Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Gender Bias in Occupation Classification from the New York Times Obituaries

New York Times Anma Yazılarından Meslek Sınıflandırmasında Cinsiyet Yanlılığı

Ceren Atik [1], Selma Tekir [2*]

[1] Dokuz Eylül Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, İzmir, Türkiye
[2] İzmir Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü, İzmir, Türkiye
Sorumlu Yazar / Corresponding Author *: selmatekir@iyte.edu.tr

Geliş Tarihi / Received: 24.04.2021 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 16.12.2021 *DOI:10.21205/deufmd.2022247109*
Atıf şekli/ How to cite: ATIK, C., TEKIR, S. (2022). Gender Bias in Occupation Classification from the New York Times Obituaries. DEUFMD, 24(71), 425-436.

Abstract

Technological developments such as artificial intelligence can strengthen social prejudices prevailing in society, regardless of the developer's intention. Therefore, researchers should be aware of the ethical issues that may arise from a developed product/solution. In this study, we investigate the effect of gender bias on occupational classification. For this purpose, a new dataset was created by collecting obituaries from the New York Times website and is provided in two different versions: With and without gender indicators. Category distributions from this dataset show that gender and occupation variables have dependence. Thus, gender affects occupation classification. To test the effect, we perform occupation classification using SVM (Support Vector Machine), HAN (Hierarchical Attention Network), and DistilBERT-based classifiers. Moreover, to get further insights into the relationship of gender and occupation in classification problems, a multi-tasking model in which occupation and gender are learned together is evaluated. Experimental results reveal that there is a gender bias in job classification.

Keywords: *Gender Bias, Occupation Classification, Multi-task Learning, Obituaries.*

Öz

Yapay zeka gibi teknolojik yenilikler, geliştiricilerin niyetlerinden bağımsız olarak toplumda mevcut olan ön yargıyı arttırabilirler. Bu sebeple, araştırmacılar geliştirilen bir ürün/çözüm ile birlikte gelebilecek etik sorunların farkında olmalıdırlar. Bu çalışmada, sosyal ön yargılardan biri olan cinsiyet yanlılığının meslek sınıflandırması üzerindeki etkisi araştırılmaktadır. Bunun için New York Times web sitesinden anma yazıları toplanarak yeni bir veri kümesi oluşturulmuş ve bu anma yazıları cinsiyet göstergeleri dahil ve hariç olmak üzere iki farklı versiyonuyla sunulmuştur. Bu veri kümesindeki sınıf dağılışları incelendiğinde cinsiyet ve meslek değişkenleri arasında bir bağımlılık ilişkisi görülmektedir. Dolayısıyla cinsiyet göstergelerinin meslek tahmini üzerinde bir etkisi olması beklenmektedir. Bu etkiyi sınamak üzere, SVM (Karar Destek Makineleri), HAN (Hiyerarşik İlgi Ağı) ve DistilBERT algoritmaları kullanılarak meslek sınıflandırması yapılmıştır. Sadece meslek sınıflandırması yapan bu modellerin yanında meslek ve cinsiyetin eş zamanlı öğrenildiği bir model de değerlendirilmiştir. Deneysel sonuçlar, meslek tahmininde cinsiyet yanlılığının etkili olduğunu ortaya koymaktadır.

Anahtar Kelimeler: *Cinsiyet Yanlılığı, Meslek Sınıflandırması, Çok Görevli Öğrenme, Anma Yazıları.*

1. Introduction

Detecting bias is becoming gradually important based on its relevance in many fields, ranging from evaluating publications to understanding political perspectives. The progress of technology day by day and its inclusion in many areas of our lives has brought new social problems. Bias is difficult to detect and evaluate because it is usually implicit. People can develop discrimination toward or against an individual, an ethnic group, and gender identity. Therefore, researchers should be aware of this ethical issue.

In the area of sentiment analysis, algorithms are not unbiased [1]. They give more accurate results when trained on female-authored data, which implies that they over-represent females' viewpoints in a gender-mixed collection. Considering the risk of increasing data bias [2], word embeddings have been analyzed first. It's shown that these word representations reflect social tendencies that exist in the data used to train them [3, 4].

Gender bias is the preference of one gender over another or approaching one gender prejudicially. In NLP literature, different methods have been used to reveal gender bias. [5, 6, 7]. Gender bias is also evaluated in classification problems. De-Arteaga et al.'s study [8] reveals gender bias on occupation classification. For each model used in the study, the performance is measured with and without gender indicators (name, gendered pronoun).

