RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification

Priyanka Bose[1], Chandra Shekhar Pandey[1], and Fraida Fund[1]

[1]New York University, New York, United States of America

Edited by
Koustuv Sinha,
Maurits Bleeker,
Samarth Bhargav

Received
04 February 2023

Published
20 July 2023

DOI
10.5281/zenodo.8173739

Reproducibility Summary

Scope of Reproducibility — We aim to reproduce a result from the paper "Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification" [1]. Our study is restricted specifically to the claim that the use of swear words impacts hate speech classification of AAE text. We were able to broadly validate the claim of the paper, however, the magnitude of the effect was dependent on the word replacement strategy, which was somewhat ambiguous in the original paper.

Methodology — The authors did not publish source code. Therefore, we reproduce the experiments by following the methodology described in the paper. We train BERT models from TensorFlow Hub [2] to classify hate speech using the DWMW17[3] and FDCL18[4] Twitter datasets. Then, we compile a dictionary of swear words and replacement words with comparable meaning, and we use this to create "censored" versions of samples in Blodgett et al.'s[5] AAE Twitter dataset. Using the BERT models, we evaluate the hate speech classification of the original data and the censored data. Our experiments are conducted on an open-access research testbed, Chameleon [6], and we make available both our code and instructions for reproducing the result on the shared facility.

Results — Our results are consistent with the claim that the censored text (without swear words) is less often classified as hate speech, offensive, or abusive than the same text with swear words. However, we find the classification is very sensitive to the word replacement dictionary being used.

What was easy — The authors used three well known datasets which were easy to obtain. They also used well-known widely available models, BERT and Word2Vec.

What was difficult — Some of the details of model training were not fully specified. Also, we were not able to exactly re-create a comparable dictionary of swear words and their replacement terms of similar meanings, following their methodology.

Communication with original authors — We reached out to the authors, but were not able to communicate with them before the submission of this report.

Correspondence should be addressed to Chandra Shekhar Pandey (cp3793@nyu.edu)
The authors have declared that no competing interests exist.
Code is available at https://github.com/indianspeedster/mlr_2022.git. – SWH swh:1:dir:e4878650f9ab13d7820684abaca84448cc13b9ff.
Open peer review is available at https://openreview.net/forum?id=MjZVx7a0KX-.

1 Introduction

African American English (AAE) text is more likely to be classified as hate speech or other type of toxic speech by hate speech classification systems and sentiment analysis, according to prior research [3, 7, 8, 9]. This paper [1] seeks to determine how AAE vernacular and the classification of hate speech relate to one another. The research aims to understand how specific characteristics of AAE text may correlate the classification of AAE text as offensive, abusive and hate speech, and inform future bias-mitigation strategies. The paper considers two common characteristics of AAE speech for such classification. One, the use of swear words and two, certain grammatical patterns. In our reproducibility report, we focus on reproducing the results which are related to the first, which is the use of swear words in AAE text. According to the original paper, some swear words are more common in AAE than other English dialects, and this can be a source of misclassification.

This report repeats the original paper's experiments and compares our reproduced results with the reported results. We trained two BERT classifiers [2], once each on the DWMW17 [3] and FDCL18 [4] datasets. Then, we used the trained models to determine how many of Blodgett et al.'s AAE twitter dataset [5] would be classified as hateful, offensive, or abusive. Next, we used a Word2Vec model to perform word replacement, to swap out swear words with words of similar meanings. We use the same BERT classifiers on this "censored" text, and calculate the reduction in hate speech classification.

We were able to broadly validate the claim that AAE text with censored words is less likely to be classified as hate speech than the original uncensored text. However, the magnitude of this effect is sensitive to the word replacement strategy used. The details of the author's word replacement strategy was not clearly specified and hence, we were not able to replicate the exact results of the original paper. We make all of our experiment code available for replication on Chameleon [6], which is an open access test bed. Other researchers can easily reproduce our experiment in the same environment.

