BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada saat ini persebaran *Corona virus Disease 2019* (COVID-19) kembali meningkat. Di Indonesia sendiri pada bulan Januari 2021 – Februari 2021 tercatat 1.217.468 orang yang terkonfirmasi positif virus *corona* berdasarkan data dari Komite Penanganan Covid-19 dan Pemulihan Ekonomi Nasional. Karena peningkatan angka tersebut, pemerintah melakukan upaya-upaya pencegahan yang salah satunya adalah dengan pendistribusian vaksin kepada masyarakat Indonesia yang telah dimulai sejak tanggal 13 Januari 2021 (Anwar, 2021).

Berdasarkan Keterangan Pers Menteri Kesehatan Budi Gunadi Sadikin bahwa dari 269 juta rakyat Indonesia, pemerintah menyiapkan 181 juta rakyat Indonesia yang akan divaksin. Vaksin yang akan di distribusikan ke masyarakat dilakukan secara bertahap dari bulan Januari 2021 – Maret 2022. Pendistribusian tahap awal yaitu berupa vaksin *Sinovac*. Terkait uji klinis vaksin *Sinovac* sendiri, BPOM telah merilis hasil evaluasi dari laporan uji klinis interim tahap III yang menunjukkan efikasi atau tingkat keampuhan vaksin covid-19 *Sinovac* sebesar 65,3 persen, dimana sudah sesuai dengan standar dan ambang batas efikasi yang ditetapkan oleh *World Human Organization* (WHO) yakni minimal 50 persen berdasarkan pernyataan Kepala BPOM Penny K Lukiot yang dilansir cnnindonesia.com (CNN, 2021).

Upaya vaksinasi covid-19 yang dilakukan oleh pemerintah tersebut memberikan pengaruh luas pada kalangan masyarakat melalui media sosial

(khususnya *Twitter*) yang kemudian memunculkan pro dan kontra di masyarakat. Walaupun BPOM sudah merilis hasil uji klinis, masih ada masyarakat yang meragukan keefektifan dan keampuhan vaksin covid-19, mengingat masih ada beberapa vaksin covid-19 masih dalam fase penelitian dan uji coba. Maka dari itu dibutuhkan analisis sentimen untuk melihat bagaimana keberhasilan upaya vaksinasi covid-19 yang dilakukan oleh pemerintah.

Algoritma *Random Forest Classifier* merupakan metode klasifikasi di bidang *Machine Learning* yang memiliki kinerja tinggi dibandingkan metode klasifikasi lainnya (Heryadi et al, 2020) dan memiliki kelemahan dalam hal tingkat keakuratan yang relatif rendah dan kestabilan datanya. Hal ini terkait dengan fungsi acak yang dibangkitkan untuk melakukan pemilihan baris data dan pemilihan kandidat atribut pemecah secara acak (Adhitya et al, 2015). Maka dibutuhkan fitur seleksi berupa *Information Gain* untuk meningkatkan akurasi dan memperbaiki kestabilan data pada algoritma tersebut.

Ada beberapa penelitian tentang analisis sentimen dengan beberapa model algoritma yang mendukung penelitian ini, seperti penelitian yang dilakukan oleh Januarsyah et al, 2019 serta Fitri, 2020 yang mengkomparasi algoritma *Random Forest* dengan algoritma *Naïve Bayes, Support Vector Machine, Decision Stump, Naïve Bayes, Bayesian Network,* dan C4.5 yang menghasilkan tingkat akurasi tertinggi ada pada algoritma *Random Forest* dengan tingkat akurasi sebesar 97,16% dan 74,2863%, serta penelitian yang dilakukan oleh Wijaya, 2017 dan Somantri et al, 2018 mengenai penerapan fitur seleksi pada algoritma SVM dan *Naïve Bayes* menggunakan *Information Gain* dan *Chi square* dan didapatkan hasil peningkatan

akurasi tertinggi yaitu sebesar 3,08% dan 1,0075% dari penerapan *Information Gain*.

