Signal, Image and Video Processing
https://doi.org/10.1007/s11760-022-02368-z

ORIGINAL PAPER

The effect of gender bias on hate speech detection

Furkan Şahinuç[1] · Eyup Halit Yilmaz[1] · Cagri Toraman[1] · Aykut Koç[2,3]

Received: 26 August 2022 / Revised: 28 August 2022 / Accepted: 18 September 2022

Abstract
Hate speech against individuals or communities with different backgrounds is a major problem in online social networks. The domain of hate speech has spread to various topics, including race, religion, and gender. Although there are many efforts for hate speech detection in different domains and languages, the effects of gender identity are not solely examined in hate speech detection. Moreover, hate speech detection is mostly studied for particular languages, specifically English, but not low-resource languages, such as Turkish. We examine gender identity-based hate speech detection for both English and Turkish tweets. We compare the performances of state-of-the-art models using 20 k tweets per language. We observe that transformer-based language models outperform bag-of-words and deep learning models, while the conventional bag-of-words model has surprising performances, possibly due to offensive or hate-related keywords. Furthermore, we analyze the effect of debiased embeddings for hate speech detection. We find that the performance can be improved by removing the gender-related bias in neural embeddings since gender-biased words can have offensive or hateful implications.

Keywords Debiased embedding · Deep learning · Gender identity · Hate speech · Language model

1 Introduction

With the expansion of online social networks, users not only socialize but also read news, and express their opinions on any subject. Unfortunately, in parallel with the increasing popularity of these platforms, such networks can exhibit undesirable usage patterns as well. For instance, some users can spread misinformation for manipulation, generate spam-type content with bot accounts, and use hate speech language against individuals or communities with different backgrounds.

✉ Cagri Toraman
ctoraman@aselsan.com.tr

Furkan Şahinuç
fsahinuc@aselsan.com.tr

Eyup Halit Yilmaz
ehyilmaz@aselsan.com.tr

Aykut Koç
aykut.koc@bilkent.edu.tr

[1] Aselsan Research Center, 06200 Ankara, Turkey

[2] Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey

[3] The National Magnetic Resonance Research Center, Bilkent University, 06800 Ankara, Turkey

The amount of hate speech is increasing daily, as reported by a popular micro-blog platform, Twitter[1]. Hateful or offensive tweets are usually directed at a particular individual or community with different background or characteristics. The domain of hate speech often includes controversial topics such as religion, gender identity (or sexual orientation), ethnicity, politics, or sports.

Various algorithms, such as word match-based [1], word vectors-based [2] or language model-based [3–5], are proposed to detect such discourses. A major drawback experienced by hate speech detection algorithms is that the datasets in the literature are not balanced in terms of hate speech instances. A possible reason can be the regulations by social media platforms on hateful posts. For this reason, studies have been carried out to meet the need for the datasets in various languages [6].

Another handicap of conducted hate speech studies so far is that different domains of hate speech are not satisfactorily considered. The dynamics behind hate speech change correspondingly in different domains. For example, an expression may be considered offensive in one domain but normal in another. Therefore, hate speech should not be dealt with independently from its domain. When the studies focusing on

[1] https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec.

Published online: 08 October 2022

gender in NLP are considered, gender identity-based hate speech reveals potential research questions worthy of attention. In particular, we focus on the impact of gender bias in language models on hate speech detection. It is previously shown that language models incorporate gender bias [7–9]. For instance, language models associate some gender-neutral occupations with males or females. The presence of biases in hateful expressions, as in occupations, may affect the performance of language models in hate speech detection tasks. Therefore, we investigate the effect of the debiasing algorithms on gender identity-based hate speech detection performance.

Moreover, hate speech detection is mostly studied for certain languages, specifically English, but not low-resource languages, such as Turkish. Therefore, this study examines gender identity-based hate speech detection for both English and Turkish. We compare the performances of state-of-the-art models using 20k tweets per language.

The contributions of this study are twofolds. First, we show that Transformer-based language models outperform bag-of-words and deep learning models in gender identity-based hate speech detection for both English and Turkish. Second, we find that gender identity-based hate speech detection can be improved by removing the gender bias in neural embeddings since gender-biased words can have offensive or hateful implications.

