Exploring Techniques to Detect and Mitigate Non-Inclusive Language Bias in Marketing Communications using a Dictionary-Based Approach

Bharathi Raja Chakravarthi[1], **Prasanna Kumar Kumaresan**[1],
Rahul Ponnusamy [1], **John P McCrae**[1],
Michaela Comerford[2], **Jay Megaro**[2], **Deniz Keles**[2], **Last Feremenga**[2]
[1] Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
[2]FMR LLC, Boston, USA,
{bharathi.raja,prasanna.kumaresan,john.mccrae}@insight-centre.org
{michaela.comerford,jay.megaro,deniz.keles,last.feremennga}@fmr.com

Abstract

We propose a new dataset for detecting non-inclusive language in sentences in English. These sentences were gathered from public sites, explaining what is inclusive and what is non-inclusive. We also extracted potentially non-inclusive keywords/phrases from the guidelines from business websites. A phrase dictionary was created by using an automatic extension with a word embedding trained on a massive corpus of general English text. In the end, a phrase dictionary was constructed by hand-editing the previous one to exclude inappropriate expansions and add the keywords from the guidelines. In a business context, the words individuals use can significantly impact the culture of inclusion and the quality of interactions with clients and prospects. Knowing the right words to avoid helps customers of different backgrounds and historically excluded groups feel included. They can make it easier to have productive, engaging, and positive communications. You can find the dictionaries, the code, and the method for making requests for the corpus at (we will release the link for data and code once the paper is accepted).

1 Introduction

Language evolves, and appropriate terminology changes as culture and society shift. Using inclusive language fosters a culture of inclusion and belonging, helps to create an environment where people of all experiences and backgrounds feel welcome, and reduces negative stereotypes[1]. It supports a customer-centric approach by assisting firms in recognizing and connecting with internal and external customers with the utmost respect and kindness.

Language has the potential to divide people and in academia, industry, and other communities, this has become intensely evident Blodgett et al. (2020). Some firmly identify with a conventional idea of gender bias, while others take a broader approach, focusing on principles of inclusivity of all bodies and genders Cao and Daumé III (2020); Lauscher et al. (2022). There's a lot to be gained from taking an aerial view, one that examines the worth of all points of view, as well as the potential harm and missed opportunities that result from a lack of regard for or value for difference. Inclusive language takes into account not only gender, but also age, race, ethnicity, culture, sexual orientation, disability, and health status Lauring and Klitmøller (2017).

We may unintentionally exclude or offend others if we lack information about and sensitivity to certain words or phrases. Being aware and mindful of our written and oral communications can help create and nurture a supportive and inclusive environment. A few main areas of preferred language and terminology include race and ethnicity, people with disabilities, gender identity, and idioms. Organizations can use preferred language and avoid non-inclusive language as a helpful tool to respond to societal shifts and deliver better products and solutions.

The work of manually reviewing the use of non-inclusive language in the material that universities, industries, and the public administration generate is too time-consuming for the equality offices that are housed inside these institutions. Natural Language Processing (NLP) technologies offer a promising way to solve the problem of non-inclusive language, saving businesses time and making inclusive language the norm in business settings. But these systems often reflect the same behaviors that businesses are trying to change through diversity and inclusion efforts Bordia and Bowman

[1] https://www.fidelity.com/about-fidelity/our-company/diversityandinclusion

Proceedings of Recent Advances in Natural Language Processing, pages 918–925
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_099

(2019); Nadeem et al. (2021); Kaneko et al. (2022); Chakravarthi (2023). On the other hand, the knowledge base shows how organized information can be used along with unorganized data.

Using techniques from NLP, we created a phrase dictionary and test sentences to automate the detection of non-inclusive sentences. The approach is intended to be applied to documents written in English.

Our work makes the following contributions:

1. We introduce an annotation scheme for labeling sentences into inclusive or non-inclusive. We create labeled data for test data.
2. We create and release the non-inclusive phrase dictionary in gender bias, age bias, disability bias, and other biases.
3. We demonstrate the ability of our non-inclusive phrase dictionary on our newly created non-inclusive data.

The best-performing model utilized the dictionary and GloVe Pennington et al. (2014) and scored a weighted F1-score of 0.62 for the binary class on a test set consisting of English sentences. The performance of the model was improved as a result of the automatic extension of the phrase dictionary. The fact that the coverage of extended dictionaries did, in fact, increase shows that the words that were automatically added to the corpus improved performance. Examples from the dataset for both binary and fine-grained labels are depicted in Figure 1.

