BAB I

PENDAHULUAN

A. Latar Belakang

Natural Language Processing (NLP) adalah sub-bidang dari *Artificial Intelligence*, *Computer Science*, dan *Linguistics* yang menggunakan algoritme *Machine Learning* untuk menafsirkan dan memanipulasi bahasa manusia [1]. NLP mengombinasikan *Computer Science* dan *Linguistics* dalam mempelajari aturan dan struktur bahasa untuk menciptakan sistem cerdas yang mampu memahami, menganalisis, dan mengekstrak makna dari data teks dan suara. NLP bekerja menggunakan *Text Vectorization* untuk mengubah data teks menjadi vektor angka, sehingga data tersebut dapat dipahami oleh komputer atau mesin. Sentimen Analisis didefinisikan sebagai salah satu aplikasi NLP yang digunakan untuk mengidentifikasi opini pada data teks. Umumnya opini tersebut dikategorikan sebagai *positive*, *negative*, *neutral*, atau apa saja di antara itu [2].

Sentimen Analisis disebut juga sebagai *Opinion Mining*, yaitu bidang studi yang menganalisis pendapat, sentimen, penilaian, penaksiran, sikap, dan emosi orang terhadap entitas seperti produk, layanan, organisasi, individu, masalah, kejadian, topik, dan atribut-atributnya. Sentimen Analisis menggambarkan sebuah lingkup subjek yang luas. Terdapat juga banyak istilah dan peran yang sedikit berbeda seperti, *Sentiment Analysis*, *Opinion Mining*, *Opinion Extraction*, *Sentiment Mining*, *Subjectivity Analysis*, *Affect Analysis*, *Emotion Analysis*, *Review Mining*, dll. Meskipun demikian, saat ini hal itu semua berada dalam satu lingkup kesatuan yaitu Sentimen Analisis atau *Opinion Mining* [3].

Sentimen Analisis telah digunakan dalam berbagai domain seperti pemasaran, politik, sosial, dll. Perusahaan menggunakan Sentimen Analisis untuk mengembangkan strategi bisnis, untuk memahami sentimen pelanggan terhadap

produk atau merek dagang mereka, serta bagaimana tanggapan orang atas kampanye atau peluncuran produk dan alasan mengapa jika konsumen tidak mau membeli produk mereka. Dalam bidang politik, digunakan untuk melacak preferensi politik, untuk mendeteksi konsistensi dan inkonsistensi antara ucapan dan tindakan pada level pemerintahan, dan juga digunakan untuk memprediksi hasil pemilihan umum. Sentimen Analisis juga digunakan dalam bidang sosial untuk mengamati dan menganalisis fenomena sosial, untuk melihat sebuah ancaman dan sebagai sistem pendukung dalam membuat keputusan. Sentimen Analisis juga memainkan peran penting dalam berbagai subjek, seperti *stakeholders*, pembuat kebijakan, dan perusahaan untuk melakukan tugas seperti memahami persepsi pelanggan, memberikan peringatan dini, memprediksi performa finansial, dll [4].

Dalam penelitian ini, Penulis berkonsentrasi mengembangkan model *Machine Learning* yang dapat melakukan tugas menganalisis sentimen pada data teks. Pengembangan model ini secara khusus digunakan untuk menganalisis sentimen pada studi kasus keterbatasan kesempatan kerja dalam bidang yang diminati bagi *Gen Z*. Menurut sumber Oxford Learner's Dictionaries, *Gen Z* atau *Generation Z* adalah sekelompok orang yang lahir antara akhir tahun 1990-an dan awal 2010-an. Penelitian Pew Research Center menyatakan *Gen Z* adalah orang yang lahir antara tahun 1997 dan 2012. Berdasarkan sensus penduduk tahun 2020, mayoritas penduduk di Indonesia didominasi oleh *Gen Z*. Jumlah *Gen Z* sebanyak 27,94% dari 270,20 juta jiwa penduduk Indonesia pada tahun 2020. Menurut Badan Pusat Statistik, *Gen Z* merupakan orang yang lahir antara tahun 1997-2012. Menurut hasil sensus penduduk tahun 2020 Indonesia masih dalam masa bonus demografi, persentase penduduk usia produktif (15-64 tahun) adalah 70,72% dari 270,20 juta jiwa. Sebagai mayoritas yang sebagian dari anggotanya akan dan telah memasuki usia kerja, dirasa penting untuk dilakukan penelitian ini untuk mengetahui sentimen *Gen Z* mengenai keterbatasan kesempatan kerja dalam bidang yang diminati. Keterbatasan kesempatan kerja atau kesempatan kerja yang kurang beragam adalah salah satu indikator dari sebuah negara ekonomi berkembang. Indonesia adalah negara berkembang, dengan jumlah

pengangguran yang terdeteksi sebanyak 7,9 juta orang pada awal 2023 dari laporan Badan Pusat Statistik.