In this study, we also focus on occupational classification. To test the effect of gender on occupation classification, we target an editorial column, which is somewhat controlled. Thus, we prepare a new dataset collecting obituaries from the New York Times (NYT). We apply the method of scrubbing to clear gender indicators. We both use traditional machine learning algorithms and deep learning models. Moreover, we investigate gender bias in a multi-tasking model where occupational and gender variables are learned together. We hypothesize that this joint learning may reduce the bias. The results confirm gender bias in occupational classification. In the case of multi-tasking, gender bias seems to be neutralized.

The aim of this work is twofold. First, we analyze whether the data collection reflects gender bias or not. Because the New York Times obituaries are an editorial column, the editors could have made the choices of individuals so that the collection would have been gender-neutral for occupation. Our study confirms the opposite. Second, we aim to raise awareness of the biases in datasets because the machine learning classifiers that rely on them will show discriminatory behavior in the decision-making processes they support.

The primary contributions of this work include:

A corpus of 5210 preprocessed NYT obituaries.

Empirical results through traditional and deep learning-based classifiers confirm gender bias in occupation classification.

Implementing a multitasking model that jointly optimizes gender and job predictions brings about a neutralizing effect on gender bias.

In the remainder of this paper, we first give the related work. In Section 3, we describe the data collection process and the models used. Afterward, we provide experimental results and their evaluation. Then, we conclude the paper.

2. Related Work

Fairness in machine learning has been a rising concern in recent years. The ever-increasing volumes of data provide deep learning systems with significant performance improvements in different classification tasks. However, on the downside, these systems encode the existing societal biases in data in their learned representations. This phenomenon poses a risk in that the classifiers trained on these data may affect the decision-making processes carrying this bias further. Thus, fairness has become a vital issue for classifiers and paved the way for bias detection and mitigation approaches.

The first practice of bias detection focused on the intrinsic evaluation task of word analogies.

When the analogy task is run using the early word embeddings such as word2vec, some undesired analogies are observed, e.g., intellectual professions are associated more with men than women [3, 4]. As a response, debiasing techniques were proposed to remove bias from the static word embeddings. By the introduction of contextual word embeddings such as Elmo [9], BERT [10], their biasing behaviour has been the subject of some work [9, 11, 12]. In GPT-3 [13], the authors perform a fairness analysis in their learned representations for the dimensions of race, gender, and religion. The results confirm the bias in the learned embeddings, thus, highlighting the necessity of bias reduction solutions for pre-trained models. Garrido-Muñoz et al. [14] provide a comprehensive survey on bias detection and correction for pre-trained language models. Webster et al. [15] analyze unintended correlations in pre-trained language models through a case study of unintended gender correlations and show that they can be reduced without major degradation in the models' accuracy.

In addition to model associations, researchers investigate unintended social stereotypes in downstream tasks to assess real-world implications. Racial and gender bias were reported in job recruitment software [8, 16], sentiment analysis [1, 17], bibliometry [6], and machine translation [18, 19].

Mitigating bias is a vital issue for the future of our digitalized society. What's more, gender bias is not the only type of bias. Age and ethnicity are the other attributes that are used for discrimination. Thus, fairness must consider these attributes and their interrelations simultaneously. In other words, a classifier should preserve its accuracy without correlating with gender, age, ethnicity, etc. Subramanian et al. [20] evaluate fairness for multiple attributes by working with a constrained model that jointly optimizes model performance and model fairness.

In addition to projecting away the undesired attribute [3], the standard overfitting techniques of dropout regularization and counterfactual data augmentation have been found helpful for bias reduction [20]. Another popular research direction in bias mitigation is adversarial losses to remove demographic information from learned representations. Chowdhury et al. [21] present an adversarial debiasing framework (Adversarial Scrubber) for scrubbing demographic information from contextual representations. Their experimental results prove that the application of Adversarial Scrubber preserves baseline performances on different text classification tasks while making the classifiers agnostic to demographic information.

3. Material and Method

This section first explains the data collection process and the preprocessing operations on the prepared dataset. Then, we describe the architecture and working principles of the models used.

3.1. Dataset

We collected the articles published between 2014 and September 2019 from the NYT obituaries using the NYT API and web scraping. The NYT API includes article id, article summary, title, author, publication date, category, number of words, and keywords information. Using web scraping, we also retrieved the article text. We extracted the name, surname, gender, age, and occupation of the person mentioned in the obituary by natural language processing techniques. We realized that some articles' subjects are not individual persons but a group of people, e.g., Apollo11 team. Thus, we removed those articles from the dataset since they were not suitable for gender and occupation prediction. To further validate the occupation information for those people, we referenced their biographies from Wikipedia.

