BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Seiring perkembangan zaman dan teknologi telekomunikasi yang semakin maju membuat masyarakat Indonesia lebih mudah untuk bisa mengakses internet. Pengguna internet di Indonesia saat ini mengalami pertumbuhan pesat. Jumlah pengguna internet di Indonesia sebesar 175.4 juta jiwa pada tahun 2020 yang meningkat sebesar 17% antara tahun 2019 dan 2020 [1]. Berbagai hal dilakukan masyarakat Indonesia dalam menggunakan internet, salah satunya menggunakan sosial media Twitter untuk berkomunikasi dan mencari hiburan. Pada tahun 2020, Twitter menduduki peringkat 10 dari *website* yang sering dikunjungi oleh masyarakat Indonesia [1]. Twitter merupakan layanan jejaring sosial dan *microblog* daring yang memungkinkan pengguna untuk mengirim dan membaca pesan berbasis teks hingga 280 karakter yang dikenal dengan sebutan kicauan (*tweet*) [2]. Twitter memiliki banyak keunggulan, salah satunya lebih bebas untuk mengunggah sesuatu daripada sosial media lainnya [2]. Hal ini tentunya dapat menimbulkan beberapa masalah, seperti adanya konten porno di Twitter. Konten porno yang dimaksud adalah konten berbentuk teks yang memiliki unsur seperti mengundang untuk melakukan kegiatan seks, *tweets* cerita seks, seks komersial, dan hal lainnya yang berhubungan dengan porno. Oleh karena itu, diperlukan teknologi informasi *Artificial Intelligence* (AI) atau teknologi kecerdasan buatan untuk melakukan analisis data agar mempermudah untuk mengetahui seberapa banyak konten-konten negatif seperti porno di Twitter.

AI adalah sistem yang dikembangkan untuk mempelajari dan menirukan kecerdasan manusia [3][4]. Salah satu bidang ilmu dari AI adalah *data science*. *Data science* merupakan bidang ilmu interdisiplin tentang metode komputasi data untuk mengubah suatu data menjadi pengetahuan [5]. Setiap hari, Twitter dibanjiri dengan banyak sekali *tweets* dari berbagai akun pengguna yang menyebabkan timbulnya *information overload*. *Information overload* adalah banyaknya informasi yang diterima oleh manusia sehingga sulit untuk mengolahnya. Karena adanya *information overload*, manusia dituntut untuk dapat mengkombinasikan berbagai

informasi yang didapatkan dari berbagai sumber sehingga menjadi satu kesatuan informasi yang utuh, akurat dan bermanfaat [6]. Untuk melakukan analisis *tweets* yang mengalami *information overload*, diperlukan teknologi canggih yang bisa melakukan teks analitik untuk memperoleh opini dari pengguna yang terdapat pada Twitter tersebut. Oleh karena itu, diperlukan *sentiment analysis* untuk melakukan teks analitik tersebut. *Sentiment analysis* merupakan proses penggunaan teks analitik untuk mendapatkan berbagai sumber data dari internet dan beragam *platform* media sosial [7]. *Sentiment analysis* menggunakan *Natural Language Processing* (NLP) untuk mengekstrak, mengubah, dan menafsirkan opini dari teks dan mengklasifikasikannya menjadi opini positif, negatif, atau netral [8]. NLP merupakan cabang ilmu komputer dan linguistik yang mengkaji interaksi antara komputer dengan bahasa (alami) manusia [9].

Pengembangan dari teknologi *data science* salah satunya adalah *web scraping*. *Web scraping* adalah teknik untuk mendapatkan informasi dari *website* secara otomatis tanpa harus menyalinnya secara manual [10]. Tujuan dari *web scraping* adalah untuk mencari informasi tertentu dan kemudian mengumpulkannya dalam web yang baru [10]. *Web scraping* berfokus dalam mendapatkan data dengan cara pengambilan dan ekstraksi [10]. Dalam penelitian yang dilakukan oleh Afrizal Aziz Maulana, Ajib Susanto, dan Desi Purwanti K [11], teknik *web scraping* digunakan untuk membuat portal pencarian produk *smartphone* dengan menggabungkan informasi dari laman-laman *web e-commerce* [11]. Contoh kasus lain dari pemanfaatan *web scraping* dilakukan oleh Ma'arif [12]. Dalam penelitian tersebut, *web scraping* digunakan untuk mengumpulkan informasi mengenai objek-objek wisata di Daerah Istimewa Yogyakarta, kemudian menggabungkannya ke dalam satu portal informasi terpadu [12].

