Leveraging Bias in Pre-Trained Word Embeddings for Unsupervised Microaggression Detection

Tolúlọpẹ́ Ògúnrèmí[1], **Nazanin Sabri**[2], **Valerio Basile**[3], **Tommaso Caselli**[4]
1. Stanford University, United States, tolulope@stanford.edu
2. Independent Researcher nazanin.sabrii@gmail.com
3. University of Turin, Italy, valerio.basile@unito.it
4. University of Groningen, Netherlands, t.caselli@rug.nl

Abstract

Microaggressions are subtle manifestations of bias (Breitfeller et al., 2019). These demonstrations of bias can often be classified as a subset of abusive language. However, not as much focus has been placed on the recognition of these instances. As a result, limited data is available on the topic, and only in English. Being able to detect microaggressions without the need for labeled data would be advantageous since it would allow content moderation also for languages lacking annotated data. In this study, we introduce an unsupervised method to detect microaggressions in natural language expressions. The algorithm relies on pre-trained word-embeddings, leveraging the bias encoded in the model in order to detect microaggressions in unseen textual instances. We test the method on a dataset of racial and gender-based microaggressions, reporting promising results. We further run the algorithm on out-of-domain unseen data with the purpose of bootstrapping corpora of microaggressions "in the wild", and discuss the benefits and drawbacks of our proposed method.

1 Introduction

The growth of Social Media platforms has been accompanied by an increased visibility of expressions of socially unacceptable language online. In a 2016 Eurobarometer survey, 75% of people who follow or participate in online discussions have witnessed or experienced abuse or hate speech. With this umbrella term, different phenomena can be identified ranging from offensive language to more complex and dangerous ones, such as hate speech or doxing. Recently, there has been a growing interest by the Natural Language Processing community in the development of language resources and systems to counteract socially unacceptable language online. Most previous work has focused on few, easy to model phenomena, ignoring more subtle and complex ones, such as microaggressions (Jurgens et al., 2019).

Microaggressions are brief, everyday exchanges that denigrate stigmatised and culturally marginalised groups (Merriam-Webster, 2021). They are not always perceived as hurtful by either party, and they can often be detected as positive statements by current hate-speech detection systems (Breitfeller et al., 2019). The occasionally unintentional hurt caused by such comments is a reflection of how certain stereotypes of others are baked into society. Sue et al. (2007) define microaggressions in the racial context, particularly when directed toward people of color, as "brief and commonplace daily verbal, behavioral, or environmental indignities", such as: "you are a credit to your race." (intended message: it is unusual for someone of your race to be intelligent) or "do you think you're ready for college?" (indented message: it is unusual for people of color to succeed). The need for moderation of hateful content has previously been explored. For instance, Mathew et al. (2019b) analyses the temporal effects of allowing hate speech on Gab, and finds that the language of users tends to become more and more similar to that of hateful users over time. Mathew et al. (2019a) further highlights that the spreading speed and reach of hateful content is much higher than with the non-hateful content. As a result, being able to remove instances of hateful language, such as microaggressions, is of great importance.

Previous work on microaggressions with com-

putational methods is quite recent. Breitfeller et al. (2019) is one of the first work to address microaggressions in a systematic way, also introducing a first dataset, SelfMA. A further contribution specifically focused on racial microaggression is Ali et al. (2020), where the authors focus on the development of machine learning systems.

In this study we introduce an unsupervised method for microaggression detection. Our method utilizes the existing bias in word-embeddings to detect words with biased connotations in the message. Although unsupervised approaches tend to be less competitive than their supervised counterparts, our method is language-independent and thus it can be applied to any language for which embedding representations exist. Furthermore, the reliance of our methods on specific lexical items and their context of occurrence makes transparent the flagging of a message as an instance of a microaggression. In addition to the usefulness of our method in languages with no labeled data, the reliance of our model on words in the sentences would make it interpretable as it allow human moderators to understand what the system has based its decision on.

