BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Hate speech atau ujaran kebencian adalah segala bentuk komunikasi yang meremehkan seseorang atau sekelompok orang atas dasar beberapa karakteristik seperti ras, budaya, agama, etnis, warna kulit, orientasi seksual, kebangsaan atau karakteristik lainnya [1]. Ujaran kebencian bertambah seiring dengan bertambahnya pengguna sosial media dan didukung dengan peraturan kebebasan berpendapat yang ada di beberapa negara salah satunya Indonesia [2]. Sosial media yang cukup banyak digunakan di Indonesia adalah Twitter, karena Twitter memberikan penggunanya kebebasan untuk beropini [3].

Deteksi *hate speech* bisa menggunakan *Natural Language Processing* (NLP), NLP adalah salah satu bidang dari ilmu komputer dimana pembelajaran mesin dan linguistik komputasional digunakan untuk berbagai macam pekerjaan secara luas [4], salah satu kegunaan dari NLP adalah *text classification*. *Text classification* bertujuan untuk memberikan label pada sebuah kalimat, paragraf, kueri, atau dokumen untuk menjawab pertanyaan, mendeteksi *spam*, sentimen analisis, dan sebagainya [5]. Jenis *language model* yang digunakan untuk melakukan *text classification* salah satunya adalah *Large Language Model* (LLM).

LLM beberapa tahun belakangan ini cukup populer karena telah menunjukan kemampuan generalisasi yang kuat dalam mengerjakan berbagai macam tugas dan *fine tuning* telah menjadi strategi umum untuk mengirimkan pengetahuan luas ke tugas-tugas hilir untuk waktu yang lama. Pada penelitian ini model dengan kategori LLM yang akan digunakan adalah BLOOM model.

BLOOM model adalah sebuah *pre-trained* model yang memiliki 176 miliar *language parameter* dalam 46 bahasa natural dan 13 bahasa pemrograman yang dibuat oleh kerjasama dari beberapa gabungan peneliti [6]. Beberapa model dengan kategori LLM lain seperti BERT pada dasarnya hanya dilatih menggunakan satu bahasa sedangkan BLOOM sendiri sudah dilatih menggunakan 46 bahasa natural dan 13 bahasa pemrograman [6][7]. Agar dapat menggunakan BLOOM model ini untuk *text classification* diperlukan *fine tuning* terhadap *pre-trained model* yang ada agar dapat memaksimalkan hasil dari model tersebut.

Penelitian tentang klasifikasi teks *hate speech* sudah pernah dilakukan

diantaranya, menggunakan jenis LLM model BERT [8], XLM-R [9]. Sedangkan untuk penelitian menggunakan dataset dari Ibrohim dan Budi [10] juga sudah pernah dilakukan dengan membandingkan model BERT dan XLM-R [11] Penelitian menggunakan model BLOOM juga pernah dilakukan [12] BLOOM terbukti memiliki performa yang lebih baik dibandingkan dengan Open Pre-trained Transformer (OPT) *Language Model* dengan dataset berupa *film review* dengan persentase akurasi BLOOM model 2% - 3% lebih besar dibandingkan OPT 175B. Terdapat juga penelitian [13] yang membuktikan BLOOM -560m terbukti memiliki performa yang lebih baik dalam hal *fine tuning* dibandingkan dengan BLOOM-1b7 dengan *training* dataset *monolingual*.

Berdasarkan masalah dan penelitian yang telah disebutkan maka penelitian ini akan mengimplementasikan klasifikasi teks dalam bahasa Indonesia menggunakan model bahasa BLOOM yang telah melalui proses *fine tuning*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan rumusan masalah dalam penelitian ”Implementasi Teks Klasifikasi Hate Speech Bahasa Indonesia dengan Fine-Tuning Model Bahasa BLOOM” adalah sebagai berikut.

1. Bagaimana cara mengimplementasikan *fine tuning* model BLOOM untuk mengklasifikasikan *hate speech* dalama bahasa Indonesia.
2. Bagaimana performa model BLOOM yang sudah di-*fine tuning* dalam melakukan klasifikasi *hate speech* dalam bahasa Indonesia.

1.3 Batasan Permasalahan

Batasan masalah dalam penelitian ini dapat dijabarkan sebagai berikut.

1. Dataset yang akan digunakan dalam penelitian ini menggunakan dataset dari Twitter yang didapat dari penelitian Ibrohim dan Budi [10].
2. Model BLOOM yang digunakan dalam penelitian ini menggunakan model versi yang lebih kecil yaitu BLOOM-560m.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini dapat dijabarkan sebagai berikut.

1. Mengimplementasi *fine tuning* model BLOOM untuk klasifikasi *hate speech* dalam bahasa Indonesia.
2. Mengukur performa untuk model BLOOM yang sudah di-*fine tuning* untuk mengklasifikasikan *hate speech* dalam bahasa Indonesia.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dalam penelitian ini dapat dijabarkan sebagai berikut.

1. Membantu memajukan riset mengenai klasifikasi *hate speech* di Indonesia.
2. Mendapatkan hasil performa yang baik dengan menggunakan BLOOM model untuk melakukan klasifikasi *hate speech* dalam bahasa Indonesia.

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, rumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian.
- Bab 2 LANDASAN TEORI
  Bab ini berisikan mengenai *text classification*, *text preprocessing*, model BLOOM, metrik evaluasi.
- Bab 3 METODOLOGI PENELITIAN
  Bab ini berisikan mengenai metodologi penelitian dan perancangan model berupa *flowchart*.
- Bab 4 HASIL DAN DISKUSI
  Bab ini berisikan mengenai spesifikasi sistem, potongan kode yang digunakan selama pembentukan model, uji coba model yang sudah dibuat dan evaluasi hasil uji.

- Bab 5 KESIMPULAN DAN SARAN
  Bab ini berisikan kesimpulan dari penelitian yang sudah dilakukan dan saran untuk penelitian selanjutnya.