BAB I

PENDAHULUAN

1.1 Latar Belakang

Media sosial merupakan sarana yang diciptakan untuk memfasilitasi seseorang membuat dan berbagi konten dengan orang lain. Secara umum, media sosial digunakan untuk keperluan personal seperti berkomunikasi dengan teman, kerabat, pasangan, dan lain-lain, atau digunakan untuk keperluan bisnis seperti membangun hubungan kerja sama antar organisasi, periklanan, hingga untuk menjual produk atau menawarkan jasa (Lewis, 2010). Mengacu pada definisi tersebut, maka contoh dari media sosial ialah seperti, Facebook, Instagram, Whatsapp, YouTube, X, dan lain-lain. Konten yang ada di media sosial bermacam-maacam, namun tidak terlepas dari gambar, audio, video, dan teks.

Pada tahun 2023, tercatat lima media sosial dengan pengguna aktif terbanyak yakni, Facebook sebanyak 2.958 juta pengguna, YouTube sebanyak 2.514 juta pengguna, Whatsapp sebanyak 2.000 juta pengguna, Instagram sebanyak 2.000 juta pengguna, dan WeChat sebanyak 1.309 juta pengguna (Kemp, 2023). Sedangkan, di Indonesia, berdasarkan laporan dan perkiraan per 14 Agustus 2023, terdapat sebanyak 228 juta pengguna media sosial dan diperkirakan akan bertambah terus hingga mencapai 267 juta pada tahun 2028 (Degenhardt, 2023). Angka tersebut merupakan angka yang relatif besar dan akan membuat lebih banyak lagi informasi dan konten yang beredar di media sosial, hal ini dapat menimbulkan banyak sekali potensi, diantaranya ialah potensi baik seperti sebuah pesan yang ditulis merupakan informasi yang mengandung kebaikan dan pembaca dapat mencerna informasi tersebut sesuai dengan apa yang diinginkan oleh penulis, di lain sisi, hal ini dapat menimbulkan potensi buruk seperti tersebarnya informasi yang salah, tersebarnya informasi yang palsu, ujaran kebencian, dan juga perundingan dunia maya atau *cyberbullying*. Maka dari itu, muncul banyak penelitian *machine learning* agar sebuah mesin yang pada dasarnya tidak mengetahui makna apa yang terkandung dalam sebuah kalimat menjadi tahu, proses inilah yang sering disebut dengan analisis sentimen.

Menurut penelitian yang dilakukan oleh Bo Pang dan Lillian Lee, Analisis sentimen menurupakan proses untuk mendeteksi sentimen dan pendapat dari teks (Pang & Lee, 2008). Terdapat banyak model yang telah dikembangkan untuk melakukan analisis sentiment, beberapa diantaranya ialah, Naïve Bayes, Deep Neural Network, Recurrent Neural Network, dan Support Vector Machine (SVM). Walapun ada banyak model yang dapat digunakan, secara umum setiap model memerlukan *input* berupa angka sebab mesin tidak dapat langsung memahami atau memproses teks tersebut. Maka dari itu, terlepas dari model yang ingin digunakan, data teks yang ingin dianalisis perlu dikonversikan menjadi numerik terlebih dahulu.

Ada beberapa metode untuk mengkonversi teks menjadi representasi numerik, dua diantaranya ialah, Term Frequency-Inverse Document Frequency (TF-IDF) dan Word2Vec. TF-IDF merupakan metode pemberian bobot pada suatu kata dengan cara mengevaluasi frekuensi kemunculan kata dan seberapa penting kata tersebut (Manning et al., 2008). Menurut penelitian yang dilakukan oleh Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Word2Vec merupakan metode untuk mengubah kata-kata menjadi vektor numerik berdimensi rendah. sehingga tercipta representasi distribusional dari sekumpulan kata, lalu, sekumpulan kata dengan makna serupa memiliki representasi vektor yang mirip. Tujuan dari medote ini untuk menghasilkan representasi vektor dari sebuah kata dengan memperhitungkan makna semantik antar kata (Mikolov et al., 2013).

Terdapat banyak penelitian analisis sentimen yang telah dilakukan menggunakan TF-IDF dan Word2Vec. Berdasarkan penelitian sebelumnya yang dilakukan oleh Imam Riadi, Abdul Fadlil, dan Murni telah dilakukan analisis sentimen menggunakan SVM dengan pembobotan kata menggunakan TF-IDF. Mereka berhasil mengumpulkan 5,000 dataset dari Twitter dengan cara menggunakan Twitter API dan Twitterscraper tool di python dengan kata kunci yang berhubungan dengan ujaran kebencian. Selanjutnya data tersebut dilakukan *preprocessing* data yakni case folding, tokenization, stopword removel, stemming. Setelah langkah *preprocessing data* selesai, dilakukan pelabelan data yang mengandung ujuran kebencian dan tidak, lalu dataset tersebut dilakukan pembobotan kata menggunakan TF-IDF, setelah itu dataset dibagi menjadi data *training* sebanyak 80% dari 5.000 dan sisanya digunakan sebagai data *testing*, setelah itu, dilakukan

