

BAB I

PENDAHULUAN

1.1. Latar Belakang

Diketahui pengguna internet di Indonesia sudah mencapai jumlah 202,6 juta pengguna atau sekitar 73,7% dari total jumlah penduduk yang ada di Indonesia. Sebagian besar dari penggunaan internet tersebut ditujukan untuk beraktivitas di media sosial. Hal ini ditunjukkan oleh jumlah pengguna media sosial yang mencapai 170 juta pengguna atau sekitar 83,9% dari total pengguna internet di Indonesia. Adapun lima media sosial yang paling sering digunakan oleh rentang pengguna dengan umur 16-64 tahun adalah YouTube, WhatsApp, Instagram, Facebook, dan Twitter[1].

Seiring dengan meningkatnya jumlah pengguna internet, interaksi di dunia maya juga mengalami peningkatan signifikan melalui berbagai platform media sosial, forum diskusi, dan situs berita. Media sosial seperti YouTube, WhatsApp, Facebook, Instagram, dan Twitter memfasilitasi kebebasan berpendapat melalui pertukaran informasi, kolaborasi, dan komunikasi secara online tanpa batasan waktu dan tempat[2].

Namun, kebebasan berpendapat di internet tidak jarang diiringi dengan munculnya komentar *Toxic* yang mengandung ujaran kebencian, pelecehan, atau hinaan. Komentar-komentar tersebut tidak hanya mengganggu kenyamanan pengguna, tetapi juga dapat menimbulkan dampak negatif seperti perundungan siber (cyberbullying) dan pelecehan online. Komentar *Toxic* biasanya berisi kata-kata ancaman, cabul, penghinaan, atau kebencian terhadap identitas tertentu, yang dapat menyebabkan pelecehan di media sosial. Akibatnya, sebagian pengguna memilih untuk berhenti berpendapat atau menghindari perdebatan, sehingga diskusi di media sosial menjadi tidak sehat dan tidak adil[3].

Oleh karena itu, diperlukan sistem otomatis yang mampu mengidentifikasi dan mengklasifikasikan komentar *Toxic* secara akurat untuk menciptakan lingkungan digital yang lebih aman dan kondusif.

Berdasarkan pemaparan fakta dan permasalahan di atas, diperlukan adanya penelitian dalam klasifikasi komentar *Toxic*. Penelitian mengenai klasifikasi komentar *Toxic* telah dilakukan sebelumnya. Salah satu penelitian menggunakan dataset sebanyak 1.500 komentar dari media sosial Facebook, yang dibagi menjadi dua kelas: *Toxic* dan *non-Toxic*. Penelitian tersebut menerapkan metode Naive Bayes dengan transformasi TF-IDF dan menghasilkan akurasi sebesar 75%, precision 63%, recall 67%, dan F-measure 64%[4]. Penelitian lain menggunakan data dari Kaggle membandingkan performa metode klasifikasi *Long Short Term Memory* (LSTM) dan Naive Bayes. Hasil penelitian menunjukkan bahwa metode LSTM lebih unggul dengan akurasi sebesar 73%, dibandingkan Naive Bayes yang mencapai 64%[5].

Penelitian ini bertujuan untuk mengembangkan model klasifikasi komentar *Toxic* dengan akurasi tinggi untuk membantu moderator platform digital dalam menyaring konten negatif secara otomatis. Berbeda dengan penelitian sebelumnya yang hanya membedakan komentar *Toxic* dan *non-Toxic*, penelitian ini akan menggunakan klasifikasi multi-label. Dengan pendekatan ini, komentar akan dikategorikan ke dalam beberapa label, yaitu *Toxic*, *Severe Toxic*, *Obscene*, *Threat*, *Insult*, dan *Identity Hate*. Selain itu, hasil penelitian ini dapat berkontribusi dalam pengembangan teknologi berbasis kecerdasan buatan yang dapat meningkatkan kualitas komunikasi di internet.

1.2. Rumusan Masalah

1.2.1. Permasalahan

Perkembangan media sosial yang sangat pesat menyebabkan meningkatnya interaksi pengguna dalam bentuk komentar teks. Namun, kondisi ini juga diiringi dengan maraknya komentar *Toxic* yang mengandung unsur penghinaan, ujaran kebencian, ancaman, dan konten tidak pantas lainnya. Komentar *Toxic* tidak hanya berdampak negatif terhadap kenyamanan

pengguna, tetapi juga dapat memicu konflik sosial, cyberbullying, serta menurunkan kualitas lingkungan digital. Permasalahan menjadi semakin kompleks karena satu komentar dapat mengandung lebih dari satu kategori toksisitas (multilabel), sehingga sulit dideteksi secara manual dan memerlukan pendekatan otomatis yang mampu memahami konteks bahasa secara mendalam.

Selain itu, dataset komentar *Toxic* umumnya memiliki distribusi kelas yang tidak seimbang, di mana komentar non-*Toxic* atau kategori tertentu jauh lebih dominan dibandingkan kategori lainnya. Ketidakseimbangan data ini berpotensi menurunkan kinerja model klasifikasi, khususnya pada kelas minoritas seperti *Severe Toxic*, *Threat*, dan *Identity Hate*. Oleh karena itu, diperlukan analisis lebih lanjut mengenai tingkat akurasi model klasifikasi teks berbasis *Long Short Term Memory* (LSTM) serta pengaruh keseimbangan data terhadap performa model dalam mengklasifikasikan komentar *Toxic* secara multilabel.

