ABSTRACT

Hate speech can be distributed not only through utterance or actions, but also texts in social media. Some social media which have many users such as Twitter are very vulnerable to the spread of hate speech. Therefore, this has attracted the attention of several researchers to develop a hate speech detection system. Several previous studies conducted a classification of hate speech texts in English. But not many researchers have developed an Indonesian language hate speech detection system.

Detection of hate speech can be done with predictions on existing text data. Prediction performance can be measured using evaluations such as accuracy, precision, recall, f1-score and others. Several factors can improve the performance of predictions such as the complexity of the model, selection of training data algorithm and the quality of the dataset. Previous research on this topic has been carried out with a combination of RFDT algorithm and BoW method with n-gram feature extraction: unigram, bigram and trigram. However, there are some weaknesses that need to be done improvisation to improve performance in detecting expressions of hate speech in Indonesian.

Two phases of development in this study, the first phase did a combination of n-gram features for feature extraction and the second phase carried out a hybrid classifier on the naïve bayes algorithm and logistic regression. The hybrid classifier process is carried out using voting techniques on the predicted results generated from each algorithm. In this study, it was found that the n-gram (2.2) and (1.3) features had the best prediction performance on hybrid classifier compared to other n-gram features. However, based on computational time, the word n-gram (2.2) has a speed of time that is twice as fast as n-gram (1.3) with a consumption time of 0.00066 seconds per tweet, in addition the evaluation value obtained is 94.11 % on accuracy and 94.59% on fi-score.

Keywords : *Features Combination, Hybrid Classifier, Text Classification, Naïve Bayes, Logistic Regression, Multi Classifier.*

INTISARI

Penyebaran ujaran kebencian tidak hanya dapat disebar lewat ucapan atau tindakan, tetapi juga bisa lewat tulisan di media sosial. Beberapa media sosial yang memiliki banyak pengguna seperti Twitter sangat rentan terhadap penyebaran ujaran kebencian. Oleh karena itu, hal ini menarik perhatian beberapa peneliti untuk mengembangkan sistem deteksi ujaran kebencian. Beberapa penelitian sebelumnya melakukan klasifikasi teks ujaran kebencian dalam Bahasa Inggris. Namun belum banyak peneliti mengembangkan sistem deteksi ujaran kebencian berbahsa Indonesia.

Deteksi ujaran kebencian dapat dilakukan dengan prediksi pada data teks yang ada. Performa prediksi dapat diukur dengan menggunakan evaluasi seperti akurasi, presisi, recall, f1-*score* dan lainnya. Beberapa factor yang dapat meningkatkan performa prediksi seperti kompleksitas model, pemilihan algortime pelatihan data dan kualitas *dataset.* Penelitian sebelumnya terkait topik ini pernah dilakukan dengan kombinasi algoritme RFDT dan metode BoW dengan ekstraksi fitur n-*gram*: *unigram, bigram* dan t*rigram*. Namun terdapat beberapa kelemahan sehingga perlu dilakukan improvisasi untuk meningkatkan kinerja dalam pendeteksian ujaran kebencian berbahasa Indonesia.

Dua fase pengembangan pada penelitian ini, fase pertama melakukan kombinasi pada fitur n-*gram* untuk ekstraksi fitur dan fase kedua melakukan *hybrid classifier* pada algoritme naïve bayes dan *logistic regression*. Proses *hybrid classifier* dilakukan dengan menggunakan teknik *voting* pada hasil prediksi yang dihasilkan dari setiap algoritme. Pada penelitian ini didapatkan bahwa fitur n-*gram* (2,2) dan (1,3) memiliki performa prediksi yang paling baik pada *hybrid classifier* dibandingkan dengan fitur n-*gram* lainnya. Akan tetapi berdasarkan waktu komputasi, *word* n-*gram* (2,2) memiliki kecepatan waktu dua kali lebih cepat dibandingkan n-*gram* (1,3) dengan konsumsi waktu prediksi 0.00066 detik per tweet, selain itu nilai evaluasi yang didapatkan sebesar 94,11% pada akurasi dan 94,59% pada fi-*score*.

Kata Kunci: Kombinasi Fitur, *Hybrid Classifier*, *Text Classification*, *Naïve Bayes*, *Logistic Regression*, *Multi Classifier*.