Bias Detection and Mitigation in Textual Data: a Study on Fake News and Hate Speech Detection

Apostolos Kasampalis, Despoina Chatzakou[0000−0002−9564−7100], Theodora Tsikrika[0000−0003−4148−9028], Stefanos Vrochidis[0000−0002−2505−9178], and Ioannis Kompatsiaris[0000−0001−6447−9020]

Information Technologies Institute, Centre for Research and Technology Hellas
{apkas,dchatzakou,theodora.tsikrika,stefanos,ikom}@iti.gr

Abstract. Addressing bias in NLP-based solutions is crucial to promoting fairness, avoiding discrimination, building trust, upholding ethical standards, and ultimately improving their performance and reliability. On the topic of bias detection and mitigation in textual data, this work examines the effect of different bias detection models along with standard debiasing methods on the effectiveness of fake news and hate speech detection tasks. Extensive discussion of the results draws useful conclusions, highlighting the inherent difficulties in effectively managing bias.

Keywords: Bias · NLP · Fake news detection · Hate speech detection

1 Introduction

Despite the undeniable benefits of Natural Language Processing (NLP) tools, the possible presence of bias (such as gender, race, cultural, and ideological bias) in such tools is a major issue with potentially negative impact on society, as it can lead to discrimination against certain social groups. Among others, such biases often emerge due to the inherent biases in the data used for developing the respective models. Therefore, bias detection in the data and bias mitigation in the derived NLP models are crucial to ensure their fairness, equity, and reliability.

Focusing on text-based data, *bias detection* methods range from rule-based heuristics [31] to neural network-based models that identify subtle biases; e.g., [10] delves into gender bias present in the English language, identifying instances of bias in naming, ordering, descriptions, metaphors, and the presence of gender-specific terms, shedding light on the deeply embedded nature of gender bias in language, while [21] develops a BERT-based text classifier for the detection of "media bias", defined as the unfair favoritism and reporting of certain ideas or viewpoints. Regarding *bias mitigation*, Sun et al. [24] provide a comprehensive review on gender bias mitigation in NLP, using e.g., gender swapping, whereby each word defined as male is swapped with its female equivalent and vice versa [11], also discussing the trade-offs and challenges associated with debiasing methods.

In this context, this work performs a comprehensive evaluation and analysis of commonly considered solutions for detecting as well as mitigating bias in

textual data, with particular interest in media and gender bias. The focus is on two case studies of NLP models for *fake news detection* and *hate speech detection*, given their significance in our society [2] as also illustrated by the various NLP models developed thus far for tackling these tasks, e.g., [28, 27].

In particular, this work examines four classification models for bias detection and two simple yet common methods for bias mitigation (namely gender swapping and data augmentation) on two datasets for fake news detection and on one dataset for hate speech detection. Overall, the results indicate that bias detection models perform well when applied to data from the same data source, but their performance drops significantly when this changes, even if the type of bias is the same (e.g. gender bias). The results also show that there was some bias in the employed datasets and its mitigation leads to a better performance for both the fake news and hate speech detection models, with data augmentation appearing to better address data bias compared to gender swapping.

2 Methodology

2.1 Bias Detection

Four bias detection models have been developed based on the following datasets:

- **MBIC** [22]: This dataset pertains to media bias; it contains 2,036 biased and 1,066 unbiased samples from online sources (Reuters, Fox News, HuffPost).
- **MBIB_workplace:** [30]: This dataset focuses on identifying gender bias in workplace-related content; it contains 624 biased and 513 unbiased samples.
- **MBIB_reddit** [30]: This dataset addresses gender bias on Reddit; it consists of 2,033 biased and 943 unbiased samples.
- **MBIB_twitter** [30]: This dataset consists of Twitter data that are potentially sexist; it contains 1,809 biased and 11,822 unbiased samples.

All datasets were divided into a train set (90%) (with 10% kept for validation) and a test set (10%).

