BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Transisi energi di sektor otomotif Indonesia saat ini tengah berada pada titik krusial, di mana pemerintah secara agresif mendorong adopsi kendaraan listrik (EV) melalui berbagai regulasi dan insentif [1]. Namun, transisi dari mobil bermesin pembakaran internal (ICE) ke mobil listrik tidaklah sederhana [2]. Masyarakat memiliki pandangan yang sangat beragam dan kontras terkait isu infrastruktur pengisian daya, nilai depresiasi harga yang tinggi, hingga dominasi produsen asal Cina yang mengganggu pasar global. YouTube telah menjadi salah satu platform di mana opini publik ini berkembang secara cepat [3]. Melalui video ulasan dari pakar otomotif seperti Fitra Eri, *podcast* yang meliputi pemilik bengkel mobil, hingga pengalaman pengguna mobil listrik yang menyebabkan ribuan komentar dari pengguna internet.

Banyaknya data komentar yang dihasilkan membuat berbagai tantangan muncul dalam pengolahan informasi. Melakukan analisis sentimen yang bersifat tidak terstruktur, menggunakan bahasa tidak baku, penuh slang, dan singkatan khas media sosial adalah hal yang sulit [4]. Sebelum era model berbasis transformator, analisis sentimen di Indonesia sering menggunakan metode berbasis leksikon atau pembelajaran mesin klasik seperti Support Vector Machine (SVM) dan Naive Bayes, yang seringkali gagal menangkap arti kontekstual atau nuansa bahasa dalam teks media sosial yang bervariasi[4].

Kinerja model IndoBERT sangat bergantung pada kualitas label data latih yang digunakan [4]. Pelabelan data yang dilakukan secara manual dapat menjadi cara yang efektif dalam meningkatkan kualitas label data [5]. Namun, cara tersebut dapat memakan banyak waktu jika memiliki ribuan sampel data [6]. Salah satu cara agar menghemat waktu dan efisiensi dalam pelabelan data adalah otomisasi pelabelan (*automated labelling*) seperti menggunakan pendekatan berbasis kamus (berbasis leksikon). Sayangnya, metode leksikon memiliki keterbatasan yaitu prosesnya mencocokkan kata per kata tanpa memahami konteks kalimat [7]. Contoh dalam ulasan otomotif adalah kata ”murah” yang dapat berarti sentimen positif (”harganya murah”) dan sentimen negatif (”materialnya murahan”). Ketidakmampuan leksikon dalam menangkap ironi dalam konteks

sering kali menghasilkan dataset latih yang bias dan kurang akurat.

Pendekatan modern yang dapat mengatasi kelemahan dari metode leksikon adalah teknik *pseudo-labelling*, yaitu memanfaatkan model *pre-trained* yang telah dilatih khusus untuk tugas klasifikasi sentimen [5]. Model tersebut dapat memahami konteks kalimat lebih baik daripada pendetakan kamus. Oleh karena itu, penelitian ini bertujuan untuk membuktikan peningkatan kinerja model dasar IndoBERT *Base Uncased* ketika menggunakan label dataset hasil *pseudo-labelling* dibandingkan dengan model yang menggunakan dataset berlabel leksikon sebagai *baseline*.

1.2 Rumusan Masalah

1. Bagaimana perbandingan performa model IndoBERT yang dilatih dengan dataset hasil pelabelan leksikon dan dataset hasil *pseudo-labelling* menggunakan metrik akurasi dan F1-Macro?
2. Bagaimana perbedaan karakteristik label yang dihasilkan oleh kedua metode dalam hal distribusi kelas sentimen (positif, netral, negatif)?
3. Bagaimana tingkat validitas dan kesesuaian hasil pelabelan otomatis terhadap data uji *ground truth* manual?

1.3 Batasan Permasalahan

1. Sumber Data: Komentar dari tujuh video YouTube bertema mobil listrik di Indonesia.
2. Volume Data: Sebanyak 4.951 komentar mentah yang kemudian difilter relevansinya terhadap topik mobil listrik, menghasilkan 2.416 komentar yang digunakan dalam penelitian.
3. Model: Menggunakan arsitektur IndoBERT *Base Uncased* `(indolem/indobert-base-uncased)` dengan 12 lapisan *transformer*.
4. Metode Automated Labeling: Membandingkan dua teknik pelabelan otomatis, yaitu metode berbasis leksikon dan *pseudo-labelling* dengan model `mdhugol/indonesia-bert-sentiment-classification`.
5. Kelas Sentimen: Klasifikasi dibatasi pada tiga kelas, yaitu Positif, Netral, dan Negatif.