2 Related work

2.1 Hate speech detection

Initial studies for detecting undesirable content in social media utilize external resources (e.g., lexicons) and are based on keyword matching [1]. In addition to using lexical sources, extracting the features of the social media users can also effectively disclose hateful and offensive patterns in the text [10–12]. However, after the emergence of the Transformer architecture [13], encoder-based language models started to outperform previous models [3,4]. On the other hand, there are other approaches to detect hateful and offensive speech. For example, in [14], graph auto-encoders (GAE) are utilized to obtain representations of the text. GAE takes both the text feature matrix and the graph matrix as input and encodes information in an unsupervised manner.

Despite the limited resources, there are important pioneering studies on Turkish hate speech detection. Berk et al. examine the incendiary news detection problem [15]. To this end, they train Linear Support Vector Machine, Naive Bayes, and Multilayer Perceptron models using fastText [16] word representations and compare them with Bag of Words based representations [15]. On the other hand, one of the most recent studies about Turkish hate speech is [17]. Authors compile a large hate speech dataset for both Turkish and English. There are 100k instances for each language, and each data instance belongs to different possible hate domains such as gender, race, religion, politics, and sports. The authors also examine the cross-domain transfer performance of the language models. Another Turkish hate speech dataset is curated in [18]. Unlike other hate speech datasets, this dataset considers fine-grained hate levels in labeling such as insult, humiliation, dehumanizing, and threat.

2.2 Bias in language models

Language models may unintentionally host several types of bias, such as gender and ethnicity. Such biases may stem from the datasets on which language models are trained. One of the initial studies that aim for removing bias from language models proposes to debias word embeddings [7]. The proposed method is based on creating a gender subspace first, then subtracting projections of the gender-neutral words on the created gender subspace. Caliskan et al. also propose Word Embedding Association Test (WEAT) to reveal different types of biases existing in word embeddings [8]. On the other hand, Gonen and Goldberg [9] assert that existing debiasing methods may not be sufficient to eliminate the bias from word embeddings. Furthermore, they show that biased word embeddings still cluster according to their bias even after debiasing operation [9]. Finally, since language models' bias affects the language models' fairness, there are also studies tackling the problem in the legal domain [19].

As the static word embeddings evolve into contextual sentence representations, measuring the bias of contextual models emerges as a new issue. To solve this problem, May et al. extend the WEAT so that bias inside sentence encoders can be quantified [20]. A similar procedure is also followed in [21]. In parallel with measuring the bias of the contextual models, there are also efforts to mitigate it. For instance, in [22], Counterfactual Data Substitution (CDS), which is an attempt to invert gender-specific words while maintaining the grammatical coherence is applied to Gendered Ambiguous Pronouns (GAP) corpus [23]. Modified GAP corpus is used in additional training of BERT [24] model to counterbalance bias. In another study, a contextualized version of the hard-debiasing algorithm of [7] is applied to sentence encoder representations [25].

3 Methodology

3.1 Conventional models

The problem of hate speech detection on social media can be addressed with supervised training algorithms. We apply a conventional classification model, SVM Classifier,

to measure the baseline detection performance. We use Bag-of-Words representation to calculate the Term Frequency-Inverse Document Frequency scores for the tweets and train the SVM model. The main strength of this approach is the ability to capture keyword-specific patterns that may expose hate speech.

3.2 Deep learning models

Another tool we use for hate speech detection is fastText embeddings [26]. The fastText model provides an embedding for each word in the vocabulary. However, we need an overall sentence representation while classifying the word sequences. To this end, we first apply mean pooling on the individual embeddings of the words in the tweet to obtain an overall sentence embedding. Then, the obtained sentence embedding is fed to the linear classifier to predict the correct hate label. For the implementation, the text classification library of fastText is utilized [16].

We also apply Bidirectional Long Short-Term Memory (BiLSTM) using the fastText embeddings. We first perform a left-to-right pass with a stack of LSTM layers and a right-to-left pass with another stack. Then, we concatenate the hidden representations at the end of each pass and apply fully connected layers. Such modeling allows us to capture sequential dependencies that might exist in two directions in the given text.