Preprocessing. All aforementioned datasets were subjected to preprocessing to remove non-informative pieces of text, such as stopwords and special characters. In addition, tokenization and lowercasing were carried out.

Embedding layer. The first layer of the models' architecture is the embedding layer. Its purpose is to map each of the words in a sequence to a layer of higher dimension; we opted for GloVe [16] pre-trained embeddings of 100 dimensions.

Deep Neural Network Models. Overall, five neural network-based architectures were examined to develop effective classification models for bias detection based on the dataset at hand (in all cases *sigmoid* is used as activation function):

1. **Bidirectional Long Short-Term Memory based [biLSTM-based]**: one bi-LSTM layer of 64, spatial dropout of p=0.4 before, and a dropout of p=0.2 after the LSTM layer to reduce/avoid overfitting.

2. **Convolutional Neural Network based [CNN-based]**: three 1D convolutional (Conv1D) layers of 128 filters and kernel size of 3, 4, 5 respectively, two 1D average pool layer, one 1D global average pooling, and two spatial dropout layers before each of the first two Conv1D layers with p=0.2.
3. **Combined biLSTM and CNN [CNN_LSTM-based]**: one Conv1D layer of 16 filters and kernel size of 2, one 1D max pooling layer of pool size of 4, a spatial dropout layer of p=0.4 and lastly a biLSTM layer of 8 units.
4. **Gated Recurrent Unit based [GRU-based]**: one GRU layer of 8 units, spatial dropout of p=0.5 before and a dropout layer of p=0.5 after the GRU.
5. **Bidirectional Encoder Representations from Transformers based [BERT-based]**: the DistilBERT (uncased) pretrained model is used [5] followed by a dense layer of 128 units.

2.2 Bias Mitigation

Overall, two methods were considered for bias mitigation:

Gender swapping. A rule-based approach where gendered words in a sentence are swapped with the opposite gender. Inspired by [12, 29, 17, 15, 20], we applied two rules: (i) **[Rule#1]** *Gendered Words*: The swapping was performed based on compiled lists of gendered words [1, 4, 31] such as he-she, male-female, etc.; (ii) **[Rule#2]** *Gender-Neutral Words and Exception Cases* [29, 11, 6]: Examples of swapping would be "fireman" to "firefighter", while exceptions would be cases like "She is pregnant" where gender swapping to "He is pregnant" would be incorrect. We then applied the rules in two different ways: (i) **[M2F_swap]** male to female swapping, and (ii) **[F2M_swap]** female to male swapping.

Data augmentation [DA_swap]. This method essentially uses gender swapping to balance the dataset with regards to gender. Particularly, the original sentences are first gender swapped and then added to the dataset, effectively creating a larger dataset, with equal numbers of male and female gendered sentences. To avoid gender-swapped sentences becoming meaningless, but to also make the method more robust, names were anonymized; names were detected based on examples from [9], [13], [26] and the Gender-guesser [7] library.

2.3 Case Studies

For an in-depth evaluation of the bias detection and mitigation methods considered in this work, we focused on two general tasks:

Fake news detection. Two fake news classification models were built based on two popular datasets: (i) WELFake [19] which consists of 72,134 news articles with around $35k$ real and $37k$ fake news, and (ii) ISOT [3] which consists of around $21k$ real and $25k$ fake news articles.

Hate speech detection. A hate speech classification model was built on a set of data collected from the 4chan website and in particular from the Politically Incorrect Board [18]. Each entry in the dataset is characterized by its toxicity level on a [0, 1] scale; we considered as hate speech those with a toxicity level

above 0.4 and randomly selected 1,592 toxic and 6,069 non-toxic samples.

In both cases, the same preprocessing steps described in Section 2.1 were followed, while also the same neural network-based architectures were examined to determine the best performing one based on the dataset under consideration.

