Debiasing a movie information retrieval system using gender bias count

Baran Polat
s4605411

Clara Amata Rus
s1064211

Josip Kopřčina
s1062758

ABSTRACT

Despite the advancement of the modern world, gender bias still exists in many fields. One example is the movie industry, where some gender stereotypes are still present. The purpose of this paper is to debias an information retrieval system, like the BM25 model, when searching through a dataset of movie scripts. In order to accomplish this, a list of biased and unbiased words was used to compute a weight of bias for each movie script. The weight was used to re-rank the output list of the BM25 model for a given query, in such a way that the least biased and most relevant movies are at the top of the list. This research can give more insight in how IR systems can be debiased to create more fair rankings.

KEYWORDS

movie, gender bias, information retrieval, BM25

ACM Reference Format:
Baran Polat, Clara Amata Rus, and Josip Kopřčina. 2020. Debiasing a movie information retrieval system using gender bias count. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the early 20th century, the suffrage movement has turned the tide of women's rights. Today, both genders are considered equal by law in most western countries, but there are many aspects of society where discrimination still exists. The movie industry is one part of society that is falling behind. The history of Hollywood is ridden with discrimination. The movie industry has not shied away from controversy throughout its lifetime [3], [11]. While some practices portray the zeitgeist of that time, the progression in representation and equal pay has been slow. This slow and stifling progression finally culminated into the '#MeToo' movement [6] and '#OscarsSoWhite' [15] protest. Since then, big production studios have changed their way of casting, giving more representation and balance the wage gap [8], [17]. While some of this new representation can unfortunately be seen as virtue signalling, steps are being made. As these biases are in movies, they can also slip into algorithms [16] and can have negative consequences. The same thing happens to language models [2]. This bias can stem from the fact that these models use human sources. One of the techniques that is known to show biases are word embeddings, that came to prominence in 2013 because of the Word2Vec model created by Mikolov et al. [14]. As fair machine learning is getting more attention and the field of Natural Language Processing (NLP) is growing, this research can contribute to this evolving issue.

Our study tries to find a way for people today who wish to watch movies to be able to choose from a list made up not only of the best movies that fit the descriptions they give, but the least biased ones. A study done by Bolukbasi et al. [2] shows using word embeddings how certain words are more connected to women while others favour men more. We will use this method to calculate the bias in movie scripts. The dataset contains 1093 movie scripts taken off IMSDs (an online movie script database). The movies here range in all genres and years of release. We will use a BM25 model [4] to get a ranking of the given movie scripts for certain queries. After that, the bias of the movie is added as a weight, changing the ranking list order. The goal of our project is to get a list of movies both relevant to the given queries and fairly placed on the list considering how biased it is. This should improve the fairness of recommendation systems used by apps such as IMDB, Netflix and Disney+.

RQ: Is it possible to make a movie information retrieval systems less biased using word embeddings?
By following previous research papers, a weighted list of biased words was constructed. The words will be used to calculate the amount of bias in movie scripts. In order to measure the method used, we will calculate the percentage of change between the ranked list generated by a basic information retrieval system, like the BM25 model, and the ranked list after the biased weights are added. If the new ranked list properly takes the calculated movie bias into account, it will be proven that it is a plausible task to add more fairness to movie information retrieval systems.

The paper is organised as follows: a short introduction about the importance of studying gender bias in a movie information retrieval system, then the relevant related work is presented in Section 2. Section 3 and Section 4 describes the approach in which we carry out our study. Lastly, in Section 5 the results of the study and the final conclusions, as well as an outlook on future work are presented.

2 RELATED WORK

Gender bias is an important matter that is investigated in various scientific papers. In one paper, Bolukbasi et al. [2] investigate the appearance of gender bias in word embeddings such as Word2Vec [14]. They found that stereotypical professions are confirmed in Word2Vec. As word embeddings are used across multiple applications, it can be dangerous to not acknowledge the existence of bias

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

towards gender. To measure how biased a word is, the authors considered words that are biased by definition, such as "she" and "he", as a baseline of comparison. If a word is equally distant from both "she" and "he" the word is considered to be neutral. A powerful example that was emphasized in the paper as biased is: "man - woman : computer programmer - homemaker". This suggests that while a man is a computer programmer, a woman should be a homemaker. This can have a negative impact on an information retrieval system as the example from [2] shows, where a user can search for people by profession. Professions can be very biased towards gender. The paper by Gerritse [9] investigates this matter by applying both cosine similarity and Principal Component Analysis (PCA) on a list of professions exemplified in [2]. Applying cosine similarity on the list, suggested that there is a bigger bias towards female professions than to men professions. The Principal Component Analysis did not reveal a significant gap between female and male professions as the cosine similarity showed.