Our contributions can be summarised as follows:

- we introduce a **new unsupervised method** for the detection of microaggressions which builds on top of pre-trained word embeddings;
- we **compare the performance** of our model using different pre-trained word embeddings (Glove, FastText, and Word2Vec) and discuss the potential reasons behind the differences;
- we **test** the proposed algorithm **on unseen data from a different domain** (i.e., Twitter), in order to qualitatively evaluate its efficacy in discovering new instances of microaggression.

The rest of this paper is structured as follows: we introduce our method in Section 2. The data and our results are reported in Section 3. We deploy our model and discuss its limitations in Section 4. Finally, we present the conclusion and future work in Section 5.

2 Use the Bias Against the Bias

Embedded representations, either from pre-trained word embeddings or pre-trained language models, have been shown to contain and amplify the biases present in the data used to generate them (Bolukbasi et al., 2016; Lauscher and Glavaš, 2019; Bhardwaj et al., 2020). As such, they often exhibit gender and racial bias (Swinger et al., 2019). Many studies have attempted to reduce this bias (Yang and Feng, 2020; Zhao et al., 2018; Manzini et al., 2019). In this work, we take a different turn by using this bias to our advantage: rather than taming the hurtfulness of the representations (Schick et al., 2021), we actively use it to promote social good. In this first study, we employ word representations derived from generic textual corpora of English, in order to capture the background knowledge needed to disambiguate instances of microaggressions in the text. Recently, however, there have been studies involving word representations created from tailored collections of social media content aimed at capturing abusive phenomena like verbal aggression (Dynel, 2021) and hate speech (Caselli et al., 2020).

We devise a simple and effective method that exploits existing bias in word embeddings and identify words in a message that are related to particular and distant semantic areas in the embedding space. Messages are analysed in three steps: first, for each token t^i we compute its relatedness to a list of manually curated seed words $s = s_1, ..., s_n$ denoting potential targets of microaggressions; second, we consider only the similarities of the pairs (t_i, s_j) above an empirical *similarity threshold* ST and compute their variance v_i; finally, we classify the token t_i as a micro aggression trigger, and consequently the message as a micro aggression, if the v_i is above an empirically determined *variance threshold* VT.

The intuitive idea behind this algorithm is that some lexical elements in a verbal microaggression are often (yet sometimes subtly) hinting at specific features of the recipient of the message, in an otherwise neutral lexical context.

In this work, we choose to focus on microaggressions related to race and gender, therefore the seed words have to be chosen accordingly. The seed word lists for race and gender are, respectively, *[white, black, asian, latino, hispanic, arab, african, caucasian]* and *[girl, boy, man, woman, male, female]* for gender. There is also a practical reasons to focus on gender and race, namely the scarcity of data available for other categories of microaggression and other idiosincrasies of the

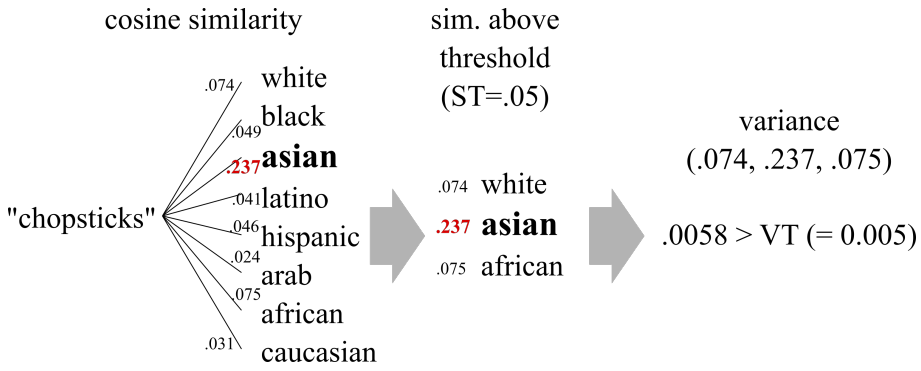

Figure 1: Worked example of unsupervised method for word "chopsticks" in the message *"Ford: Built With Tools, Not With Chopsticks"*

available datasets — the religion class was specific to different religions, therefore hard to generalise, sexuality and gender presented a large overlap, and so on.