2 Scope of reproducibility

The paper explores the role of grammar and word choice in bias toward AAE in hate speech classification. The authors consider two research questions (reproduced verbatim here from the original paper), and then based on their experiment results they made one claim for each research question:

- **RQ 1:** "How strongly does use of swear words or "offensive language" impact the hate speech classification of AAE text?" **Claim 1:** There is a considerable reduction in hate speech, offensive speech, and abusive speech when tweets are censored for "swear" words.
- **RQ 2:** "How do grammatical patterns of AAE tweets impact the hate speech classification of AAE text?" **Claim 2:** The classification of hate speech is independent of the four grammar subcategories that the authors examined.

In this report, we attempt to reproduce and validate only the first claim:

- **Claim 1:** There is a considerable reduction in hate speech, offensive speech, and abusive speech when tweets are censored for "swear" words.

To validate this claim, we trained a BERT classifier[2] on the DWMW17 [3] and FDCL18 [4] hate speech training datasets. Then, we generated censored and uncensored versions of a subset of the AAE text in the Blodgett AAE [5] dataset. We validate that the censored AAE text is less likely than the uncensored text to be classified as abusive, offensive, or hateful. Since there was no original code provided, all the code that is written is our own, using the description in the paper as a guideline.

3 Methodology

To replicate the experiments in the original paper, we retrieve the same datasets as used by the original authors, train comparable BERT and Word2Vec models, generate word replacement dictionaries, and then classify censored and uncensored text samples using those BERT models. Here, we elaborate more on each of those steps, including the challenges we encountered in trying to follow the authors' instructions.

3.1 Datasets

There were 3 Twitter datasets used in this paper: DWMW17 [3], FDCL18 [4], and Blodgett [5]. Because Twitter does not allow researchers to redistribute the text of the tweets, it is not always clear what data the authors used originally and if there are any missing samples in the dataset that we used for our experiment. This is a common and well-known problem with using Twitter data. The authors did not specify tweet IDs. However, they did specify the number of tweets in each case, so we were able to verify whether we had a similar number.
DWMW17 and FDCL18 are Twitter hate speech datasets that categorize tweets by various hate speech terms. These are used to train the BERT models. For DWMW17 [3], 24,783 tweets are classified as "hate speech", "offensive speech" or "neither". For FDCL18 [4], 50,487 tweets are classified as "abusive", "hate", "spam" or "normal". For these datasets, we found sources online including 24,783 tweets and 99,996 tweets, respectively, which is consistent with the original for DWMW17 but not for FDCL18.
The Blodgett dataset [5] is used to evaluate the impact of swear word replacement, by comparing the number of censored and uncensored AAE tweets classified as "abusive" or "hateful", "offensive" or "hate" . This dataset is also used to train a Word2Vec model to generate a swear word replacement dictionary. In the original paper, the authors say that 50,000 tweets in this dataset had high likelihood (.9 or above) of being AAE. However, we found 548,516 tweets in the data with high likelihood of being AAE. It is not clear which 50,000 tweets were used in the original paper. For the evaluation, we used the first 50,000 tweets that were available via the Twitter API. The tweet IDs we used are shared in the supplementary materials.

3.2 Models

The original paper uses two types of models:

- **Offensive, abusive, or hateful speech classification**: the authors train two BERT models, one on the DWMW17 [3] dataset and one on the FDCL18 [4] dataset, to classify text.
- **Word replacement**: the authors train a Word2Vec model on AAE texts in the Blodgett [5] dataset, to find replacement words for each swear word.

The authors did not specify hyperparameters or other specific details related to model training. Since our primary goal is to validate the broad claim, and not necessarily precise numeric results, we did not do an extensive hyperparameter search.
To train the BERT model, we used a BERT classifier from the Tensorflow Hub, specifically bert_multi_cased_L-12_H-768_A-12 TF2.0 Saved Model (v4) [2], and followed the process described in this article [10] to train it on each of the two datasets[3, 4]. In the paper [1], the authors mention that "First we split the dataset appropriately and train a BERT classifier on the data. Then we test the model and use the best performing model". Since more specific instructions were not given, we used a split strategy of 70-30 for training and testing on the DWMW17[3] and FDCL18[4] datasets, with a further 80-20 split of the training data into training and validation. The seed for the random split is given in

the supplementary materials. We used the AdamW optimizer with an initial learning rate of 3e-5, and a decay schedule specified in the code in the supplementary materials. We trained each model for 30 epochs, achieving a training accuracy of 98.5% and 97% accuracy and validation accuracy of 90.3% and 91.2% for DWMW17 [3] and FDCL18 [4], respectively.
For the word replacement model, the authors say they train a Word2Vec model on AAE tweets from the Blodgett dataset. We used AAE tweets from the same dataset (the tweet IDs are given in the supplementary materials) and trained a word2vec-google-news-300 model from Gensim[11] (we used `window=10`, `min_count=2`, `workers=4`, and left other settings at their default values).

