

BAB I

PENDAHULUAN

1.1. Latar Belakang

Pada era saat ini sosial media diiringi *information technology* (IT) yang maju menjadi tempat informasi publik. Ini dapat terlihat dari meningkatnya jumlah penggunaan sosial media setiap tahunnya. Hal ini sangat menarik untuk diteliti, karena sebagian besar data yang ada pada sosial media tersebut mengandung opini sentimen. Sebagai contoh topik yang ramai dibicarakan beberapa saat yang lalu adalah vaksinasi Covid-19 yang merupakan penanggulangan virus yang menular dan menyebar ke seluruh dunia sejak akhir 2019 hingga saat ini masih sangat menarik perhatian dunia.

Sebagai langkah untuk dapat mengurangi penularan virus Covid-19, pemerintah mengeluarkan program vaksinasi sehingga tercapainya kekebalan tubuh pada masyarakat. Namun dengan adanya masalah-masalah yang dialami pada program vaksinasi sebelumnya yang dilakukan oleh pemerintah, membuat sebagian masyarakat menolak adanya vaksinasi kali ini. Untuk mendapatkan kembali kepercayaan masyarakat, pemerintah mencoba menyebarluaskan informasi vaksinasi lewat sosial media (*Instagram*), kemudian inilah awal mula yang menjadi daya tarik peneliti untuk mengolah dan meneliti lebih lanjut.

Melalui media sosial masyarakat dapat menjalin hubungan sosial antar pengguna secara jarak jauh, serta berbagi informasi, peristiwa, ataupun pengalaman yang dapat dibagikan pada media sosial mereka. Dengan

terbangunnya hubungan sosial tersebut, maka menghasilkan suatu data yang berbentuk sentimen. Dari banyaknya data tersebut masih terdapat beberapa hal yang masih sulit ditemukan, sebab itulah perlunya analisis teks. Analisis dilakukan untuk menghasilkan informasi spesifik (Negara, Andryani, & Saksono, 2016).

Dengan melakukan penelitian lebih lanjut terhadap sentimen masyarakat terhadap program vaksinasi yang dilakukan pemerintah melalui sosial media, akan didapatkan informasi yang baik terhadap respon dari masyarakat. Apakah program tersebut dapat sambutan yang baik dari masyarakat ataukah tidak. Juga akan diperoleh informasi yang lebih lengkap dan tepat, isu apa saja yang diterima baik dan isu apa saja yang tidak diterima oleh masyarakat.

Analisis sentimen atau penggalian opini merupakan sebuah studi komputasi opini, sentimen, evaluasi, sikap, suasana hati, dan emosi seseorang (Reyhana, 2018). Analisis sentimen merupakan salah satu area penelitian paling aktif dalam pemrosesan bahasa alami, penambangan data, pengambilan informasi, dan juga penambangan web (Gifari, Adha, Freddy, & Durrand, 2022). Analisis sentimen bagian dari penambangan opini (*opinion mining*), terutama menggunakan pemrosesan bahasa alami dan teknik ekstraksi informasi untuk melakukan penambangan dan analisis teks. Kecenderungan teks tertentu dinilai berdasarkan konteks dan polaritas yang diperoleh dan dapat menjadi argumen, opini, atau keadaan sentimental teks yang potensial (Septianingrum, Jaman, & Enri, 2021). Apabila semakin banyak data tekstual yang dikumpulkan dalam penelitian analisis sentimen, maka semakin mudah juga menemukan korelasi yang signifikan antara

teks dan jenis sentimen. Jenis sentimen tersebut berupa sentimen positif, negatif dan netral.