Fokus pengembangan model dalam penelitian ini adalah untuk membuat sebuah sistem yang dapat melakukan klasifikasi dan dapat menafsirkan opini yang disampaikan dalam bentuk data teks, kemudian mengembalikan hasil berupa kategori sentimen yaitu *positive* atau *negative*. Sehingga, diharapkan model dapat melakukan tugas klasifikasi secara otomatis dan dapat mengekstrak *output* berupa informasi yang akurat. Penelitian ini mengimplementasikan algoritme *Naive Bayes* untuk melakukan tugas klasifikasi.

Dalam mengembangkan model *Machine Learning* untuk tugas Sentimen Analisis diperlukan *tool* dan bahasa pemrograman. Python adalah salah satu bahasa pemrograman tercanggih untuk melakukan tugas analisis pada data tekstual. Python juga dikenal sebagai bahasa pemrograman tingkat tinggi, digunakan untuk tujuan yang luas, bersifat *open-source*, dan mudah digunakan. Python efektif digunakan untuk kebutuhan *Machine Learning*, seperti memproses *dataset* yang besar dan melakukan komputasi matematik [5]. Pengembangan model ini menggunakan *tool* Jupyter Notebook untuk menulis dan mengeksekusi kode program.

Berdasarkan uraian yang telah dijelaskan, maka dapat disimpulkan bahwa Sentimen Analisis adalah sebuah teknik dan aplikasi dari NLP yang digunakan untuk memahami nada emosional pada teks atau pun suara. Sentimen Analisis dapat digunakan untuk mengidentifikasi sentimen atau opini pada sebuah teks, pada umumnya adalah sentimen *positive, negative, neutral*, atau apa saja di antara itu. Informasi atau *output* yang didapatkan dari hasil analisis sentimen sangat berharga, terutama dalam bidang bisnis menggunakan informasi tersebut untuk memahami opini pelanggan terhadap produk atau layanan perusahaan. Sentimen Analisis telah banyak diimplementasikan dalam berbagai domain dan telah banyak memberikan manfaat, sehingga dirasa sangat penting untuk diterapkan dan dikembangkan. Lebih khususnya, mengembangkan model *Machine Learning* yang dapat melakukan tugas

Sentimen Analisis untuk menyelesaikan berbagai permasalahan dalam kehidupan sehari-hari dalam Bahasa Indonesia.

B. Rumusan Masalah

Berdasarkan latar belakang masalah dari penelitian ini, maka diperoleh beberapa rumusan masalah sebagai berikut :

1. Apa langkah-langkah yang diperlukan untuk mengembangkan sebuah model *Machine Learning* untuk melakukan tugas Sentimen Analisis?
2. Berapa tingkat akurasi model dalam mengklasifikasi sentimen dengan menggunakan algoritme *Naive Bayes* ?
3. Apa hasil dari analisis sentimen dengan studi kasus keterbatasan kesempatan kerja dalam bidang yang diminati bagi *Gen Z*, apakah sentimen *Positive* atau *Negative*?
4. Apa kelebihan dan kekurangan model?

C. Batasan Masalah

Berdasarkan studi kasus dan rumusan masalah dalam penelitian ini, berikut adalah beberapa batasan-batasan masalah yang perlu diperhatikan :

1. Penelitian ini hanya menggunakan data dalam bentuk teks Bahasa Indonesia.
2. Pengumpulan data dilakukan dengan mengambil data dari Twitter, halaman blog, dan survei *online* melalui Google Form.
3. Penelitian ini menggunakan algoritme pengklasifikasi *Naive Bayes* untuk membuat pemodelan data.
4. Menggunakan *tool* Jupyter Notebook untuk menulis kode program.
5. Menggunakan bahasa pemrograman Python.

6. Data yang digunakan untuk melatih dan menguji model adalah data opini atau sentimen *Gen Z* mengenai fenomena keterbatasan kesempatan kerja dalam bidang yang diminati.

D. Tujuan Penelitian

Tujuan dilakukan penelitian ini adalah sebagai berikut :

1. Mengetahui langkah-langkah dalam proses pengembangan model *Machine Learning* untuk tugas Sentimen Analisis.
2. Mengetahui tingkat akurasi model dalam mengklasifikasi sentimen menggunakan algoritme *Naive Bayes*.
3. Mengetahui hasil dari analisis sentimen berdasarkan studi kasus keterbatasan kesempatan kerja dalam bidang yang diminati bagi *Gen Z*.
4. Mengetahui kelebihan dan kekurangan model.