2 Related Work

Formal theories of inclusive language have been asserted as an essential objective for the future development of society, yet there needs to be concrete guidance for their implementation. In the domains of NLP and machine learning, empirical studies have provided evidence for techniques that are effective in recognizing and minimizing the presence of bias, vagueness, and exclusion in datasets and models. Moreover, there needs to be more literature on the practical application of these methods within downstream applications Dinan et al. (2020).

While there are several works in NLP on gender inclusion Lauscher et al. (2022) and gender bias Bolukbasi et al. (2016); Bordia and Bowman (2019); Kaneko et al. (2022), more research is needed. Rudinger et al. (2018) introduce Winogender schemas and assess rule-based, statistical, and neural coreference resolution algorithms. They discover that the professional forecasts of these algorithms greatly favor one gender over the other. Bolukbasi et al. (2016) presented a strategy to eliminate gender prejudice by analyzing the degree to which words are gendered based on the extent to which they point in a particular gender direction. WEAT, which stands for ”the association between two sets of target words and two sets of attribute words,” was a metric that was established by Caliskan et al. (2017) in order to quantify the bias that exists between attributes and targets.

Blodgett et al. (2020) surveyed 146 papers analyzing different kinds of bias in NLP systems. In their study, it was discovered that (a) most work objectives are frequently imprecise, inconsistent, and devoid of normative reasoning, and (b) most proposed quantitative methodologies for assessing or reducing “bias” are poorly matched to their goals and do not engage with the relevant literature outside of NLP. To assist researchers and practitioners in avoiding these problems, Blodgett et al. (2020) presented three guidelines for analyzing “bias” in NLP systems, along with a number of specific study topics for each. These recommendations are predicated on a greater knowledge of the connections between language and social hierarchies, a crucial step in defining a road ahead in our view.

How the insensitivity of annotators to dialect differences might contribute to other biases in computerized hate speech detection models, thereby exacerbating harm to minority populations, was studied by Davidson et al. (2019). In particular, African American English (AAE) and annotators’ assessments of toxicity in current datasets are highly correlated. This bias in annotated training data and the tendency of machine learning models to exacerbate it cause existing hate speech classifiers to frequently mislabel AAE material as abusive/offensive/hate speech (high false positive rate) Davidson et al. (2019); Xia et al. (2020).

A number of methods have been proposed for evaluating and addressing biases that exist in datasets and the models that use them Blodgett et al. (2020). All the above research deals with only one dimension of the problem but we deal with all the biases ranging from age bias, disability bias, gender bias, and other biases. In our research, we created a phrase dictionary and fine-grained test set to cover these non-inclusive categories.

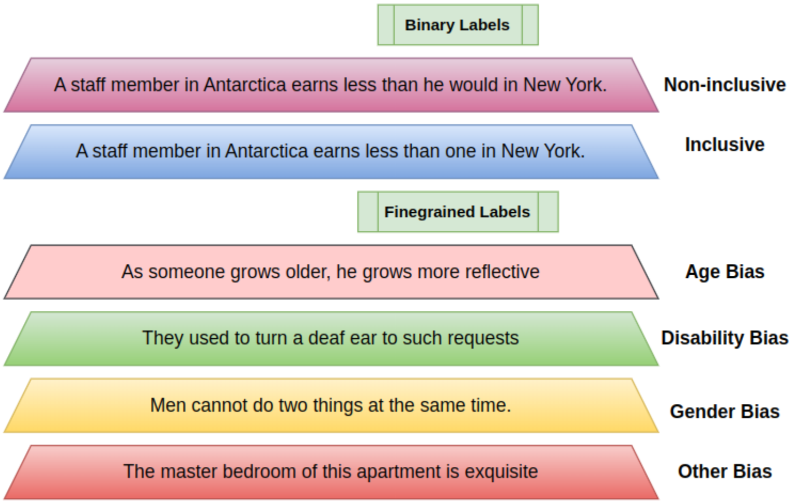
Binary Labels
A staff member in Antarctica earns less than he would in New York.
Non-inclusive
A staff member in Antarctica earns less than one in New York.
Inclusive
Finegrained Labels
As someone grows older, he grows more reflective
Age Bias
They used to turn a deaf ear to such requests
Disability Bias
Men cannot do two things at the same time.
Gender Bias
The master bedroom of this apartment is exquisite
Other Bias