3 Experiments

3.1 Experimental Setup

For our experiments we use Keras [8] with TensorFlow [25] on a server equipped with one NVIDIA GeForce RTX 3080 of 10GB memory. For training, we use binary cross-entropy as loss function and AdamW [14] as optimizer. A maximum of 200 epochs is allowed, while also a validation set is used for early stopping; training is interrupted if the validation loss does not drop in 5 consecutive epochs. For evaluation purposes, standard metrics are considered, namely accuracy (Acc), precision (Prec), recall (Rec), and weighted area under the ROC curve (AUC). We repeated each experiment 5 times and report the average (AVG) values. Finally, we highlight the best results with bold.

3.2 Experimental Results

We first select the best performing neural network-based architecture for bias detection in each of the four datasets described in Section 2.1. Table 1 presents the results obtained with the best performing model in the test set of each of the four datasets; these are BERT-based for MBIC, biLSTM-based for MBIB_workplace, CNN-based for MBIB_reddit, and biLSTM-based for MBIB_twitter. Overall, we observe that a quite satisfactory performance is achieved in all cases, with the best results achieved for the most general MBIC dataset.

We then examine how well a model built to detect media bias in general can detect a more specific type of bias, namely gender bias, in the same context by applying the MBIC-based model on the three MBIB datasets. The results in Table 2, reporting also the standard deviation (STD) values, are particularly poor. An exception is the recall value of the true class, i.e. Rec (T), with a score above 82% in all cases, indicating that the general model detects the gender-biased documents quite successfully. Similar behavior was observed when a gender-biased model trained on data from platform x was tested on data collected from platform z (e.g., trained on data collected from Reddit and tested on Twitter-based

Table 1. Performance of the Bias detection models (average).

	MBIC	**MBIB_workplace**	**MBIB_reddit**	**MBIB_twitter**
Prec	91.26	83.56	82.23	85.41
Rec	89.06	83.41	79.28	88.61
Acc	91.00	83.33	83.56	93.65
AUC	89.06	83.41	79.28	88.61

Table 2. Performance of the media bias detection MBIC model on the MBIB datasets.

	MBIB_workplace		**MBIB_reddit**		**MBIB_twitter**	
	AVG	**STD**	**AVG**	**STD**	**AVG**	**STD**
Prec (T)	55.45	1.00	67.54	0.59	13.04	0.36
Rec (T)	82.50	12.09	90.92	7.57	95.72	6.44
Prec (F)	49.04	6.25	19.57	5.68	83.92	7.48
Rec (F)	19.30	12.30	5.90	5.91	2.42	3.60
Accuracy	53.99	1.95	63.98	3.38	16.87	5.04
AUC	50.90	1.63	48.41	1.14	49.07	1.42

Table 3. MBIC & MBIB_reddit on the Fake news classification task.

	Original	**M2F_swap**	**F2M_swap**	**DA_swap**	**Original**	**M2F_swap**	**F2M_swap**	**DA_swap**
	WELfake				**ISOT**			
	MBIC							
Prec	96.23	96.38	96.10	**97.20**	98.72	97.64	98.73	**99.25**
Rec	94.27	94.57	94.39	**95.65**	98.77	98.30	98.89	**99.55**
Acc	97.85	97.98	97.72	**99.10**	98.69	98.95	98.82	**99.09**
AUC	98.50	98.79	98.37	**99.44**	99.12	98.65	99.31	**99.75**
	MBIB_reddit							
Prec	96.23	96.23	96.19	**97.06**	98.72	97.64	99.04	**99.36**
Rec	94.27	94.77	94.51	**95.36**	98.77	98.30	98.43	**99.47**
Acc	97.85	98.18	97.96	**98.65**	98.69	98.95	98.82	**99.10**
AUC	98.50	98.99	98.21	**99.04**	99.12	98.65	98.78	**99.13**

data); the results are omitted due to space limits. This observation highlights the difficulty of this task, meaning that creating models that could be applied to detect similar types or the exact same type of bias, but on data collected from different sources (possibly with different inherent characteristics, such as text length, formal vs. informal way of writing, use of slang) is not always possible.