E. Metode Pengembangan Model

Metode pengembangan model merupakan tahapan-tahapan teknik yang digunakan dalam proses pengembangan model *Machine Learning*. Adapun penjelasan dari metodologi yang digunakan dalam penelitian ini, sebagai berikut:

1. Studi Literatur

Penulis melakukan studi literatur, yaitu dengan mengumpulkan informasi yang berkaitan dengan topik penelitian. Informasi tersebut dapat berupa jurnal penelitian terdahulu, artikel, video pembelajaran, dan dokumentasi kode program. Studi literatur ini, digunakan sebagai landasan berpikir dan referensi untuk melakukan penelitian ini.

2. Pengumpulan Data (*Data Collection*)

Data Collection adalah tahap pengumpulan data untuk kebutuhan penelitian. Pengumpulan data dilakukan dengan menggunakan beberapa teknik, berdasarkan dari kebutuhan data yang sudah diidentifikasi. Beberapa teknik yang digunakan, yaitu media kuesioner *online*, mengambil dari media sosial, dan mengambil data secara langsung dari artikel web atau blog, dll [6].

3. Analisis Data Eksploratif (*Exploratory Data Analysis*)

Exploratory Data Analysis adalah sebuah pendekatan yang digunakan untuk menganalisis data dan menemukan tren, pola, dan memeriksa asumsi di dalam data dengan bantuan ringkasan statistik dan representasi grafis [7]. *Exploratory Data Analysis* juga dapat membantu mengidentifikasi data yang hilang, eror, *outliers*, anomali, dan data yang tidak konsisten, yang dapat memiliki dampak yang signifikan dalam analisis data. Adapun, *Exploratory Data Analysis* penting untuk dilakukan untuk menyiapkan data untuk analisis lebih lanjut dalam pengembangan model *Machine Learning*.

4. *Preprocessing Data*

Preprocessing Data adalah proses menyiapkan data mentah dan mengubahnya menjadi format yang dapat dipahami oleh mesin. Dalam database dunia nyata umumnya data mengandung, *noise*, nilai yang hilang, data yang tidak konsisten, dan mungkin dalam format yang tidak dapat digunakan [8]. Dikarenakan ukurannya yang biasanya sangat besar dan kemungkinan berasal dari berbagai sumber yang heterogen.

5. Visualisasi Data

Visualisasi merupakan proses menggunakan elemen visual seperti diagram, grafik, atau peta untuk merepresentasikan data. Visualisasi data dapat membantu menerjemahkan data yang kompleks dan bervolume tinggi menjadi representasi visual yang lebih mudah dipahami [9]. Metode ini dapat membantu Penulis untuk mengomunikasikan data dalam bentuk visual, agar dapat mengekstraksi wawasan tentang data dengan lebih mudah.

6. Pelatihan Model

Merupakan proses melatih model *Machine Learning* yang melibatkan algoritme *Machine Learning* dan data pelatihan untuk dipelajari. Istilah model *Machine Learning* merujuk pada artefak model yang dibuat dari proses pelatihan. Data pelatihan harus meliputi jawaban yang benar, yang dikenal sebagai variabel target atau label. Algoritme akan mempelajari pola dalam data pelatihan yang memetakan atribut data *input* ke label (jawaban yang ingin diprediksi), dan menghasilkan model *Machine Learning* yang menangkap pola-pola tersebut. Sehingga, menghasilkan model yang dapat digunakan untuk melakukan prediksi pada data baru yang tidak diketahui target atau labelnya [10].

7. Evaluasi Model

Tujuan dari model *Machine Learning* adalah mempelajari pola dan menggeneralisasi data baru yang belum pernah dilihat sebelumnya. Setelah model dilatih, penting untuk memeriksa apakah model dapat berkinerja dengan baik pada data baru yang belum pernah dilihat sebelumnya atau yang belum pernah digunakan untuk melatih model. Evaluasi model adalah proses yang menggunakan beberapa metrik untuk membantu mengukur kinerja model dalam melakukan prediksi. Terdapat sejumlah metrik yang digunakan untuk mengevaluasi kinerja atau kualitas

model, metrik-metrik tersebut disebut sebagai metrik evaluasi atau metrik performa [11].

F. Sistematika Penulisan

Untuk mempermudah pembaca dalam memahami, berikut adalah sistematika penulisan laporan dalam penelitian ini.

1. BAB 1 berisi penjelasan mengenai pendahuluan, yang terdiri dari latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, dan metode pengembangan model.
2. BAB 2 berisi tinjauan pustaka yang menjelaskan penelitian-penelitian terdahulu yang berkaitan dengan penelitian ini beserta dengan tabel perbandingannya.
3. BAB 3 berisi landasan teori, yaitu menjelaskan teori-teori yang digunakan dalam penelitian ini.
4. BAB 4 berisi penjelasan mengenai *dataset* dan pengembangan model.
5. BAB 5 berisi hasil dan pembahasan.
6. Kemudian, pada BAB 6 dijelaskan kesimpulan dari penelitian ini.