There are some papers that focus on solving this problem by 'debiasing' the word embeddings. The approach of [2] is to change the word embeddings of the gender-neutral words (e.g. nurse) so that they are equally distant from both male and female words. Professions such as a nurse are not bound to men or women, but some gender-specific words such as brother and sister are by definition gendered. The paper showed that by removing the gender bias from word embeddings, the performance of Word2Vec on standard evaluation metrics is not affected. They also argue that bias in Word2Vec comes from the fact that society is still biased. This is due to the fact that Word2Vec is trained on a corpus of Google News texts. Even if the news is written by professional journalists, they reflect the events in society, which are still affected by gender bias. Gonen & Goldberg [10] claim that some debiasing methods such as the one presented by Bolukbasi et al. [2] do not remove the bias entirely. The experiments showed that even after the bias is removed, most of the previously biased words still cluster together by gender. For example, the word "nurse" is in the same cluster with many female-related words. This shows that there still exists a biased component in the word representations, besides the gender direction and that words that are semantically related also preserve the bias. Ignoring this matter can be dangerous in classification problems, as it can learn to favour men over women or vice versa. Even in state-of-the-art language models like BERT [7] this bias is still present [1]. In newer language models like BERT, context can be learned better than models of the older generation like Word2Vec. As BERT is relatively newer than Word2Vec and the bias in BERT is less well documented, sticking to Word2Vec for this research is the choice that was made.

3 EXPERIMENTAL SETUP

The data used in this study was gathered through IMSDs, which is an online collection of movie scripts written in HTML format. The scripts were taken indirectly from *https://osf.io/zytmp/*[1]. The dataset consists of 1093 movie scripts in txt format. They were 'web crawled' off the site in 2017 and consists of multiple genres and dates of release up to 2017. It is acknowledged that typos or crawling errors may exist in the extracted movie scripts. The text files were preprocessed by removing punctuation and unnecessary white spaces. Each movie script was turned into a list of tokens, by means of splitting according to spaces. The next step was to compose a list of queries. We came up with a list of 60 queries that were produced by taking into consideration the nature of the problem, meaning choosing queries that do not reflect a certain movie title, but a movie "type" or better said, something that is often a movie trope or characteristic. Examples of such tropes or characteristics could be "stranded on an island" or "woman with a sword". An implementation of the BM25 model [4] was imported and tested on the dataset. This implementation is based on the algorithms presented in [18]. For this study, the Okapi BM25 algorithm was used. In order to evaluate the resulted ranked list, we checked if the short descriptions of the movies, on the top of every list, suits the given query. To compute the bias weight of each movie script, a list of words considered negatively biased and a list of words considered positively biased was extracted from [5]. As there was no scientific proof to the fact that these words are indeed biased, the next step would be to test this fact using the method described in Bolukbasi et al. [2]. By creating word embeddings, the results shown in Figure 1 and 2 were generated. The plots show results similar to those explained under the Related Work section. The list of bias words was then changed so that the words that are considered biased were the ones clustered around words such as "he" and "she" and those that were far away were considered non-biased. The clusters are not completely optimal as the ones that deal with biased words are not too close to the words we chose as naturally biased. Contrary, we see in Figure 2 where we plot the list of biased words, that the words are more closely clustered to "he" and "she" than those on Figure 1. The list of biased and non-biased words can be found in the GitHub repository [2] associated with this paper.

4 METHOD

In order to quantify the bias of each movie, we used a simple word token count. Working through all tokens for all scripts, the program counted the number of words from the biased and non-biased list they contained. This presented a weight of how biased each text is. Do note that fewer points were given to words after a repeated presence, in case some words with different meanings were used, as this would make them very common in the script. An example of such a word is "master" used as a nickname in the animated movie "Megamind". The list of results generated by a BM25 model was re-ranked considering also the bias weight of each movie script, for all 60 queries. An example of the resulted re-ranked list is shown in Table 2. It can be observed that the movie "Mulan" receives an upgrade in rank by 1 position. Given the opinions found online [13], that consider this movie in the category of movies that defy the gender stereotypes, we could consider this result successful in combating the gender bias, given the query "sword girl". Analysis of bias is subjective, as it is difficult to measure if the change is good (more balanced representation), or bad (over-correction). The movie titles that are present in Table 1 and no longer in Table 2, are downgraded by the bias weight as they contain much more biased words than the top 5 results presented.

