BAB I

PENDAHULUAN

1.1. Latar Belakang

Media sosial sudah menjadi media komunikasi utama, termasuk bagi masyarakat Indonesia. Belum lagi di tahun 2020, pandemi COVID-19 membuat tingkat penetrasi *online* semakin bertumbuh dan bahkan telah menggeser gaya hidup masyarakat Indonesia semua ke arah berbasis digital. Hasil Survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) di Tahun 2018 menunjukkan bahwa media sosial menjadi layanan internet peringkat kedua yang paling sering diakses oleh masyarakat Indonesia (APJII, 2018). Media sosial membawa keuntungan seperti meningkatkan koneksi dan menjalin inter-komunikasi tanpa dibatasi jarak dan waktu. Tetapi di satu sisi, seperti yang kita juga tahu bahwa banyak fenomena yang cenderung negatif di dunia media sosial. Hari ini, media sosial sedikit banyak digunakan sebagai alat kekerasan (Shahbaz & Funk, 2019) – dipenuhi dengan banyak konten cerita, informasi, dan opini dalam bentuk kata-kata atau kalimat yang bersifat ofensif, entah itu menyerang suatu organisasi atau individu. Hal tersebut kemudian dikenal dengan istilah *Digital Aggression* atau *Online Harassment* (Wong-Lo & Bullock, 2011).

Di Indonesia, Plan Indonesia melaporkan dari 500 remaja perempuan Indonesia berusia 15-20 tahun, 32A persen mengatakan pernah mengalami kekerasan/pelecehan seksual di media sosial, dan 56 persen pernah menjadi saksi kekerasan tersebut (Plan International, 2020). Pada tahun 2018 APJII juga

melaporkan dari 5.900 responden, 49 persen menyatakan bahwa mereka pernah mengalami *cyberbullying* di media sosial. 31,6 persen di antaranya membiarkan saja hal tersebut dan hanya 3,6 persen yang melaporkan ke pihak berwajib. Presentase yang melapor justru lebih kecil dibanding presentase orang yang membalas *bullying* yang menimpanya, yaitu 7,9 persen yang notabene sama saja melanggar aturan. Lalu juga lebih kecil dibandingkan 5,2 persen yang memilih untuk menghapus bukti dari tindakan *cyberbullying* yang menimpanya (APJII, 2018).

Angka persentase di atas menunjukkan bahwa belum adanya kesadaran yang serius dari masyarakat internet di Indonesia dalam menanggapi tindakan kekerasan di media sosial dan akibat yang dapat ditimbulkan dari hal tersebut. Dikarenakan pembatasan akses internet bukan menjadi solusi yang efektif, maka yang bisa kita lakukan adalah memaksimalkan fungsi pengawasan dan pencegahan (Stewart, 2013).

Dalam melaksanakan fungsi tersebut, tidak cukup jika hanya dari pihak pemerintah atau instansi terkait yang melakukan pengawasan terhadap setiap tindakan *cyberbullying*/*hate-speech* yang terjadi di masyarakat (*top-down*). Diperlukan juga respon dari masyarakat untuk memberikan pelaporan (*bottom-up*) sehingga tindak penanganan kasus *cyberbullying*/*hate-speech* dapat lebih cepat dieksekusi. National Domestic Violence Hotline menyediakan jasa layanan *call center* yang pada tahun 2019 melaporkan peningkatan volume pelaporan terhadap kekerasan atau pelecehan digital sebesar 101% di Amerika Serikat (National Domestic Violence Hotline, 2019). Pemerintah Australia melalui lembaga *eSafety*

Commisioner telah membuka wadah bagi para korban *cyberbullying* untuk melakukan pelaporan dan berbagi tips kepada para korban bagaimana tindakan-tindakan yang harus diambil ketika mengalami diserang oleh pelaku *cyberbullying* (eSafety Commisioner, 2020b). Ditch the Label, organisasi global yang khusus menangani isu generasi muda termasuk *cyberbullying*/*hate speech* juga telah menyediakan *platform* pelaporan serupa[1]. Beberapa contoh terebut dapat menjadi solusi bagi pemerintah maupun instansi terkait dalam menangani momok kekerasan/pelecehan yang menguasai dunia media sosial di Indonesia.

Beberapa solusi di atas masih mengandalkan sepenuhnya evaluasi dari manusia. Jika kasus ini terus meningkat, kapasitas manusia pun tidak cukup untuk menangani seluruh isu yang ada. Suatu sistem cerdas dengan metode dan algoritma yang handal, misalnya dengan pendekatan Kecerdasan Artifisial – dibutuhkan agar mampu mendeteksi konten kekerasan atau pelecehan di media sosial seakurat mungkin (Rosa et al., 2019; Van Hee et al., 2018; Van Royen et al., 2017), sehingga dapat membantu pemerintah atau instansi terkait dalam melakukan evaluasi tindakan demi menekan angka kejadian serta mencegah sedini dan sebanyak mungkin, namun tetap tepat sasaran.

Screenshot atau tangkapan layar merupakan cara tercepat dan paling efektif untuk merekam bukti dari tindakan *cyberbullying* (eSafety Commisioner, 2020a). Tangkapan layar ini bisa berasal dari *platform* apapun di mana korban mendapat perlakuan yang mengusik bahkan merugikan dirinya, sehingga dibutuhkan model

[1] https://www.ditchthelabel.org/report

analisis data yang mampu mengeneralisasi berbagai *platform* media sosial. Generalisasi masih menjadi tantangan penelitian pada topik *automated cyberbullying*/*hate-speech detection* hingga saat ini (Emmery et al., 2019; Radford et al., 2018; Salminen et al., 2020).

