

## ABSTRAK

Indonesia merupakan salah satu negara dengan pengguna media sosial terbanyak. Hal tersebut dapat memicu munculnya *cyberbullying*. *Cyberbullying* adalah tindakan berulang yang melecehkan, mempermalukan, mengancam, atau mengganggu orang lain melalui komputer, ponsel, dan perangkat elektronik lainnya, termasuk situs web jejaring sosial *online*. Twitter merupakan salah satu media sosial yang sering digunakan untuk melakukan *cyberbullying*. Penelitian ini mendeteksi *cyberbullying* yang berasal dari tweet berbahasa Indonesia dengan menggunakan metode *Random Forest* sebagai pengklasifikasi. Seleksi fitur *Information Gain* juga digunakan untuk menyeleksi fitur yang berupa atribut. *Random Forest* adalah salah satu teknik *ensemble learning* yang cara kerjanya berdasarkan metode *Decision Tree*, yang menggunakan aturan untuk membagi data dalam mode *biner*. *Information Gain* adalah salah satu metode seleksi fitur berbasis filter yang cara kerjanya melakukan perangkingan berdasarkan atribut. Deteksi *cyberbullying* pada penilitan ini tersusun atas tahapan: *preprocessing*, pembobotan kata, seleksi fitur dan yang terakhir klasifikasi. *Dataset* yang digunakan pada tahapan klasifikasi merupakan *dataset* yang berupa matrix dimana kata hasil pembobotan kata sebagai kolom dan telah diurutkan berdasarkan hasil perhitungan nilai entropi. Pengurutan dimulai dari nilai entropi yang tertinggi hingga ke rendah. Kemudian *K – Fold Cross Validation* yang menggunakan *fold* sebanyak 10 digunakan sebagai metode evaluasi. *Accuracy* tiap *fold* dihitung lalu dirata – rata. Setelah perhitungan *K – Fold Cross Validation* pertama atribut dikurangi satu lalu dilakukan lagi perhitungan *K – Fold Cross Validation* kedua. Hal tersebut dilakukan hingga jumlah atribut tersisa satu. Hasil akhir untuk skenario pengujian ditentukan dengan memilih rata – rata *Accuracy* tertinggi dari perhitungan pengurangan atribut. Dari hasil pengujian, didapatkan nilai *Accuracy* tertinggi sebesar 72.1% dengan atribut berjumlah 1295 dari 2277 atribut. Hal ini berarti, pemilihan fitur yang baik dapat meningkatkan performa dari metode *machine learning*.

Kata kunci: *Cyberbullying*, *Random Forest*, *Tweet*, *Information Gain*

## ABSTRACT

*Indonesia is one of the countries with the most social media users. This can trigger the emergence of cyberbullying. Cyberbullying is the repeated act of harassing, humiliating, threatening or harassing another person via computers, cell phones and other electronic devices, including online social networking websites. Twitter is one of the social media that is often used for cyberbullying. This study detects cyberbullying originating from Indonesian-language tweets using the Random Forest method as a classifier. Feature selection Information Gain is also used to select features in the form of attributes. Random Forest is an ensemble learning technique that works based on the Decision Tree method, which uses rules to divide data in binary mode. Information Gain is a filter-based feature selection method that performs ranking based on attributes. Cyberbullying detection in this research is composed of stages: preprocessing, word weighting, feature selection and finally classification. The dataset used at the classification stage is a dataset in the form of a matrix where words are weighted as columns and have been sorted based on the entropy value calculation results. Sorting starts from the highest entropy value to the lowest. Then K – Fold Cross Validation which uses 10 folds is used as an evaluation method. Accuracy for each fold is calculated and then averaged. After the first K-Fold Cross Validation calculation, the attribute is reduced by one, then the second K-Fold Cross Validation calculation is performed. This is done until the number of attributes remains one. The final result for the test scenario is determined by selecting the highest average accuracy from the attribute reduction calculation. From the test results, the highest accuracy value was 72.1% with 1295 attributes out of 2277 attributes. This means, good feature selection can improve the performance of machine learning methods.*

**Keyword:** Cyberbullying, Random Forest, Tweet, Information Gain