Figure 1: Examples for the Binary and Fine-grained Labels

3 Dataset

The most frequent approach to the problem posed by NLP text classification tasks such as sentiment analysis, hate speech detection, and offensive language identification uses specialized dictionaries, sometimes known as lexicons, in which each word is assigned a proportional weight (positive or negative) based on the attitude it communicates. Negation, irony, ambiguity, idioms, and neologisms are just a few examples of the common linguistic subtleties that can make it challenging to exactly create a model for text classification problems. For these procedures to be effective, therefore, the specialists must have access to the raw texts and are often watched during the process. In our work, we create training, testing dataset, and dictionaries to improve the models' performances.

For our current research, we collected sentences and phrases from government and other organization guidelines documents and websites. For the test sentences, we gathered sentences from these websites and two annotators manually checked the validity of the sentences.

3.1 Annotation Style

We collected a set of comments from the websites. Our annotation schema proposes a hierarchical modeling of inclusive/non-inclusive languages. It classifies each example using the following two-level hierarchy. Level A- Inclusive/Non-inclusive, that is the text is inclusive or non-inclusive.

1. **Inclusive:** Sentences/phrases contain that recognize diversity and communicate respect for all individuals, including enthusiastic words, phrases, and expressions. Those sentences avoid using male pronouns or nouns for mixed-gender groups.

2. **Non-inclusive:** Sentences/phrases reinforce negative stereotypes or phrases, assimilate, or minimize groups of individuals, exclude specific groups of individuals, and assume the historically dominant groups to be the norm, for instance. It may cause emotional upset or offense.

We annotated the Level B- fine-grained to four classes in the non-inclusive category including age bias, gender bias, disability bias, and other biases.

1. **Age bias:** Ageism is present in our day-to-day language and is so deeply rooted in our culture that many ageist comments are often not noticed, missed, or accepted. Being elderly is often associated with undesirable characteristics and wrong opinions, such as dependency and the societal role in the capability to gain new knowledge in the workplace. Sentences containing the above ageism are considered as age bias sentences.

2. **Disability bias:** This is a wide range of physical, psychological, intellectual, and socio-emotional impairments. Different groups of people with disabilities categorize themselves in different ways. To demonstrate professional awareness and solidarity, we must recognize and respect the language choices of

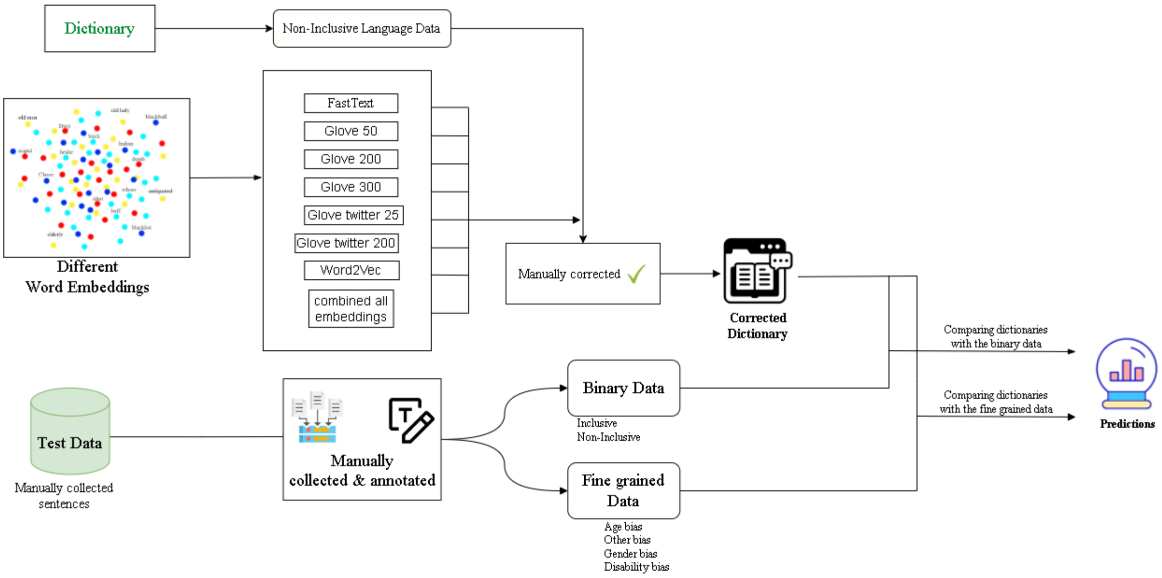

Figure 2: Visualization of the proposed methods

these groups[2]. For example, more inclusive of using the term "blind and low vision" instead of "visually impaired" Dunn and Andrews (2015).