1.2.2. Metode Usulan

Metode yang diusulkan dalam penelitian ini adalah penerapan model *Long Short Term Memory* (LSTM) untuk melakukan klasifikasi komentar *Toxic* secara multilabel, dengan tujuan mendeteksi berbagai kategori toksisitas seperti *Toxic*, *Severe Toxic*, *Obscene*, *Threat*, *Insult*, dan *Identity Hate*. LSTM dipilih karena kemampuannya dalam memproses data teks berbasis urutan (sequence data) dan mempertahankan informasi konteks jangka panjang, sehingga lebih efektif dalam memahami makna dan pola bahasa pada komentar media sosial.

Tahapan metode dimulai dengan proses *preprocessing* data yang meliputi *Data Cleaning*, tokenisasi, dan pembentukan pipeline TensorFlow. Selanjutnya dilakukan eksperimen menggunakan dua kondisi dataset, yaitu data asli tanpa penyeimbangan dan data yang telah diseimbangkan menggunakan metode *Undersampling*. Performa model kemudian dievaluasi menggunakan metrik classification report untuk mengetahui tingkat akurasi

dan pengaruh keseimbangan data terhadap hasil klasifikasi. Dengan pendekatan ini, diharapkan model LSTM dapat berkontribusi dalam mendeteksi komentar *Toxic* secara otomatis sebagai upaya menciptakan lingkungan digital yang lebih aman dan sehat.

1.2.3. Pertanyaan Penelitian

Seberapa akurat dan optimal performa model Long Short-Term Memory (LSTM) dalam melakukan klasifikasi komentar *Toxic* multi-label dengan penerapan metode *Undersampling* ?

1.3. Tujuan dan Manfaat

Adapun tujuan dari penulisan penelitian ini adalah sebagai berikut :

1. Mengembangkan model klasifikasi multi-label komentar *Toxic* pada media sosial menggunakan metode *Long Short Term Memory* (LSTM).
2. Meningkatkan akurasi deteksi komentar *Toxic* berdasarkan kategori *Toxic*, *Severe Toxic*, *Obscene*, *Threat*, *Insult*, dan *Identity Hate*.
3. Memberikan solusi teknologi berbasis kecerdasan buatan yang dapat membantu moderator platform digital dalam menyaring konten negatif secara otomatis.

Adapun manfaat dari penulisan penelitian ini adalah sebagai berikut :

1. Akademis: Memberikan kontribusi dalam pengembangan metode klasifikasi teks menggunakan LSTM untuk deteksi komentar *Toxic*.
2. Praktis: Membantu pengelola media sosial dalam konten secara otomatis sehingga menciptakan lingkungan digital yang lebih sehat dan kondusif.
3. Sosial: Mengurangi dampak negatif dari komentar *Toxic* seperti cyberbullying dan pelecehan online.
4. Teknologi: Mengembangkan teknologi AI yang dapat diimplementasikan dalam sistem konten digital.

1.4. Batasan Masalah

1. Dataset diambil mandiri dari proses crawling dan scraping dari twitter dan instagram dengan topik sosial politik.

2. Dataset sebanyak 10.146 data, yang digunakan sebagai bahan utama dalam proses pelatihan, pengujian, dan evaluasi model klasifikasi komentar *Toxic*.
3. Dataset diklasifikasikan menjadi 6 kelas multi-label: *Toxic*, *Severe Toxic*, *Obscene*, *Threat*, *Insult*, dan *Identity Hate*.

1.5. Metodologi Penulisan Skripsi

1. Studi Literatur
Melakukan kajian pustaka terkait media sosial, komentar *Toxic*, teknik klasifikasi teks, dan penerapan model *Long Short Term Memory* (LSTM).
2. Pengumpulan Data
Mengumpulkan dataset publik yang berisi komentar dengan label multi-kategori *Toxic*.
3. Pra-pemrosesan Data
Melakukan pembersihan data, seperti menghapus tanda baca, angka, stopwords, serta melakukan tokenisasi dan padding agar data siap dimasukkan ke dalam model LSTM.
4. Pengembangan Model
Membangun dan melatih model klasifikasi multi-label menggunakan arsitektur *Long Short Term Memory* (LSTM).
5. Evaluasi Model
Mengukur performa model menggunakan metrik evaluasi seperti akurasi, precision, recall, dan F1-score untuk menilai efektivitas klasifikasi.
6. Analisis dan Pembahasan
Menganalisis hasil evaluasi model, membahas kelebihan dan kekurangan model LSTM dalam klasifikasi komentar *Toxic*.
7. *Deploy Model*
Mengaplikasikan model ke dalam framework berbasis web.
8. Kesimpulan dan Saran
Menyimpulkan hasil penelitian dan memberikan rekomendasi untuk pengembangan model dan penelitian selanjutnya.