Gender bias detection on hate speech classification: an analysis at feature-level

NASCIMENTO, Francimaria R. S., CAVALCANTI, George D. C. and COSTA-ABREU, Marjory Da <http://orcid.org/0000-0001-7461-7570>

Appendix A Supplementary results

This Section presents the mean and the standard deviation of all results described in Section Experimental Results.

Table A1: Results obtained using FNED bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.041 ± 0.000	0.223 ± 0.000	0.025 ± 0.000	0.000 ± 0.000	0.204 ± 0.000	0.111 ± 0.000
TF-IDF	0.142 ± 0.000	0.227 ± 0.000	0.154 ± 0.000	0.000 ± 0.000	0.231 ± 0.000	0.122 ± 0.000
GloVe	0.175 ± 0.000	0.034 ± 0.000	0.139 ± 0.000	0.105 ± 0.000	0.158 ± 0.000	0.052 ± 0.000
FastText	0.132 ± 0.000	0.065 ± 0.000	0.137 ± 0.000	0.082 ± 0.000	0.214 ± 0.000	0.062 ± 0.000
BERT	0.131 ± 0.000	0.037 ± 0.000	0.069 ± 0.000	0.042 ± 0.000	0.125 ± 0.000	0.026 ± 0.000
RoBERTa	0.053 ± 0.000	0.037 ± 0.000	0.069 ± 0.000	0.031 ± 0.000	0.115 ± 0.000	0.016 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.084 ± 0.010	0.215 ± 0.058	0.098 ± 0.021	0.000 ± 0.000	0.167 ± 0.037	0.181 ± 0.039
TF-IDF	0.193 ± 0.006	0.209 ± 0.047	0.128 ± 0.021	0.000 ± 0.000	0.238 ± 0.016	0.204 ± 0.021
GloVe	0.250 ± 0.007	0.069 ± 0.009	0.198 ± 0.009	0.128 ± 0.006	0.138 ± 0.008	0.067 ± 0.005
FastText	0.200 ± 0.012	0.084 ± 0.011	0.159 ± 0.008	0.142 ± 0.007	0.174 ± 0.015	0.085 ± 0.012
BERT	0.131 ± 0.007	0.051 ± 0.010	0.057 ± 0.007	0.028 ± 0.004	0.113 ± 0.031	0.017 ± 0.002
RoBERTa	0.083 ± 0.006	0.051 ± 0.012	0.078 ± 0.004	0.036 ± 0.002	0.098 ± 0.032	0.020 ± 0.003

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.237 ± 0.026	0.107 ± 0.048	0.130 ± 0.026	0.000 ± 0.000	0.245 ± 0.012	0.081 ± 0.018
TF-IDF	0.132 ± 0.062	0.075 ± 0.059	0.215 ± 0.025	0.000 ± 0.000	0.217 ± 0.022	0.079 ± 0.023
GloVe	0.152 ± 0.010	0.058 ± 0.010	0.133 ± 0.006	0.084 ± 0.008	0.129 ± 0.010	0.106 ± 0.010
FastText	0.067 ± 0.002	0.057 ± 0.007	0.078 ± 0.007	0.060 ± 0.004	0.097 ± 0.014	0.074 ± 0.007
BERT	0.105 ± 0.003	0.036 ± 0.006	0.065 ± 0.005	0.037 ± 0.002	0.081 ± 0.016	0.050 ± 0.004
RoBERTa	0.101 ± 0.009	0.051 ± 0.007	0.103 ± 0.007	0.068 ± 0.004	0.124 ± 0.022	0.041 ± 0.002