[1] *https://osf.io/zytmp/*

[2] *https://github.com/jkoprcina/IRProjectMovieScriptGenderBiasRetrieval*

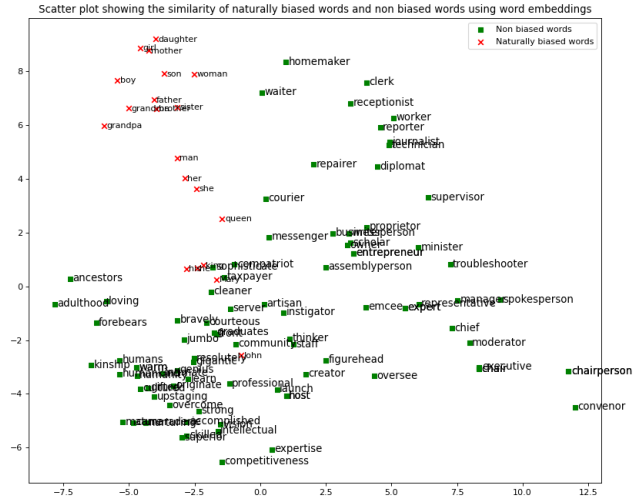

Figure 1: Ensuring words are not biased by comparing word embeddings with naturally gender biased words

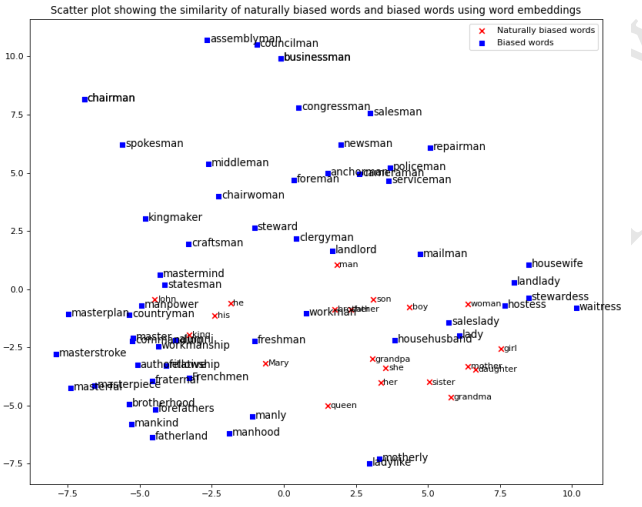

Figure 2: Ensuring words are biased by comparing word embeddings with naturally gender biased words

5 RESULTS AND DISCUSSION

The results presented in Table 3 and Table 4 show the least biased movies and most biased movies, according to our method. A negative weight represents a large number of biased words in the script. This means that the movie will be downgraded in the ranked list generated by BM25 for a given query. A positive weight will increase the position in the ranking. In order to measure if our method has an effect on the ranked list, all 60 queries were used to retrieve movies from the dataset using first the basic BM25 model

Rank	Title	Score
1	Kill Bill Volume 1, 2	7.085868
2	Mulan	7.076451
3	Conan the Barbarian	7.017045
4	Robin Hood Prince of Thieves	6.988501
5	Scott Pilgrim vs the World	6.980803

Table 1: Output list of the BM25 model using the query "sword girl".

Rank	Title	Bias Score
1	Mulan	7.118910
2	Conan the Barbarian	6.705281
3	Scott Pilgrim vs the World	6.701241
4	Kill Bill Volume 1, 2	6.661396
5	Book of Eli, The	6.596988

Table 2: Output list of the BM25 model re-ranked considering the bias weight, using the query "sword girl" .