3.3 Word replacement

For a list of swear words to replace, the authors report using LIWC2007 and a hand-curated list of swear words. However, they did not provide their curated list of additional words. We retrieved an LIWC2007 swear word list from the dataset associated with [12], and used only that list.
Next, they used the Word2Vec model trained on Blodgett AAE tweets to create a dictionary of replacements for the swear words, replacing each swear word with the word that is closest according to cosine similarity. We followed a similar approach. However, sometimes, we observed that the replacement words were also swear words, or contained a swear word. The authors of the original paper [1] did not specify how they address this. In our implementation, we iteratively replaced swear words until the replacement was no longer a swear word and did not contain any swear word.
The authors claim that, on manual inspection of a set of 50 tweets, 78% were successfully reworded (i.e. "removing all swear words and having identical meaning to the original tweet"). However, it is not clear what qualifies as "identical meaning". In our experiment, we observed words in the replacement dictionaries that made no sense, words noted as swear words that weren't actually swear words, and replacement words that were sometimes just the asterisked versions of the words.
We also considered two other word replacement strategies, in addition to the strategy described in the original paper:

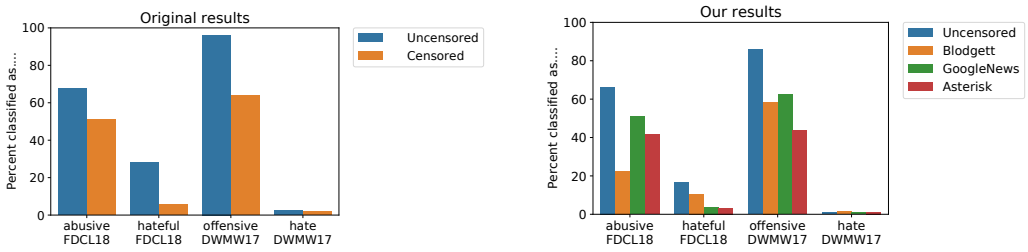
- Standard Word2Vec on Google News [11]: We used a Word2Vec model trained on the Google News 300 dictionary to get the censored tweets and ran our BERT classifier on these new tweets. We thought that this model might find replacement words that were closer in meaning to the original.
- We created our own asterisk dictionary, where the replacement word for the swear word was a word of same length with all its characters as asterisks.

3.4 Classifying censored and uncensored text

After creating the swear word replacement dictionary, the dictionary is used to create a censored version of each of the tweets in the Blodgett dataset. Then, the BERT models trained earlier are used to classify the censored text.

3.5 Computational requirements

We used Chameleon [6] to provision all our resources required for the experiments. Chameleon [6] is an experimental platform for systems research in the computer sciences. On Chameleon, users have bare metal servers and complete control over the software stack, including root rights, kernel customisation, and console access. We ran our experiments using the RTX 6000 GPU nodes. The training of our model took approximately 12h using the original approach. For our entire experiment, approximately 7610



(a) Original results reported in [1]. **(b)** Our results.

Figure 1. Original results reported in [1], and our results after reproducing their experiments.

service units were used up. One Service Unit (SU) on Chameleon is equivalent to one hour of usage of one allocatable resource (physical host, network segment, or floating IP). We have included instructions to provision resources on Chameleon in the supplementary materials so that our work can be reproduced easily.

4 Results

We were able to reproduce the general claim that there is a significant reduction in hate speech classification of AAE tweets when the swear words are censored. However, the magnitude of the effect was different in our case, compared to what was mentioned in the paper. Fig. 1a shows the summary results of the original paper, in terms of the percentage of hate speech classification when the speech is uncensored versus when it is censored. Our reproduced results in Fig. 1b show the percentages for the uncensored text and for censored text using each of the three word replacement strategies we considered.