(c) DV dataset

Table A2: Results obtained using FPED bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.028 ± 0.000	0.228 ± 0.000	0.009 ± 0.000	0.000 ± 0.000	0.173 ± 0.000	0.103 ± 0.000
TF-IDF	0.095 ± 0.000	0.237 ± 0.000	0.096 ± 0.000	0.000 ± 0.000	0.190 ± 0.000	0.118 ± 0.000
GloVe	0.182 ± 0.000	0.053 ± 0.000	0.103 ± 0.000	0.072 ± 0.000	0.108 ± 0.000	0.036 ± 0.000
FastText	0.125 ± 0.000	0.052 ± 0.000	0.097 ± 0.000	0.068 ± 0.000	0.174 ± 0.000	0.049 ± 0.000
BERT	0.094 ± 0.000	0.035 ± 0.000	0.028 ± 0.000	0.018 ± 0.000	0.067 ± 0.000	0.015 ± 0.000
RoBERTa	0.021 ± 0.000	0.047 ± 0.000	0.031 ± 0.000	0.025 ± 0.000	0.064 ± 0.000	0.012 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.092 ± 0.011	0.212 ± 0.054	0.103 ± 0.018	0.000 ± 0.000	0.167 ± 0.026	0.166 ± 0.032
TF-IDF	0.184 ± 0.012	0.209 ± 0.046	0.126 ± 0.024	0.000 ± 0.000	0.217 ± 0.009	0.194 ± 0.018
GloVe	0.251 ± 0.005	0.078 ± 0.006	0.215 ± 0.009	0.117 ± 0.006	0.138 ± 0.015	0.065 ± 0.007
FastText	0.232 ± 0.012	0.094 ± 0.009	0.162 ± 0.012	0.151 ± 0.011	0.171 ± 0.027	0.083 ± 0.009
BERT	0.112 ± 0.011	0.043 ± 0.004	0.024 ± 0.003	0.016 ± 0.002	0.095 ± 0.018	0.013 ± 0.001
RoBERTa	0.070 ± 0.007	0.055 ± 0.015	0.070 ± 0.007	0.039 ± 0.007	0.088 ± 0.038	0.023 ± 0.001

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.243 ± 0.017	0.115 ± 0.051	0.123 ± 0.017	0.000 ± 0.000	0.263 ± 0.011	0.077 ± 0.020
TF-IDF	0.138 ± 0.058	0.076 ± 0.060	0.212 ± 0.022	0.000 ± 0.000	0.232 ± 0.019	0.077 ± 0.023
GloVe	0.206 ± 0.009	0.072 ± 0.014	0.160 ± 0.017	0.106 ± 0.004	0.151 ± 0.016	0.115 ± 0.007
FastText	0.084 ± 0.005	0.062 ± 0.007	0.105 ± 0.010	0.065 ± 0.006	0.120 ± 0.006	0.078 ± 0.009
BERT	0.117 ± 0.004	0.045 ± 0.006	0.081 ± 0.008	0.048 ± 0.006	0.102 ± 0.018	0.064 ± 0.005
RoBERTa	0.128 ± 0.010	0.064 ± 0.007	0.127 ± 0.009	0.077 ± 0.004	0.163 ± 0.018	0.047 ± 0.002

(c) DV dataset

Table A3: Results obtained using Subgroup AUC bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.508 ± 0.000	0.518 ± 0.000	0.511 ± 0.000	0.502 ± 0.000	0.519 ± 0.000	0.499 ± 0.000
TF-IDF	0.529 ± 0.000	0.517 ± 0.000	0.541 ± 0.000	0.502 ± 0.000	0.531 ± 0.000	0.517 ± 0.000
GloVe	0.554 ± 0.000	0.498 ± 0.000	0.547 ± 0.000	0.520 ± 0.000	0.572 ± 0.000	0.518 ± 0.000
FastText	0.552 ± 0.000	0.532 ± 0.000	0.553 ± 0.000	0.538 ± 0.000	0.565 ± 0.000	0.528 ± 0.000
BERT	0.515 ± 0.000	0.533 ± 0.000	0.530 ± 0.000	0.533 ± 0.000	0.533 ± 0.000	0.516 ± 0.000
RoBERTa	0.523 ± 0.000	0.516 ± 0.000	0.528 ± 0.000	0.515 ± 0.000	0.546 ± 0.000	0.508 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.497 ± 0.002	0.508 ± 0.006	0.491 ± 0.007	0.500 ± 0.000	0.493 ± 0.011	0.492 ± 0.006
TF-IDF	0.505 ± 0.007	0.507 ± 0.011	0.502 ± 0.003	0.500 ± 0.000	0.499 ± 0.013	0.498 ± 0.004
GloVe	0.498 ± 0.007	0.502 ± 0.008	0.490 ± 0.001	0.513 ± 0.006	0.480 ± 0.017	0.501 ± 0.002
FastText	0.518 ± 0.006	0.507 ± 0.009	0.504 ± 0.002	0.503 ± 0.005	0.517 ± 0.014	0.501 ± 0.002
BERT	0.510 ± 0.006	0.510 ± 0.008	0.515 ± 0.002	0.508 ± 0.002	0.513 ± 0.016	0.504 ± 0.002
RoBERTa	0.508 ± 0.002	0.504 ± 0.008	0.501 ± 0.003	0.499 ± 0.003	0.505 ± 0.004	0.497 ± 0.001