Untuk mendapatkan *dataset* di Twitter, diperlukan interaksi dengan *Application Programming Interface* (API) Twitter. Oleh karena itu diperlukan *Tweepy* untuk bisa mengakses API Twitter agar bisa mendapatkan akses untuk mengambil *dataset* agar bisa diekstraksi. *Tweepy* (*Twitter for Pyhthon*) adalah sebuah paket atau *library* bahasa pemrograman *Python* untuk melakukan interaksi dengan API yang telah disediakan oleh Twitter dengan mudah [13]. API adalah sebuah software yang memungkinkan para *developer* untuk mengintegrasikan dan

mengizinkan dua aplikasi yang berbeda secara bersamaan untuk saling terhubung satu sama lain [14].

Algoritma yang digunakan untuk melakukan *sentiment analysis* data *tweets* yaitu algoritma *Bidirectional Encoder Representation from Transformers* (BERT). Algoritma BERT yaitu teknologi *open source* berbasis jaringan *neural* untuk *pre-training* NLP [15]. Algoritma ini akan membuat sistem lebih mudah untuk memahami konteks pencarian yang dimaksud oleh *user* dengan menelaah korelasi dari setiap kata kunci yang diketik [15]. Algoritma BERT memiliki model arsitektur *multi-layer bidirectional transformer encoder* [15][16]. Dalam penelitian yang dilakukan oleh Jacob Devlin, Ming-Wei Chang, Kenton Lee, dan Kristina Toutanova [15], algoritma BERT dan algoritma lainnya dilatih untuk menganalisis *dataset* dari berbagai sumber seperti *General Language Understanding Evalution* (GLUE), *Stanford Question Answering Dataset* (SquAD), dan *Situations With Adversarial Generations* (SWAG). Algoritma BERT memiliki persentase keakuratan tertinggi dibanding algoritma lain dengan hasil keakuratan F1-*Score* sebesar 91.0%, %, 93.2 %, dan 86.3 % dari masing-masing sumber *dataset*.

Tugas Akhir ini menganalisis performansi tingkan keakuratan algoritma BERT dalam sistem *sentiment analysis*. Algoritma BERT digunakan untuk mengklasifikasikan jenis *tweets* yang diolah menjadi opini positif atau negatif. Pengambilan *dataset tweets* menggunakan *Tweepy* yang memiliki kelebihan yaitu bisa mencari *dataset* melalui *hashtag*, nama akun pengguna, dan bisa mengakses API Twitter [13]. Ketika *dataset* yang dikumpulkan melalui *Tweepy* akan diolah, *dataset* tersebut harus melalui tahap *pre-processing*. *Pre-processing* ini melibatkan penghapusan *hashtag*, penghapusan *link* akun pengguna, mengkompres kata, dan penghapusan kata berhenti [13]. Sistem yang dirancang akan mengambil *dataset* dari Twitter berupa *tweets* dari akun pengguna. Kata kunci yang digunakan saat proses pencarian adalah *username* pengguna dan *hashtag*.

1.2 Rumusan Masalah

Berdasarkan latar belakang terkait, dapat dirumuskan beberapa permasalah sebagai berikut:

1. Bagaimana cara kerja metode BERT dalam melakukan *sentiment analysis*?
2. Bagaimana cara agar metode BERT mendapatkan performansi yang tinggi?

3. Bagaimana cara kerja teknologi *Tweepy* dalam melakukan *web scraping*?
4. Bagaimana cara mengukur dan menganalisis parameter performansi akurasi, presisi, *recall*, dan F1-*Score* untuk kasus *sentiment analysis*?