Next, we evaluate the effect of the bias detection and mitigation process on the fake news and hate speech detection tasks. In particular, the following steps are applied: (1) Apply classification models to detect fake news and hate speech in textual data (see Section 2.3); (2) Apply the four best performing bias detection models (Table 1) on the datasets used by the fake news and hate speech classification models; (3) For texts identified as bias-related, apply the two debiasing methods (Section 2.2); and (4) Re-build the fake news and hate speech classification models based on the updated sets of data: when gender swapping is used as the debiasing method, the updated dataset consists of the original texts detected as non-biased, along with the debiased texts, while for data augmentation the dataset includes both original and debiased texts.

For the fake news detection, Table 3 presents the results for both the WELfake and ISOT datasets. Due to space limitations, we only present results for the MBIC and MBIB_reddit bias detection models; for gender bias, a similar behavior is observed for the rest of the models (MBIB_workplace and MBIB_twitter). The results indicate that the overall performance improves in most cases, even slightly, when debiasing methods are applied, with the best results observed when data augmentation is used as the debiasing method. Regarding gender swapping,

Table 4. MBIC & MBIB_reddit on the Hate speech classification task.

	Original	**M2F_swap**	**F2M_swap**	**DA_swap**
	MBIC			
Prec	88.41	88.22	88.41	**89.10**
Rec	93.65	93.57	93.65	**94.58**
Acc	93.09	92.96	93.09	**93.68**
AUC	93.66	93.57	93.66	**94.24**
	MBIB_reddit			
Prec	88.41	88.56	88.56	**89.38**
Rec	93.65	93.95	93.95	**94.84**
Acc	93.09	93.22	93.22	**94.07**
AUC	93.66	93.95	93.95	**94.45**

we observe that M2F swapping yields better performance compared to F2M in the case of the WELfake dataset, indicating that this particular dataset is probably more female-biased. An opposite behavior is observed in the case of the ISOT dataset, indicating that its data is probably more biased towards males. Table 4 presents how the performance of the hate speech classification model changes after data debiasing. Overall, similar performance is observed to the fake news classification task, with data augmentation resulting in improved performance.

Discussion. Overall, classification models are widely used to detect bias in textual data as they constitute an easy-to-understand approach, requiring though a sufficient amount of properly annotated data that reflect the bias we want to detect. Thus, although they appear to be a promising solution, they can only help in an effective bias detection if the annotated datasets used are large enough, trustworthy, and of good quality in terms of syntax/content. But even in such a case, their generalization is not always possible as highlighted above.

The application of bias detection models along with the data debiasing methods leads overall to improved results for both fake news and hate speech classification tasks. According to Park et al. [15] and Zhao et al. [31], *gender-swapping* has been found to be a quite effective solution, leading to improved classification results after the models' retraining. The better performance of the *data augmentation* could be attributed to different factors. First, as the dataset is augmented by including data presenting the same content as a function of both genders, the retrained model is forced to focus on more gender-neutral features and avoid making predictions based on gender-related information. Moreover, the mere fact of increasing the size of the dataset has been shown to lead to improved performance [23], particularly for neural network-based solutions. Finally, we should mention that even if bias may be addressed at the data level, it may remain to some extent in the classification models, due to the fact that pretrained embeddings were used, where bias is inherent. Therefore, more holistic solutions are required to address bias more effectively in NLP-based models.

4 Conclusions

This work performed a comprehensive evaluation of (media/gender) bias detection and mitigation in textual data on NLP models for fake news and hate speech detection. Overall, the proposed bias detection and mitigation process appears to have a positive effect on the effectiveness of fake news and hate speech classification models, although the generalization of specific bias detection models to other types of bias is rather poor. These findings could act as a stepping stone for future bias management studies, including using larger datasets, further optimizing the models, and debiasing the word embeddings themselves.

Acknowledgements This project has received funding from the European Union's H2020 research and innovation programme as part of the STARLIGHT (GA No 101021797) project.