4.1 Results reproducing original paper

Table 1 contains the results of our experiment while Table 2 contains the results of the original paper. We observe that our result broadly validate the claim of the paper that there is a notable decrease in hate speech classification of AAE text when swear words are censored. However, the magnitude of the decrease is different than the original.

Table 1. Results from our experiment using Word2Vec model trained on Blodgett

	DWMW17 - Hate	FDCL18 - Hate	DWMW17 - Offensive	FDCL18 - Abusive
Original Sample	0.95	16.52	85.86	66.22
Edited Sample (Blodgett)	1.00	10.60	56.64	22.18
Difference (Original v/s Blodgett)	0.05	-5.92	-29.22	-44.04

Table 2. Results from the original paper

	DWMW17 - Hate	FDCL18 - Hate	DWMW17 - Offensive	FDCL18 - Abusive
Original Sample	2.46	27.89	96.18	67.77
Edited Sample (Asterisk)	1.93	5.496	64.07	51.04
Difference (Original v/s Asterisk)	-0.53	-22.394	-32.11	-16.73

4.2 Results beyond original paper

In addition to the word replacement strategy with the Word2Vec model trained on the Blodgett dataset as specified in the paper, we also considered two other strategies:

1. Swear word replacement using a Word2Vec model on Google News 300, without further training on AAE text.
2. Replacing swear words with words of same length but all characters as asterisks.

Additional Result 1 — Using the Google News 300 Word2Vec model without further training on AAE text, we obtained the results as mentioned in Table 3.

Table 3. Results for the Google News 300 Word2Vec without training on AAE text

	DWMW17 - Hate	FDCL18 - Hate	DWMW17 - Offensive	FDCL18 - Abusive
Original Sample	0.95	16.52	85.86	66.22
Edited Sample (Google)	0.82	3.81	62.36	51.00
Difference (Original v/s Google)	-0.13	-12.71	-23.5	-15.22

With this word replacement strategy, our results are more similar in magnitude to the original results in Table 2.

Additional Result 2 — With the asterisk word replacement strategy, we obtained the results given in Table 4.

Table 4. Results for the Asterisks word replacements

	DWMW17 - Hate	FDCL18 - Hate	DWMW17 - Offensive	FDCL18 - Abusive
Original Sample	0.95	16.52	85.86	66.22
Edited Sample (Asterisk)	1.066	3.27	43.63	41.611
Difference (Original v/s Asterisk)	0.116	-13.25	-42.23	-24.609

These results are also more similar in magnitude to the original results in Table 2, than our results for the experiment that tried to reproduce the original methodology as closely as possible (Table 1).

Swear Word Replacement Examples — We also share examples of certain swear words and their replacement words with similar meanings according to the different dictionaries in Table 5.
For our Word2Vec model trained on the Blodgett dataset, we found certain words were replaced with other words that did not have similar meanings, or that were themselves offensive. For example, the replacement word for “bitches” was replaced with the word

Table 5. Comparison of replacement words as generated by the different dictionaries

Word	Replacement word according to Blodgett Dictionary	Replacement word according to Google News 300 Dictionary	Replacement word according to Asterisk Dictionary
ass	ahh	butt	***
nigga	boy	boy	*****
bitches	hoes	girls	*******
hell	usual	h_*	****
fucker	hismain_concerned	f_**ker	******
shit	shyt	sh_*_t	****
WTF	<Not available in Blodgett>	OMFG	***
PISSED	<Not available in Blodgett>	DAMMIT	******
dick	neck	d_*_ck	****

"hoes" which in itself is an offensive word. The pre-trained Word2Vec model on the Google News 300 [11] has a larger vocabulary, which improved the results somewhat. Taking the last example again, the word "bitches" here is replaced with the word "girls" which is of comparable meaning and is not a swear word itself. The asterisk replacement strategy is also shown, for completeness.