Moreover, to form the class labels for occupations, "SOC 2018" and "O*NET" occupation dictionaries were used [22]. With the help of keyword search on these dictionaries, the obtained results were assigned to their main categories to reduce the number of classes. However, the dataset is imbalanced since the main occupation categories are not evenly distributed. The final dataset consists of 5210 articles, where 4002 of them belong to males, and the remaining belong to females. Appendix A includes the final occupation labels with their distribution.

3.2. Preprocessing

As part of preprocessing, we removed the first sentences of articles as they contain the name, surname, and occupational information of the person, and the name usually indicates gender inherently.

We performed lowercasing next. Then, we removed special characters and numbers from the text. Afterward, we applied tokenization using NLTK's word tokenization function. At the next step, we removed stop words based on the NLTK's stop words' set for English (Appendix B). To see the effect of gender pronouns in occupation prediction, a different set of stop words was formed and used for filtering. As shown in the second figure in Appendix B, we removed gendered pronouns and possessive pronouns from the NLTK's stop words in a second version stop word list. Thus, there are two versions for the article collection: Regular articles and scrubbed articles. The former set was filtered using the default stop word set from NLTK, while in the latter, we cleaned gendered and possessive pronouns. Finally, stemming is applied as the last preprocessing step.

3.3. Models

We selected Support Vector Machines (SVM), Hierarchical Attention Network (HAN), DistilBERT, and Multi-task Learning (MTL) as the models. Our reasoning behind this choice is provided below:

Textual data is typically high-dimensional. With its ability to generalize well over high-dimensional feature spaces, SVM significantly simplifies the application of text categorization by eliminating the need for feature selection. However, classical machine learning models such as SVM do not have contextual understanding and do not preserve word order. For this reason, we apply HAN and DistilBERT models. HAN can process long documents and has an attention mechanism to understand which sentences and words are essential to capture meaning. DistilBERT is a distilled version of BERT, which is a pre-trained language model and has multi-head attention. DistilBERT is chosen for this project due to its lighter memory footprint and its faster inference speed.

Support Vector Machines (SVM)

SVM is by default for binary classification. Since occupation classification is a multi-class classification problem, we choose the one-vs-one (OVO) strategy in using SVM. OVO is not a specific feature of SVM. This method aims to develop an expert binary classifier for each possible class pair and build an ensemble. If the multi-class problem has N classes, the OVO ensemble will be composed by (N*(N-1))/2. The majority voting assigns the labels.

Hierarchical Attention Network (HAN)

HAN is a deep neural network for document classification. It embraces the idea that not every word in the sentence is equally important to capture meaning. So is every sentence in a document. For this purpose, their proposed architecture (Figure 1) is composed of word and sentence level encoders with attention layers on top to learn the importance weights of contexts. Finally, a feed-forward layer with a softmax on top performs the classification.

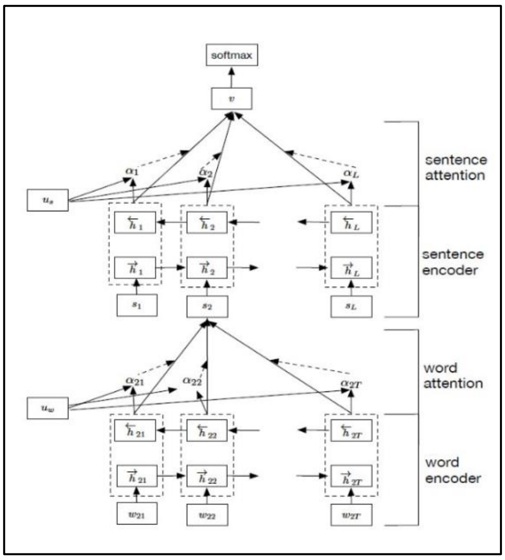


Figure 1. Illustration of HAN Architecture [23]

DistilBERT

DistilBERT [24] is a smaller language model under the supervision of BERT [10]. The DistilBERT network architecture is a transformer encoder model and has half the number of BERT layers while keeping the hidden representation dimension the same.

Multi-task Learning (MTL)

A hard parameter sharing architecture (Figure 2) is used as a multi-task learning model in this work. In hard parameter sharing, a separate loss is computed per task, and those loss terms are combined into the general loss of the network through weighting. It allows having a single model for all of the tasks. Additionally, the model has to find a representation that captures all the tasks in increasing its learning level. This characteristic prevents overfitting on the original task.

