Quantifying Bias in Contextualized Embeddings

Darius Irani, JungMin Lee, Beatriz Medeiros, David Francisco
{dirani2, jlee581, bmedeir1, dfranc25}@jhu.edu

I. INTRODUCTION

Word embeddings are learned representations of written language that capture distributional similarities between words. Because language describes the world we live in, which is biased and filled with racial and gender disparities, word embeddings also learn these biases (Bolukbasi et al.; Garg et al.). This is concerning because the biases in these embeddings have adverse effects on downstream NLP tasks like Named Entity Recognition (Mehrabi et al.) or Coreference Resolution (Rudinger et al.; Liu; Kurita et al.).

In recent years, contextualized embeddings extracted from language representation models like BERT (Devlin et al.) and ELMo (Peters et al.) have replaced the use of context-free word embeddings in state-of-the-art NLP systems. Embeddings learned by these models capture contextualized meaning, which makes them powerful for NLP tasks. While bias exists in these embeddings, identifying where it occurs in the models is challenging since the subspace is not linear (Karve et al.).

In this work, we explore and extend several methods that identify and quantify bias within BERT. Using the *Sentence Encoder Associate Test (SEAT)* (May et al.), we identify racial bias that exist within pre-trained BERT embeddings. We also expand upon similar methods used for identifying gender bias (Babaeianjelodar et al.; Zhao et al.; Sun et al.) and illustrate the prevalence of racial bias in various corpora by extracting sentence-level embeddings from BERT models fine-tuned on these corpora. The contributions of our work are two-fold:

- We demonstrate that BERT captures biases present in the dataset used for a given task. We show this by fine-tuning four BERT models on different corpora.
- We measure the implicit bias captured by the contextualized embeddings extracted from the fine-tuned BERT models using SEAT and evaluate how the corpora reflect racial bias.

II. METHODS

For the first stage of our experiments, we fine-tuned BERT on four corpora. We chose three corpora that we believed would exacerbate racial bias and one GLUE corpus that we did not expect to introduce excessive bias[1]:

- MSRParaphrase Corpus (GLUE)
- Jigsaw Unintended Bias in Toxicity Classification
- COVID-news-QuestionAnswering
- Twitter Hate Speech Offensive Language

Each corpus had a different task associated with its learning objective. For all four tasks, the BERT pre-trained model used was `bert-base-uncased` with a final linear classification layer provided by HuggingFace's `BertForSequenceClassification` extension. For fine-tuning, we generally followed the training parameters recommended by Devlin et al. and only made modifications to prevent over-fitting and meet computational resource limitations. For training, we used the AdamW optimizer with a learning rate of 2×10^{-5} and a decay rate of 1×10^{-8}. All data was tokenized with the BERT tokenizer and was split into 90% training and 10% validation. The models were saved and checkpointed only when there was a strict decrease in the validation loss.

[1]Examples from each corpus are included in Appendix A.1

A. MSRParaphrase and Validating Sentence Paraphrasing

The task described by the MSRParaphrase Corpus is to check whether two input sentences are paraphrases of each other Shah et al.. For pre-processing, numerical and non-ASCII characters were removed to maintain UTF-8 encodings. This left 5,682 sentences for training and validation. Fine-tuning was performed using a batch size of 32 for 4 epochs. After fine-tuning, the model achieved an accuracy of 85% on validation data.

B. Jigsaw Unintended Bias in Toxicity Classification

The Jigsaw toxic comment dataset (AI) contains 1,780,823 comments from the Civil Comments platform. Each comment has a toxicity label ranging from 0-1 and additional identity labels ranging from 0-1. Identity labels are those mentioned in the comment, such as male, female, black, and white. Each label represents the fraction of annotators who believed the label fit the comment. For our task, only comments with either a "black" or "white" label of 1 (all annotators agreed that the comment included black/white identities) were extracted for a total of 18520 comments. A binary toxicity label was assigned where toxicity ≥ 0.5 was considered to be in the positive (toxic) class. 5961 was labeled as toxic, 12559 as non-toxic. Due to memory limitations, comment length was limited to 256 tokens. After fine-tuning using a batch size of 16 for 4 epochs, the model was able to achieve a 79% accuracy on validation data.

C. COVID News Question Answering

The dataset includes 481 questions from news articles each with a correct and an incorrect answer for a total of 962 Q&A pairs (Lu). For each question, two sentences were added to the training data. One was the question concatenated with the corresponding correct answer separated by a '[SEP]' token. The other was the same except it used the incorrect answer instead of the correct one. Due to memory limitations, Q&A sentences were limited to only those with a length of 300 tokens leaving 778 examples. Binary labels were associated with each row and the task of the model was to determine whether the answer for the question was correct or incorrect. After fine-tuning using a batch size of 4 for 8 epochs, the model was able to achieve an 88% accuracy on validation data.

