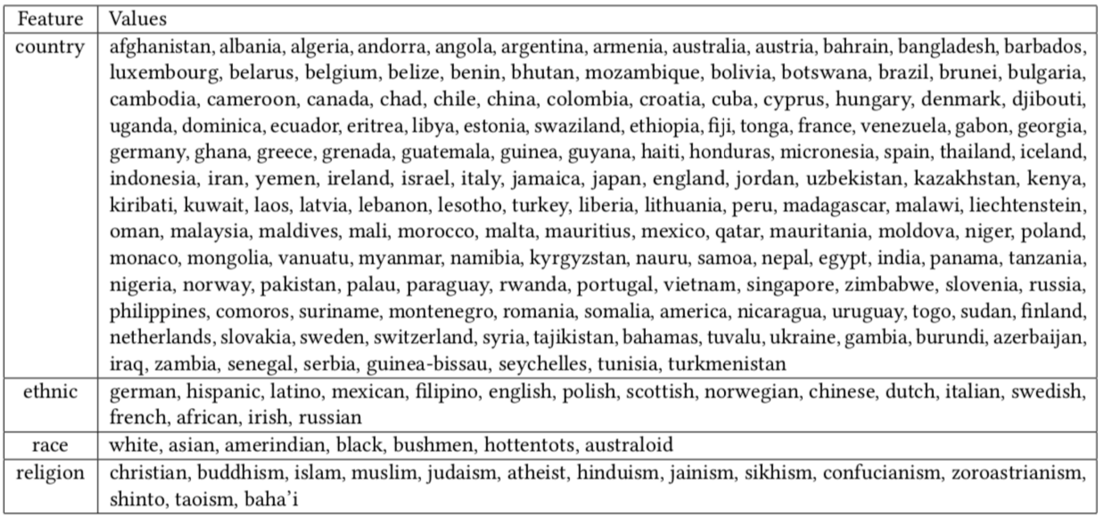
A APPENDIX

A.1 Sensitive Feature

Table 5 lists the value domain of sensitive features, country, ethnic, race, religion, in our experiment.

Table 5: The value domain of sensitive features.

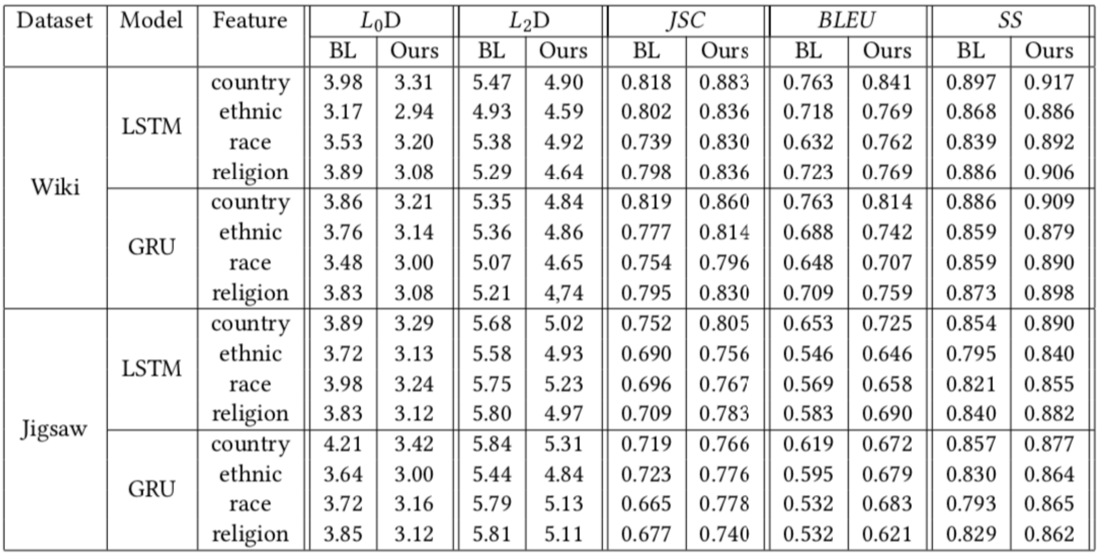


Feature	Values
country	afghanistan, albania, algeria, andorra, angola, argentina, armenia, australia, austria, bahrain, bangladesh, barbados, luxembourg, belarus, belgium, belize, benin, bhutan, mozambique, bolivia, botswana, brazil, brunei, bulgaria, cambodia, cameroon, canada, chad, chile, china, colombia, croatia, cuba, cyprus, hungary, denmark, djibouti, uganda, dominica, ecuador, eritrea, libya, estonia, swaziland, ethiopia, fiji, tonga, france, venezuela, gabon, georgia, germany, ghana, greece, grenada, guatemala, guinea, guyana, haiti, honduras, micronesia, spain, thailand, iceland, indonesia, iran, yemen, ireland, israel, italy, jamaica, japan, england, jordan, uzbekistan, kazakhstan, kenya, kiribati, kuwait, laos, latvia, lebanon, lesotho, turkey, liberia, lithuania, peru, madagascar, malawi, liechtenstein, oman, malaysia, maldives, mali, morocco, malta, mauritius, mexico, qatar, mauritania, moldova, niger, poland, monaco, mongolia, vanuatu, myanmar, namibia, kyrgyzstan, nauru, samoa, nepal, egypt, india, panama, tanzania, nigeria, norway, pakistan, palau, paraguay, rwanda, portugal, vietnam, singapore, zimbabwe, slovenia, russia, philippines, comoros, suriname, montenegro, romania, somalia, america, nicaragua, uruguay, togo, sudan, finland, netherlands, slovakia, sweden, switzerland, syria, tajikistan, bahamas, tuvalu, ukraine, gambia, burundi, azerbaijan, iraq, zambia, senegal, serbia, guinea-bissau, seychelles, tunisia, turkmenistan
ethnic	german, hispanic, latino, mexican, filipino, english, polish, scottish, norwegian, chinese, dutch, italian, swedish, french, african, irish, russian
race	white, asian, amerindian, black, bushmen, hottentots, australoid
religion	christian, buddhism, islam, muslim, judaism, atheist, hinduism, jainism, sikhism, confucianism, zoroastrianism, shinto, taoism, baha'i