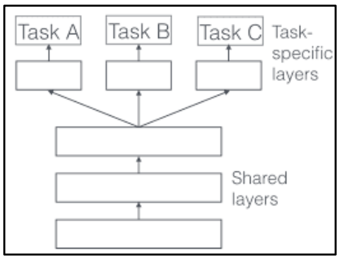


Figure 2. Hard Parameter Sharing [25]

4. Results

This section reports the classification results for the models where occupation and gender are learned together and the models that solely perform occupation classification to analyze gender bias. Since the variables of gender and profession have a dependency relationship based on the ground-truth categorical data, the classification results are expected to confirm this. The independence hypothesis states that the models' occupation prediction using regular and scrubbed articles for individuals should be the same. On the contrary, if there is a significant difference between these two results, the hypothesis is broken; we can say that gender information plays a role in the estimation. Therefore, there is gender bias.

To test the dependence relationship between gender and occupation ground-truth categories, we applied the χ^2 test. The null hypothesis for this test is that gender and occupation are independent. The obtained Chi-squared statistic value is 128.07 with 22 degrees of freedom. The associated p-value is far smaller than 0.01. It means we can reject the null hypothesis in a statistically significant way. In other words, there is a dependency between gender and occupation.

SVM Results

The SVM model uses TF-IDF weighting on word unigrams, bigrams, and trigrams with the following parameter settings: Minimum document frequency is 2, and maximum document frequency is 1.0, where terms that occur in all documents will be ignored. We used the TruncatedSVD function from the scikit-learn library for dimensionality reduction, where the reduced dimension is 300.

As mentioned before, our dataset is not class balanced. The "class_weight" parameter of SVM is set as balanced as proposed in working with imbalanced datasets. We used the OVO strategy to perform multi-class classification and reported the results in micro-averaged F1-Score. We evaluated the accuracy of SVM using 10-fold cross-validation. The dataset has 80% training and 20% test set partitioning.

Table I gives True Positive Rates (TPR) for occupation classification for female obituaries and male obituaries in regular and scrubbed versions.

Table I. SVM TPRs for Occupation Classification for Obituaries in Regular and Scrubbed Version.

	Articles	Scrubbed Articles
TPR_{female}	0,46261	0,45421
TPR_{male}	0,48592	0,46308
TPR_{gap}	0,02330	0,00886

TPR_{gap} value is obtained by taking the difference of TPR_{male} and TPR_{female} values. As shown in the table, there is a decreased gap between TPR_{male} and TPR_{female} in scrubbed articles. It means gender indicators affect the results.

HAN Results

Neither every sentence in a document nor every word in a sentence is of equal importance. Based on this idea, HAN performs well in document classification. After preprocessing, sentences are tokenized, and tokens are vectorized using GloVe pre-trained embeddings in 100 dimensions. We treat Articles and Scrubbed Articles as two separate corpora. The maximum number of sentences is 279, and the maximum number of words for sentences is 127 for the Articles Corpus.

On the other hand, the maximum number of sentences is 278, and the maximum number of words for sentences is 119 for the Scrubbed Articles Corpus. Afterward, padding was applied to ensure that inputs are of equal length. The dataset is divided into 60% training set, %20 validation set, and 20% test set. We set the model's dropout rate to 0.5 and trained it with the Adam optimizer by the learning rate 0.0005, in 7 epochs and using a batch size of 50.

Table II. TPRs of Articles and Scrubbed Articles from HAN for Occupation Classification.

	Articles	Scrubbed Articles
TPR_{female}	0,32235	0,19980
TPR_{male}	0,26135	0,18791
TPR_{gap}	0,06100	0,01188

As shown in Table II, the TPR results obtained for HAN also show differences between regular and scrubbed articles. These results show that the model pays attention to gender indicators. Therefore, gender bias occurs.

DistilBERT Results

BERT expects input data in a specific format, with unique tokens for the beginning ("[CLS]") and the "[SEP]" tag to separate or end sentences. Since DistilBERT is a distilled version of BERT, it uses the same input format as BERT.

Again we use a dataset of 60% training set, 20% validation set, and 20% test set. We set the learning rate as 5e-5, the dropout rate as 0.1, the epoch number as 3, and the batch size as 50. Table III shows a lower gender gap for scrubbed articles than regular articles. The decrease in the gap in the absence of gender indicators indicates that there is gender bias.

Table III. TPRs of Articles and Scrubbed Articles from DistilBERT for Occupation Classification.

	Articles	Scrubbed Articles
TPR_{female}	0,38741	0,33819
TPR_{male}	0,35806	0,33205
TPR_{gap}	0,02934	0,00614

We include in Appendix D the original output images of DistilBERT highlighting the words contributing to the final label in green and the detracting words in red. These visualizations clearly show that gender indicators have a role in the final estimation in the regular articles.

