BAB 1 PENDAHULUAN

1.1 Latar Belakang

Seiring dengan pertumbuhan teknologi yang berkembang sangat pesat dan kemudahan menggunakan internet menimbulkan pertumbuhan pengguna di internet khususnya di indonesia. Menurut survei pengguna aktif di indonesia berkembang hingga penggunanya mencapai 170 juta pengguna [1]. Namun demikian, tidak semua individu pengguna Twitter bersikap bijak dalam memilih kata-kata dalam cuitannya. Tidak sedikit netizen menuliskan kata-kata yang berbau SARA (Suku, Agama, Ras, dan Antar golongan) atau bahkan mengungkapkan ekspresi dengan menuliskan kata-kata kasar dan bersifat ofensif. Kata-kata kasar biasanya diucapkan atau dituliskan untuk menyerang pihak tertentu seperti ke perorangan atau sebuah instansi, untuk mengungkapkan kekesalan, kekecewaan, atau untuk meluapkan emosi terhadap peristiwa tertentu [2].

Menurut laporan terbaru dari *We Are Social*, perusahaan asal Inggris bekerja sama dengan Hootsuite yang melakukan survei per Januari 2021, mendapatkan kesimpulan bahwa pengguna aktif media sosial bertambah 6.3 % atau mengalami kenaikan dari januari 2020 mencapai 10 juta pengguna [3]. Jika dilihat dari jumlah populasi di dunia ada di sekitar 274.9 juta jiwa, maka 61.8 % di antaranya adalah pengguna aktif di media sosial. Sementara itu, diketahui, pengguna internet Indonesia terkini mencapai 202.6 juta dan 170 juta diantaranya pengguna aktif. *We Are Social* dan Hootsuite mengungkapkan jumlah pengguna Twitter di Indonesia adalah 14.05 juta [3].

Twitter Merupakan sebuah situs *mikroblog* yang mana situs *mikroblog* berevolusi menjadi sebuah sumber informasi karena *mikroblog* tempat orang bisa memposting pesan secara *real time* tentang pendapat mereka tentang berbagai topik. Maka dari itu para pengguna Twitter *men-tweet* informasi dengan berbagai topik secara *real time*, para pengguna Twitter sering berdiskusi tentang permasalahan yang baru atau yang sedang hangat dan mengungkapkan opini dengan sentimen positif atau negatif tentang topik tersebut yang dekat dengan aktivitas mereka [4].

Seleksi fitur sangat penting untuk klasifikasi text dan bagus untuk mempengaruhi performa klasifikasi. *Support Vector Machine* akan bermasalah ketika memilih parameter atau fitur yang sesuai untuk memilih fitur dalam SVM sangat berpengaruh hasil akurasi klasifikasi maka dari itu seleksi fitur itu sangat penting

untuk text *classification* dan mempengaruhi performa dan seleksi fitur yang cocok untuk kasus ini adalah *Information Gain* [8]. Hasil penelitian yang dilakukan mendapatkan akurasi SVM sebesar 83.05% dan ketika menggunakan *Information Gain* menghasilkan akurasi sebesar 85.65% mengalami peningkatan dari SVM [8]. Pada Jurnal penelitian lain juga membuktikan *Information Gain* merupakan model yang lebih baik dibandingkan *Chi Squared Statistic* dalam meningkatkan tingkat akurasi SVM, yaitu dengan menghasilkan rata-rata kenaikan tingkat akurasi sebesar 2.514% dengan tingkat akurasi terbaik sebesar 72.45% [9].

Berdasarkan penelitian terdahulu yang sudah dibahas secara singkat sebelumnya, pada penelitian ini akan mencoba untuk memaksimalkan hasil pendeteksian. Peneliti akan membandingkan pengambilan fitur dengan menggunakan seleksi fitur *Information Gain* (IG) dan ekstraksi fitur TF-IDF untuk meningkatkan akurasi dari pendeteksian dengan *Support Vector Machine*(SVM) untuk mendeteksi suatu konten yang menggunakan kata kasar di media Twitter dengan bahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas dirumuskan masalah sebagai berikut:

1. Apakah menggunakan *stemming* pada *preprocessing* dapat mempengaruhi akurasi?
2. Teknik pengambilan fitur apa yang memiliki akurasi paling tinggi dalam melakukan pendeteksian kata kasar di Twitter?
3. Berapa nilai parameter paling optimal setiap fungsi *kernel Support Vector Machine*?

