detection biases: (1) gender biases, and (2) racial biases.

2.1 Gender Biases

Gender bias is entangled with grammatical gender information in word embeddings of languages with grammatical gender. For example, "You are a good woman" was considered "sexist" when trained on an existing dataset, which is not robust for practical use. Bolukbasi et al. [2] is one of the first works to point out the gender stereotypes inside word2vec [10] and constructed a gender neutralization framework based on cosine similarity and orthogonal vector projections to remove gender bias. Lu et al. [9] showed that gender bias in coreference resolution and language modeling can be mitigated through a data augmentation technique that expands the corpus by swapping the gender pairs like he and she, or father and mother. They called this Counterfactual Data Augmentation (CDA). We could make improvements by combining these two methods together for pre-trained embeddings to achieve a better performance. However, some limitations of the mentioned studies include the infeasibility of data augmentation for gender neutral languages and the subjectivity posed by less agreeable gendered words for use in gender subspace removal. To avoid these limitations, Dinan et al. [4] introduced a general framework that decomposes gender bias in text along several pragmatic and semantic dimensions: bias from the gender of the person being spoken about, bias from the gender of the person being spoken to, and bias from the gender of the speaker [4]. This provides us with more options to find the gender bias contents in the text. We will improve our bias detection algorithm by identifying bias with different approaches and comparing them.

2.2 Racial Biases

Speakers' identity or dialect is another major bias in toxic language detection. For example, the n word is the common term which is understood as non-offensive by African American English (AAE) speakers but is labeled as toxic by most available toxicity detection tools, regardless of the racial identity of speakers. Sap et al. [14] showed that dialect and race priming factors significantly decreased the likelihood of AAE tweets being labeled as offensive in a small controlled experiment on Amazon Mechanical Turk workers. However, they did not test the significance of race factor in any NLP models. On the other hand, Xia et al. [18] used adversarial training at a model level to demote AAE dialect The majority of existing works used false-positive rate (FPR) [1] as the only criteria for correcting models' biases against AAE users. In fact, AAE samples in popular toxic-language datasets are mainly annotated as toxic, leaving a very small sample of AAE instances that are true negatives(non-toxic). Hence, Halevy et al. [6] employed another two fairness metrics to measure the

[1]FPR: the probability of classifying non-toxic samples as toxic conditional on the samples being non-toxic

performance of detection models, along with FPR: disparate impact (DI) metric [2] and false negative rates (FNR). Inspired by these existing works, we will consider racial identity and dialect of speakers in our language reasoning models, and also evaluate the performance of our models using several fairness metrics.

2.3 Visualization

We examined various approaches of visualizing text mining and natural language models. Word clouds are commonly used to showcase analytical information through words. Concentri Cloud is an innovative tool that uses concentric layout to showcase the association between multiple documents [8]. Document placement determined by similarities, while most frequent words are closer to the center. We expect words and phrases to have strong association with the output, so it would be helpful to represent our samples with common words. The layout, however, limits the number of samples displayed at once. The greedy algorithm leaves out the information between samples that are less similar. A study in social network thematic similarity uses bubble chart and network to visualize the common themes they extract through clustering [12]. While this method allows us to display all features, potentially many of them may not be as legible enough to relate sensible information to the user. Efforts have also been made to visualize text through interactive linking and annotation. Latif et al. [7] proposed a framework which is suitable for marking up generated feedback on specific sections of paragraphs, associating text with graphic visualization. However, this method requires screen space to display the samples themselves and may not be space efficient to become viable for a larger dataset.

3 INNOVATION, RISK, PAYOFF, COST & CHECKPOINTS

Social stereotypes are nuanced connotation hidden under the surface of human language. Previous works annotating social stereotypes relies on crowd-sourcing methods are inherently hard to control the annotation quality due to the subjective nature of this topic and unintuitive annotation quality inspection process.

[2]DI evaluates the predictive parity ratio to compare predicted outcomes across AAE group and Standard American English group

Our visualization tool has the promise to significantly alleviate the situation by providing the visualized frame of the annotated/predicted social stereotypes to the practitioners. Furthermore, we intend to group the frames with different dimensions which can help the audience understand the dataset better from a macroscope.

We also assume that risks may occur during development. The language model expects sentences as input and may not take into account the context of the language used. Thus it may not work as well for paragraphs where the author's intention is less clear in single sentences. We might also find the model to overfit with the training data, producing less desirable results when processing more complex real-world data. We would also like to acknowledge the ethical concerns regarding the usage of our tool. The predictions are imperfect and cannot be used as only guidance to circumventing inappropriate speech.

We will evaluate our method through user studies and gather feedback on how they retrieve information from our visualization tool. We will also measure the scalability of our code and identify any bottlenecks to supporting large input.

We propose to use the AWS platform to host our service, the usage of which should be covered by our student credits. We will measure our progress by checking off tasks in time on our plan of activities. At mid point, we should have most of the components in the language model and visualization functional. At final stage, we should have the full pipeline deployed and ready to use.

4 DETAILED METHOD DESCRIPTION

We illustrate our methods from both framework and modeling perspectives.

4.1 Framework

We adopt the framework demonstrated in social bias frame [15]. Specifically, the categories that we are interested in when visualizing the social biases and stereotypes in the text data are:

- **Offensiveness** denotes the overall toxicity of a certain sentence.

