BAB I
PENDAHULUAN

1.1 Latar Belakang

Cabang dari kecerdasan buatan yang menjadi tren saat ini salah satunya yaitu pembelajaran mesin (*machine learning*), dimana penerapan *machine learning* dalam beberapa tahun terakhir berkembang di mana-mana. Bersamaan dengan hal tersebut, meningkatnya penggunaan sosial media sebagai sarana untuk berkomunikasi, informasi dan sebagai media hiburan bagi penggunanya tidak bisa dibendung keberadaannya. Dengan meningkatnya penggunaan situs *microblogging* seperti twitter dan sebagainya, dapat dimanfaatkan sebagai data untuk penerapan *machine learning* dengan membuat analisis sentimen, karena pada umumnya *machine learning* tidak akan bekerja tanpa adanya data.

Terdapat dua macam teknik dalam *machine learning* yaitu *supervised learning* dan *unsupervised learning*. Perbedaan antara kedua teknik tersebut adalah keberadaan label di subset data latih. *Supervised learning* melibatkan atribut *output* yang telah ditentukan selain penggunaan atribut *input*. Sebaliknya, *unsupervised learning* melibatkan pengenalan pola tanpa keterlibatan atribut target[1].

Konsep dasar dalam analisis sentimen itu adalah *text mining* dimana data teks yang tidak terstruktur harus diubah menjadi data semi terstruktur agar dapat ditemukan pola pada suatu dokumen. Proses pengubahan data teks menjadi data semi terstruktur tersebut merupakan tahap *text preprocessing*. Tahapan yang penting dalam *text preprocessing* yaitu *term weighting* (pembobotan kata) dimana tahapan ini dilakukan untuk memberikan suatu bobot pada *term*/kata yang terdapat

pada suatu dokumen. Setelah tahap *text preprocessing* dilakukan, barulah pengaplikasian algoritma klasifikasi dapat dilakukan.

Support Vector Machine (SVM) merupakan salah satu metode *supervised learning* yang populer digunakan untuk analisis sentimen, dimana metode ini mampu menemukan *hyperplane* terbaik sehingga menjadikannya sebagai algoritma dengan akurasi terbaik dibanding algoritma lainnya. Meskipun SVM memiliki akurasi yang baik dalam analisis sentimen, pemilihan fungsi kernel sangatlah penting dan berpengaruh terhadap akurasi yang dihasilkan. Ada beberapa kernel yang dapat digunakan diantaranya *Linear Kernel, Radial Basis Function* (RBF), dan *Polynomial Kernel*.

Dalam penelitian[2] di bidang *text mining*, mencoba melakukan identifikasi dan analisis terhadap perbandingan performa kernel pada algoritma SVM terhadap data *Child Autisme Disease database* untuk melihat seberapa efektif penerapan kernel pada algoritma SVM. Hasil yang didapat dari penelitian tersebut menunjukkan bahwa SVM dengan kernel *Polynomial* memperoleh hasil terbaik dengan akurasi 100%. Penelitian lainnya yang dilakukan [3] juga membandingkan performa kernel SVM terhadap dataset *movie review*, dimana hasil akurasi dari kernel RBF lebih tinggi dibanding dengan kernel lainnya yaitu sebesar 97,20%. Selain itu, pada penelitian[4] juga melakukan hal yang sama, hasil yang didapat algoritma SVM dengan kernel RBF lebih unggul dibanding kernel *Polynomial*. Menurut penelitian yang dilakukan oleh [5] bahwa pemilihan suatu kernel pada algoritma SVM sangat mempengaruhi akurasi yang dihasilkan.

Dalam penelitian ini penulis mencoba membandingkan 2 macam kernel dari algoritma *Support Vector Machine* yang merupakan kernel dengan akurasi terbaik dan sering digunakan untuk analisis sentimen yaitu kernel *Polynomial* dan *Radial Basis Function* (RBF) yang dipadukan dengan salah satu *term weighting* yang populer digunakan yaitu *Term Frequency Inverse Document* (TF-IDF) dan *Term Frequency Relevance Frequency* (TF-RF). Kedua teknik *term weighting* tersebut merupakan penggabungan metode dengan *Term Frequency* (TF) guna memperoleh performansi yang lebih baik[6].

Berdasarkan latar belakang di atas, maka peneliti melakukan penelitian dengan judul "Perbandingan Performa Kernel pada Algoritma *Support Vector Machine* dan *Term Frequency* terhadap Analisis Sentimen".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

a. Apakah penerapan kernel Polynomial dan RBF serta pembobotan kata TF-IDF dan TF-RF pada algoritma SVM berpengaruh terhadap analisis sentimen?
b. Bagaimana akurasi yang diperoleh algoritma SVM dengan menerapkan kombinasi antara kernel dan pembobotan kata dalam analisis sentimen?

1.3 Batasan Masalah

Dalam penelitian ini diperlukan batasan-batasan agar tujuan penelitian dapat tercapai. Adapun batasan masalah yang dibahas dalam penelitian ini adalah sebagai berikut.

a. Algoritma yang digunakan dalam penelitian ini adalah algoritma *Support Vector Machine* (SVM).
b. Kernel yang digunakan pada algoritma SVM adalah kernel *Polynomial* dan *Radial Basis Function* (RBF).
c. Pembobotan kata (*term weighting*) yang digunakan adalah *Term Frequency Inverse Document Frequency* (TF-IDF) dan *Term Frequency Relevance Frequency* (TF-RF).
d. Data diambil secara bertahap (maret-mei 2021) dari *crawling* data twitter berbahasa Indonesia dengan *keyword* Corona dan Covid sebanyak 16000 data.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

a. Mengetahui pengaruh penerapan kernel *Polynomial* dan RBF serta pembobotan kata TF-IDF dan TF-RF pada algoritma SVM terhadap analisis sentimen.
b. Mencari kombinasi terbaik dari kernel *Polynomial* dan RBF dengan pembobotan kata TF-IDF dan TF-RF pada algoritma SVM.

c. Mengetahui performa (akurasi, presisi, recall, dan f1) tiap kombinasi antara kernel dengan pembobotan kata pada algoritma SVM.