1.3 Tujuan Penelitian

Tujuan dari tugas akhir ini adalah menganalisis sebuah konten berbahasa indonesia di Twitter, yang menggunakan kata kasar atau tidak dan membandingkan pengaruhi penggunaan seleksi fitur atau tidak, Seleksi fitur yang digunakan dengan metode *Information Gain*(IG) dan menggunakan *Support Vector Machine* 4 fungsi *gamma* yang berbeda, maka tujuan yang ingin dicapai dari penelitian ini adalah:

1. Mengetahui proses *stemming* diperlukan atau tidak untuk identifikasi kata kasar.
2. Menerapkan metode pengambilan fitur seperti metode TF-IDF dan *Information Gain* agar dapat disimpulkan pengambilan fitur yang paling baik dalam penelitian deteksi kata kasar di Twitter.

3. Mengetahui pengaruh *gamma* dengan parameter *gamma* dan C dan yang paling baik untuk objek penelitian ini.

1.4 Batasan Masalah

Untuk mempersempit masalah yang diteliti dan lebih terarah antara lain:

1. Data yang digunakan pada penelitian adalah *tweet* yang berbahasa indonesia.
2. Data set berupa konten pada Twitter yang di ambil dari kaggle.
3. Klasifikasi akan menganalisis ada, atau tidaknya suatu*tweet* yang menggunakan kata kasar.

1.5 Kontribusi Penelitian

Kontribusi yang diberikan dari penelitian ini adalah percobaan menggunakan *Support Vector Machine* dan mencoba apakah dengan seleksi fitur *Information Gain* akan mempengaruhi akurasi, untuk mengetahui suatu konten di Twitter menggunakan kata kasar atau tidak.

1.6 Metodologi Penelitian

Metodologi pada penelitian ini diawali dengan studi literatur dari beberapa sumber pustaka, baik sumber tertulis maupun artikel-artikel yang bersumber dari internet. Tugas Akhir ini akan diselesaikan dengan Metodologi berikut:

1. Studi Literatur
   Melakukan studi literatur mengenai konsep dari mengidentifikasi kata kasar sebagai dasar dari penelitian tugas akhir ini berdasarkan hasil penelitian dari para ahli maupun dari sumber-sumber lainnya, sehingga mendapatkan sebuah gambaran mengenai masalah yang ada dan melakukan analisis mengenai mendeteksi kata kasar.
2. Data sampling
   Data sampling yang akan digunakan berupa data yang diambil dari penyedia data terbuka di internet mengenai kata kasar di Twitter
3. Analisis Masalah
   Pada tahap ini dilakukan analisis permasalahan yang ada, batasan yang dimiliki dan kebutuhan yang diperlukan.
4. Perancangan dan Implementasi Algoritma
   Melakukan perancangan dan mengimplementasikan fungsi seleksi fitur

Information Gain dan ekstraksi fitur TF-IDF untuk pengambilan fitur lalu akan diklasifikasi dengan metode *Support Vector Machine* .

5. Pengujian
   Melakukan pengujian terhadap sistem yang sudah dibangun.

6. Dokumentasi
   Pada tahap ini dilakukan pendokumentasian hasil analisis dan implementasi secara tertulis dalam bentuk laporan skripsi

1.7 Sistematika Pembahasan

Laporan tugas akhir ini disusun berdasarkan sistematika penulisan sebagai berikut:

Bab 1: PENDAHULUAN
Bagian ini berisi latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, kontribusi penelitian, serta metode penelitian.

Bab 2: LANDASAN TEORI
Bagian ini menjelaskan teori-teori dasar yang akan digunakan dalam tugas akhir ini, diantaranya mengenai dataset, definisi kata kasar, pembobotan seleksi fitur *Information Gain* ,pembobotan dengan TF-IDF dan klasifikasi menggunakan *Support Vector Machine*.

Bab 3: ANALISIS DAN PERANCANGAN
Bagian ini menguraikan mengenai penggunaan dataset, pembobotan seleksi fitur *Information Gain* ,pembobotan dengan ektraksi fitur TF-IDF dan klasifikasi menggunakan *Support Vector Machine*. Sebagai acuan dalam penulisan tugas akhir ini.

Bab 4: IMPLEMENTASI DAN PENGUJIAN
Bagian ini akan akan membahas hasil dari hasil simulasi sistem yang telah diimplementasikan dan diuji oleh penulis.

Bab 5: KESIMPULAN DAN SARAN
Bagian ini yang berisi kesimpulan dari penelitian dan saran untuk penelitian lebih lanjut di masa mendatang.