An example of how the proposed method works is illustrated in Figure 1. In the example, consider the word "chopsticks" in the message *"Ford: Built With Tools, Not With Chopsticks"* (from the SelfMA dataset, described in Section 3). The target word exhibits a much higher relatedness to the word *asian* (0.237) than any other seed words. Even just considering the seed words with a similarity above a fixed threshold (*white*, *asian* and, *african*), the variance of their similarity score with respect to *chopsticks* is still higher than the variance threshold, and therefore this target word, in this context, triggers a microaggression according to the algorithm. This process is repeated for all the words in the message in order to detect microaggressions. Some categories of words are bound to exhibit a high relatedness to all the seed words, e.g., "people" or "human". This is the reason to introduce the variance threshold in the final step of our algorithm, to filter out these cases when classifying a given message, and instead focus on words that are related to different races (or genders) unevenly, with a skewed distribution of similarity scores.

An important by-product of this algorithm is that the output is one or more trigger words, in addition to the microaggression label — in the example, the trigger word is indeed *chopsticks* — therefore enabling a more informative and interpretable decision process.

Source	Number of posts
SelfMA Gender	1,314
SelfMA Racial	1,278
Tumblr	2,021

Table 1: Statistics of the two subsets of the SelfMA dataset used in this paper, and the extra data downloaded to balance the dataset.

3 Experiments

To test our method, we use two subsets of the *SelfMA: microaggressions.com* dataset (Breitfeller et al., 2019), comprised of 1,314 and 1,278 Tumblr posts respectively[1]. The posts in SelfMA are all instances of microaggressions, manually tagged with one of four categories: race, gender, sexuality and religion. These posts can be tagged with more than one form of microaggressions, meaning certain instances can appear in both subsets of race and gender used for the purposes of this study. The dataset consists of first and second hand accounts of microaggressions, as well as direct quotes of phrases or sentences said to the person posting. In order to reduce linguistic perturbation introduced by accounts of a situation, we only take direct quotes found in the dataset as instances of microaggressions that we can detect with our unsupervised method. For training, we pull out direct quotes from the gender (561) and racial (519) dataset to test the algorithm. In order to balance the dataset, we scraped 2,021 random Tumblr posts, for a total of 4,612 instances. Table 1 summarises the composition of our dataset.

It is important to note that a microaggression can have multiple tags, so there is an overlap of

[1]Tumblr is a popular American microblogging platform https://www.tumblr.com

instances. However, the seed words used to detect microaggression types in the method are different for each target phenomenon (e.g., race, gender).

We ran the algorithm on the *SelfMA* dataset, empirically optimising the two thresholds on the training split, for each word embedding type and each microaggression category, filtering by the seed words listed in Section 2. We test the algorithm with three pre-trained word embedding models for English, namely `FastText` (Joulin et al., 2016) (trained on Wikipedia and Common Crawl), `word2vec` (Mikolov et al., 2013) (trained on Google News), and `GloVe` (Pennington et al., 2014) (trained on Wikipedia, GigaWord corpus, and Common Crawl). The optimization is performed by exhaustive grid search over the hyperparamter space.

The results, shown in Table 2, indicate that `FastText` has a better F1 score on Racial microaggressions while `word2vec` performs better on Gender microaggressions. The difference in performance between `FastText` and `word2vec` is not major, and we attribute this to the difference between the corpora on which the two models were trained (i.e., web crawl and Wikipedia for `FastText` vs. news data for `word2vec`). The `GloVe` pretrained model, trained on a combination of newswire texts, encyclopedic entries and texts from the Web, underperforms in both experiments. In general, the absolute figures are encouraging, especially considering the simplicity of this unsupervised approach.