3. **Gender bias:** Gender bias[3] involves unjust favoritism toward one gender due to stereotypes, leading to unequal treatment in areas like pay and leadership. It's evident in language, attitudes, and actions implying one gender's superiority. This issue perpetuates inequality and is recognized as a key factor in maintaining gender disparities, often unintentionally.

4. **Other bias:** Individuals' connection to their racial group shapes their self-perception, varying based on their grasp of psychological, sociopolitical, and cultural aspects tied to the group. Racial identification is fluid due to socially constructed definitions, evolving with context[4]. Worrell (2015) proposed cultural influence could supplant racial and ethnic identity, seen as psychological and social reflections of these concepts. This research encompasses LGBTIQ+ biases and anticipates adding a dedicated category for them in future studies.

3.2 Phrase Dictionary Creation

In the initial phase of dictionary methods, a set of keywords is formed for subsequent document analysis. These keywords should be pertinent to the classification, offering insight into the subject matter and tone. This dictionary is created through an extensive literature review, identifying crucial terms from government and organizational guidelines. Manual collection from various sources refines the keywords, which are then incorporated for use in the subsequent stage.

3.3 Phrase Dictionary Expansion

To expand the scope of our lexicons, we performed dictionary expansion on all four non-inclusion categories using pre-trained word embeddings such as Word2Vec Mikolov et al. (2013), fastText Bojanowski et al. (2017), and GloVe Pennington et al. (2014). We used a total of six sub-pre-trained embeddings from the above, such as fasttext-wiki-news-subwords-300, Word2Vec-Google-News-300, GloVe-wiki-gigaword-300, GloVe-wiki-gigaword-200, GloVe-wiki-gigaword-50, and GloVe-Twitter-200, to collect similar words from Wikipedia, News, Twitter, and Google News. Word embeddings are a collection of models that can capture the semantic similarity of words based on the context in which the words are found. It does this by mapping words onto an n-dimensional space and then

[2] https://t.ly/uUrpP
[3] https://rb.gy/v4zlc
[4] https://rb.gy/u3h1m

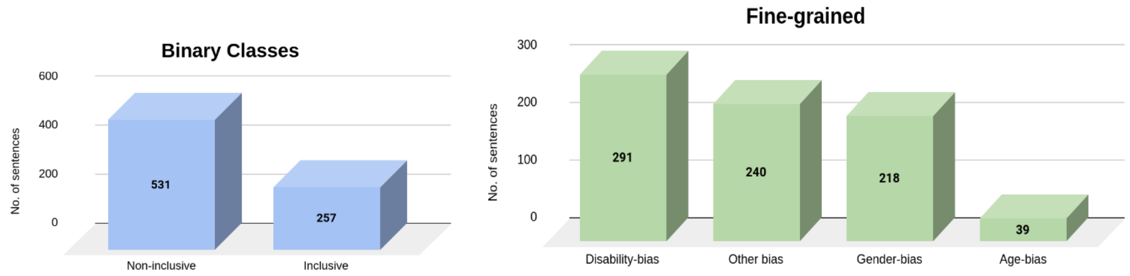

Figure 3: No. of sentences in each label

placing words in this space at locations within the space that are analogous to the circumstances in which the words were found. So, words that are more comparable to one another are those that are closer to one another in the cosine distance. We noticed a significant semantic variance across all of the non-inclusive words in our corpus, which leads us to believe that expanding the dictionary by using word embeddings will lead to the extraction of non-inclusive words that have not been found before. This is based on the assumption that similar, non-inclusive words are used in contexts that are analogous to one another.