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.533 ± 0.009	0.530 ± 0.014	0.530 ± 0.011	0.508 ± 0.000	0.535 ± 0.014	0.518 ± 0.011
TF-IDF	0.516 ± 0.005	0.508 ± 0.013	0.531 ± 0.015	0.508 ± 0.000	0.534 ± 0.011	0.500 ± 0.005
GloVe	0.524 ± 0.007	0.515 ± 0.012	0.538 ± 0.004	0.525 ± 0.004	0.529 ± 0.009	0.531 ± 0.006
FastText	0.500 ± 0.003	0.505 ± 0.008	0.511 ± 0.007	0.495 ± 0.004	0.515 ± 0.007	0.507 ± 0.005
BERT	0.534 ± 0.003	0.525 ± 0.007	0.526 ± 0.005	0.523 ± 0.005	0.551 ± 0.020	0.521 ± 0.003
RoBERTa	0.495 ± 0.001	0.509 ± 0.006	0.497 ± 0.002	0.507 ± 0.003	0.526 ± 0.019	0.500 ± 0.003

(c) DV dataset

Table A4: Results obtained using AUC for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.584 ± 0.000	0.536 ± 0.000	0.616 ± 0.000	0.502 ± 0.000	0.588 ± 0.000	0.555 ± 0.000
TF-IDF	0.626 ± 0.000	0.538 ± 0.000	0.638 ± 0.000	0.518 ± 0.000	0.591 ± 0.000	0.548 ± 0.000
GloVe	0.599 ± 0.000	0.562 ± 0.000	0.626 ± 0.000	0.599 ± 0.000	0.611 ± 0.000	0.647 ± 0.000
FastText	0.622 ± 0.000	0.555 ± 0.000	0.648 ± 0.000	0.623 ± 0.000	0.647 ± 0.000	0.646 ± 0.000
BERT	0.626 ± 0.000	0.559 ± 0.000	0.638 ± 0.000	0.624 ± 0.000	0.605 ± 0.000	0.624 ± 0.000
RoBERTa	0.632 ± 0.000	0.529 ± 0.000	0.631 ± 0.000	0.590 ± 0.000	0.617 ± 0.000	0.618 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.901 ± 0.003	0.761 ± 0.017	0.887 ± 0.005	0.832 ± 0.008	0.878 ± 0.005	0.891 ± 0.008
TF-IDF	0.899 ± 0.005	0.770 ± 0.009	0.898 ± 0.005	0.809 ± 0.004	0.887 ± 0.005	0.901 ± 0.007
GloVe	0.862 ± 0.007	0.659 ± 0.009	0.885 ± 0.006	0.840 ± 0.010	0.871 ± 0.008	0.833 ± 0.010
FastText	0.864 ± 0.007	0.653 ± 0.012	0.885 ± 0.007	0.841 ± 0.005	0.884 ± 0.006	0.829 ± 0.010
BERT	0.867 ± 0.005	0.630 ± 0.017	0.870 ± 0.007	0.828 ± 0.009	0.864 ± 0.006	0.813 ± 0.011
RoBERTa	0.871 ± 0.006	0.652 ± 0.009	0.873 ± 0.007	0.840 ± 0.003	0.875 ± 0.006	0.834 ± 0.005