3.3 Transformer-based language models

Transformer architecture consists of encoder and decoder parts for seq2seq training schemes [13]. However, these parts can also be utilized separately. The encoder part is more suitable for text classification tasks in general. This study focuses on the BERT [24] model consisting of a bidirectional encoder structure. The model consists of an embedding layer, an output projection layer, and several encoder blocks. Important steps of the architecture are given as follows:

$$\begin{aligned} X &= EmbeddingLayer(Tokenized\ Input), \\ \hat{X}_i &= LayerNorm(X_i + Attention(X_i)), \\ X_{i+1} &= LayerNorm(\hat{X}_i + FeedForw(\hat{X}_i)), \\ Y &= OutputProj(X_{N+1}), \end{aligned} \tag{1}$$

where X represents the input embeddings and the input of the first encoder block. The input of the first encoder is the tokenized version of the given sentence or sequence. These input tokens can be either words or subwords depending on the tokenization algorithm. An encoder block consists of an attention layer, a feed-forward layer, and two layer normalization after these layers. There are N encoder blocks in the model. X_i represents the input of the ith encoder block or the output of the $(i-1)$th encoder block. Y stands for the output representations for given input tokens.

In general, training of the transformer-based language models consists of two stages, namely, pretraining and fine-tuning [24]. While fine-tuning BERT, we use the sequence classification scheme. We need a representation of the overall sequence in such a training strategy. To this end, we use $C \in Y$ embedding, corresponding to the $[CLS]$ token of the input sequence. In order to use this embedding, an additional linear layer whose weights are $W^{K \times H}$ is employed. Here, K stands for the number of labels, and H represents the size of the hidden state (i.e., size of C). Then, the standard classification loss is calculated with $log(softmax(CW^T))$.

3.4 Debiased models

To observe the effect of debiasing language models on hate speech detection in the gender domain, we compare the biased and debiased versions of one deep learning and one transformer-based language model. In Fig. 1, a general scheme for debiasing and hate speech detection is presented.

Since fastText consists of the conventional word embeddings, we apply hard-debiasing proposed in [7] on fastText

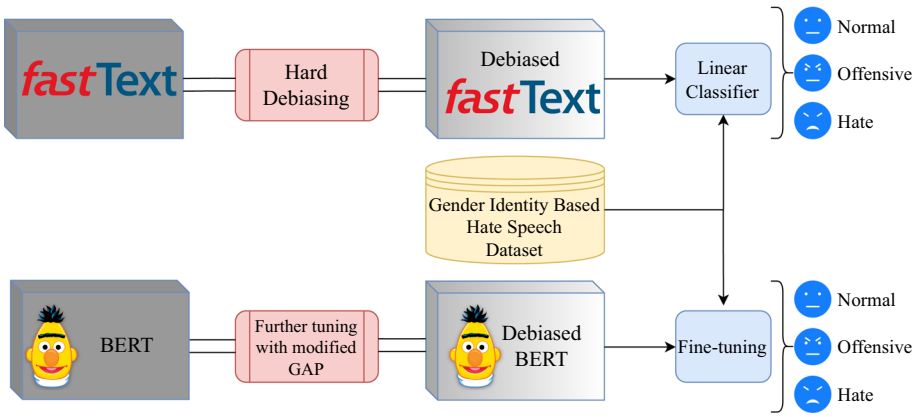

Fig. 1 Illustration of the methodology followed for how debiased language models are used in hate speech detection

word embeddings [26]. In the hard-debiasing scheme, a gender direction vector (e.g., $\overrightarrow{she} - \overrightarrow{he}$) is used as a gender subspace. In the neutralization step of the hard-debiasing, words that are supposed to be gender neutral are made orthogonal to the gender subspace. In the equalize phase, the gender-neutral words are made equidistant to the pairs in the equality sets. For example, the distance between the word "doctor" and the words of the pair $\{man, woman\}$ becomes equal.