Sudah terdapat beberapa penelitian terkait deteksi kekerasan di media sosial meliputi *cyberbullying* atau *hate speech*. Tetapi masih terdapat beberapa isu yang ditemui. Pertama, ketersebaran data. Beberapa penelitian membangun datasetnya sendiri-sendiri dengan teknik pengumpulan yang berbeda-beda sehingga menyebabkan kualitas data yang juga beragam. Ada yang memberikan akses kepada datasetnya untuk publik, ada juga yang tidak. Penelitian yang dilakukan dari waktu ke waktu menggunakan dataset yang terus berubah sehingga tidak ada tolak ukur yang jelas untuk keberlanjutan penelitian pada topik ini. Lalu, sejauh ini pun penelitian terkait tindakan kekerasan di media sosial dalam teks Bahasa Indonesia pun baru menganalisis satu media sosial saja (Aluru et al., 2020; Andriansyah et al., 2018; Febriana & Budiarto, 2019; Ibrohim & Budi, 2018, 2019; Nurrahmi & Nurjana, 2018; Okky Ibrohim et al., 2019; Pratiwi et al., 2019; Sazany & Budi, 2018). Data dari media sosial yang beragam dianggap lebih komprehensif untuk generalisasi dari model yang akan dibangun, seperti yang sudah dibahas sebelumnya.

Penelitian ini akan melakukan pemodelan *Machine Learning* dengan metode berbasis *Bidirectional Encoder Representations from Transformers* (BERT) untuk mengklasifikasi teks berbahasa Indonesia yang mengandung konten kekerasan/pelecehan pada media sosial. Dataset yang digunakan merupakan

kompilasi dari dataset yang digunakan oleh penelitian-penelitian sebelumnya. Dataset tersebut meliputi komentar atau pesan dari media sosial Instagram dan Twitter, bebas duplikasi, dan yang terkini. Model yang sudah dibangun kemudian akan diuji kembali melalui implementasinya dalam bentuk sistem pelaporan *online* berbasis web, untuk mengklasifikasi teks dengan konten kekerasan/pelecehan yang terkandung dalam objek gambar tangkapan layar.

1.2. Rumusan Masalah

1. Bagaimana performa dari pemodelan algoritma berbasis BERT untuk mengklasifikasi teks media sosial berbahasa Indonesia yang mengandung kekerasan atau pelecehan?
2. Bagaimana performa dari hasil pemodelan algoritma berbasis BERT yang sudah dibangun dalam mengklasifikasi data teks baru untuk mendeteksi teks media sosial berbahasa Indonesia yang mengandung kekerasan atau pelecehan dari gambar tangkapan layar (*screenshot*)?

1.3. Tujuan Penelitian

1. Mengetahui performa dari pemodelan algoritma berbasis BERT untuk mengklasifikasi teks media sosial berbahasa Indonesia yang mengandung kekerasan atau pelecehan.
2. Mengetahui performa dari hasil pemodelan algoritma berbasis BERT yang sudah dibangun dalam mengklasifikasi data teks baru untuk mendeteksi teks media sosial berbahasa Indonesia yang mengandung kekerasan atau pelecehan dari gambar tangkapan layar (*screenshot*).

1.4. Batasan Penelitian

Adapun batasan masalah untuk mengetahui ruang lingkup dari penelitian ini adalah sebagai berikut.

1. Hanya menganalisis konten kekerasan/pelecehan dari teks tunggal pada media sosial. Menurut penelitian terkini memang pesan/komentar tunggal belum bisa menggambarkan tindakan mengusik, melecehkan, atau merundung secara menyeluruh karena tindakan tersebut bersifat berkelanjutan dan melibatkan banyak pihak (Rosa et al., 2019). Tetapi dengan metode yang tepat, pada dasarnya penelitian ini sudah bisa menjadi dasar bagi penelitian pada topik *automated online harassment detection*. Selain itu, memang penelitian yang dilakukan juga hanya dapat diimplementasikan untuk mendeteksi satu objek tunggal, yaitu gambar *screenshot* yang diinput oleh pengguna pada sistem pelaporan *online* berbasis web yang akan dibangun.
2. Menggunakan dataset yang sudah dibangun oleh penelitian-penelitian sebelumnya tanpa membangun dataset baru. Hal ini dilakukan sebagai upaya pembangunan dataset *benchmark* untuk topik deteksi kekerasan di media sosial berbahasa Indonesia.
3. Hanya satu algoritma pembelajaran yang digunakan, yaitu *Bidirectional Encoder Representations from Transformers* (BERT). Model yang digunakan adalah model *pre-trained*, yaitu model yang sudah dilatih sebelumnya dan untuk melakukan tugas spesifik hanya perlu melakukan *fine-tuning* dengan dataset yang sudah disiapkan untuk melakukan tugas yang diinginkan.

1.5. Manfaat Penelitian

1. Menyumbangkan gagasan model untuk pengembangan penelitian selanjutnya dalam membangun sistem *automated online harassment detection* dalam Bahasa Indonesia.
2. Kompilasi dataset untuk upaya standarisasi dataset demi keberlanjutan penelitian pada topik deteksi kekerasan di media sosial berbahasa Indonesia.
3. Dapat menjadi referensi *tool* bagi lembaga pengawas atau penindak lanjut tindakan kekerasan di media sosial serta korban kekerasan untuk mengidentifikasi kekerasan pada media sosial melalui teks.