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

a. Memahami penerapan kernel *Polynomial* dan RBF serta pembobotan kata TF-IDF dan TF-RF pada algoritma SVM terhadap analisis sentimen.
b. Mengetahui hasil performa dari kombinasi tiap kernel dan pembobotan kata pada algoritma SVM.
c. Memperoleh kombinasi terbaik tiap kernel dan pembobotan kata dari hasil performa yang didapat.

1.6 Metode Penelitian

Penelitian ini berjenis penelitian eksperimen. Berikut merupakan tahapan-tahapan eksperimen yang dilakukan pada penelitian ini.

1.6.1 Pengumpulan Data

1.6.1.1 Studi Pustaka

Peneliti mengumpulkan literatur atau jurnal yang berkaitan dengan *text mining*, kernel pada algoritma *Support Vector Machine*, metode *term weighting* - seperti TF-IDF dan TF-RF.

1.6.1.2 Text Mining

Mengumpulkan data *twitter* dengan memanfaatkan *Twitter API* yang telah disediakan oleh *Twitter*.

1.6.2 Metode Analisis

Dalam metode analisis terdapat tahapan-tahapan sebagai berikut.

1.6.2.1 Text Preprocessing

Tahapan ini bertujuan untuk membersihkan data agar lebih mudah diproses pada tahap pembobotan kata serta tahap lainnya dan diharapkan hasil akurasi dari klasifikasi menjadi lebih akurat. Terdapat Langkah-langkah yang dilakukan pada saat proses *text preprocessing* yaitu *labeling*, *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*.

1.6.2.2 Pembobotan Kata

Tahapan pembobotan bertujuan untuk mendapatkan nilai/bobot dari tiap kata dasar yang berhasil diekstrak pada tahap sebelumnya. Pada penelitian ini pembobotan kata yang digunakan yaitu TF-IDF dan TF-RF.

1.6.2.3 Klasifikasi SVM

Pada analisis sentimen, tahapan klasifikasi merupakan tahapan yang utama. Pada tahap ini data diklasifikasi menggunakan algoritma *Support Vector Machine* dengan kernel *polynomial* dan *radial basis function* (RBF). Input yang digunakan adalah dari masing-masing pembobotan kata, maka akan ada empat kali proses atau kombinasi yaitu TF-IDF dan kernel *Polynomial*, TF-IDF dan kernel RBF, TF-RF dan kernel *Polynomial*, serta TF-RF dan kernel RBF.

1.6.3 Metode Pengujian

Untuk mengukur kinerja dari keempat kombinasi, metode yang digunakan adalah *confusion matrix*. Dari keempat kombinasi tersebut akan dihitung presisi, *recall*, f1 skor, dan juga akurasinya, agar dapat dibandingkan performa tiap kombinasi.

1.6.3.1 Akurasi

Pengujian ini menggambarkan seberapa akurat model dapat mengklasifikasi dengan benar. Untuk memastikan seberapa akurat model tersebut, peneliti menggunakan *K-Fold Cross Validation*.

1.6.3.2 Presisi

Presisi digunakan untuk mengukur tingkat keakuratan antara data dengan hasil prediksi yang diberikan oleh model.

1.6.3.3 Recall

Pengujian ini dapat menggambarkan seberapa besar keberhasilan model dalam menemukan kembali sebuah informasi.

1.6.3.4 F1 Skor

Sering terjadi dilema antara presisi dan *recall* yang disebabkan adanya *trade off* di antara keduanya. F1 skor merupakan *harmonic mean* dari presisi dan *recall*. Untuk menghindari dilema tersebut, maka skor yang digunakan adalah f1 skor.

1.7 Sistematika Penulisan

Sistematika penulisan skripsi ini dibagi dalam beberapa bab dengan pokok permasalahan sebagai berikut.

BAB I PENDAHULUAN

Pada bab ini dijelaskan mengenai latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan dalam penjabaran penelitian dan metode penelitian.

BAB II LANDASAN TEORI

Pada bab ini dijelaskan mengenai landasan teori dan kajian pustaka dari berbagai penelitian yang memiliki keterkaitan dengan penelitian ini. Kajian pustaka berguna untuk memperkuat dasar dan alasan dilakukanya penelitian ini. Selain Kajian Pustaka, Pada bab ini juga dijelaskan mengenai teori-teori terkait yang bersumber dari buku, jurnal, ataupun website yang berfungsi sebagai dasar dalam melakukan penelitian agar dapat memahami konsep atau teori penyelesaian permasalahan yang ada.

BAB III METODE PENELITIAN

Bab ini berisi penjelasan mengenai tahap pengumpulan data, analisis data, pembobotan kata, hingga tahap klasifikasi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil penelitian beserta pembahasannya.

BAB V PENUTUP

Bab ini berisi simpulan dari penelitian dan saran bagi penelitian mendatang yang berasal dari kekurangan maupun temuan dari penelitian ini.