BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Hate speech adalah ungkapan atau pidato yang menghina individu atau kelompok berdasarkan (dugaan) anggotaan dalam suatu kelompok sosial yang dikenali melalui ciri-ciri seperti ras, etnis, jenis kelamin, orientasi seksual, agama, usia, disabilitas fisik atau mental, dan faktor-faktor lainnya [1]. Dikarenakan pertumbuhan pesat internet yang memungkinkan untuk terjadinya komunikasi dari berbagai belahan dunia dengan mudah tanpa ada batasan jarak, interferensi bahasa lain ke dalam bahasa Indonesia pun tidak dapat dihindari, baik yang berasal dari daerah Indonesia maupun dari luar Indonesia secara terus menerus [2], sehingga terjadilah fenomena *code-mixed*. *Code-mixed* adalah suatu Fenomena linguistik yang mencampurkan beberapa bahasa dalam komunikasi, dipengaruhi oleh faktor-faktor sosiolinguistik [3]. Salah satu interferensi bahasa di Indonesia adalah penggunaan bahasa Inggris di kehidupan sehari-hari dilakukan pengguna media sosial (khususnya gen Z) di Indonesia untuk mengekspresikan dirinya dengan bilingual bahasa Indonesia-Inggris [4]. Mengingat Indonesia merupakan salah satu negara dengan pengguna aktif Twitter/X terbanyak di dunia [5], dan tingginya tingkat ujaran kebencian pada media sosial di Indonesia yang dapat merusak persatuan bangsa Indonesia[6] [7]. Diperlukannya metode yang dapat menanggulangi hal tersebut.

Proses deteksi *hate speech* bilingual yang akan dilakukan termasuk dalam *Natural Language Processing* (NLP). NLP adalah sebuah subbidang dari *artificial intelligence* (AI) yang memiliki fokus dalam pemahaman dan memproses bahasa manusia. Pendeteksian *hate speech* sendiri dapat dilakukan dengan adanya *text classification* [8]. *Text classification* merupakan sebuah prosedur untuk menetapkan label atau nilai yang telah ditetapkan sebelumnya untuk teks seperti kalimat, paragraf atau dokumen, bahkan *hate speech* [9]. Proses *text classification* di dalam *Natural Language Processing* pada penelitian ini menggunakan *large language models* (LLM), sehingga NLP pada penelitian ini menggunakan teknik *deep learning*. LLM sendiri menggunakan *transformer models* yang dilatih menggunakan dataset yang besar sehingga memungkinkan untuk mengenali, memprediksi, atau menghasilkan teks yang menyerupai buatan manusia.

BLOOM adalah sebuah model bahasa yang memiliki 176 miliar parameter, dilatih dalam 46 *natural languanges* dan 13 bahasa pemrograman yang ditemukan pada 2022. Model bahasa ini dikembangkan dan dirilis oleh kolaborasi ratusan peneliti. Komputasi untuk *training* BLOOM disediakan melalui hibah publik Prancis dari GENCI dan IDRIS, dengan memanfaatkan superkomputer Jean Zay milik IDRIS [10]. Pada penelitian ini BLOOM dengan versi parameter lebih kecil akan digunakan, yaitu BLOOM-560M dengan jumlah 560 juta parameter. Hal tersebut dilakukan karena adanya sebuah penelitian dimana BLOOM-560M memiliki performa lebih baik dalam proses NLP [11].

Sebelumnya pada penelitian terdahulu sudah dilakukan pembangunan model pengenalan *Hate speech* satu bahasa dengan menggunakan algoritma BERT, namun perbedaan yang signifikan adalah BERT hanya dilatih menggunakan satu *natural languanges* sedangkan BLOOM sudah dilatih menggunakan 46 *natural languanges*, yang membuat BLOOM lebih sesuai untuk proses pendeteksian *hate speech* bilingual [12]. Lalu pada penelitian terdahulu yang sudah menggunakan algoritma BLOOM Filter untuk mendeteksi *hate speech* berbahasa Indonesia sudah mendapatkan hasil metrik evaluasi *accuracy* 88,35%, *precision* 84,25%, *recall* 87,93%, dan *f1-score* 88,17% [13]. Terdapat juga penelitian terdahulu yang menggunakan mBERT untuk mendeteksi bilingual *hate speech*, dengan hasil *accuracy* 73% dan *f1-score* 56%.