4 Discovering Microaggressions

To better understand the performance of our unsupervised model, we performed an additional experiment. Our goal is to understand the false positive results and the potential harm the model could cause. To do so, we use our unsupervised model to label unseen instances from another domain (Twitter) than the SelfMA dataset (Tumblr) in order to see how the model would perform in detecting microaggressions.

We begin by performing keyword searches on Twitter (using Twitter's official API) and collect a new dataset of of 3M tweets with seven keywords potentially containing race and gender expressions.Next, we set the threshold values ST and VT in our model in order to obtain the highest Precision scores, rather than the highest F1 value. This step is performed exactly like the optimization described in Section 2 with the only difference of the target metric. The aim of this step is to only label tweets as microaggressions with the highest possible degree of confidence. We set $ST = 0.12$ and $VT = 0.014$ for racial microaggressions leading to Precision of .931 and $ST = 0.13$ and $VT = 0.019$ for gender-based microaggressions leading to a Precision of .912. Precision has been measured on the original SelfMA dataset used as a validation set.

We then run the unsupervised model on the new Twitter dataset by automatically labelling 256,843 tweets for gender and 373,631 tweets for race. After the data is labeled, we manually explore the positive instances in order to evaluate the performance of the model. The algorithm tuned for high precision found in this dataset 6,306 gender-related microaggression candidates, 13,004 race-related microaggression candidates.

We find that while the model does detect actual instances of microaggression, there is a noticeable amount of false positive instances. These tweets discuss race or gender in some manner. However, they do not necessarily contain microaggressions towards these groups. While the model does learn to detect discussions of these topics, it seems to sometimes confuse these discussions with microaggressions towards the aforementioned groups. Some examples follow, paraphrased to avoid tracking the original messages.

> *Saying "Arrested Development isn't funny" in an office full of women just to feel something*

> *"Men have moustaches, women have oversized bracelets"*

The humorous attempts in this tweets hinge on gender stereotypes, and therefore in some contexts it could be perceived as offensive by some recipients. The high relatedness in the word embedding space between some words (moustaches and bracelets) and gender-related seed words (men and women) triggers the detection algorithm.

The automatic detection of racial microaggressions "in the wild" is more challenging than gender-based ones, according to our manual exploration of this automatically labeled dataset. This may be due to the difficulty of crafting a list of seed words that is sufficiently race-related, but at the same time avoids generating too many false positives. We indeed found many of them,

Target	Model	Class	Precision	Recall	F1-Score
Gender	FastText	not-MA	.609	.746	.671
		MA	.714	.570	.634
		macro avg.			.680
	GloVe	not-MA	.692	.380	.491
		MA	.603	.848	.705
		macro avg.			.598
	word2vec	not-MA	.659	.789	.718
		MA	.769	.634	.694
		macro avg.			**.706**
Race	FastText	not-MA	.659	.875	.654
		MA	.814	.547	.752
		macro avg.			**.702**
	GloVe	not-MA	.765	.371	.500
		MA	.611	.896	.726
		macro avg.			.613
	word2vec	not-MA	.640	.814	.747
		MA	.776	.584	.667
		macro avg.			.692

Table 2: Results of the experiment on the Gender and Racial subset of SelfMA, in terms of Precition (P), Recall (R), and F1-score (F1) on the positive class (MA), on the negative class (not-MA), and their macro-average. Best scores per microagression category are in bold.

mainly due to named entities and multi-word expressions such as "White House", or simply because of the polysemy of color words, e.g. "black" and "white". We, however, still found instances of messages containing different extent of racial stereotyping.

> *"why are you being so dramatic? just say I'm not originally arab, you don't have to fight about it"*

> *"I will need to explain that to the chinese old lady who works at my school's administrative office"*

In summary, running the unsupervised microaggression detection algorithm on unseen data seems to represent a promising intermediate step towards the semi-automatic creation of language resources for this phenomenon. While the accuracy is not ideal, and lists of seed words have to be handcrafted carefully in order to avoid false positives, these drawbacks are balanced by the fairly cheap computational cost and the ease of application in a multilingual scenario.