and then the model with the bias weights added. We compared the two ranked list for each query and calculated how many positions each movie was downgraded or upgraded in the list. The results showed that on average, a movie changed its rank by a number of 348 positions for a given query. Therefore, the method presented in this paper has an effect on the ranked list, however the movies that were found to be the least biased from Table 3, do not correlate positively with the opinions found online [13]. For example, the movie "Mulan" is found on position 583 in the least biased movie list with a bias weight of 0.042, but according to [13], it should have been at the top of the list. The opinions found online are indeed subjective and do not reflect the general opinion of the population, but more of an example of which movies can be considered to combat the gender bias. It can be concluded that the presented method has an effect in changing the content of the usually received list of movies, but to which extent this decreases the gender bias is unknown. For future improvements, the opinions of annotators in the field of linguistics could be used in order to validate or expand our list of biased words, as there is no objective way to determine how biased a word is. We use 77 gender biased unigrams and bigrams and 122 non-biased or positively biased unigrams and bigrams. In the world of text features this is a very small amount, and it could be that the method would have better results simply by increasing these numbers. In this paper, the bias weight is quantified considering only one feature, the count of biased words that are present. A consideration for the future could also be to calculate the bias weight as an average of multiple feature functions in detecting the gender bias. Some other functions that can be used are dialogue counting, actor counting in important roles, wage information etc. Dialogue counting would be simply counting how many words are said by female and how many by male actors, as well as how many times each gender spoke. Information about roles could give us insight on which gender has more often a main role in a movie [12]. Finally, while BERT is a new language model and thus is not yet fleshed out on research regarding bias, it might yield better results over

Rank	Title	Bias Weight
1	Hudsucker Proxy	4.225
2	Kafka	2.844
3	Schindler's List	2.779
4	The Island	2.618
5	The Hebrew Hammer	2.55

Table 3: Top least biased movies

Rank	Title	Bias Weight
1	The Master	-6.610
2	Megamind	-5.135
3	Thunderbirds	-4.620
4	The Distinguished Gentleman	-4.433
5	The Producer	-4.279

Table 4: Top most biased movies

Word2Vec, since BERT has had strong positive results regarding contextualisation of words.

REFERENCES

[1] Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2020. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*.

[2] Tolga Bolukbasi, K 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*, 4349–4357.

[3] Tom Brook. 2015. When white actors play other races. (2015). https://www.bbc.com/culture/article/20151006-when-white-actors-play-other-races.

[4] Dorian Brown. [n. d.] Implementation of the bm25 model. (). https://github.com/dorianbrown/rank_bm25.

[5] Compiled by Service-Growth Consultants Inc. 2003. Examples of gender-sensitive language. (2003). https://www.servicegrowth.net/documents/Examples%20of%20Gender-Sensitive%20Language.net.pdf.

[6] Shelley Cobb and Tanya Horeck. 2018. Post weinstein: gendered power and harassment in the media industries. *Feminist Media Studies*, 18, 3, 489–491. DOI: 10.1080/14680777.2018.1456155.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[8] Jamie Doward and Tali Fraser. 2020. Hollywood's gender pay gap revealed: male stars earn $1m more per film than women. (2020). https://www.theguardian.com/world/2019/sep/15/hollywoods-gender-pay-gap-revealed-male-stars-earn-1m-more-per-film-than-women.

[9] Emma Gerritse. 2019. Impact of debiasing word embeddings on information retrieval. In *FIDA*.

[10] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of North American Chapter of the Association for Computational Linguistics*. Volume 1, 609–614.

[11] Wil Haygood. 2019. Why won't blackface go away? it's part of america's troubled cultural legacy. (2019). https://www.nytimes.com/2019/02/07/arts/blackface-american-pop-culture.html.

[12] Dima Kagan, Thomas Chesney, and Michael Fire. 2020. Using data science to understand the film industry's gender gap. *Palgrave Communications*, 6, 1. DOI: 10.1057/s41599-020-0436-1.

[13] [n. d.] List of movies that defy gender stereotypes. (). https://www.commonsensemedia.org/lists/movies-that-defy-gender-stereotypes.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

[15] Isabel Molina-Guzmán. 2016. #oscarssowhite: how stuart hall explains why nothing changes in hollywood and everything is changing. *Critical Studies in Media Communication*, 33, 5, 438–454. DOI: 10.1080/15295036.2016.1227864.

[16] Cathy O'Neil. 2017. *Weapons of math destruction*. Penguin Books.

[17] Reuters Staff. 2020. (2020). http://www.reuters.com/news/picture/gender-pay-gap-narrows-only-marginally-i-idUSKBN27M1KP.

[18] Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In (November 2014), 58–65. DOI: 10.1145/2682862.2682863.