In order to follow the same debiasing strategy in Turkish, some adjustments are necessary. For example, $\overrightarrow{she} - \overrightarrow{he}$ vector is used as a gender direction vector. However, Turkish pronouns do not contain any gender information. This problem is tackled in [27], and gender-specific words for measuring and debiasing Turkish word embeddings are proposed. For instance, $\overrightarrow{kadın} - \overrightarrow{adam}$ ($\overrightarrow{woman} - \overrightarrow{man}$ in English) is used as the gender direction vector. We follow the same procedure for debiasing Turkish fastText embeddings.

To mitigate the gender bias in the BERT model, we follow the procedure proposed in [22]. This method is based on counterbalancing the existing bias by additional training with a balanced dataset regarding male and female entities. To this end, the GAP dataset is utilized. This dataset includes sentences to resolve gendered ambiguous pronouns. Gender entities in the GAP dataset are swapped by Counterfactual Data Substitution (CDS). In other words, female pronouns are swapped with male pronouns and vice versa. Then, the BERT model is tuned according to Masked Language Modeling (MLM) with the modified GAP dataset to counterbalance the gender bias. After this additional tuning operation, the resulting BERT model is used in the hate speech detection task in the gender domain.

To observe the effect of the debiasing operation on hate speech detection, we need to compare the debiased and fine-tuned model with the biased fine-tuned model. At this point, one may think that a direct comparison between the biased and debiased models would not be a fair comparison. Since debiasing operation includes an additional training phase, the debiased model learns more knowledge via MLM than the biased (regular BERT base) model. This may affect the performance of the debiased model. In order to make a fair comparison, we subject the BERT-base model to the same additional training scheme as the original (unmodified) GAP dataset. In other words, both models can learn additional information from the GAP dataset only with the difference in CDS processing. Hate speech fine-tuning of the biased model is implemented as explained above.

The methodology we follow for debiasing the English Transformer encoder is not entirely compatible with Turkish. There is no Turkish version of the GAP corpus. Furthermore, translation of the GAP corpus is not suitable for this task.

Table 1 The dataset statistics

Statistics	Turkish	English
# of Normal Tweets	10,699	12,256
# of Offensive Tweets	6521	6431
# of Hate Tweets	2780	1313
Total	20,000	20,000
Avg. # of Words	25.26	28.50
Longest Tweet Length	121	67
Shortest Tweet Length	5	5
# of Tweets w/Hashtag	7870	2733

Since modification of the GAP corpus via CDS is based on gendered pronouns, the structure of Turkish does not allow us to implement the same additional debiasing training. Since the main objective of this study is to observe the effect of gender debiasing on hate speech detection, not to develop a debiasing algorithm for Turkish models, we leave this task to future work.

4 Experiments

4.1 Dataset

We conduct experiments on a subset of the dataset from [17] that only contains tweets under the "gender" topic. The dataset consists of 20,000 English and 20,000 Turkish tweets labeled hate, offensive, or normal by five annotators. We focus on the gender topic to assess potential bias's effect on the downstream hate speech detection performance. The dataset statistics are given Table 1.

To prevent the dataset from being biased by a certain opinion or point of view, the number of tweets from a single user does not exceed 1% of the total number of tweets. Similarly, there is at most %80 cosine similarity between TF-IDF vectors of the tweets to avoid duplicates. All tweets contain at least five words without hashtags and URLs.

4.2 Experimental design

Since the gender identity-based hate speech dataset from [17] has no default train and test splits, we prepare our train and test sets. In order to validate our experimental results, we construct ten different train and test splits. In each split, the number of the training instances constitutes 90% of all instances in the dataset. The remaining 10% is used for testing. There is no intersection among the different test splits. Therefore, in each experiment, the same ten splits are used.

In the implementation of models, the SVM radial basis function kernel is applied with the C value as 1 and squared

Table 2 The average of the 10-fold weighted $F1$ scores for each method

Lang	BERT	fastText	LSTM	SVM
English	**0.797**	0.712	0.766	0.753
Turkish	**0.769**	0.655	0.654	0.663

The bold that indicates the highest score for each language

L2 regularization. The hyperparameters for the LSTM model are as follows. The hidden dimension is chosen as 512. The number of layers is 4; a stack of 4 LSTM layers is used. There are two fully connected layers following LSTM output, which are of size 128. The LSTM output is passed to a GELU activation function, and a ReLU activation function follows the fully connected layers.

