1. Pendahuluan

Latar Belakang

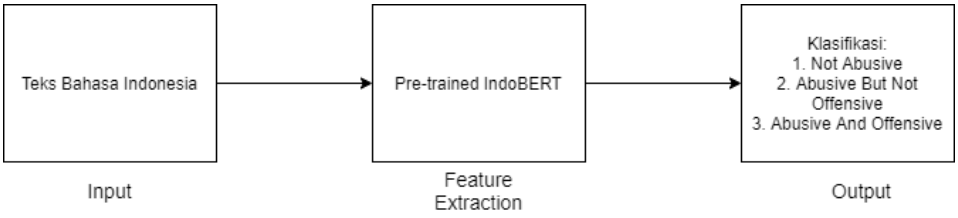
Keberadaan internet dan munculnya berbagai jenis jejaring sosial adalah salah satu hasil dari perkembangan sebuah teknologi[1]. Asosisasi Penyelenggara Jasa Internet Indonesia (APJII) pada Laporan Survei Internet APJII 2019 – 2020 (Q2) menuliskan jumlah pengguna internet di Indonesia mencapai 196,7 juta pengguna atau setara dengan 73,7% dari total populasi RI yang mencapai 266,9 juta penduduk[2]. Sehingga dengan penggunaan internet di Indonesia itu semakin meningkat dapat mengakibatkan juga penyalahgunaan dalam internet itu sendiri, khususnya media sosial. Banyak yang masih menggunakan media sosial dengan tidak bertanggung jawab dengan dalih 'media sosial saya ini, ya terserah saya mau ngomong apa saja'. Kalimat-kalimat yang melecehkan seringkali digunakan untuk menyerang partai politik tertentu atau bahkan hanya untuk bercanda. Jika kalimat *abusive* dapat dengan mudah ditemukan di internet, dapat mempengaruhi kesehatan mental pihak yang dituju jika pihak dituju masih tergolong remaja[3].

Kalimat *abusive* merupakan ungkapan yang mengandung kata-kata kasar dan frasa yang kotor baik dalam bahasa lisan maupun tulisan, dikarenakan kalimat *abusive* mungkin akan melibatkan sarkasme, pelecehan dll[4]. Alasan banyaknya penggunaan kalimat yang melecehkan di internet atau jejaring sosial adalah kurangnya alat yang efektif untuk menyaring penggunaan kalimat yang melecehkan, kurangnya empati di antara pengguna Internet, dan kurangnya pengawasan orang tua[5]. Sebagai contoh, *Hate Speech* atau ujaran kebencian yang berisi kalimat-kalimat *abusive* biasanya memicu konflik sosial karena akan membawa emosi kepada pihak atau pembaca yang dituju[6]. Contoh di *twitter* banyak sekali cuitan atau *tweet* yang mengandung *hate speech* baik secara individu (*cyberbullying)* atau juga kelompok tertentu (LGBT, Agama, Jenis Kelamin)[7], dan untuk kalimat yang mengarah pada tindakan *cyberbullying* yang mana tingkat depresi yang akan dialami korban dapat lebih tinggi daripada depresi yang didapat dari kekerasan secara fisik. Oleh karena itu penting mendeteksi adanya kalimat *abusive* atau *hate speech* seperti itu untuk menganalisis sentimen publik dari sekelompok pengguna terhadap kelompok yang lain[8].

Dalam mengatasi masalah tersebut, perlu dilakukan tindakan preventif dan penegakan hukum secara tegas menurut hukum yang berlaku. Salah satu tindakan yang bisa dilakukan ialah mendeteksi kalimat *abusive* pada media sosial, akan tetapi akan memakan waktu lama untuk mendeteksi secara manual. Oleh karena itu hal ini mendorong para peneliti untuk menciptakan cara – cara otomatis dengan cara membangun sistem yang dapat mendeteksi kalimat *abusive* pada media sosial. Penelitian sebelumnya ada penelitian yang mengenai kalimat *abusive* pada media sosial di Indonesia dengan dataset yang memiliki multi-label dengan menggunakan Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifier and Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) sebagai data transformasinya[6]. Dalam beberapa tahun terakhir model bahasa menggunakan *pre-trained* telah menunjukkan terobosan besar dalam NLP dan pada Devlin dkk membuat BERT yaitu sebuah arsitektur untuk melatih model bahasa yang lebih cepat yang mehilangkan *recurrences* dengan menambahkan *multi-head attention layer*[9]. Sehingga pada penelitian kali ini digunakan metode IndoBERT untuk mengklasifikasikan kalimat inputan menjadi label *Not Abusive*, *Abusive But Not Offensive*, dan *Abusive And Offensive* serta akan dinilai akurasi sistem menggunakan *F1 Score*. IndoBERT itu sendiri merupakan *transformers-based model* yang bergaya BERT itu sendiri[10]. Akan tetapi terdapat *imbalance* atau ketidakseimbangan jumlah antar label pada dataset yang digunakan sehingga terdapat kelas mayoritas dan minoritas. Untuk itu peneliti juga ingin mengetahui pengaruh terhadap adanya penambahan jumlah dataset pada arsitektur yang dibangun.

Topik dan Batasannya

Topik penelitiannya adalah mendeteksi penggunaan kalimat-kalimat *abusive* melalui input berupa teks bahasa Indonesia kemudian, menggunakan metode IndoBERT untuk memproses bahasa Indonesia untuk mengklasifikasikan jenis kalimat. Ada 3 jenis label dalam kategori keluaran, seperti yang ditunjukkan di bawah ini:



Gambar 1 Gambaran Sistem Secara Umum

Batasan masalah dalam penelitian ini ialah jumlah dataset yang digunakan terhadap penelitian ini merupakan dataset yang bersifat imbalance atau ketidakseimbangan dari tiap labelnya.

Tujuan

Tujuan dari penilitian ini yaitu mengevaluasi penerapan metode IndoBERT untuk melakukan deteksi penggunaan kalimat *abusive* pada teks bahas Indonesia, perhitungan nilai akurasi nilai *F1 Score* untuk masing masing kelas dan pengaruh adanya penambahan data pada dataset kelas minoritas terhadap nilai akurasi *F1 Score*.

Organisasi Tulisan

Pada bagian 2 menjelaskan dasar teori yang digunakan sebagai pedoman penelitian. Pada bagian 3 menjelaskan alur penelitian yang dilakukan. Pada bagian 4 menjelaskan hasil yang didapat melalui algoritma dan sistem yang sudah dibangun. Pada bagian 5 menjelaskan kesimpulan yang didapat dan saran yang bisa dikembangkan untuk penelitian selanjutnya