MTL Results

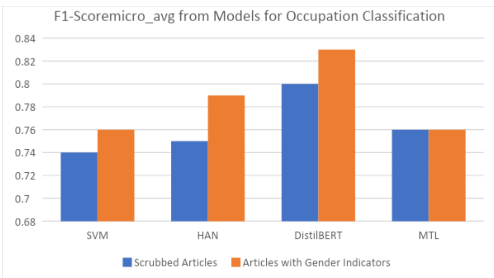
The MTL model used is based on the architecture of HAN with a hard parameter sharing. After shared layers, there is a task-specific layer to get predictions for gender and occupation. We compute binary cross-entropy loss for gender and categorical cross-entropy for the profession. The system calculates a weighted sum of individual losses as the final loss value. The weight hyper-parameter was tuned using Adam optimizer. We tried different weight values in the calculation of loss. The best result among the tried weights is obtained when it is "1." for occupation and ".9" for gender.

In our case, the occupation loss has a higher weight than the loss term for gender. We use a dataset of 60% training set, 20% validation set, and 20% test set. Same as HAN's, we set the MTL model's dropout rate to 0.5. We train our model with Adam optimizer and learning rate 0.0005, in 8 epochs and using a batch size of 50.

Table IV. TPRs of Articles and Scrubbed Articles from MTL for Occupation Classification.

	Articles	Scrubbed Articles
TPR_{female}	0,17429	0,21636
TPR_{male}	0,23031	0,19929
TPR_{gap}	0,05601	0,01706

Table IV also shows a performance gap between regular and scrubbed articles. For males, the MTL model produced higher estimates when we included gender indicators in the input. However, in the female TPR row, we see an opposite effect. Thus, the results differ between genders having an overall neutralizing behavior.



Graph 1. F1-$Score_{micro\_avg}$ from Models for Occupation Classification.

As micro-averaged (biased by class frequency) F1 score is a robust metric to measure classification performance in multi-class classification, we computed the F1-Score$_{micro_avg}$ of all models (Graph 1). As for general classification performance, deep learning models HAN and DistilBERT perform better than SVM. DistilBERT seems superior because it learns contextual representations and puts a softmax layer on top to perform classification. We give accuracy, precision, recall, and F1-Score metrics values for the regular and scrubbed articles in Table V and VI respectively.

To test the impact of scrubbing on the classification performance for each classifier, we applied a paired-sample t-test. The associated null hypothesis is that there is no difference in performance between regular and scrubbed articles. Among the classifiers, only HAN showed a statistically significant effect with a p-value of 0.04406, which means that scrubbing introduced a significant change in the classification performance within the 95% confidence interval.

Table V. Performance of Models with Articles for Occupation Classification.

	Articles			
	Acc	**Precision** weighted_avg	**Recall** weighted_avg	**F-Score** weighted_avg
SVM	0.77	0.79	0.77	0.77
HAN	0.79	0.75	0.79	0.76
Distil BERT	0.83	0.79	0.83	0.80
MTL	0.77	0.71	0.77	0.73

Table VI. Performance of Models with Scrubbed Articles for Occupation Classification.

	Scrubbed Articles			
	Acc	**Precision** weighted_avg	**Recall** weighted_avg	**F-Score** weighted_avg
SVM	0.75	0.78	0.75	0.76
HAN	0.76	0.68	0.76	0.70
Distil BERT	0.81	0.78	0.81	0.78
MTL	0.76	0.71	0.76	0.72

As shown in Graph 1, in MTL, scrubbing does not introduce a change in the score. Here, the dependency between gender and occupation is tuned inside the system, which could have caused a neutralizing effect on the gender bias for the profession classification.

5. Discussion and Conclusion

In this work, we released a new dataset of the NYT obituaries to reveal gender bias. The dataset has two versions: Regular articles and scrubbed articles. The scrubbed articles are gender proof, which means that we removed the first sentences and gender indicators from them. We tested the dataset with different models and examined the TPR gender gaps and overall F1-Scoremicro_avg from occupation estimation to check whether there was gender bias. Based on the F1-Scoremicro_avg scores, the HAN model gives the highest t statistic value to confirm the dependency between gender and occupation. In MTL, gender indicators for females behave differently compared to males for classification, having an overall neutral result. Thus, gender indicators play a role in predicting the occupation, but a neutralizing effect was observed in multi-tasking where gender and occupation classification performed simultaneously.

References

[1] Thelwall, M. 2018. Gender Bias in Sentiment Analysis: Online Information Review. Vol. 42, p. 7, DOI: https://doi.org/10.1108/OIR-05-2017-0139

[2] Bölükbaşı, T., Chang, K.-W., Zou, J., Saligrama, V., Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embedding. Proceedings of the 30th International Conference on Neural Information Processing Systems, 5-10 December, Barcelona, Spain, 4356–4364.