For the conventional fastText classification model, the learning rate and the number of epochs are 0.01 and 10, respectively. At the initial step, each language's corresponding pre-trained embeddings are used. For transformer-based models, BERT-base-uncased and BERTurk-base-uncased models are used for English and Turkish, respectively. These models are trained along three epochs with $1e-5$ learning rate. The same hyperparameters are applied for biased and debiased models. Model performances are measured by weighted precision, recall and $F1$ scores.

4.3 Experimental results

4.3.1 Model comparison

We assess each method's hate speech detection performance in terms of the weighted $F1$ score and report them in Table 2. BERT achieves the highest score in both English and in Turkish. Bag-of-words (BOW) representation (used in SVM Classifier) achieves strong results such that it is the second best method in Turkish and better than fastText in English. Since BOW picks up on keyword structures, hate speech might correlate with certain keywords, with a higher predominance in Turkish.

4.3.2 Effect of debiased models

The performances of the biased and debiased models are given in Table 3 for English and Turkish in terms of weighted precision, recall, and $F1$ score metrics. Debiasing operation increases the performance of the English debiased BERT model and Turkish fastText model. However, hard-debiasing results in a slight decrease in the performance of the English fastText model. One possible reason for the different outcomes of hard-debiasing the word embeddings can be the difference between the construction of gender subspaces in Turkish and English. In debiasing English word embeddings,

Table 3 The average of weighted precision, recall and $F1$ scores of biased and debiased models for English and Turkish hate speech detection task

Lang	Model	Pre	Rec	$F1$
EN	fastText	0.706	0.718	0.712
EN	Debiased fastText	0.706	0.717	0.709
EN	BERT	0.797	0.797	0.797
EN	Debiased BERT	0.809	0.806	0.807
TR	fastText	0.650	0.663	0.655
TR	Debiased fastText	0.662	0.673	0.663
TR	BERTurk	0.769	0.774	0.769
TR	Debiased BERTurk	–	–	–

the gender subspace consists of $\overrightarrow{she} - \overrightarrow{he}$ vector. As mentioned, Turkish pronouns do not include any information about masculinity or femininity. Therefore, $\overrightarrow{kadın} - \overrightarrow{adam}$ ($\overrightarrow{woman} - \overrightarrow{man}$ in English) is used for gender direction as a correspondence to $\overrightarrow{she} - \overrightarrow{he}$. Similar cases also exist in constructing the equality pairs for equalizing step of the hard-debiasing.

On the other hand, the debiasing operation for the BERT model enhances the model performance in detecting gender-based hate speech. This emphasizes the value of a balanced and unbiased corpus. When the MLM training scheme is considered, the gender-neutral concepts must equally occur in the context of words with gender information. Another point emphasizing the importance of unbiased data is the structure of the current state-of-the-art language models. Generally, transformer-based language models use encoder blocks to learn the contextual information. These blocks consist of the attention-layer, feed-forward layer and the normalization between these layers. Removing the bias by manipulating such nested and sizable models may not be a feasible solution. In such cases, feeding the model with balanced and unbiased data rather than post-training debiasing may be a more suitable solution.

To have a better insight about how and in which way debiasing operation affects hate speech detection, we scrutinize some sentences whose labels are incorrectly predicted by biased models and correctly predicted by debiased models. For the debiased Turkish fastText model, the sentence *"Kadın mücadelesini örgütlü hale getirmeyi amaçlayan konferansta, feminist kadın grevinin hazırlığı yapılıyor"* [*At the conference, which aims to organize the women's struggle, preparations for the feminist women's strike are being made*] is predicted as normal speech while the biased model mispredicts it as hate speech. One possible reason for this situation is that some gender-specific word vectors are biased toward some hateful and offensive word vectors due to the large cor-