D. Hate Speech and Offensive Language Detection

For our final task, BERT was fine-tuned on a Twitter dataset with the goal of classifying tweets as containing hate speech, offensive language, or neither. The dataset was compiled by Davidson et al. using the Twitter API and accumulated 84.4 million tweets from 33,458 users, from which 25,000 were labeled using a crowd-sourcing platform. The final class label indicates the majority annotator rating for that tweet. The dataset is imbalanced (there are 1430 tweets labeled as hate speech, 19190 as containing offensive

language, and 4163 as neither). For pre-processing the dataset, we followed the procedure outlined in Mozafari et al. and replaced user handles, numbers, hashtags, and URLs with corresponding tokens. Additionally, elongated words were converted to their standard formats, all text was made lower case, and emoticons and punctuation were removed from the tweets. Since sentences were tweets, we set the max sentence length to 64 tokens. After fine-tuning using a batch size of 32 for 3 epochs, the final test accuracy achieved was 91% on the held-out test data.

III. BIAS EVALUATION

A. Extracting Sentence Embedding from BERT

BERT encodes a representation of the input sentences it is trained on. Sentence embeddings can be extracted from BERT's 12 hidden Tranformer layers using a number of strategies including averaging the output of all layers, averaging the last four layers, and taking just the last hidden layer (Devlin et al.). For our experiments, we used the token-level average of the last two hidden layers as the sentence embedding of length 768.

B. Sentence Encoder Associate Test (SEAT)

SEAT is an extension of the Word Embedding Association Test (WEAT) which uses cosine similarity of word embeddings as a statistical analogue to reaction time in the Implicit Association Test (Caliskan et al.). While WEAT is applied to sets of words, SEAT inserts these words into sentence templates to perform the same calculations on sentence embeddings. It is important to note that while SEAT can confirm the existence of bias, low scores do not imply that there is no bias (May et al.; Liang et al.).

Two sets of target concepts (X, Y) and attribute concepts (A, B) are first chosen. In our case, the target sets were "white" and "black", and the attribute sets were positive/negative emotions. The SEAT score $s(X, Y, A, B)$ can be calculated using the following equations (May et al.).

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$s(w, A, B) = mean_{a \in A} \cos(w, a) - mean_{b \in B} \cos(w, b)$$

A normalized difference of means of $s(w, A, B)$ - the effect size - is used to measure the magnitude of the association between (A, B) and (X, Y).

$$d = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{stddev_{w \in X \cup Y} s(w, A, B)}$$

C. Sentence Templates

To quantify biases present in the fine-tuned BERT models, we designed a set of *bleached* template sentences. These template sentences are semantically bleached because the sentence context does not contain information about the bias (May et al.). We defined target and attribute template sentences (see Appendix A.2). The target sentences were created by filling the slots with each target word (i.e. "black" and "white"). Finally, attribute sentences were created by filling the slots with each attribute word defined in the Equity Evaluation Corpus (Kiritchenko and Mohammad).

IV. DISCUSSION/CONCLUSION

The Jigsaw Toxicity dataset produced the most statistically significant bias, so we will use these results to frame our discussion. Figure 1 shows the cosine similarity between target and attribute sentences (see Appendix A.3 for the other plots). This plot shows that there is a trend where the "black" sentence embeddings are more similar to the negative emotion attributes and are less similar

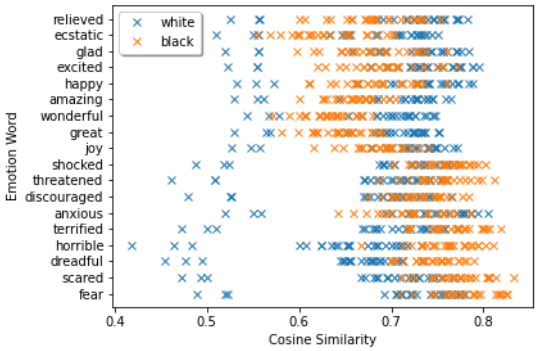

Fig. 1: Cosine similarities between embeddings of sentences with the terms "black" and "white" and those of sentences with each emotion word. The embeddings were taken from the Jigsaw model.

to the positive emotions. This indicates that there is bias towards negative emotions for "black". The opposite trend was observed for sentence embeddings with "white", indicating more bias towards positive emotions.