5 Conclusion and Future Work

In this paper we introduce a novel algorithm that exploits the existing bias in pre-trained word embeddings to detect subtly abusive language phenomena such as microagressions. While supervised methods of detection in the field of natural language processing are plentiful, these methods are only viable for languages and topics with available labeled datasets. That is however not the case for many languages. As a result, the unsupervised method of detection introduced in this study could help address the need for the moderation of microaggressions in languages other than English. This is further helped by the availability of multilingual word-embeddings as they would allow the method to be used in any of the languages supported by the embedding.

The method is unsupervised and only needs a small list of seed words. Considering its simplicity, the results obtained from an experiment on a dataset of manually annotated microaggressions are very promising. Further, the method is transparent, explicitly identifying the words triggering a microaggression, and thus paving the way for explainable microaggression detection.

Although the preliminary results are promising, an experiment on unseen data from a different domain shows that there is leeway for improvement. Given that we are looking at the explicit words used in each message, our method is not sensitive

to implicit expressions like "you people" or "your kind", often occurring in microaggressions. We would have to add further steps to our algorithm to catch expressions like these.

Polysemy is another known issue, e.g., in words like "black" and "white" whose relatedness to certain identified trigger words could not necessarily be due to race. While a careful composition of the seed word lists helps to minimize this issue, a systematic approach to polysemy would certainly be desirable. The seed word list may also be expanded, either manually or exploiting existing lexicons such as HurtLex (Bassignana et al., 2018) for offensive terms (including stereotypes for several categories of individuals) or specialized lists of identity-related terms[2].

In future work, we plan on improving our model to account for lexical ambiguity, and the complexity derived from the interference between pragmatic phenomena and aggression, e.g., in humorous and ironic messages, following the intuition in recent literature (Frenda, 2018) about the interconnection between irony or sarcasm and abusive language online. Our current plan is to apply the algorithm presented in this paper to bootstrap the creation of a multilingual resource of online verbal microaggressions and release it to the research community.

Acknowledgements

This work of Valerio Basile is partially funded by the project "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call).

References

Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. 2020. Automated detection of racial microaggressions using machine learning. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2477–2484. IEEE.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2020. Investigating gender bias in bert. *arXiv preprint arXiv:2009.05021*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. HateBERT: Retraining BERT for Abusive Language Detection in English. *arXiv preprint arXiv:2010.12472*.

Marta Dynel. 2021. Humour and (mock) aggression: Distinguishing cyberbullying from roasting. *Language & Communication*, 81:17–36.

Simona Frenda. 2018. The role of sarcasm in hate speech. a multilingual perspective. In *e Doctoral Symposium of the XXXIVInternational Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 13–17. Lloret, E.; Saquete, E.; Martínez-Barco, P.; Moreno, I.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy, July. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. *arXiv preprint arXiv:1904.11783*.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019a. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh

[2]See for instance this compendium of LGBTQIA+ terminology: https://www.umass.edu/stonewall/sites/default/files/documents/allyship_term_handout.pdf

Mukherjee. 2019b. Temporal effects of unmoderated hate speech in gab. *arXiv preprint arXiv:1909.10966*.

Merriam-Webster. 2021. Merriam-webster's definition of microaggression. https://www.merriam-webster.com/dictionary/microaggression. Accessed: 2021-03-08.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Jeffrey Pennington, R. Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *arXiv preprint arXiv:2103.00453*.

Derald Sue, Christina Capodilupo, Gina Torino, Jennifer Bucceri, be Aisha, Kevin Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: Implications for clinical practice. *The American psychologist*, 62:271–86, 05.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.

Zekun Yang and Juan Feng. 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9434–9441.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.