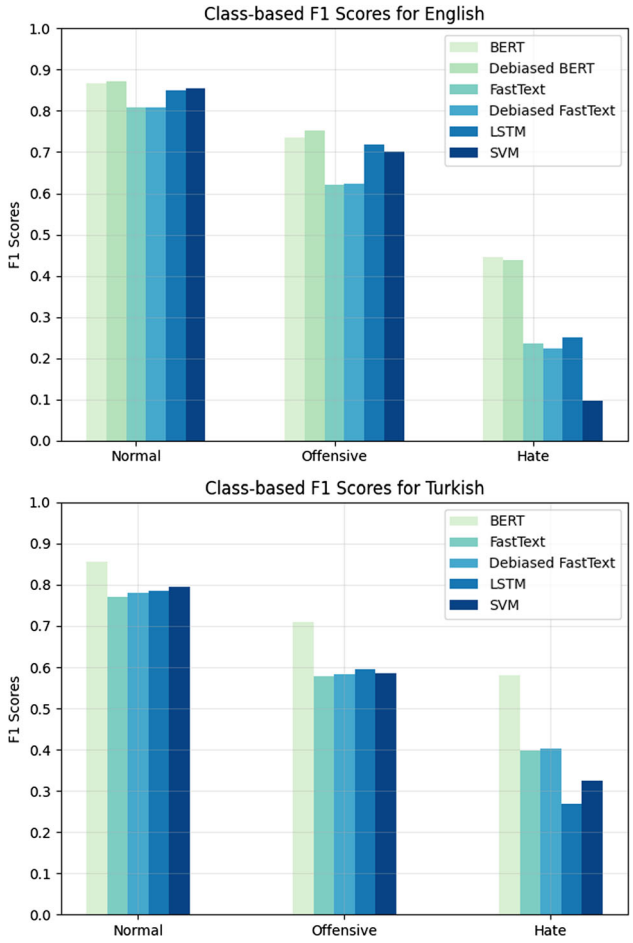

Fig. 2 The average $F1$ scores of 10-folds according to each class. Hate class has an overall lower detection score than other classes. Debiased BERT is not available in Turkish

pora on which the initial fastText model is trained. Recall that gender-specific words are made equidistant to gender-neutral words in equalizing step of hard-debiasing. This may increase the distance between some gender-specific words (e.g., feminist) and hateful (or offensive) words. Therefore, the model would be more fairly initialized.

4.3.3 Performance on hate classes

We analyze the performance of the models according to each class, given in Fig. 2. We observe that the overall performance of the hate class is lower than other classes for every model. This might be attributed to the fact that the number of hate instances is smaller than other classes in the dataset. All models except LSTM achieve a higher $F1$ score for the hate class in Turkish. The higher detection performance might hint that hate speech in Turkish might be easier, perhaps due to the higher number of hate instances in the training data. The debiased versions of BERT and fastText increase performance with respect to each class.

5 Conclusion

In this study, we examine gender identity-based hate speech for English and Turkish languages by comparing the performances of state-of-the-art models using 20k tweets per language. We observe that transformer-based language models outperform bag-of-words and deep learning models. In contrast, the conventional bag-of-words model has surprising performances, possibly due to offensive or hate-related keywords. Furthermore, we find that the performance can be improved by removing the gender-related bias in neural embeddings since gender-biased words can have offensive or hateful implications. In future work, we plan to extend our experiments to different languages and develop novel algorithms that remove bias in transformer-based language models, specifically in low-resource languages such as Turkish.

Author Contributions FŞ came up with the main idea. Experiments were implemented by FŞ and EHY. CT and AK played significant role in determining the general structure of the article. All authors reviewed the manuscript.

Funding No funding is received for this study.

Data Availability This declaration is not applicable for this study.

Declarations

Conflict of interest The authors have no competing interests.

Ethical approval This declaration is not applicable for this study.