Berdasarkan uraian di atas, maka akan dilakukan sentimen analisis menggunakan algoritma *Random Forest Classifier* dengan melakukan seleksi fitur yaitu *Information Gain* untuk meningkatkan akurasi serta mengoptimalkan atribut pada data dengan mengambil judul " Analisis Sentimen Terhadap Opini Masyarakat Terkait Vaksinasi Covid-19 Pada *Twitter* dengan Algoritma *Random Forest Classifier* dan *Information Gain*"**.**

1.2 Rumusan Masalah

Berdasarkan pada latar belakang, maka rumusan masalah penelitian ini adalah sebagai berikut :

1) Bagaimana hasil performansi dari algoritma *Random Forest Classifier* sebelum dan setelah penerapan *Information Gain* ?
2) Bagaimanakah hasil prediksi sentimen terhadap opini masyarakat terkait Vaksinasi Covid-19 pada *Twitter* dengan menggunakan Algoritma *Random Forest Classifier* setelah penerapan *Information Gain*?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut :

1) Algoritma yang digunakan dalam penelitian ini, yaitu algoritma *Random Forest Classifier* untuk menganalis polaritas sentimen terhadap opini masyarakat terkait Vaksinasi Covid-19 pada *Twitter*.
2) Menerapkan *Information Gain* pada algoritma *Random Forest Calssifier* untuk meningkatkan akurasi dan mengoptimalkan atribut.

3) Data yang digunakan berasal dari media sosial *Twitter* dengan rentang waktu bulan Mei-Juni dan bulan Juli untuk melihat perbedaan persentase kelas sentimen yang dihasilkan sebelum dan setelah masyarakat umum mendapatkan vaksin Covid-19.
4) Menggunakan bahasa Pemrograman *Python* untuk mengimplementasikan algoritma yang digunakan dalam menganalisis dan memprediksi sentimen terhadap opini masyarakat terkasit Vaksinasi Covid-19 pada *Twitter*.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1) Menguji performansi dari algoritma *Random Forest Classifier* sebelum dan setelah penerapan *Information Gain*,
2) Memprediksi opini masyarakat terkait Vaksinasi Covid-19 pada *Twitter* dengan menggunakan algoritma *Random Forest Classifier* setelah penerapan *Information Gain* berdasarkan nilai *accuracy*, *precision, recall* dan *f1-score* yang dihasilkan.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut :

1) Mampu menguji performansi dari algoritma *Random Forest Classifier* sebelum dan sesudah penerapan *Information Gain*.
2) Mampu memprediksi respon ataupun opini masyarakat terkait kebijakan Vaksinasi Covid-19 yang dilakukan oleh pemerintah
3) Dapat menghasilkan informasi mengenai respon ataupun opini masyarakat terkait kebijakan vaksinasi Covid-19 yang dilakukan oleh pemerintah

sehingga dapat dimanfaatkan sebagai tolak ukur keberhasilan kebijakan tersebut dan dari respon masyarakat tersebut dapat dijadikan bahan evaluasi pada kebijakan tersebut agar pemerintah lebih memperhatikan keamanan dan efektivitas vaksinasi Covid-19.

1.6 Metodologi Penelitian

Terdapat empat proses utama dalam penelitian ini, yaitu proses pada tahapan awal, *data processing*, *text classification,* dan analisis *result*. Berikut adalah penjelasan dari proses-proses tersebut:

1.6.1 Tahapan Awal

Pada tahapan awal terbagi menjadi dua proses yang akan dilakukan sebagai berikut:

1) Pengumpulan Data

Penelitian ini diawali dengan pengumpulan data dengan menggunakan teknik *data crawling* pada *twitter* untuk mengumpulkan data berupa *tweet* yang akan dipakai pada penelitian ini dengan menggunakan akses dari pihak *twitter* bagi pengguna dengan memanfaatkan *Twitter API*.

2) *Data Correction*

Proses koreksi ini dilakukan di *MS.Excel* dengan mengimport *file CSV*, dengan ini format data akan berubah menjadi *xslx* dari format *csv* untuk mempermudah pengoreksian data.

1.6.2 *Data Processing*

Data processing terdiri enam tahap *text preprocessing* serta proses *data labelling* dan *polarity* yang akan dijelaskan sebagai berikut:

1.6.2.1 *Text Preprocessing*

1) *Data Cleaning*, tahap ini akan melakukan pembersihan kalimat dan menghilangkan tanda baca dari kalimat
2) *Case Folding*, tahap ini akan melakukan pemeriksaan ukuran setiap karakter dari awal sampai akhir karakter dan jika ditemukan karakter menggunakan huruf kapital (*uppercase*), maka huruf tersebut akan diubah menjadi huruf kecil (*lowercase*).
3) *Tokenization*, tahap ini dilakukan pembagian kalimat menjadi perkata (merubah kalimat dan teks menjadi sebuah token-token) untuk menghapus kata-kata yang tidak penting pada tahap *stop removal*.
4) *Spell Checking*, tahap ini akan melakukan penghilangan kata-kata yang menggangu pada teks dan juga membenahi kata-kata yang berbentuk singkatan ataupun *typo* menjadi kata yang sesuai dengan kamus KBBI.
5) *Stop Removal*, tahap ini akan melakukan penghapusan kata-kata yang terlalu umum dan kurang penting yang memiliki frekuensi kemunculan yang jumlahnya cukup banyak dibandingkan dengan kata yang lainnya. Tahap ini juga membuang kata-kata yang tidak deskriptif.
6) *Stemming*, tahap ini akan melakukan perubahan kata menjadi ke bentuk dasarnya yaitu dengan menghilangkan semua imbuhan kata pada kata turunannya.