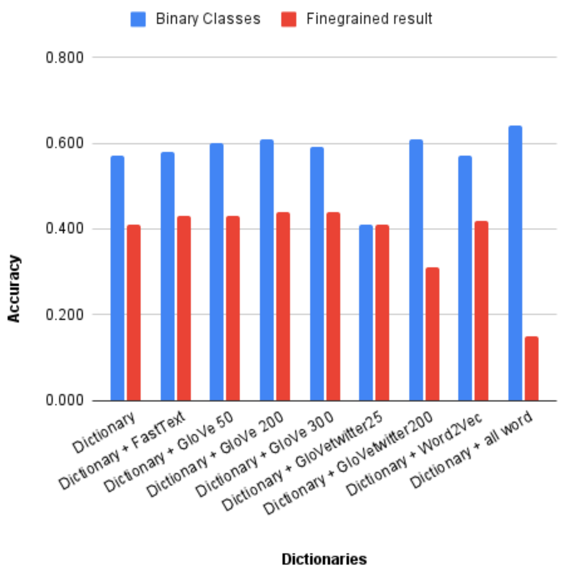

Figure 4: Accuracy for Binary and Fine-grained classes

3.4 Dataset Creation

We were able to collect only 788 sentences/comments from the website and guidelines documents; they are very small in size compared to other datasets created for similar classification tasks. To improve our corpus, we used the keywords from our dictionary which is collected using the embeddings. We manually annotated the sentences which are collected from the websites.

3.5 Annotation

All the annotators that contributed to the annotation of the corpus were of comparable age and had comparable educational backgrounds. The first annotation stage was done by two research assistants called A and B, who took the training in equality, diversity, and inclusion. They are also provided with guidelines links from web-pages[5] [6] [7] [8]. To obtain labels that matched the gold-standard criterion, a third annotator, marked by the letter C, was used as a tiebreaker. The consistency of the annotation system is measured with Inter-annotator agreement (IAA) and yielded values of 0.898 for binary classes and 0.811 for fine-grained classes. Cohen's kappa coefficient Krippendorff (1970) is used for computing IAA.

3.6 Corpus Statistics

Table 1 displays the corpus statistics of the dataset, providing insight into its size, complexity, and composition. Specifically, the number of characters in the text is 89,400, the number of words is 17,649, the number of sentences is 906, and the number of comments is 788. Furthermore, these statistics can be used to gain a better understanding of the texts' structure, vocabulary, and overall composition, thereby allowing for more informed decisions to be made. Additionally, these statistics can be compared to other datasets to determine how the text in the current dataset compares to other texts, helping to identify any differences or similarities.

Corpus statistics	
Number of characters	89400
Number of words	17649
Number of sentences	906
Number of comments	788

Table 1: Corpus statistics

4 Methodology

Our main aim of the dictionary-based approach is to analyze the binary and fine-grained classes from the sentences that are collected and annotated manually and establish the benchmark for this problem. The overall process is shown in Figure 2. The binary types are inclusive & non-inclusive, and the fine-grained classes are age bias, disability bias, Other bias & gender bias. These sentences were taken as a test set for our phrase dictionary approach. Testing these sentences using the phrase dictionary with different approaches. We created nine different lexicons with dictionary data and word embeddings.

We used seven sets of word embeddings to collect more words related to the keywords and bias with the help of gensim downloader[9]. Using this gensim downloader, we expanded the keywords to create more keywords for the lexicon approaches. We combined each word embedding to the original keywords and predicted with the test set for binary and fine-grained classes.

Firstly, we took the phrase dictionary and predicted them by comparing them with the sentence in the test data. Secondly, we collected the words in word embeddings such as fastText, Word2vec, and GloVe, and the prediction was made with each embedding add-on with the keywords. Lastly, we combined all the words collected from the word embeddings with the keywords and made the prediction.

5 Results

We have tested several different combinations of methods discussed in the previous sections across the test set sentences with lexicon-based sentences. As an evaluation measure, macro and weighted scores for precision, recall, and F1 scores are reported. We have used a phrase dictionary approach and test set sentences in English texts. These tasks are still crucial when dealing with the lexical method. We evaluated the lexical approach classified with test sentences in the previous section and briefly discussed their performance. We used nine different lexical-based dictionaries to classify the English sentences that are named as the test set. For all these experiments, predictions are given in the below Table 2 and Table 3. Also, we show the accuracy in Figure 4 for all nine experiments in both binary and fine-grained classes. Some dictionaries delivered similar results in both tasks.