References

1. Sood, S., Antin, J., Churchill, E.: Profanity use in online communities. In: Proceedings of SIGCHI Conference on Human Factors in Computer System, pp. 1481–1490 (2012). https://doi.org/10.1145/2207676.2208610
2. Nobata, C., et al.: Abusive language detection in online user content. In: Proceedings of WWW, pp. 145–153 (2016). https://doi.org/10.1145/2872427.2883062
3. Liu, P., Li, W., Zou, L.: NULI at SemEval-2019 Task 6: transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of 13th International Workshop on Semantic Evaluation, pp. 87–91 (2019). https://doi.org/10.18653/v1/S19-2011
4. Caselli, T., Basile, V., Mitrović, J., Granitzer, M.: HateBERT: retraining BERT for abusive language detection in English. In: Proceedings of 5th Workshop on Online Abuse and Harms, pp. 17–25 (2021). https://doi.org/10.18653/v1/2021.woah-1.3
5. Mathew, B., et al.: HateXplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14867–14875 (2021)
6. Poletto, F., et al.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**(2), 477–523 (2021). https://doi.org/10.1007/s10579-020-09502-8

7. Bolukbasi, T., et al.: Man is to computer programmer as woman is to homemaker? In: NeurIPS, Debiasing word Embeddings (2016)
8. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017). https://doi.org/10.1126/science.aal4230
9. Gonen, H., Goldberg, Y.: Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: NAACL, pp. 609–614 (2019). https://doi.org/10.18653/v1/N19-1061
10. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: NAACL Student Research Workshop, pp. 88–93 (2016). https://doi.org/10.18653/v1/N16-2013
11. Chatzakou, D., et al.: Mean birds: Detecting aggression and bullying on Twitter. In: Proceedings of the ACM Web Science Conference, pp. 13–22 (2017). https://doi.org/10.1145/3091478.3091487
12. Unsvåg, E.F., Gambäck, B.: The effects of user features on Twitter hate speech detection. In: Proceedings of the Second Workshop on Abusive Language Online, pp. 75–85 (2018). https://doi.org/10.18653/v1/W18-5110
13. Vaswani, A., et al.: Attention is all you need. In: NeurIPS, vol. 30, pp. 5998–6008 (2017)
14. De la Peña Sarracén, G.L., Rosso, P.: Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. In: LREC, pp. 2196–2204 (2022)
15. Berk, E.A., Filatova, E.: Incendiary news detection. In: International FLAIRS Conference (2019)
16. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: EACL, pp. 427–431 (2017)
17. Toraman, C., Şahinuç, F., Yilmaz, E.: Large-scale hate speech detection with cross-domain transfer. In: LREC, pp. 2215–2225 (2022)
18. Beyhan, F., et al.: A Turkish hate speech dataset and detection system. In: LREC, pp. 4177–4185 (2022)
19. Sevim, N., Şahinuç, F., Koç, A.: Gender bias in legal corpora and debiasing it. Nat. Lang. Eng. (2022). https://doi.org/10.1017/S1351324922000122
20. May, C., et al.: On measuring social biases in sentence encoders. In: NAACL, pp. 622–628 (2019). https://doi.org/10.18653/v1/N19-1063
21. Kurita, K., et al.: Measuring bias in contextualized word representations. In: Proceedings of the Third Workshop on Gender Bias in Natural Language, pp. 166–172 (2019). https://doi.org/10.18653/v1/W19-3823
22. Bartl, M., Nissim, M., Gatt, A.: Unmasking contextual stereotypes: measuring and mitigating BERT's gender bias. In: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pp. 1–16 (2020)
23. Webster, K., Recasens, M., Axelrod, V., Baldridge, J.: Mind the GAP: a balanced corpus of gendered ambiguous pronouns. Trans. ACL **6**, 605–617 (2018). https://doi.org/10.1162/tacl_a_00240
24. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
25. Liang, P.P., et al.: Towards debiasing sentence representations. In: ACL, pp. 5502–5515 (2020). https://doi.org/10.18653/v1/2020.acl-main.488
26. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. ACL **5**, 135–146 (2017). https://doi.org/10.1162/tacl_a_00051
27. Sevim, N., Koç, A.: Investigation of gender bias in Turkish word embeddings. In: Proceedings of SIU, pp. 1–4 (2021). https://doi.org/10.1109/SIU53274.2021.9477774