Klasifikasi ujaran kebencian kekerasan seksual pada Twitter Indonesia = Classification of sexual violence hate speech: a case study of Indonesian Twitter

Muammar Nota Reza Ramadhan, author

Abstrak

Saat ini media sosial merupakan sarana komunikasi yang tidak terlepas dari penyebaran ujaran kebencian yang cukup meresahkan penggunanya. Sejak tahun 2018 KOMINFO telah menangani sebanyak 3.640 ujaran kebencian yang tersebar di berbagai media sosial. Selain itu SafeNet telah menangani kasus Kekerasan Berbasis Gender Online (KBGO) pada tahun 2021 sebanyak 677 aduan yang didominasi dengan kasus pelecehan seksual. Disisi lain Sejak tahun 2020 Komnas Perempuan mencatat kasus kekerasan yang terjadi dalam komunitas dan ranah publik Indonesia sebesar 21 % (1.731 kasus) dengan kasus kekerasan seksual yang paling mendominasi. Banyaknya jenis ujaran kebencian yang berbeda-beda menyebabkan banyak tantangan dalam mendeteksi ujaran kebencian termasuk dalam domain kekerasan seksual. Tujuan dari penelitian ini adalah menghasilkan model klasifikasi ujaran kebencian kekerasan seksual dengan performa dan tingkat akurasi yang baik sehingga dapat dimanfaatkan secara teori bagi akademisi dan praktikal bagi lembaga seperti KOMINFO, SafeNet, LBH APIK Jakarta, Komnas Perempuan, POLRI. Data yang digunakan pada penelitian ini adalah hasil crawling media sosial twitter pada bulan Desember 2021 hingga Januari 2022. Dengan menggunakan pendekatan Machine Learning, dataset diolah dengangan teknik ekstraksi fitur Term Frequency-Inverse Document Frequency (TF-IDF), beberapa teknik sampling seperti Random Over Sampling (ROS), Random Under Sampling (RUS), Synthetic Minority Over-sampling Technique (SMOTE), dan Adaptive Synthetic (ADASYN) serta beberapa algoritma klasifikasi seperti Nave bayes (NB), Support Vector Machine (SVM), Logistic Regresion (LR), Decition Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM) dan Extreme Gradient Boosting (XGBoost). Penelitian ini menghasilkan akurasi tertinggi sebesar 0.9239 dimana Algoritma terbaik didominasi oleh SVM dan RF. Implikasi penelitian ini secara teori adalah perbandingn hasil klasifikasi 35 model klasifikasi dan secara praktik dapat diimplementsikan pada Lembaga yang memiliki sistem pendeteksi ujaran kebencian.
.....Currently, social media is a means of communication that cannot be separated from the spread of hate speech which is quite disturbing for its users. Since 2018, KOMINFO has handled 3,640 hate speech spread across various social media. SafeNet has handled cases of Online Gender-Based Violence (KBGO) in 2021 as many as 677 complaints, which were dominated by cases of sexual harassment. In 2020 Komnas Perempuan has recorded 21% of cases of violence occurring in the Indonesian community/public sphere (1,731 cases) with the most prominent case being sexual violence. Different types of hate speech cause many challenges in detecting such hate speech. The purpose of this study is to produce a classification model of sexual violence hate speech with good performance and accuracy so that it can be used theoretically for academics and practically for institutions such as KOMINFO, SafeNet, LBH APIK Jakarta, Komnas Perempuan, and POLRI. The data used in this study is the result of crawling social media twitter from December 2021 to January 2022. By using a Machine Learning approach, the dataset is processed using the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction technique, several sampling techniques such as Random Over Sampling (ROS), Random Under Sampling (RUS), Synthetic

Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic (ADASYN) as well as several classification algorithms such as Nave Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGBoost). This research produces the highest accuracy of 0.9239 where the best algorithm is dominated by SVM and RF. The theoretical implication of this research is the comparison of the classification results of 35 classification models and practically it can be implemented in institutions that have a hate speech detection system.