Salah satu masalah yang dikenali dalam proses analisis sentimen adalah jumlah data atau properti teks yang sangat besar (Az-Zahra, 2021). Dalam proses analisis sentimen, setiap kata atau disebut *term*, dikumpulkan menjadi properti atau dimensi, membentuk table data. Dikarenakan jumlah *term* yang sangat besar, hal ini menyebabkan proses yang terlalu lama dan memerlukan komputer dengan kekuatan atau kemampuan yang sangat besar. Selain itu, hal ini dapat menyebabkan turunnya kualitas dari model yang akan dibentuk, dikarenakan data yang terlalu besar, akan memberikan nilai bias yang besar juga (Kustiyo, Firqiani, & Giri, 2008). Tidak semua *term* mempunyai kontribusi atau relasi kepada keputusan atau *label* yang berupa nilai positif, negatif dan netral. Untuk itu, dalam penelitian ini akan digunakan metode pemilihan fitur, untuk memilih fitur atau *term* yang lebih memberikan kontribusi kepada keputusan atau *label*. Juga diharap hal ini dapat memberikan peningkatan dari kualitas model prediksi yang akan dibentuk (Hasibuan, 2019).

Sudah banyak dilakukan penelitian tentang proses pemilihan fitur dalam proses klasifikasi yang datanya berupa data diskrit, tetapi masih belum banyak adanya penelitian tentang pemilihan fitur untuk proses analisis sentimen. Sedangkan pada proses analisis sentimen, fitur yang terbentuk relatif jauh lebih banyak berbanding fitur yang ada pada proses klasifikasi data diskrit. Sehingga diharapkan terjadi pengurangan proses waktu yang signifikan dan hasil akurasi yang jauh meningkat.

Beberapa metode pemilihan fitur yang dapat diterapkan pada analisis sentimen, seperti yang dijelaskan dalam penelitian yang dilakukan oleh (Ahmad, Bakar, & Yaakub, 2019), diantaranya adalah metode *syntactic*, *semantic*, *lexico-structural*, *implicit*, *explicit* dan *frequent*. Di antara metode tersebut, yang akan digunakan dalam penelitian ini adalah metode *information gain* yang termasuk ke dalam metode *frequent* dan satu metode lainnya, yaitu metode *wrappers*.

Dalam penelitian ini, penulis akan melanjutkan hasil dari penelitian yang dilakukan oleh Anggraini, Negara Harahap, & Kurniawan (2021). Dalam hal ini penulis akan menggunakan dataset yang sama yang digunakan dalam penelitian sebelumnya, hanya saja dalam penelitian ini akan dilanjutkan dengan menerapkan pemilihan fitur dan lalu membandingkan hasil prosesnya dalam hal waktu proses dan tingkat akurasi dari model yang dibentuk, maka dari itu penulis mengajukan penelitian dengan judul **Analisis Metode Pemilihan Fitur pada Analisis Sentimen terkait Isu Vaksinasi Covid-19**.

1.2. Identifikasi Masalah

Adapun identifikasi masalah yang dapat diangkat dalam penelitian ini adalah:

1. Banyaknya data yang perlu diolah di dalam proses analisis sentimen.
2. Akurasi yang relatif rendah yang disebabkan banyak fitur yang tidak mempunyai relasi yang kuat kepada keputusan atau label, sehingga menjadi *noise* atau gangguan.

3. Belum banyaknya penelitian yang menerapkan pemilihan fitur pada analisis sentimen, sehingga peneliti merasa perlu untuk melakukan penelitian lebih lanjut untuk mempercepat waktu proses dan meningkatkan nilai akurasi.

1.3 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka hal yang dapat dirumuskan dalam penelitian ini secara garis besar adalah

1. Bagaimana cara mengelola data opini terkait isu vaksin covid-19 di Indonesia dan kemudian mengklasifikasikan data sentimen tersebut dengan menggunakan beberapa algoritma, seperti *Naive Bayes Classifier*, *Logistic Regression*, *Random Forest* dan *Support Vector Machine (SVM)*?
2. Bagaimana menambahkan proses pemilihan fitur yang sesuai dengan analisis sentimen?
3. Bagaimana cara menganalisis dan membandingkan hasil dari proses-proses klasifikasi dan pemilihan fitur untuk mengetahui metode mana yang paling baik dalam hal kecepatan waktu proses dan nilai akurasi yang diperoleh?

1.4 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Data sentimen yang digunakan adalah data yang digunakan pada penelitian terdahulu yang dilakukan oleh Anggraini, Negara Harahap, & Kurniawan (2021).

