Klasifikasi Toxic Comment Pada Sosial Media Menggunakan SVM, Information Gain dan TF-IDF

Muhammad Ilham Maulana[1], Dr. Kemas Muslim L., S.T., M.ISD.[2], Mahendra Dwifebri S.Kom., M.Kom[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung

[4]Divisi Digital Service PT Telekomunikasi Indonesia

[1]ilhaammaulana@student.telkomuniversity.ac.id,

[2]kemasmuslim@telkomuniversity.ac.id,

[3]mahendrap@telkomuniversity.ac.id

Abstrak

Sosial media merupakan suatu bentuk perantara interaksi sosial secara *online*. Aplikasi media sosial pun sudah dalam banyak bentuk dan di dalam sosial media ini meskipun banyak hal positif yang dapat diambil, ada beberapa juga hal-hal negatif contoh nya *toxic comment*. *Toxic comment* sendiri tidaklah mudah untuk dideteksi secara manual, maka penelitian berencana untuk mengklasifikasikan toxic comment tersebut menggunakan machine learning. Beberapa penelitian untuk klasifikasi *toxic comment* sudah dilakukan, dalam beberapa penelitian tersebut digunakan metode *Support Vector Machine*. Dalam penelitian ini metode yang digunakan *adalah Support Vector Machine* (SVM) sebagai *classifier*, *Information Gain* sebagai *feature selection* dan TF-IDF sebagai *feature extraction*. Data-data yang dikumpulkan adalah melalui cuitan *twitter* beberapa pengguna di media sosial tersebut. Komentar-komentar tersebut dikumpulkan menjadi satu lalu diklasifikasikan menggunakan metode-metode yang sudah disebutkan.

1. Pendahuluan

Latar Belakang

Dalam perkembagan teknologi yang pesat di era modern, sosial media sangat populer di kalangan masyarakat. Berdasarkan survey yang diambil dari tahun 2005 – 2015 di AS, penggunaan sosial media telah meningkat secara drastis [1]]. Di Indonesia sendiri pada tahun 2020 ini, pengguna internet sudah mencapai 175,4 juta pada bulan Januari, sedangkan pengguna media sosial sudah mencapai 160,0 juta dengan peningkatan sebanyak 12 juta (8,1%) dari tahun 2019 sampai Januari 2020 [2].

Salah satu hal yang menjadi semakin marak karena populernya media sosial adalah *cyberbullying*. Menurut Patchin dkk. [3] *cyberbullying* adalah kekerasan yang secara disengaja dan berulang kali dilakukan melalui medium berupa teks elektronik. Beberapa studi yang dilakukan juga menemukan bahwa *cyberbullying* sering ditemukan melalui pesan teks dan sosial media seperti Facebook atau Instagram, di sosial media sendiri itupun lebih banyak ditemukan pada bagian komentar yang dimungkinkan karena anonimitas yang dirasakan orang-orang di kolom komentar [4]. Maka *cyberbullying* yang dilakukan pada kolom komentar tersebut disebut dengan *toxic comment*. *Toxic Comment* adalah komentar yang kasar, tidak menghargai ataupun komentar yang membuat seseorang merasa tidak nyaman sehingga mereka meninggalkan diskusi [5].

Karena banyaknya *cyberbullying* melalui komentar *toxic* ini banyak dilakukan, maka dilakukan klasifikasi teks menggunakan *machine learning* untuk mengidentifikasi komentar-komentar yang terdapat di sosial media. Beberapa penelitian telah dilakukan untuk mengklasifikasikan komentar *toxic* menggunakan beberapa metode yang berbeda [6].

Dalam penelitian ini, metode classifier yang dipilih adalah *Support Vector Machine* (SVM). Metode SVM untuk mengklasifikasikan teks ini telah dibandingkan dengan metode lain, salah satunya adalah Naive Bayes. Penelitian tersebut menunjukkan bahwa metode SVM memiliki hasil *f-measure* yang lebih baik daripada metode *Naive Bayes* [7].

Topik dan Batasannya
Pada penelitian ini penulis melakukan klasifikasi komentar *toxic* dengan menggunakan metode *Support Vector Machine* dan *feature selection Information Gain. Dataset* yang digunakan berjumlah 711 diambil dari sosial media *Twitter* dengan label *Toxic* dan *Non-Toxic.*

Tujuan
Tujuan dari tugas akhir ini adalah mengidentifikasi *toxic commen*t di sosial media dan mengetahui akurasi klasifikasi teks *toxic comment* menggunakan metode SVM dan *Information Gain*.

Organisasi Tulisan
Struktur penulisan dari tugas akhir ini disusun sebagai berikut: Bagian pertama berisi pendahuluan terkait tugas akhir ini. Bagian kedua menjelaskan studi yang terkait dengan tugas akhir ini. Bagian ketiga akan menjelaskan pemodelan dan performansi dari sistem yang dibangun. Bagian keempat menjelaskan hasil dan evaluasi hasil pengujian yang telah dilakukan pada bagian ketiga. Kemudian, pada bagian terakhir menjelaskan kesimpulan dan saran berdasarkan hasil pengujian yang dilakukan pada tugas akhir ini.

2. Studi Terkait

2.1. Klasifikasi Single Label
Klasifikasi *Single Label* adalah teknik klasifikasi yang digunakan dalam menentukan label kelas untuk data dengan suatu label L di mana jumlah L adalah lebih dari 1. Kalau L sama dengan 2 maka disebut sebagai *single label classifcation* atau bisa disebut juga dengan *binary classification problem* [8].

2.2. Support Vector Machine (SVM)
Support Vector Machine adalah metode pengklasifikasian *supervised* dengan linear kernel. Sebagai *classifier* dengan linear kernel, SVM bertujuan untuk mencari *hyperplane* yang memisahkan nilai data dengan *margin* maksimal di antara dua batas nilai tersebut. SVM dapat mereduksi kemungkinan *overfitting* karena SVM juga bertujuan untuk meminimalisasi *error* secara *general* [9].
Pada *Support Vector Machine* kernel digunakan untuk memetakan data masukan ke ruang dimensi yang lebih tinggi di mana batas *decision* dapat dibangun. Fungsi *decision* dapat ditulis sebagai berikut.

$$\boldsymbol{D(x) = w\emptyset(x) + b} \tag{1}$$

Dimana w dan b adalah parameter SVM sedangkan $w\emptyset$ adalah fungsi kernel yang memetakan data masukan ke dimensi M yang baru.

2.3. Infomation Gain
Information Gain adalah salah satu *feature selection* untuk klasifikasi yang cukup populer yang dapat mengukur bagus atau tidaknya suatu atribut. *Information Gain* mengukur reduksi dari entropi dengan memisahkan dataset menurut nilai yang diberikan kepada variabel acak [10]. Rumus *Information Gain* adalah sebagai berikut.

$$IG(t) = -\sum_{i=1}^{|C|} P(c_i) log P(c_i) + P(t) \sum_{i=1}^{|C|} P(c_i|t) log P(c_i|t) + P(t) \times \sum_{i=1}^{|C|} P(c_i|t\bar{}) log P(c_i|t \tag{2}$$

di mana c_i adalah kategori ke-i, $P(c_i)$ adalah probabilitas dari kategori ke-i, $P(t)$ dan $P(\bar{t})$ adalah probabilitas muncul atau tidaknya istilah t di dalam dokumen, $P(c_i|t)$ adalah probabilitas muncul istilah t, sedangkan $P(c_i|\bar{t})$ adalah probabilitas tidak munculnya istilah t [11].