Klasifikasi multi label untuk identifikasi ujaran kebencian dan ujaran kasar pada Twitter berbahasa Indonesia = Multi-label classification to identify hate speech and abusive language on Indonesian Twitter

Muhammad Okky Ibrohim, author

Abstrak

ABSTRAK
Penyebaran ujaran kebencian dan ujaran kasar di media sosial merupakan hal yang harus diidentifikasi secara otomatis untuk mencegah terjadinya konflik masyarakat. Selain itu, ujaran kebencian mempunyai target, golongan, dan tingkat tersendiri yang juga perlu diidentifikasi untuk membantu pihak berwenang dalam memprioritaskan kasus ujaran kebencian yang harus segera ditangani. Tesis ini membahas klasifikasi teks multi label untuk mengidentifikasi ujaran kasar dan ujaran kebencian disertai identifikasi target, golongan, dan tingkatan ujaran kebencian pada Twitter berbahasa Indonesia. Permasalahan ini diselesaikan menggunakan pendekatan machine learning menggunakan algoritma klasifikasi Support Vector Machine (SVM), NaÃ¯ve Bayes (NB), dan Random Forest Decision Tree (RFDT) dengan metode transformasi data Binary Relevance (BR), Label Power-set (LP), dan Classifier Chains (CC). Jenis fitur yang digunakan antara lain fitur frekuensi term (word n-grams dan character n-grams), fitur ortografi (tanda seru, tanda tanya, huruf besar/kapital, dan huruf kecil), dan fitur leksikon (leksikon sentimen negatif, leksikon sentimen positif, dan leksikon kasar). Hasil eksperimen menunjukkan bahwa secara umum algoritma klasifikasi RFDT dengan metode transformasi LP memberikan akurasi yang terbaik dengan waktu komputasi yang cepat. Algoritma klasifikasi RFDT dengan metode transformasi LP menggunakan fitur word unigram memberikan akurasi sebesar 66,16%. Jika hanya mengidentifikasi ujaran kasar dan ujaran kebencian (tanpa disertai identifikasi target, golongan, dan tingkatan ujaran kebencian), algoritma klasifikasi RFDT dengan metode transformasi LP menggunakan gabungan fitur word unigram, character quadgrams, leksikon sentimen positif, dan leksikon kasar mampu memberikan akurasi sebesar 77,36%. <hr /> Hate speech and abusive language spreading on social media needs to be identified automatically to avoid conflict between citizen. Moreover, hate speech has target, criteria, and level that also needs to be identified to help the authority in prioritizing hate speech which must be addressed immediately. This thesis discusses multi-label text classification to identify abusive and hate speech including the target, category, and level of hate speech in Indonesian Twitter. This problem was done using machine learning approach with Support Vector Machine (SVM), NaÃ¯ve Bayes (NB), and Random Forest Decision Tree (RFDT) classifier and Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) as data transformation method. The features that used are term frequency (word n-grams and character n-grams), ortography (exclamation mark, question mark, uppercase, lowercase), and lexicon features (negative sentiment lexicon, positif sentiment lexicon, and abusive lexicon). The experiment results show that in general RFDT classifier using LP as the transformation method gives the best accuracy with fast computational time. RFDT classifier with LP transformation using word unigram feature give 66.16% of accuracy. If only for identifying abusive language and hate speech (without identifying the target, criteria, and level of hate speech), RFDT classifier with LP transformation using combined fitur word unigram, character quadgrams, positive sentiment lexicon, and abusive lexicon can gives 77,36% of accuracy.