1.3 Tujuan

Berdasarkan rumusan masalah yang telah dijelaskan, maka tujuan dari Tugas Akhir ini adalah:

1. Menganalisis performansi teknologi *sentiment analysis* dengan menggunakan algoritma BERT.
2. Mengeksploitasi metode BERT agar mendapatkan performansi akurasi lebih tinggi.
3. Mengukur dan menganalisis parameter performansi akurasi, presisi, *recall*, dan F1-*Score*.
4. Mengklasifikasikan konten negatif yang terdapat di Twitter menggunakan *sentiment analysis*.

1.4 Batasan Masalah

Adapun batasan yang digunakan dalam Tugas Akhir ini antara lain:

1. *Sentiment analysis* yang digunakan merupakan hasil training menggunakan metode BERT.
2. *Dataset* yang diambil merupakan hasil *web scraping*.
3. Sistem *web scraping* menggunakan *Tweepy*.
4. *Dataset* berupa kumpulan *tweet* yang dicari dengan kata kunci yang berindikasi porno dan kata kunci yang bersifat umum.
5. Parameter evaluasi yang didapatkan adalah akurasi, presisi, *recall*, dan F1-*Score*.
6. Spesifikasi *tools* yang digunakan dalam Tugas Akhir ini sebagai berikut:
    a. *Python* 3.9.0
    b. *Google Colab* dengan spesifikasi: GPU Tesla, RAM 12GB dan *Disk* 300GB.
    c. *Visual Studio Code*

1.5 Metode Penelitian

Metode penelitian yang digunakan pada Tugas Akhir ini adalah sebagai berikut:

1. Studi Literatur

   Melakukan studi literatur dengan mencari, mengumpulkan, dan memahami jurnal, *paper*, artikel, buku, *website* dan referensi lain yang berkaitan dengan *data science, web scraping, deep learning, sentiment analysis,* dan BERT.

2. Pengambilan *Dataset*

   Tahap ini dilakukan dengan mengambil *tweets* secara acak di Twitter sehingga menghasilkan 840 *dataset* yang akan diolah.

3. Perancangan Sistem

   Sistem dirancang berdasarkan *flowchart* yang telah dibuat. Untuk merealisasikan sistem tersebut dibutuhkan teks editor *Visual Studi Code* dan *Google Colab* dengan bahasa pemrograman *Python*.

4. Analisis Hasil Pengujian

   Tahap ini menganalisis kinerja sistem berdasarkan parameter yang telah ditentukan yaitu akurasi, *precision*, F1-*Score* dan *recall*.

5. Kesimpulan

   Tahap ini diselesaikan dengan menyimpulkan hasil dari proses pengujian dan analisis data serta menyusun laporan dari kesimpulan yang telah dibuat.

1.6 Sistematika Penulisan

Sistematika penulisan dalam Tugas Akhir ini adalah sebagai berikut:

1. BAB I PENDAHULUAN

   Bab ini berisi latar belakang, rumusan masalah, tujuan dan manfaat, batasan masalah, metode penelitian dan sistematika penulisan.

2. BAB II DASAR TEORI

   Bab ini membahas terkait konsep dasar dan tinjauan Pustaka yang digunakan dalam penelitian seperti teknologi *sentiment analysis*, *web scraping, Tweepy,* algoritma BERT, dan bahasa pemrograman *phython*.

3. BAB III MODEL DAN PERANCANGAN SISTEM

   Bab ini menjelaskan desain sistem yang telah dirancang, perancangan sistem, parameter performansi, dan spesifikasi perangkat yang digunakan.

4. BAB IV ANALISIS HASIL PENGUJIAN

   Bab ini berisi data hasil pengujian sistem yang dilakukan dan analisis hasil pengujian yang didapat yaitu akurasi, *precision*, *recall*, dan F1-*Score*.

5. BAB V KESIMPULAN DAN SARAN

   Bab ini berisi kesimpulan dari hasil analisis terhadap pengujian sistem dan saran untuk penelitian selanjutnya agar dapat meningkatkan performansi sistem *sentiment analysis*.