1.6.2.2 *Data Labelling dan Polarity*

1. *Data Labelling*

Proses ini dilakukan menggunakan *library python* yaitu *Textblob* untuk memberi label secara otomatis dengan memberikan *score* di setiap cuitan yaitu sebagai berikut:

a) Label positif memiliki *score* > 0
b) Label netral memiliki *score* == 0
c) Label negatif memiliki *score* < 0

2. *Polarity*

Proses ini dilakukan untuk mempermudah proses selanjutnya yaitu proses *text classification*. Metode yang digunakan pada *text classification* merupakan metode *machine learning* dimana metode tersebut tidak dapat melatih teks secara langsung sehingga data teks harus diubah menjadi numerik. Maka dilakukan proses *polarity* untuk mengkonvert label ke polaritas sebagai berikut:

a) Label positif memiliki nilai *polarity* = 1
b) Label netral memiliki nilai *polarity* = 0
c) Label negatif memiliki nilai *polarity* = -1

1.6.3 *Text Classification*

Tahap klasifikasi teks yang terbagi menjadi dua proses yakni:

1) *Random Forest Classifier* tanpa *Information Gain*

Proses ini dilakukan langkah-langkah bagaimana *Random Forest Classifier* untuk mengklasifikasi *dataset* sehingga dapat mengetahui cara

kerja dari metode *Random Forest Classifier* dan untuk mendapatkan hasil terakhir (pada tahap analisis hasil).

2) *Random Forest Classifier* dengan *Information Gain*

Proses *Random Forest Classifier* dengan *Information Gain* dilakukan untuk melihat kinerja dari *feature selection* berupa *Information Gain* pada pengoptimalan tingkat akurasi klasifikasi dari algoritma *Random Forest Classifier*.

1.6.4 Analisis Result

Pada tahap ini akan dilakukan prediksi sentimen serta perbandingan performa dari metode *Random Forest Classifier* dengan *Information Gain* dan metode *Random Forest Classifier* tanpa *Information Gain* dengan membagi *dataset* menjadi 20% *data testing* dan 80% *data training*. Hasil yang didapatkan akan ditampilkan berupa *confusion matrix*, dari nilai *confusion matrix* ini di hitunglah nilai *accuracy, precission, recall, dan F1 score*.

1.7 Sistematika Penulisan

Laporan Tugas Akhir ini dibagi menjadi 5 bab yang saling berhubungan untuk memberikan gambaran yang jelas mengenai pembahasan Tugas Akhir. Sistematika penulisan yang digunakan adalah sebagai berikut:

BAB I : PENDAHULUAN

Bab ini membahas tentang garis besar keseluruhan laporan. Bab ini berisi tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan.

BAB II : LANDASAN TEORI

Bab ini berisi tentang uraian penelitian-penelitian terkait serta dasar teori yang menjadi rujukan dalam penelitian. Sumber referensi yang menjadi acuan adalah buku, jurnal, dan media elektronik. Dasar teori yang menjadi rujukan dalam penelitian adalah penjelasan *Twitter*, Analisis Sentimen, *Random Forest Classifier*, *Information Gain*, dan *Literature Review*

BAB III : METODOLOGI

Bab ini berisi tentang metode yang digunakan dalam pembahasan serta langkah-langkah penyelesaian masalah selama melakukan penelitian. Langkah-langkah tersebut memuat konsep dari metode yang digunakan, analisis kebutuhan, serta uraian-uraian lainnya yang berkaitan dengan penelitian yang akan dibahas.

BAB IV : HASIL DAN PEMBAHASAN

Bab ini berisi tentang hasil dari proses analisis yang telah dilakukan serta pembahasan yang meliputi penjelasan mengenai data yang akan di analisis. Bab ini juga dipaparkan data-data pendukung dalam pelaksanaan penelitian serta penerapan metode yang digunakan.

BAB V : KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil analisis sentimen opini masyarakat terhadap Vaksinasi COVID-19 pada *Twitter* dengan algoritma *Random Forest Classifier* dan *Information Gain* serta beberapa saran yang bermanfaat untuk penelitian di masa mendatang.