1.4 Tujuan Penelitian

Tujuan yang hendak dicapai dalam penelitian ini adalah sebagai berikut:

1. Mengimplementasikan pelabelan otomatis (*automated labeling*) pada data komentar YouTube terhadap mobil listrik menggunakan metode berbasis leksikon dan *pseudo-labelling*.
2. Menganalisis perbedaan karakteristik label yang dihasilkan oleh kedua metode dalam hal distribusi kelas sentimen (positif, netral, negatif).
3. Mengevaluasi tingkat validitas dan kesesuaian hasil pelabelan otomatis terhadap data uji *ground truth* manual

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat Akademis: Memberikan bukti empiris mengenai pengaruh kualitas pelabelan otomatis terhadap performa model *transformer* untuk analisis sentimen bahasa Indonesia. Penelitian ini juga memperkaya literatur tentang perbandingan metode berbasis leksikon dan *pseudo-labelling* berbasis model *pre-trained* pada *domain* media sosial.
2. Manfaat Praktis: Memberikan gambaran nyata tentang pandangan masyarakat terhadap mobil listrik berdasarkan komentar YouTube, yang dapat menjadi bahan pertimbangan bagi pelaku industri otomotif maupun pemerintah dalam memahami sentimen publik.
3. Manfaat Metodologis: Menghasilkan sebuah *pipeline* analisis sentimen yang tidak hanya menangani kebersihan teks informal dan slang otomotif, tetapi juga menerapkan filter relevansi topik dan dua pendekatan *automated labeling* yang siap diadaptasi untuk penelitian serupa di domain lain.

1.6 Sistematika Penulisan

Laporan penelitian ini disusun secara sistematis agar memberikan gambaran yang jelas mengenai proses dan temuan penelitian. Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN

  Berisikan latar belakang pemilihan topik analisis sentimen mobil listrik di YouTube, rumusan masalah yang berfokus pada perbandingan dua metode *automated labeling*, batasan penelitian, tujuan, manfaat, serta sistematika laporan.

- Bab 2 LANDASAN TEORI

  Menjelaskan teori pendukung seperti arsitektur BERT dan IndoBERT, konsep analisis sentimen, pendekatan pelabelan otomatis berbasis leksikon dan *pseudo-labelling* dengan model *transformer*, serta literatur terkait penerapan NLP pada data media sosial dan domain otomotif.

- Bab 3 METODOLOGI PENELITIAN

  Menguraikan alur kerja penelitian mulai dari pengumpulan komentar YouTube, tahap *preprocessing* dan penyaringan relevansi topik, proses pelabelan otomatis dengan dua metode (leksikon dan *pseudo-labelling*), pembagian data menjadi train, validation, dan test dengan menjaga keseimbangan proporsi kelas, hingga konfigurasi pelatihan model IndoBERT menggunakan *Trainer* dengan penanganan ketidakseimbangan data melalui bobot kelas.

- Bab 4 HASIL DAN DISKUSI

  Menguraikan hasil eksperimen yang meliputi distribusi label dari masing-masing metode, riwayat pelatihan, evaluasi performa akhir pada data uji (akurasi dan F1-Macro), analisis *confusion matrix* dan *classification report*, serta visualisasi *Word Cloud* untuk setiap kelas sentimen. Bab ini juga membahas perbandingan performa dan kualitas label antara kedua pendekatan.

- Bab 5 KESIMPULAN DAN SARAN

  Menampilkan ringkasan hasil penelitian mengenai keunggulan relatif *pseudo-labelling* terhadap metode leksikon dalam meningkatkan kinerja IndoBERT pada domain otomotif, serta memberikan rekomendasi untuk pengembangan metode pelabelan otomatis di masa depan, seperti eksplorasi *threshold* leksikon yang adaptif atau pemanfaatan model *zero-shot* yang lebih besar.