We can confirm the observation from above by calculating the effect size of the SEAT score for the Jigsaw model. Table I shows the effect size and p-values of the SEAT scores across the baseline and fine-tuned BERT models. An effect size of higher magnitude indicates more racial bias, and a negative score indicates that "black" is more associated with "negative" sentiments. For the Jigsaw model we can clearly see that fine-tuning had a significant impact on the prevalence of racial bias in the sentence embeddings, and that the association between "black" and negative sentiments have been strengthened.

	BERT-base	MSRP	Jigsaws	COVID	Hate
Effect size	-0.541	0.198	-1.762	0.573	0.537
p value	0.145	0.347	0.000078	0.132	0.147

TABLE I: Effect size and p-values of SEAT scores for all models.

All other models except the Jigsaws model have relatively low effect size and larger p-values. As we noted above, a low effect size in SEAT does not imply an absence of bias. Large p-values indicate that we cannot draw strong conclusions about the existence of bias in these corpora. While a negative effect size for BERT-base may hint that the pre-trained embeddings in BERT contain negative bias towards African Americans, other bias evaluation metrics must be explored to make any conclusive claims.

We were surprised that the Hate Speech and COVID-QA fine-tuned models did not exhibit any significant changes in the racial bias present in sentence embeddings, but we believe this may be due to the nature of the target concepts we chose. For a fine-tuned model to learn any bias associated with the target concepts, the target concept must actually appear in the corpus. While such examples were abundant in the Jigsaws dataset as we explicitly selected examples relating to "black" or "white", such explicit mentions may not have been present in the other corpora.

For future work, more diverse target sets that can capture racial identities should be explored. In addition, other evaluation metrics such as those discussed in Kurita et al. which offers a more robust log-probability score that measures the association of attribute and target words in a `BERTMaskedLM` model should be explored. We have seen that while there are many different ways to identify and qualify the bias within BERT, due to the nature of the architecture and the way it processes input, developing a way to debias BERT is be a formidable task for the future.

V. Code

The code used to obtain our results can be found in our GitHub repository at https://github.com/dr-irani/Quantifying-Bias-Contextualized-Embeddings.

VI. Acknowledgements

We are very grateful to Dr. João Sedoc, who provided great feedback and suggestions for scoping our project and whose research on debiasing context-free word embeddings inspired our work. We are also grateful to Dr. Silvio Amir, whose feedback during presentations was helpful for understanding the results we were obtaining.

References

[1] Jigsaw/Conversation AI. Jigsaw unintended bias in toxicity classification: Detect toxicity across a diverse range of conversations. 7 2019. URL http://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/.

[2] Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 752–759, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3383559. URL https://doi.org/10.1145/3366424.3383559.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [cs, stat]*, July 2016. URL http://arxiv.org/abs/1607.06520. arXiv: 1607.06520.

[4] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL http://arxiv.org/abs/1608.07187. arXiv: 1608.07187.

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

[7] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, April 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1720347115. URL http://arxiv.org/abs/1711.08412. arXiv: 1711.08412.

[8] Saket Karve, Lyle Ungar, and João Sedoc. Conceptor Debiasing of Word Representations Evaluated on WEAT. *arXiv:1906.05993 [cs]*, June 2019. URL http://arxiv.org/abs/1906.05993. arXiv: 1906.05993.

[9] Svetlana Kiritchenko and Saif Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL http://aclweb.org/anthology/S18-2005.

[10] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring Bias in Contextualized Word Representations. *arXiv:1906.07337 [cs]*, June 2019. URL http://arxiv.org/abs/1906.07337. arXiv: 1906.07337.

[11] Paul Pu Liang, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards Debiasing Sentence Representations. page 12.

[12] Bo Liu. Anonymized BERT: An Augmentation Approach to the Gendered Pronoun Resolution Challenge. *arXiv:1905.01780 [cs]*, June 2019. URL http://arxiv.org/abs/1905.01780. arXiv: 1905.01780.

[13] Xing Han Lu. Covid-qa: A collection of covid-19 qa pairs and transformer baselines for evaluating question-answering models. 3 2020. URL https://github.com/xhlulu/covid-qa.

[14] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://www.aclweb.org/anthology/N19-1063.

[15] Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. *arXiv:1910.10872 [cs]*, October 2019. URL http://arxiv.org/abs/1910.10872. arXiv: 1910.10872.

[16] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks 2019: 8th International Conference on Complex Networks and their Applications*, volume Studies in Computational Intelligence book series (SCI, volume 881) of *Complex Networks and Their Applications VIII : Volume 1, Proceedings of the Eighth International Conference on Complex Networks and Their Applications*, pages 928–940, Lisbonne, Portugal, December 2019. Springer. doi: 10.1007/978-3-030-36687-2_77. URL https://hal.archives-ouvertes.fr/hal-02344806.