- **Intent to offend** indicates the author's intention in making the sentence offensive. This variable has four possible answers (yes,probably, probably not, no).
- **Lewd** suggests whether sentence contains sexual references.
- **Group implications** capture whether the sentence targets a specific group.
- **Targeted group** is a text slot following the the positive prediction of **Group implications** indicating which group is being offended.
- **Implied statement** is *free-text answers* demonstrating the power dynamic or stereotype that is referenced in the sentence.
- **In-group language** captures whether the sentence belongs to certain group since in-group language is often perceived differently, such as reclaimed slurs.

By utilizing this novel framework, we have the theoretical foundation to make our visualizing content pertinent to the analyzing of social biases in the documents.

4.2 Modeling

The backbone of our visualization of stereotypes in certain long text data is a customized GPT-2 model [13], which outputs the aforementioned categorical or free-text variables synchronously. This supports real-time stereotype analysis of the documents.

GPT-2 is a 1.5B parameter Transformer model [16] that trained over large amount of text data on the internet. With GPT-2, we predict the framework of sentences in the documents in a hybrid classification and language generation mode. We linearize the variables following the frame hierarchy in social bias frames [15].

4.3 Innovations

- We use the social bias frame as the framework to analyze the stereotypes in different kinds of text datasets and documents. To the best of our knowledge, this is the first time the framework is applied for text analysis.
- We provide a competitive model to perform real-time stereotype analysis for the text data and visualize the analysis for better human understanding. GPT-2 is innovatively used as the backbone of text stereotype analysis.

5 DESIGN OF UPCOMING EXPERIMENTS

We are currently focused on cleaning data for model evaluation. After getting the output of the model, we will visualize the distribution of different biases such as gender biases and racial biases through an user-interactive web app containing node graphs, pie charts, and histograms.

6 PLAN OF ACTIVITIES

All team members contribute a similar amount of effort. Following is the old plan:

TASK	ASSIGNED	PROGRESS	START	END
Phase 1				
Writing Proposal	All	100%	2/28/2022	3/5/2022
Model building	Xuhui Zhou	0%	3/7/2022	3/20/2022
Phase 2 Title				
Training Data	Renjun Lu	0%	3/21/2022	3/27/2022
Node graph	Chengyu Bao	0%	3/28/2022	4/6/2022
Histogram/Pie chart	Zhili Luo	0%	3/28/2022	4/6/2022
Phase 3 Title				
Final draft	All	0%	4/7/2022	4/11/2022
Presentation	All	0%	4/12/2022	4/16/2022
Final paper	All	0%	4/17/2022	4/20/2022

Following is the revised plan:

TASK	ASSIGNED	PROGRESS	START	END
Phase 1				
Writing Proposal	All	100%	2/28/2022	3/5/2022
Model building	Xuhui Zhou	100%	3/7/2022	3/20/2022
Phase 2 Title				
Training Data	Renjun Lu	50%	3/21/2022	4/3/2022
Node graph	Chengyu Bao	0%	4/4/2022	4/10/2022
Histogram/Pie chart	Zhili Luo	0%	4/4/2022	4/10/2022
Phase 3 Title				
Final draft	All	0%	4/9/2022	4/14/2022
Presentation	All	0%	4/15/2022	4/22/2022
Final paper	All	0%	4/15/2022	4/22/2022

REFERENCES

[1] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1004–1015.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NIPS*.

[3] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

[4] Emily Dinan, Angela Fan, Ledell Yu Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *EMNLP*.

[5] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *ArXiv* abs/1802.00393 (2018).

[6] Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna M. Howard. 2021. Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2021).

[7] Shahid Latif, Diao Liu, and Fabian Beck. 2018. Exploring Interactive Linking Between Text and Visualization. In *EuroVis 2018 - Short Papers*, Jimmy Johansson, Filip Sadlo, and Tobias Schreck (Eds.). The Eurographics Association. https://doi.org/10.2312/eurovisshort.20181084

[8] Steffen Lohmann, Florian Heimerl, Fabian Bopp, Michael Burch, and Thomas Ertl. 2015. Concentri Cloud: Word Cloud Visualization for Multiple Text Documents. In *19th International Conference on Information Visualisation, IV 2015, Barcelona, Spain, July 22-24, 2015*, Ebad Banissi, Mark W. McK. Bannatyne, Fatma Bouali, Remo Burkhard, John Counsell, Urska Cvek, Martin J. Eppler, Georges G. Grinstein, Wei dong Huang, Sebastian Kernbach, Chun-Cheng Lin, Feng Lin, Francis T. Marchese, Chi-Man Pun, Muhammad Sarfraz, Marjan Trutschl, Anna Ursyn, Gilles Venturini, Theodor G. Wyeld, and Jian J. Zhang (Eds.). IEEE, 114–120. https://doi.org/10.1109/iV.2015.30

[9] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. *Gender Bias in Neural Natural Language Processing*. Springer International Publishing, Cham, 189–202. https://doi.org/10.1007/978-3-030-62077-6_14

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.

[11] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5356–5371.

[12] Albert Park, Mike Conway, and Annie T. Chen. 2018. Examining Thematic Similarity, Difference, and Membership in Three Online Mental Health Communities from Reddit. *Comput. Hum. Behav.* 78, C (jan 2018), 98–112. https://doi.org/10.1016/j.chb.2017.09.001

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[14] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *ACL*.

[15] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social Bias Frames: Reasoning about Social and Power Implications of Language. (Nov. 2019). arXiv:cs.CL/1911.03891

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[17] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL*.

[18] M. Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *SOCIALNLP*.