[3] Caliskan, A., Bryson, J. J., Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases: Science, Vol. 356, p. 183-186, DOI: 10.1126/science.aal4230

[4] Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes: Proceedings of the National Academy of Sciences, Vol. 115, p. E3635-E3644. DOI: 10.1073/pnas.1720347115

[5] Buolamwini, J. and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 23-24 February, New York, USA, 77-91.

[6] Caplar, N., Tacchella, S., Birrer, S. 2017. Quantitative Evaluation of Gender Bias in Astronomical Publications from Citation Counts: Nature Astronomy, Vol. 1, p. 8, DOI: https://doi.org/10.1038/s41550-017-0141

[7] Fu, L., Danescu-Niculescu-Mizil, C., Lee, L. 2016. Tie Breaker: Using Language Models to Quantify Gender Bias in Sports Journalism. Proceedings of IJCAI workshop on NLP meets Journalism, 10 July, New York, USA.

[8] De-Artega, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Couldechova, A., Geyik, S., Kenthapadi, K., Kalai, A. T. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Settings. ACM Conference on Fairness, Accountability, and Transparency, 29-31 January, New York, USA, 120-128.

[9] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. 2018. Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018), 1-6 June, New Orleans, USA, 2227-2237. DOI: 10.18653/v1/N18-1202

[10] Devlin, J., Chang, M., Lee, K., Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), 2-7 June, Minneapolis, USA, 4171-4186. DOI: 10.18653/v1/N19-1423

[11] Basta, C., Costa-Jussa, M., Casas, N. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 28 July-2 August, Florence, Italy, 33-39. DOI: 10.18653/v1/W19-3805

[12] Tan, Y.C., Celis, Y.E., 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. Advances in Neural Information Processing Systems (NEURIPS 2019), 8-14 December, Vancouver, Canada, 13209-13220.

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NEURIPS 2020), 1877-1901.

[14] Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., Ureña-López, L.A. 2021. A Survey on Bias in Deep NLP, Appl. Sci, Vol. 11, p. 3184. DOI: 10.3390/ app11073184

[15] Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Petrov, S. 2020. Measuring and Reducing Gendered Correlations in Pre-trained Models. https://arxiv.org/abs/2010.06032

[16] Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Rumshisky, A., Kalai, A. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), 2-7 June, Minneapolis, USA, 4187-4195. DOI: 10.18653/v1/N19-1424

[17] Kiritchenko, S., Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, June 5-6, New Orleans, USA, 43-53. DOI: 10.18653/v1/S18-2005

[18] Stanovsky, G., Smith, N.A., Zettlemoyer, L. 2019. Evaluating Gender Bias in Machine Translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 28 July-2 August,

Florence, Italy, 1679-1684. DOI: 10.18653/v1/P19-1164

[19] Prates, M.O.R., Avelar, P.H., Lamb, L.C. 2020. Assessing gender bias in machine translation: a case study with Google Translate, Neural Comput & Applic, Vol. 32, p. 6363-6381. DOI: 10.1007/s00521-019-04144-6

[20] Subramanian, S., Han, X., Baldwin, T., Cohn, T., Frermann, L. 2021. Evaluating Debiasing Techniques for Intersectional Biases. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), 2492-2498.

[21] Basu Roy Chowdhury, S., Ghosh, S., Li, Y., Oliva, J., Srivastava, S., Chaturvedi, S. 2021. Adversarial Scrubbing of Demographic Information for Text Classification. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), 550-562.

[22] Bureau of Labor Statistics. https://www.bls.gov/soc/2018. (Access Time: 20 September 2019).

[23] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. Hierarchical Attention Networks for Document Classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June, San Diego, USA, 1480-1489. DOI: 10.18653/v1/N16-1174

[24] Sanh, V., Debut, L., Chaumond, L., Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. http://arxiv.org/abs/1910.01108

[25] Ruder, S. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. https://ruder.io/multi-task/. (Access Date: December 2019).

Appendix A.

Major Titles and Their Distribution

Major Titles	Male	Female
Arts, Design, Entertainment, Sports, and Media Occupations	2211	799
Management Occupations	555	99
Life, Physical, and Social Science Occupations	393	87
Community and Social Service Occupations	150	37
Legal Occupations	120	34
Healthcare Practitioners and Technical Occupations	108	28
Architecture and Engineering Occupations	82	8
Military Specific Occupations	67	4
Protective Service Occupations	55	7
Computer and Mathematical Occupations	47	10
Business and Financial Operations Occupations	40	13
Food Preparation and Serving Related Occupations	33	16
Educational Instruction and Library Occupations	30	25
Transportation and Metarial Moving Occupations	29	15
Protective Service Occupations	55	7
Sales and Related Occupations	22	1
Personal Care and Service Occupations	11	8
Office and Administrative Support Occupations	10	10
Farming, Fishing, and Forestry Occupations	6	2
Construction and Extraction Occupations	5	0
Installation, Maintenance, and Repair Occupations	3	1
Building and Ground Cleaning and Maintenance Occupations	2	0
Healthcare Support Occupations	0	1

Appendix B.