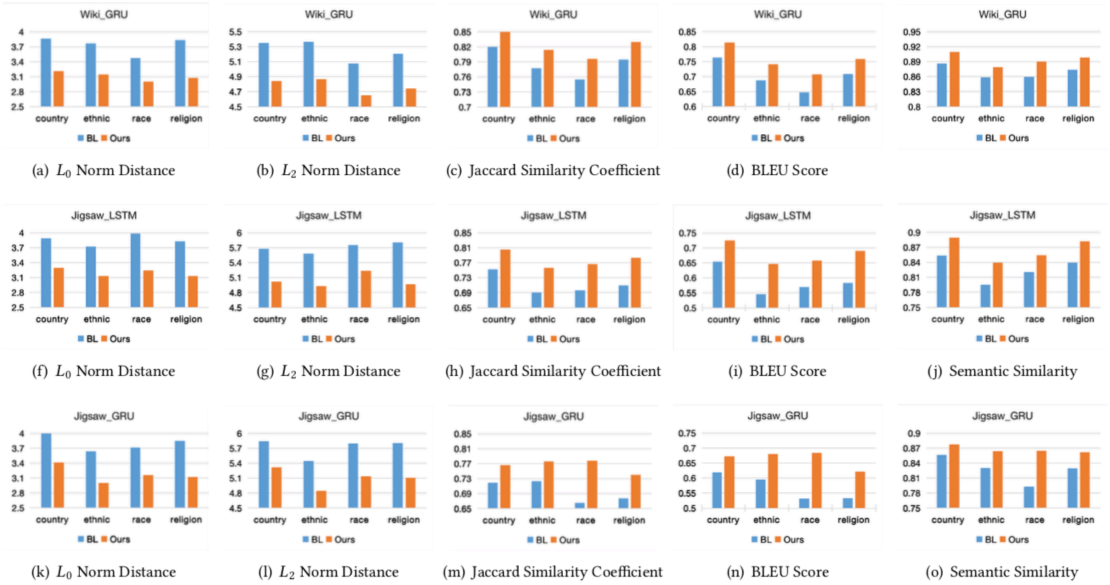
A.2 Validity Analysis

Table 6 summarize the results of five metrics among 16 benchmarks. And Figure 6 shows the differences among all the benchmarks between random and global-guided perturbation more intuitively, except for the LSTM on Wiki which is presented in Section 4.2.

Table 6: Metrics.



Dataset	Model	Feature	L_0D		L_2D		*JSC*		*BLEU*		*SS*	
			BL	Ours	BL	Ours	BL	Ours	BL	Ours	BL	Ours
Wiki	LSTM	country	3.98	3.31	5.47	4.90	0.818	0.883	0.763	0.841	0.897	0.917
		ethnic	3.17	2.94	4.93	4.59	0.802	0.836	0.718	0.769	0.868	0.886
		race	3.53	3.20	5.38	4.92	0.739	0.830	0.632	0.762	0.839	0.892
		religion	3.89	3.08	5.29	4.64	0.798	0.836	0.723	0.769	0.886	0.906
	GRU	country	3.86	3.21	5.35	4.84	0.819	0.860	0.763	0.814	0.886	0.909
		ethnic	3.76	3.14	5.36	4.86	0.777	0.814	0.688	0.742	0.859	0.879
		race	3.48	3.00	5.07	4.65	0.754	0.796	0.648	0.707	0.859	0.890
		religion	3.83	3.08	5.21	4,74	0.795	0.830	0.709	0.759	0.873	0.898
Jigsaw	LSTM	country	3.89	3.29	5.68	5.02	0.752	0.805	0.653	0.725	0.854	0.890
		ethnic	3.72	3.13	5.58	4.93	0.690	0.756	0.546	0.646	0.795	0.840
		race	3.98	3.24	5.75	5.23	0.696	0.767	0.569	0.658	0.821	0.855
		religion	3.83	3.12	5.80	4.97	0.709	0.783	0.583	0.690	0.840	0.882
	GRU	country	4.21	3.42	5.84	5.31	0.719	0.766	0.619	0.672	0.857	0.877
		ethnic	3.64	3.00	5.44	4.84	0.723	0.776	0.595	0.679	0.830	0.864
		race	3.72	3.16	5.79	5.13	0.665	0.778	0.532	0.683	0.793	0.865
		religion	3.85	3.12	5.81	5.11	0.677	0.740	0.532	0.621	0.829	0.862



(a) L_0 Norm Distance

(b) L_2 Norm Distance

(c) Jaccard Similarity Coefficient

(d) BLEU Score

(f) L_0 Norm Distance

(g) L_2 Norm Distance

(h) Jaccard Similarity Coefficient

(i) BLEU Score

(j) Semantic Similarity

(k) L_0 Norm Distance

(l) L_2 Norm Distance

(m) Jaccard Similarity Coefficient

(n) BLEU Score

(o) Semantic Similarity

Figure 6: Validity analysis.