5 Discussion

To summarise, our experiments do validate the high-level claim made in the paper[1] regarding their research question 1: How strongly does use of swear words or "offensive language" impact the hate speech classification of AAE text? We observe that the results are qualitatively similar to what has been said in the paper. However, we find that the magnitude of the effect is highly dependent on the details of the word replacement strategy, which was somewhat ambiguous in the original paper. We feel that the experiment could have been replicated better if we had access to the actual word replacement dictionary that the authors used for their experiment.
Also, due to the nature of Twitter data, there is a high possibility that some tweets from the original dataset are no longer available. This might also be a reason in some of the minor differences in the results we observed.

5.1 What was easy

The datasets used by the authors in the paper are easily obtainable. Having the number of samples they used at each step of the experiment was a good reference point for us to follow while validating the claim. The steps required for each experiment were well documented. Finally, they used models that are widely available.

5.2 What was difficult

The most challenging part about reproduction of the paper was that the details of some experiments were not fully specified. There were certain parts of the paper that were ambiguous, and we have listed the specifics below:

Datasets: The authors acknowledge that many tweets in the original dataset that could not be accessed by them. This is a well known issue with Twitter datasets and we had similar issues - for two of the three datasets, the size of the data that we compiled was different from the size reported by the original authors. Moreover, the authors state that they split the data "appropriately", however no clarity is provided around how they actually split the data into training and testing sets.

Models: While easily available models were used, there was still a lack of clarify around how exactly the experiments were run by the authors. For instance, they specified that they test the model and choose the best performing model. However, there is no clarification provided around the criteria for choosing the "best performing model" or how this was done.

Word Replacement Dictionary: Details were also missing in the discussion of how swear words were mapped to replacement words. For example, the authors do not specify what should happen if the replacement word is also a swear word or a variation of a swear word, which happened often in our observation.

5.3 Communication with original authors

The original code was not published by the authors and we were not able to communicate with the authors before publication time.

References

1. C. Harris, M. Halevy, A. Howard, A. Bruckman, and D. Yang. "Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification." In: **2022 ACM Conference on Fairness, Accountability, and Transparency**. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 789–798. DOI: 10.1145/3531146.3533144. URL: https://doi.org/10.1145/3531146.3533144.
2. TensorFlowHub. **bert_en_uncased_L-12_H-768_A-12 TF2.0 Saved Model (v4)**. https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4.
3. T. Davidson, D. Bhattacharya, and I. Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In: **Proceedings of the Third Workshop on Abusive Language Online**. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–35. DOI: 10.18653/v1/W19-3504. URL: https://aclanthology.org/W19-3504.
4. A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." In: **11th International Conference on Web and Social Media, ICWSM 2018**. AAAI Press. 2018.
5. S. L. Blodgett, J. Wei, and B. O'Connor. "Twitter Universal Dependency Parsing for African-American and Mainstream American English." In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1415–1425. DOI: 10.18653/v1/P18-1131. URL: https://aclanthology.org/P18-1131.
6. K. Keahey, J. Anderson, Z. Zhen, P. Riteau, P. Ruth, D. Stanzione, M. Cevik, J. Colleran, H. S. Gunawi, C. Hammock, et al. "Lessons learned from the chameleon testbed." In: **Proceedings of the 2020 USENIX Conference on Usenix Annual Technical Conference**. 2020, pp. 219–233.
7. S. Kiritchenko and S. Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." In: **Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics**. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 43–53. DOI: 10.18653/v1/S18-2005. URL: https://aclanthology.org/S18-2005.
8. M. Halevy, C. Harris, A. Bruckman, D. Yang, and A. Howard. "Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework." In: **Equity and Access in Algorithms, Mechanisms, and Optimization**. EAAMO '21. NY, USA: Association for Computing Machinery, 2021. DOI: 10.1145/3465416.3483299. URL: https://doi.org/10.1145/3465416.3483299.
9. M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. "The Risk of Racial Bias in Hate Speech Detection." In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1668–1678. DOI: 10.18653/v1/P19-1163. URL: https://aclanthology.org/P19-1163.
10. TensorFlow. **Classify text with Bert**. https://www.tensorflow.org/text/tutorials/classify_text_with_bert.
11. R. Rehurek and P. Sojka. "Gensim–Python framework for vector space modelling." In: **NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic** 3.2 (2011), p. 2.
12. K. H. Kwon and A. Gruzd. "Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos." In: **Internet Research** 27.4 (2017), pp. 991–1010.