References

1. Biased words: `https://github.com/gregology/biased-words`, Accessed: 2023
2. Blanco-Herrero, D., Sánchez-Holgado, P.: Fake news and hate speech: who is to blame? Study of the perceptions of Spanish citizens about the actors responsible for the production and spread of fake news and hate speech. In: Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21). pp. 448–451 (2021)
3. Clement Bisaillon: `https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset`, Accessed: 2023
4. Debiaswe: try to make word embeddings less sexist: `https://github.com/tolga-b/debiaswe/tree/master/data`, Accessed: 2023
5. DistilBERT base model (uncased): `https://huggingface.co/distilbert-base-uncased`, Accessed: 2023
6. Doughman, J., Khreich, W., El Gharib, M., Wiss, M., Berjawi, Z.: Gender bias in text: Origin, taxonomy, and implications. In: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing. pp. 34–44 (2021)
7. Gender Guesser: `https://github.com/lead-ratings/gender-guesser`, Accessed: 2023
8. Keras: `https://keras.io/` (2020)
9. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for Named Entity Recognition. arXiv preprint arXiv:1603.01360 (2016)
10. Leavy, S.: Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In: Proceedings of the 1st international workshop on gender equality in software engineering. pp. 14–16 (May 2018)
11. Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A.: Gender bias in neural natural language processing. Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday pp. 189–202 (2020)
12. Manzini, T., Lim, Y.C., Tsvetkov, Y., Black, A.W.: Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. arXiv preprint arXiv:1904.04047 (2019)
13. NER Tagger: `https://github.com/glample/tagger`, Accessed: 2023

14. OverLordGoldDragon: Keras adamw. GitHub. Note: https://github.com/OverLordGoldDragon/keras-adamw/ (2019)
15. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231 (2018)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
17. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with Google Translate. Neural Computing and Applications **32**, 6363–6381 (2020)
18. Raiders of the Lost Kek: `https://zenodo.org/records/3606810#.YH2TYCXivIU`, Accessed: 2023
19. Saurabh Shahane: `https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification`, Accessed: 2023
20. Seaborn, K., Chandra, S., Fabre, T.: Transcending the "male code": Implicit masculine biases in NLP contexts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2023)
21. Spinde, T., Plank, M., Krieger, J.D., Ruas, T., Gipp, B., Aizawa, A.: Neural media bias detection using distant supervision with BABE – bias annotations by experts. arXiv preprint arXiv:2209.14557 (2022)
22. Spinde, T., Rudnitckaia, L., Sinha, K., Hamborg, F., Gipp, B., Donnay, K.: MBIC – a media bias annotation dataset including annotator characteristics. arXiv preprint arXiv:2105.11910 (2021)
23. Stylianou, N., Chatzakou, D., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Domain-aligned data augmentation for low-resource and imbalanced text classification. In: European Conference on Information Retrieval. pp. 172–187. Springer (2023)
24. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Wang, W.Y.: Mitigating gender bias in Natural Language Processing: Literature review. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1630–1640 (July 2019)
25. TensorFlow: `https://www.tensorflow.org/`, Accessed: 2023
26. Transition-based NER system: `https://github.com/clab/stack-lstm-ner`, Accessed: 2023
27. del Valle-Cano, G., Quijano-Sánchez, L., Liberatore, F., Gómez, J.: SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles. Expert Systems with Applications **216**, 119446 (2023)
28. Verma, P.K., Agrawal, P., Amorim, I., Prodan, R.: WELFake: word embedding over linguistic features for fake news detection. IEEE Transactions on Computational Social Systems **8**(4), 881–893 (2021)
29. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., Shieber, S.: Causal mediation analysis for interpreting neural nlp: the case of gender bias (2020). CoRR arXiv (2004)
30. Wessel, M., Horych, T., Ruas, T., Aizawa, A., Gipp, B., Spinde, T.: Introducing MBIB - the first media bias identification benchmark task and dataset collection. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2765–2774 (2023)
31. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876 (2018)