[17] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[18] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender Bias in Coreference Resolution. *arXiv:1804.09301 [cs]*, April 2018. URL http://arxiv.org/abs/1804.09301. arXiv: 1804.09301.

[19] Hemal Shah, Calton Pu, and Rajesh Madukkarumukumana. High performance sockets and rpc over virtual interface (vi) architecture. pages 91–107, 01 1999. doi: 10.1007/10704826_7.

[20] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding,

Kai-Wei Chang, and William Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv:1906.08976 [cs]*, June 2019. URL http://arxiv.org/abs/1906.08976. arXiv: 1906.08976.

[21] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Contextualized Word Embeddings. *arXiv:1904.03310 [cs]*, April 2019. URL http://arxiv.org/abs/1904.03310. arXiv: 1904.03310.

VII. APPENDIX

A.1: Examples of Dataset

Sentence Pair	Label
Sentence 1: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. **Sentence 2**: Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.	1
Sentence 1: Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for $2.5 billion. **Sentence 2**: Yucaipa bought Dominick's in 1995 for $693 million and sold it to Safeway for $1.8 billion in 1998.	0

TABLE II: Examples from MSRParaphrase Corpus

Question	Answer	Label
When should I get tested?	Your doctor will tell you if you need to get tested.	1
When should I get tested?	If you have been in contact with a person with a COVID-19 infection while they were ill, you need to isolate yourself for 14 days after you last saw that person. In general, stay home as much as you can.	0
What are the rules of self-isolation?	Remember, the vast majority of Australians aren't being told to self-isolate. It only applies to those who have recently come back from overseas...	1
What are the rules of self-isolation?	Your doctor will tell you if you need to get tested.	0

TABLE III: Examples from COVID-news-Q&A

Comment	Label
Just what we need.....more white Republicans telling African American men to behave.	0
Black people appear to be the most racist and take full advantage of it. Blacklivesmatter is a joke. Twice as many white people are killed by cops than black people yet we never hear about them.	1

TABLE IV: Examples from Jigsaw Unintended Bias in Toxicity

Tweet	Label
<user> those god damn ch**ks <url>	0
sold weed to a b**ch whos gonna be a lawyer bet shell have a ni**as back	1
just a young ni**ah wit a promised future <user>	1
<user> yu forgot we beat yall thats like a garbage man calling a janitor trash	2

TABLE V: Examples from Twitter Hate Speech Corpus

A.2: Target and Attribute Template Sentences

The target words used in our experiments were "black" and "white". The attribute word lists consisted of positive and negative emotions:

- Positive: joy, great, wonderful, amazing, happy, excited, glad, ecstatic, relieved
- Negative: fear, scared, dreadful, horrible, terrified, anxious, discouraged, threatened, shocked

Target sentences used the following templates:

- This person is [TARGET].
- This is a [TARGET] person.
- A [TARGET] person is here.
- The [TARGET] person is walking.
- The [TARGET] person is brushing their teeth.
- The [TARGET] person is running late.
- The teacher was speaking to the [TARGET] student.
- I have a [TARGET] neighbor.

Attribute sentences used the following templates:

- They make me feel [ATTRIBUTE].
- They feel [ATTRIBUTE].
- The situation makes them feel [ATTRIBUTE].

A.3: SEAT Score and Effect Size

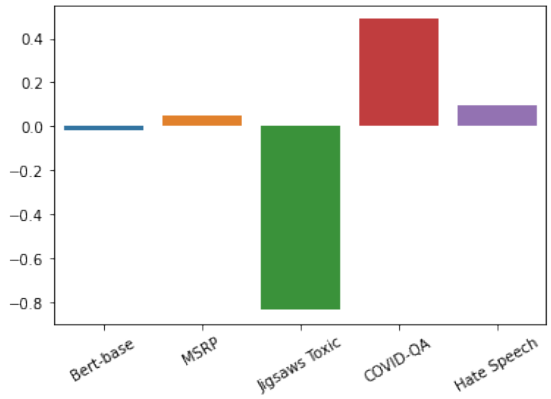

Fig. 2: This shows the raw seat scores of each fine-tines model and base BERT model

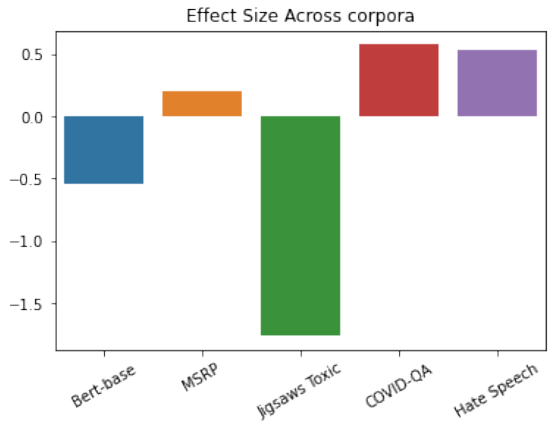

Fig. 3: Effect sizes for each task. A higher magnitude indicates more bias in the model.

