Deteksi Ujaran Kebencian dan Bahasa Kasar pada Blog Mikro Berbahasa Indonesia = Detection of Hate Speech and Abusive Language on Indonesian Microblogs

Nabila Khansa, author

Abstrak

Ujaran kebencian dan bahasa kasar mempermudah penyebaran kekerasan di kehidupan nyata, sehingga muncul urgensi adanya pendeteksian secara otomatis. Untuk melanjutkan pekerjaan yang sudah dilakukan oleh Ibrohim dan Budi (2019), penelitian ini membahas dua isu terkait deteksi ujaran kebencian dan bahasa kasar pada mikroblog berbahasa Indonesia. Isu pertama adalah kajian terkait effect size fitur dan pengembangan model menggunakan fitur-fitur tersebut. Metode Analysis of Variance f-test, Logistic Regression Analysis, dan nilai Shapley digunakan untuk melakukan kajian effect size pada fitur-fitur yang dirancang secara manual. Kemudian, digunakan beberapa algoritma pemelajaran mesin untuk mengembangkan model prediksi berbasis fitur-fitur tersebut. Isu kedua adalah kajian bias dalam pengembangan model terkait keberadaan kata-kata bersifat netral pada data yang merupakan ujaran kebencian atau bahasa kasar. Kajian terkait bias dilakukan dengan menggunakan dataset uji bias. Dataset ini dikembangkan dengan menggantikan kata-kata yang dideteksi memiliki potensi adanya bias pada model yang dilatih menggunakan dataset hasil pekerjaan Ibrohim dan Budi (2019). Penelitian ini menunjukkan bahwa keberadaan kata-kata tertentu berpengaruh terhadap hasil deteksi ujaran kebencian dan bahasa kasar. Di antara kata-kata tersebut, terdeteksi beberapa kata-kata yang berpotensi bias, karena memiliki pengaruh terhadap pendeteksian padahal secara sendiri kata-kata yang dideteksi sebagai potensi bias tidak memiliki unsur kebencian atau bersifat kasar. Hasil evaluasi pengambilan sampel bootstrap menunjukkan Logistic Regression dan XGBoost sebagai model dengan akurasi terbaik dalam pendeteksian ujaran kebencian dan bahasa kasar. Namun, ketika model yang sudah dikembangkan digunakan untuk memprediksi dataset sintetis, didapatkan penurunan akurasi dalam pendeteksian ujaran kebencian. Hasil ini menandakan adanya bias pada model yang dikembangkan. Hasil tersebut didukung juga oleh hasil prediksi dengan akurasi rendah ketika model digunakan untuk melakukan pendeteksian ujaran kebencian pada dataset yang dikembangkan secara manual, tetapi ketika kata-kata bias digantikan dari data, akurasi model meningkat. Kontribusi yang diberikan oleh penelitian ini adalah pengembangan dataset uji bias secara otomatis dari dataset yang dikembangkan oleh Ibrohim dan Budi (2019) dan juga dataset uji bias yang dikembangkan secara manual.

.....Hate speech and abusive language facilitate the spread of violence in real life, hence the urgency of automatic detection. To continue the work done by Ibrohim dan Budi (2019), this research addresses two issues related to the detection of hate speech and abusive language on Indonesian-language microblogs. The first issue is a study on the effect size of features and the development of models using these features. Analysis of Variance f-test, Logistic Regression Analysis, and Shapley values are used to investigate the effect size of manually designed features. Several machine learning algorithms are then employed to develop prediction models based on these features. The second issue involves studying bias in model development concerning the presence of neutral words in data that constitute hate speech or abusive language. The study related to bias is conducted by using a bias test dataset. This dataset is developed by

replacing words that are detected to have the potential for bias in models trained using the dataset resulting from the work of Ibrohim dan Budi (2019). This research demonstrates that certain words significantly influence the detection of hate speech and abusive language. Among these words, some are identified as potentially biased, as they affect detection despite not inherently containing hate or abusive elements. The results of bootstrap sampling evaluation indicate that Logistic Regression and XGBoost are the models with the highest accuracy in detecting hate speech and abusive language. However, when the developed models are used to predict synthetic datasets, a significant decrease in accuracy is observed in hate speech detection. This finding indicates the presence of bias in the developed models. This result is further supported by low-accuracy predictions when the models are used to detect hate speech in manually developed datasets. However, when biased words are replaced in the data, the model's accuracy significantly improves. The contributions of this research include the development of an automatically generated bias test dataset from the dataset created by Ibrohim dan Budi (2019), as well as a manually developed bias test dataset.