Penelitian ini menggunakan BLOOM-560M untuk mendeteksi apakah sebuah teks bilingual Indonesia-Inggris yang dimasukkan merupakan *hate speech* atau non-*hate speech*, pembangunan model akan dilakukan menggunakan Google Notebook dengan versi Python 3. Penelitian ini akan dilakukan bertahap, dengan mengumpulkan dan membangun data, data *pre-process*, *data trainning*, dan *data testing*. Pengumpulan dan pembangunan data akan dilakukan secara manual dengan melakukan pencarian pada aplikasi Twitter untuk data bilingual, serta penggunaan data Inggris [14] dan Indonesia [15] dari penelitian sebelumnya yang melakukan pendeteksian *hate speech*.

1.2 Rumusan Masalah

Berdasarkan latar belakang dari penelitian "Implementasi Pendeteksi Hate Speech bilingual Bahasa Indonesia dan Inggris Dengan Model Fine-Tuning BLOOM Filter" adalah sebagai berikut.

1. Bagaimana implementasi BLOOM Filter untuk mendeteksi *hate speech*

dalam bahasa bilingual Indonesia dan Inggris?

2. Bagaimana akurasi, presisi, *recall*, dan F1-Score dari BLOOM Filter yang sudah di *Fine-Tuning* dalam melakukan pendeteksian *hate speech* bilingual?

1.3 Batasan Permasalahan

Batasan masalah di dalam penelitian ini adalah.

1. Menggunakan *dataset* bahasa Indonesia [15], Inggris [14], dan *dataset* bilingual yang dikumpulkan, dengan masing-masing berjumlah 13.000 data, 1.000 data, dan 700 data. Kolom yang dipakai dari *dataset* yaitu teks, dan nilai dari teks tersebut apakah sebuah *hate speech* atau non-*hate speech*. Ketiga *dataset* akan digabungkan, lalu di-*split* dengan 80:20. Sehingga dengan total 11.000 baris data akan digunakan untuk pelatihan model. *Dataset* bilingual akan digunakan untuk evaluasi model dengan jumlah 140 baris data, berisikan 42 data *hate speech* dan 98 data non-*hate speech*.
2. Penggunaan *dataset* yang tidak seimbang akan mengakibatkan bias pada hasil evaluasi, maka nilai evaluasi yang akan lebih difokuskan adalah presisi, *recall*, dan F1-Score.
3. Pendeteksian *bilingual hate speech* menggunakan BLOOM tidak bersifat *real time*, dan akan dilakukan pada lingkungan tertutup yaitu menggunakan *dataset* yang sudah tersedia.

1.4 Tujuan Penelitian

Tujuan dari penelitian yang akan dilakukan adalah.

1. Implementasi BLOOM Filter untuk mendeteksi *hate speech* dalam bahasa bilingual Indonesia dan Inggris.
2. Mendapatkan nilai akurasi, presisi, *recall*, dan F1-Score dari BLOOM Filter yang sudah di *Fine-Tuning* dalam melakukan pendeteksian *hate speech* bilingual.

1.5 Manfaat Penelitian

Manfaat yang diharapkan akan diperoleh dari penelitian yang ingin dilakukan adalah.

1. Membantu memajukan riset mengenai pendeteksi *hate speech* di Indonesia dengan mengikuti perkembangan yang cukup pesat dari segi bahasa yang digunakan.
2. Mendapatkan hasil evaluasi dengan memanfaatkan BLOOM Model untuk melakukan deteksi *hate speech* dalam bahasa Indonesia dan Inggris.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini berisikan mengenai latar belakang dilakukannya penelitian, rumusan masalah yang akan diselesaikan, batasan masalah yang akan diselesaikan di dalam penelitian, tujuan penelitian yang ingin dicapai, dan manfaat dari penelitian yang sudah dilakukan.
- Bab 2 LANDASAN TEORI
  Bab ini berisikan penjelasan mengenai *Hate Speech* yang akan dijadikan bahan penelitian, model BLOOM Filter yang akan digunakan dalam membuat model, dan metrik evaluasi yang akan digunakan untuk menilai model.
- Bab 3 METODOLOGI PENELITIAN
  Bab ini berisikan mengenai metodologi penelitian berupa pengumpulan dan pembangunan data, perancangan model disertai flowchart, implementasi, dan pengujian yang akan dilakukan.
- Bab 4 HASIL DAN DISKUSI
  Bab ini berisikan penjelasan mengenai spesifikasi dari sistem yang digunakan untuk membuat model, potongan kode yang digunakan untuk pembentukan model, pengujian model yang sudah dibuat serta evaluasi hasil pengujian.

- Bab 5 KESIMPULAN DAN SARAN

  Bab ini berisikan pembahasan mengenai hasil yang didapatkan setelah penelitian, dan saran untuk penelitian selanjutnya.