Dictionary + all word embedding (combined) provided a more accurate prediction of 0.640 in the binary classes, which is predicting as an inclusive and non-inclusive, and Fidelity + GloVe_200 also provided a more accurate prediction of 0.440 in the fine-grained task which is finding as an age bias, other bias, disability bias, and gender bias. Other dictionaries also predicted better accuracy, similar to the best-performed method. The accuracy is used to evaluate these results because it calculates the critical metric when assessing the effects of all processes or models. It is a metric that measures how close the predicted values are to the actual values. This is an easy-to-understand metric that compares different methods in the NLP domain. Additionally, we can use accuracy to compare different algorithms or methods and data sets with each other. It is also widely used in the evaluation of supervised learning models.

6 Conclusion

We introduce a new phrase dictionary and dataset for non-inclusive sentences at binary and fine-grained levels of classification. This pioneering release combines word embedding-derived keywords with government and organizational guideline sentences, annotated for binary and fine-grained categorization. Our experiments utilize dictionary-based methods to set performance benchmarks. Notably, in binary classification, the Dictionary and GloVe200 combo achieves a high macro F1 score of 0.390. Similarly, the fine-grained task sees promise with the Dictionary and fastText fusion, yielding a top macro F1 score of 0.360. Moving forward, we plan to enhance our lexicon-based approach by integrating machine learning and few-shot learning techniques for more extensive appli-

[5] https://shorturl.at/aHQSZ

[6] https://shorturl.at/lrHOR

[7] https://shorturl.at/fgimB

[8] https://shorturl.at/koxH7

[9] https://radimrehurek.com/gensim/downloader.html

Binary Classes						
Dict_dataset	**MP**	**MR**	**MF1**	**WP**	**WR**	**WF1**
Dictionary	0.720	0.390	0.370	0.640	0.570	0.580
Dictionary + fastText	0.720	0.390	0.380	0.630	0.580	0.590
Dictionary + GloVe 50	0.710	0.380	0.370	0.620	0.600	0.610
Dictionary + GloVe 200	0.720	0.390	**0.390**	0.640	0.610	**0.620**
Dictionary + GloVe 300	0.710	0.380	0.380	0.630	0.590	0.600
Dictionary + GloVeTwitter25	0.640	0.460	0.320	0.830	0.410	0.480
Dictionary + GloVeTwitter200	0.680	0.340	0.340	0.580	0.610	0.590
Dictionary + Word2Vec	0.710	0.390	0.370	0.630	0.570	0.580
Dictionary + all word embeddings	0.690	0.350	0.340	0.590	0.640	0.600

Table 2: Result for Binary classes

Finegrained result						
Dict_dataset	**MP**	**MR**	**MF1**	**WP**	**WR**	**WF1**
Dictionary	0.650	0.460	0.320	0.850	0.410	0.490
Dictionary + fastText	0.640	0.490	**0.360**	0.810	0.430	0.490
Dictionary + GloVe 50	0.490	0.470	0.300	0.740	0.430	0.460
Dictionary + GloVe 200	0.500	0.480	0.310	0.740	0.440	0.480
Dictionary + GloVe 300	0.520	0.480	0.320	0.770	0.440	**0.500**
Dictionary + GloVeTwitter25	0.640	0.460	0.320	0.830	0.410	0.480
Dictionary + GloVeTwitter200	0.410	0.410	0.210	0.600	0.310	0.310
Dictionary + Word2Vec	0.600	0.470	0.330	0.800	0.420	0.490
Dictionary + all word embeddings	0.160	0.400	0.080	0.300	0.150	0.170

Table 3: Result for Fine-grained classes

cations.

7 Ethical Implication/Limitations

This work presents a dictionary and dataset which will be available only for the industry. Our dataset contains well-processed data annotated by experts in this field. The annotators are paid according to the University of Galway regulations. The details of our data collection and characteristics are introduced in the above section. Even though we have taken care of all the ethical problems, there might be cases in the near future the terms might change to inclusive/non-inclusive. We will be ready to update the terms in the dictionary then and there.

Acknowledgements

This work has been funded by FMR LLC. Researchers at the Data Science Institute are supported by Science Foundation Ireland as part of Grant Number SFI/12/RC/2289 P2, Insight SFI Centre for Data Analytics.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Dana S Dunn and Erin E Andrews. 2015. Person-first and identity-first language: Developing psychologists' cultural competence using disability language. *American Psychologist*, 70(3):255.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Jakob Lauring and Anders Klitmøller. 2017. Inclusive language use in multicultural business organizations: The effect on creativity and performance. *International Journal of Business Communication*, 54(3):306–324.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Frank C Worrell. 2015. Culture as race/ethnicity.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.