Analisis dan Mitigasi Religion Bias pada Dataset dan Embedding NLP Berbahasa Indonesia = Analysis and Mitigation of Religion Bias in Indonesian NLP Datasets and Embeddings

Muhammad Arief Fauzan, author

Abstrak

Riset terdahulu menunjukkan adanya misrepresentasi identitas agama pada media Indonesia. Menurut studi sebelumnya, misrepresentasi identitas marjinal pada dataset dan word embedding untuk natural language processing dapat merugikan identitas marjinal tersebut, dan karenanya harus dimitigasi. Riset ini menganalisis keberadaan bias agama pada beberapa dataset dan word embedding NLP berbahasa Indonesia, dampak bias yang ditemukan pada downstream performance, serta proses dan dampak debiasing untuk dataset dan word embedding. Dengan menggunakan metode uji Pointwise Mutual Information (PMI) untuk deteksi bias pada dataset dan word similarity untuk deteksi bias pada word embedding, ditemukan bahwa dua dari tiga dataset, serta satu dari empat word embedding yang digunakan pada studi ini mengandung bias agama. Model machine learning yang dibentuk dari dataset dan word embedding yang mengandung bias agama memiliki dampak negatif untuk downstream performance model tersebut, yang direpresentasikan dengan allocation harm dan representation harm. Allocation harm direpresentasikan oleh performa false negative rate (FNR) dan false positive rate (FPR) model machine learning yang lebih buruk untuk identitas agama tertentu, sedangkan representation harm direpresentasi oleh kesalahan model dalam mengasosiasikan kalimat non-negatif yang mengandung identitas agama sebagai kalimat negatif. Metode debiasing pada dataset dan word embedding mampu memitigasi bias agama yang muncul pada dataset dan word embedding, tetapi memiliki performa yang beragam dalam mitigasi allocation dan representation harm. Dalam riset ini, akan digunakan lima metode debiasing: dataset debiasing dengan menggunakan sentence templates, dataset debiasing dengan menggunakan kalimat dari Wikipedia, word embedding debiasing dengan menggunakan Hard Debiasing, joint debiasing dengan sentence templates, serta joint debiasing menggunakan kalimat dari Wikipedia. Dari lima metode debiasing, joint debiasing dengan sentence templates memiliki performa yang paling baik dalam mitigasi allocation harm dan representation harm.
.....Previous research has shown the existence of misrepresentation regarding various religious identities in Indonesian media. Misrepresentations of other marginalized identities in natural language processing (NLP) resources have been recorded to inflict harm against such marginalized identities, and as such must be mitigated. This research analyzes several Indonesian language NLP datasets and word embeddings to see whether they contain unwanted bias, the impact of bias on downstream performance, the process of debiasing datasets or word embeddings, and the effect of debiasing on them. By using the Pointwise Mutual Test (PMI) test to detect dataset bias and word similarity to detect word embedding bias, it is found that two out of three datasets and one out of four word embeddings contain religion bias. The downstream performances of machine learning models which learn from biased datasets and word embeddings are found to be negatively impacted by the biases, represented in the form of allocation and representation harms. Allocation harm is represented by worse false negative rate (FNR) and false positive rate (TPR) of models with respect to certain religious identities, whereas representation harm is represented by the misprediction of non-negative sentences containing religious identity terms as negative sentences. Debiasing at dataset and

word embedding level was found to correctly mitigate the respective biases at dataset and word embedding level. Nevertheless, depending on the dataset and word embedding used to train the model, the performance of each debiasing method can vary highly at downstream performance. This research utilizes five debiasing methods: dataset debiasing using sentence templates, dataset debiasing using sentences obtained from Wikipedia, word embedding debiasing using Hard Debiasing, joint debiasing using sentence templates, as well as joint debiasing using sentences obtained from Wikipedia. Out of all five debiasing techniques, joint debiasing using sentence templates performs the best on mitigating both allocation and representation harm.