NLTK Stop Words Set

{'further', 'had', "won't", 'should', 'he', 'but', 'of', 'most', "wasn't", 'down', "wouldn't", 'and', 'll', 'doing', 'are', 'weren', 'theirs', 'them', 'all', 'why', 'any', 'what', 'off', 'below', 'his', 'was', "it's", 'under', 'him', 'don', 'her', 'wouldn', 're', "couldn't", 'will', "didn't", 'an', 'up', 'it', 'couldn', "shouldn't", 'yourself', 'isn', 'd', 'yours', 'ours', 'some', 'if', 'hadn', 'your', "mustn't", 'during', 'ma', 'mightn', 'has', "aren't", 'on', 'whom', 'didn', 'hers', 'myself', 'now', 'ourselves', 'each', 'into', 'ain', 'or', 'same', "hadn't", 'through', 'until', 'can', "you'd", 'at', "you're", 'too', 'more', 'how', 'me', 'in', 'mustn', 'after', "should've", 'where', 'were', 'between', 'my', 'both', 'their', 'just', 'is', 'with', 'then', 's', "weren't", 'its', 'they', 'herself', 'be', 'did', 'who', 'because', "don't", 'do', "you've", "she's", 'shan', 'before', 'few', "you'll", 'about', "that'll", 'only', 'does', 'other', 'our', 'himself', 'out', 'she', 'which', 'i', 'here', 'than', 'we', 'while', 'again', 'having', 'have', 'that', 'not', 'a', 'no', 'am', 'to', 'against', 'y', 'o', 'over', "doesn't", 'these', 'when', 'so', 've', 'those', 'won', 'being', 'itself', "isn't", 'yourselves', 'once', 'shouldn', 'been', 'from', 'there', 'for', 'very', 'own', 'you', 'this', 'such', "shan't", 'nor', 'by', 'as', 'wasn', "mightn't", 'the', 'hasn', 'above', 'haven', 'm', 'doesn', "haven't", "hasn't", 'aren', 'themselves', "needn't", 'needn', 't'}

Stop Words without Gender Indicators

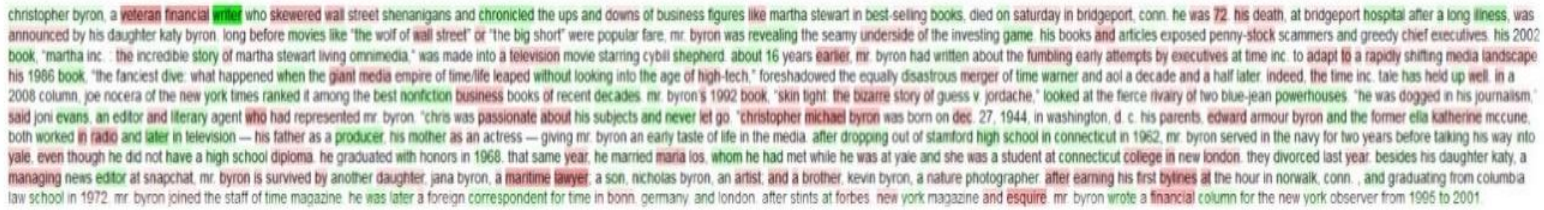
{'had', 'of', "wouldn't", 'and', 'll', 'doing', 'are', 'weren', 'theirs', 'any', 'what', 'off', 'below', 'was', 'wouldn', 'an', 'up', "shouldn't", 'yours', 'ours', 'some', 'if', 'your', 'during', "aren't", 'didn', 'now', 'ourselves', 'each', 'until', 'can', 'at', "you're", 'too', 'me', 'mustn', 'were', 'is', 'with', 'then', 's', "weren't", 'its', 'be', 'did', "you've", 'before', "you'll", 'about', 'other', 'out', 'we', 'again', 'having', 'that', 'not', 'am', 'o', 'over', "doesn't", 'when', 'so', 've', 'those', 'won', 'being', 'itself', 'once', 'shouldn', 'there', 'from', 'very', 'own', 'this', 'nor', 'as', 'wasn', 'the', 'hasn', 'above', 'haven', 'm', 'doesn', 'needn', 'further', "won't", 'should', 'but', 'most', "wasn't", 'down', 'them', 'all', 'why', 'under', "it's", 'don', 're', "couldn't", 'will', "didn't", 'it', 'couldn', 'yourself', 'isn', 'd', 'hadn', "mustn't", 'ma', 'mightn', 'has', 'on', 'whom', 'myself', 'ain', 'into', 'or', 'same', "hadn't", 'through', "you'd", 'more', 'how', 'in', 'after', "should've", 'where', 'between', 'my', 'both', 'their', 'just', 'they', 'who', 'because', "don't", 'do', 'shan', 'few', "that'll", 'only', 'does', 'our', 'which', 'i', 'here', 'than', 'while', 'have', 'a', 'no', 'to', 'y', 'against', 'these', "isn't", 'yourselves', 'been', 'for', 'you', 'such', "shan't", 'by', "mightn't", "haven't", "hasn't", 'aren', 'themselves', "needn't", 't'}