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.924 ± 0.006	0.800 ± 0.006	0.906 ± 0.009	0.899 ± 0.007	0.903 ± 0.008	0.915 ± 0.004
TF-IDF	0.933 ± 0.004	0.783 ± 0.006	0.917 ± 0.005	0.899 ± 0.007	0.900 ± 0.006	0.916 ± 0.008
GloVe	0.913 ± 0.005	0.672 ± 0.007	0.908 ± 0.004	0.859 ± 0.005	0.911 ± 0.009	0.856 ± 0.008
FastText	0.903 ± 0.005	0.655 ± 0.004	0.899 ± 0.007	0.863 ± 0.003	0.915 ± 0.003	0.850 ± 0.005
BERT	0.875 ± 0.007	0.610 ± 0.005	0.862 ± 0.005	0.806 ± 0.012	0.870 ± 0.008	0.785 ± 0.009
RoBERTa	0.890 ± 0.007	0.630 ± 0.007	0.877 ± 0.005	0.833 ± 0.010	0.900 ± 0.006	0.827 ± 0.006

(c) DV dataset

Table A5: Results obtained using macro F1-score for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.489 ± 0.000	0.435 ± 0.000	0.446 ± 0.000	0.421 ± 0.000	0.487 ± 0.000	0.409 ± 0.000
TF-IDF	0.495 ± 0.000	0.462 ± 0.000	0.475 ± 0.000	0.420 ± 0.000	0.504 ± 0.000	0.420 ± 0.000
GloVe	0.525 ± 0.000	0.541 ± 0.000	0.539 ± 0.000	0.544 ± 0.000	0.527 ± 0.000	0.579 ± 0.000
FastText	0.566 ± 0.000	0.538 ± 0.000	0.555 ± 0.000	0.571 ± 0.000	0.517 ± 0.000	0.589 ± 0.000
BERT	0.500 ± 0.000	0.535 ± 0.000	0.500 ± 0.000	0.532 ± 0.000	0.512 ± 0.000	0.541 ± 0.000
RoBERTa	0.502 ± 0.000	0.500 ± 0.000	0.493 ± 0.000	0.515 ± 0.000	0.460 ± 0.000	0.520 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.749 ± 0.003	0.698 ± 0.013	0.741 ± 0.003	0.701 ± 0.007	0.721 ± 0.006	0.742 ± 0.019
TF-IDF	0.730 ± 0.011	0.709 ± 0.008	0.747 ± 0.010	0.700 ± 0.005	0.725 ± 0.009	0.762 ± 0.012
GloVe	0.661 ± 0.019	0.556 ± 0.013	0.703 ± 0.012	0.622 ± 0.014	0.707 ± 0.016	0.615 ± 0.016
FastText	0.640 ± 0.019	0.551 ± 0.012	0.704 ± 0.011	0.623 ± 0.010	0.726 ± 0.017	0.606 ± 0.017
BERT	0.702 ± 0.012	0.515 ± 0.020	0.679 ± 0.014	0.604 ± 0.015	0.692 ± 0.011	0.577 ± 0.014
RoBERTa	0.684 ± 0.003	0.545 ± 0.009	0.695 ± 0.009	0.629 ± 0.007	0.688 ± 0.008	0.598 ± 0.010

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.707 ± 0.009	0.693 ± 0.007	0.706 ± 0.008	0.703 ± 0.005	0.698 ± 0.015	0.715 ± 0.011
TF-IDF	0.702 ± 0.006	0.681 ± 0.008	0.681 ± 0.016	0.696 ± 0.011	0.691 ± 0.010	0.682 ± 0.008
GloVe	0.642 ± 0.005	0.536 ± 0.010	0.606 ± 0.005	0.569 ± 0.005	0.690 ± 0.013	0.579 ± 0.006
FastText	0.574 ± 0.016	0.515 ± 0.007	0.581 ± 0.008	0.546 ± 0.012	0.682 ± 0.016	0.547 ± 0.008
BERT	0.593 ± 0.010	0.458 ± 0.006	0.512 ± 0.011	0.478 ± 0.010	0.611 ± 0.035	0.483 ± 0.008
RoBERTa	0.571 ± 0.015	0.487 ± 0.008	0.543 ± 0.005	0.499 ± 0.007	0.629 ± 0.030	0.489 ± 0.008

(c) DV dataset