A.4: Cosine Similarity Graphs

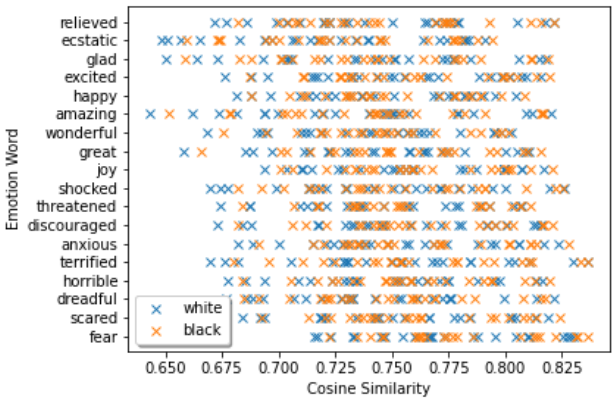

Fig. 4: Cosine similarities between embeddings of sentences with the terms "black" and "white" and those of sentences with one emotion word. These sentence embeddings were taken from the base BERT model.

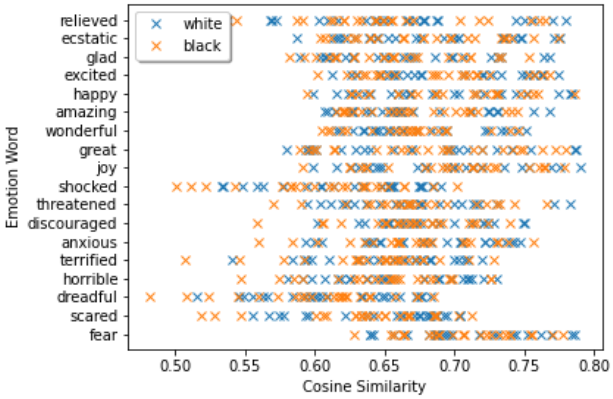

Fig. 5: Cosine similarities between embeddings of sentences with the terms "black" and "white" and those of sentences with one emotion word. These sentence embeddings were taken from the MSRPC model.

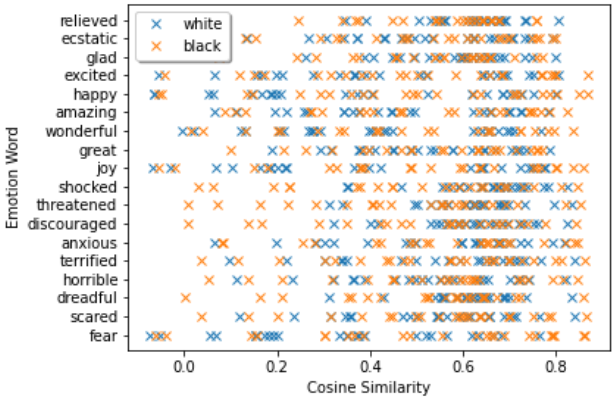

Fig. 6: Cosine similarities between embeddings of sentences with the terms "black" and "white" and those of sentences with one emotion word. These sentence embeddings were taken from the COVID-news-QuestionAnswering model.

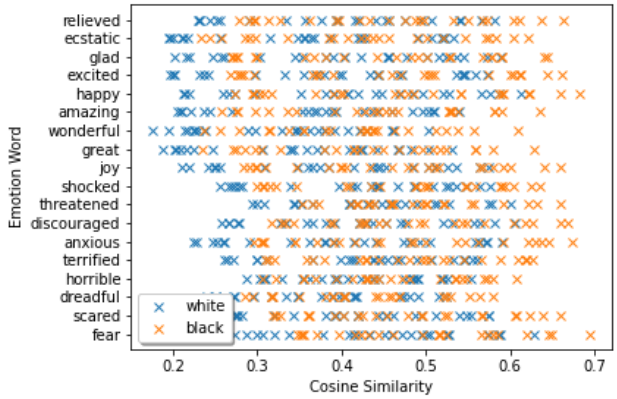

Fig. 7: Cosine similarities between embeddings of sentences with the terms "black" and "white" and those of sentences with one emotion word. These sentence embeddings were taken from the Hate Speech and Offensive Language Detection model.