Appendix C.

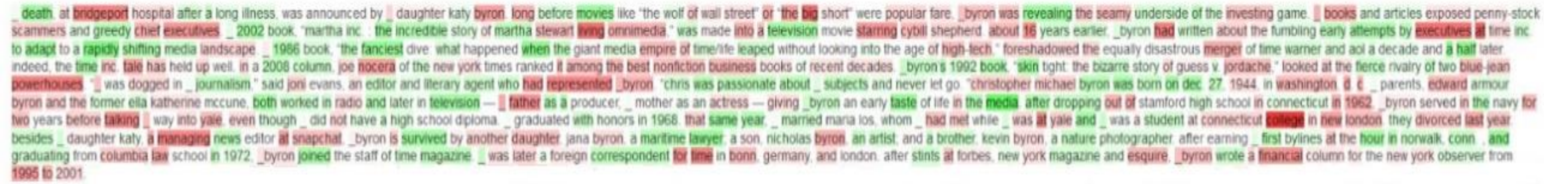
Distribution of Major Titles in Train and Test Sets

Major Titles	SVM		HAN & DistilBERT		
	Train	Test	Train	Val	Test
Architecture and Engineering Occupations	69	21	51	23	16
Arts, Design, Entertainment, Sports, and Media Occupations	2404	607	1806	595	609
Building and Grounds Cleaning and Maintenance Occupations	3	0	1	2	0
Business and Financial Operations Occupations	37	16	31	7	15
Community and Social Service Occupations	149	38	119	31	37
Computer and Mathematical Occupations	40	17	34	9	14
Construction and Extraction Occupations	4	1	3	1	1
Educational Instruction and Library Occupations	36	11	22	15	10
Farming, Fishing, and Forestry Occupations	9	3	6	3	3
Food Preparation and Serving Related Occupations	37	12	28	8	13
Healthcare Practitioners and Technical Occupations	108	28	77	25	34
Healthcare Support Occupations	0	1	0	0	1
Installation, Maintenance, and Repair Occupations	3	0	3	0	0
Legal Occupations	117	37	92	28	34
Life, Physical, and Social Science Occupations	398	82	285	102	93
Management Occupations	527	127	402	133	119
Military Specific Occupations	61	10	37	17	17
Office and Administrative Support Occupations	13	3	7	6	3
Personal Care and Service Occupations	16	3	13	3	3
Production Occupations	26	7	20	7	6
Protective Service Occupations	55	7	42	13	7
Sales and Related Occupations	26	4	18	9	3
Transportation and Material Moving Occupations	30	8	29	5	4

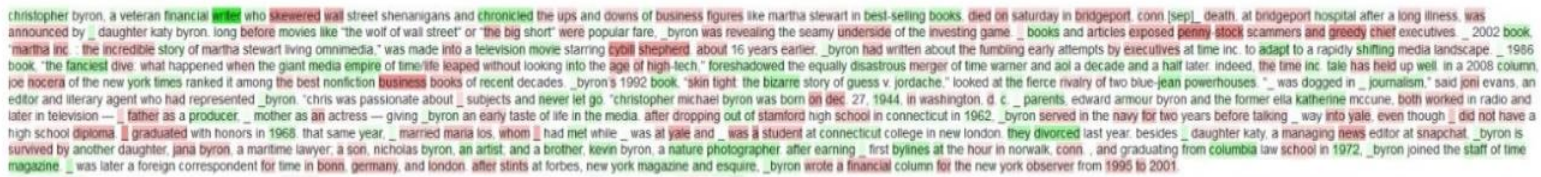
Appendix D.



Visualization of the Weight of Words for the Sample Article



Visualization of the Weight of Words for Scrubbed Article



Visualization of the Weight of Words for Scrubbed Article with its First Line