klasifikasi menggunakan SVM. Penelitian ini menghasilkan bahwa akurasi tertinggi ialah 84% menggunakan RBF karnel dengan parameter C=10 dan gamma=0.1. Namun nilai presisi yang dihasilkan tertinggi diperoleh dari linear kernel dengan nilai 86% sementara RBF karnel bernilai 85%. Nilai recall yang dihasilkan tertinggi ialah 97% menggunakan RBF karnel dan F1-score tertinggi iala 91% menggunakan RBF karnel pula (Riadi et al., 2023). Pada penelitian lain yang dilakukan oleh Rafly Indra Kurnia, Yoshua Daniel Tangkuman, Abba Suganda Girsang berhasil melakukan analisis sentimen menggunakan SVM dengan pembobotan Word2Vec. Tahapan penelitian dilakukan dari pengumpulan data ulasan yang ada di website Google Play Store yang telah berhasil dilakukan oleh peneliti dengan teknik *scrapping* menggunakan program berbahasa python. Setelah data didapatkan, data teks dilakukan tahap preprocessing yang terdiri dari cleansing, stopwords removal, stemming, tokenization. Setelah itu, pembobotan kata dilakukan menggunakan Word2Vec. Setelah itu, dilakukan klaisifikasi menggunakan SVM yang menghasilkan lima kelas yakni angka 1 (satu) sampai 5 (lima) sebagai representasi rating yang ada di Google Play Store. Peneliti melakukan empat skenario uji, yang pertama, terdapat 5 kelas yang dihasilkan dan menggunakan stopword removal, yang kedua, terdapat 5 kelas yang dihasilkan dan tidak menggunakan stopword removal, yang ketiga terdapat 3 kelas yang dihasilkan menggunakan stopword removal, dan yang terkahir, terdapat 3 kelas yang dihasilkan tidak menggunakan stopword removal. Penelitian menghasilkan bahwa nilai F1 tertinggi ialah scenario ketiga yang menghasilkan 79.5% (Kurnia, 2020).

Berdasarkan penjelasan sebelumnya, masifnya pengguna media sosial menimbulkan ledakan teks yang luar biasa berpotensi terjadinya kebaikan dan keburukan. Satu diantara beberapa keburukan dari ledakan ini ialah *cyberbullying*. *Cyberbullying* berpotensi memiliki dampak buruk bagi psikologis, psikososial, akademik, bahkan fisik korban (Sukmawati et al., 2020). Maka dari itu, permasalahan ini dapat dijadikan sebagai penelitian untuk membandingkan performa metode pembobotan kata TF-IDF dan Word2Vec untuk analisis sentimen *cyberbullying* menggunakan SVM. Urgensi membandingkan dua metode pembobotan kata ini ialah berdasarkan dua penelitian yang menggunakan TF-IDF dan Word2Vec pada paparan sebelumnya, analisis sentimen dengan metode pembobotan kata TF-IDF

mendapatkan hasil akurasi yang lebih baik daripada penelitian yang menggunakan Word2Vec. Padahal, menurut penelitian yang dilakukan oleh Said A. Salloum, Rehan Khan, dan Khaled Shaalan, menyatakan bahwa semantik diperlukan dalam pengembangan *Natural Language Processing* (NLP) sebab dengan adanya semantik dapat mengatasi permasalahan ambiguitas dan tantangan lainnya yang berkaitan dengan NLP (Salloum et al., 2020). Namun, pembobotan kata menggunakan metode TF-IDF tidak secara langsung memperhitungkan makna atau semantik antar kata, metode ini melakukan pembobotan dengan cara menganalisis frekuensi kemunculan dan seberapa penting sebuah kata, sedangkan pembobotan kata menggunakan Word2Vec menghasilkan representasi vektor yang dapat mengetahui hubungan antar kata.

Maka dari itu, peneliti akan melakukan penelitian dengan judul "Perbandingan Performa TF-IDF Dan Word2Vec Untuk Analisis Sentimen *Cyberbullying* Menggunakan Metode Support Vector Machine (SVM)"

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, rumusan masalah umum dari peneltian ini ialah "Bagaimana perbandingan performa metode pembobotan kata TF-IDF dan Word2Vec untuk analisis sentimen *cyberbullying* menggunakan SVM". Berdasarkan rumusan masalah umum tersebut, dapat dirincikan rumusan masalah detail sebagai berikut:

1. Bagaimana perbandingan akurasi, presisi, dan *recall* untuk analisis sentimen cyberbullying menggunakan metode SVM dengan pembobotan kata TF-IDF dengan Word2Vec?

1.3 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah yang telah dipaparkan sebelumnya, tujuan secara umum dari penelitian ini ialah untuk mengetahui perbandingan performa untuk analisis sentimen menggunakan SVM dengan pembobotan kata menggunakan metode TF-IDF dan Word2Vec. Bedasarkan tujuan umum tersebut, tujuan penelitian dapat dirincikan sebagai berikut:

1. Mengetahui perbandingan performa metode TF-IDF dan Word2Vec untuk melakukan analisis sentimen menggunakan SVM dilihat dari tingkat akurasi, presisi, dan *recall*

1.4 Manfaat Penelitian

Dengan dilakukannya penelitian ini, dapat diambil manfaat yakni mengetahui performa yang lebih baik antara metode pembobotan kata TF-IDF dengan Word2Vec untuk analisis sentimen *cyberbullying* menggunakan SVM.

1.5 Batasan Masalah

Agar ketidakpastian penelitian yang timbul dapat dikurangi, maka terdapat beberapa batasan masalah karena terdapat banyak algoritma klasifikasi dan pembobotan kata yang dapat digunakan untuk analisis sentimen yang dapat digunakan. Adapun batasan masalah yang dimaksudkan ialah sebagai berikut:

1. Kernel yang digunakan pada metode SVM ialah linear, RBF, dan polynomial.
2. Window yang digunakan pada skenario Word2Vec ialah 5 (lima).