2. Algoritma klasifikasi yang akan dibandingkan dalam penelitian ini adalah *Naive Bayes Classifier*, *Logistic Regression*, *Random Forest* dan *Support Vector Machine (SVM)*.
3. Metode pemilihan fitur yang diusulkan adalah metode filter berupa *chi-square* dan *information gain* dan metode *wrappers* yaitu Algoritma Genetika (*Genetic Algorithm, GA*).
4. Hasil yang akan dibandingkan adalah berupa waktu proses dan nilai akurasi model dari masing-masing algoritma beserta masing-masing metode pemilihan fitur yang diusulkan.

1.5 Tujuan Penelitian

Tujuan dari penelitian ini adalah

1. Mengimplementasikan algoritma *Naive Bayes Classifier*, *Logistic Regression*, *Random Forest* dan *Support Vector Machine (SVM)*.
2. Mengimplementasikan metode pemilihan fitur pada masing-masing algoritma klasifikasi di atas.
3. Melakukan perbandingan hasil dalam hal kecepatan waktu proses dan nilai akurasi dari masing-masing kombinasi algoritma klasifikasi dan metode pemilihan fitur, untuk mengetahui algoritma dan metode mana yang memberikan hasil paling baik.

1.6 Manfaat Penelitian

Manfaat penelitian ini yang dilakukan oleh peneliti secara akademisi adalah, diharapkan:

1. Sebagai sumber pengetahuan, agar memberikan informasi gambaran mengenai sentimen positif dan negatif terkait isu vaksin covid-19 di Indonesia
2. Sebagai sumber referensi, penelitian ini dapat digunakan sebagai tambahan informasi referensi dan rujukan untuk penelitian selanjutnya sehingga dapat menemukan kajian penelitian yang baru dan berguna bagi kalangan akademisi

1.7 Susunan dan Struktur Proposal Tesis

Susunan dan struktur proposal tesis dibuat dengan maksud dapat memberikan garis-garis besar dari penulisan sehingga hubungan antara bab satu dengan bab yang lain dapat terlihat dengan jelas. Adapun susunan dan struktur proposal tesis adalah sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini penulis membahas mengenai latar belakang, identifikasi masalah, batasan masalah, perumusan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup penelitian, serta susunan dan struktur proposal tesis.

BAB II KAJIAN PUSTAKA

Pada bab ini membahas tentang landasan teori yang berhubungan dengan penelitian terdahulu yaitu Tinjauan Pustaka yang berisi pembahasan Data Mining, Text Mining, Algoritma klasifikasi seperti *Naïve Bayes Classifier*, *Logistic Regression*, *Random Forest* dan *Support Vector Machine (SVM)* dan pembahasan penelitian terdahulu yang terkait dengan pemilihan fitur pada analisis sentimen.

BAB III METODOLOGI PENELITIAN

Pada bab ini pembahasan terdiri dari desain dan jadwal penelitian, data penelitian, kemudian konsep dan metode penelitian yang digunakan, metode pengumpulan data serta teknik pengelompokan data.

BAB IV TEKNIK ANALISIS DATA

Pada bab ini, akan dibahas permasalahan yang akan diselesaikan dalam penelitian. Bagaimana meningkatkan akurasi dan mengurangi jumlah data atau fitur yang perlu di proses. Terdapat beberapa proses untuk mencapai hasil yang maksimal.

BAB V HASIL DAN PEMBAHASAN

Pada bab ini, akan dibahas tentang hasil dari pembentukan model untuk tiap-tiap algoritma, lalu dibandingkan dengan hasil setelah diterapkan pemilihan fitur dengan menggunakan variabel jumlah fitur yang dipilih berbeda-beda. Berikutnya, dari hasil perbandingan, dilakukan pembahasan mengapa hasil tersebut dapat diperoleh.

BAB VI KESIMPULAN

Pada bab ini, akan dibahas tentang kesimpulan yang dapat diambil dalam penelitian ini.

LAMPIRAN

Berisi lampiran pendukung dari penelitian