BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Hate speech atau ujaran kebencian umumnya merujuk pada ekspresi kebencian terhadap individu atau kelompok berdasarkan atribut seperti ras, agama, jenis kelamin, atau gender, dengan tujuan menjelekkan, mengintimidasi, atau memicu kebencian [1, 2]. Ujaran kebencian tidak hanya berbentuk kata-kata, tetapi juga dapat berbentuk simbol, gambar, dan berbagai perilaku yang merugikan, menyinggung, atau memicu kebencian terhadap kelompok tertentu [3]. Ujaran kebencian berbeda berdasarkan karakteristik yang dilindungi dan kemampuan mereka untuk merusak atau memicu kebencian [4]. Setiap negara memiliki hukumnya sendiri seperti hasutan untuk kekerasan atau genosida, hukuman ini difokuskan pada kerugian yang lebih luas, sementara hukum yang lebih umum difokuskan pada diskriminasi dan kerugian yang lebih luas [5].

Regulasi ujaran kebencian tidak seimbang dengan kebebasan berekspresi, terutama di internet dan di seluruh global [6]. X (twitter) merupakan salah satu platform yang digunakan untuk berbagi informasi dan berekspresi secara *online*. X memungkinkan pembaruan dan diskusi terkini tentang berbagai topik karena *tweet* dibagikan secara instan. Melalui fitur *retweet* dan *hashtag*, konten agresif dapat menyebar dengan cepat karena basis pengguna yang besar dan interaksi yang intens [6, 7]. Kebijakan penegakan hukum platform X dikritik oleh publik karena gagal mencegah penyalahgunaan dan ujaran kebencian serta tidak membagi tanggung jawab antara platform dan pengguna [8]. Hal ini menyoroti kebutuhan akan kebijakan yang lebih jelas dan efektif.

Penelitian sebelumnya menunjukkan jumlah ujaran kebencian digital di Indonesia telah meningkat hingga sepuluh kali lipat dalam dua tahun terakhir [9]. Sebagian besar ujaran kebencian ditujukan kepada kelompok minoritas agama dan etnis. Keadaan ini memicu kekhawatiran besar tentang perlindungan privasi dan keamanan internet, terutama bagi kaum muda yang paling aktif menggunakan media sosial [10, 11]. Selain itu, lingkungan komunikasi digital yang penuh ujaran kebencian akan menghambat kebebasan berekspresi yang sehat. Kebebasan berekspresi merupakan salah satu pilar utama dalam sistem demokrasi [12]. Dalam konteks ruang digital, kebebasan ini mencakup hak setiap individu untuk

menyampaikan pendapat, gagasan, maupun kritik secara terbuka tanpa rasa takut akan intimidasi atau pembalasan [12].

Indonesia memiliki hukum yang mengatur tentang ujaran kebencian yaitu Pasal 28 ayat (2) UU ITE 2024. Dalam pasal tersebut pelaku ujaran kebencian akan dipidana penjara paling lama enam tahun dan/atau denda sampai 1 miliar rupiah [13]. Salah satu upaya untuk mencegah perilaku tersebut adalah dengan mendeteksi kalimat yang mengandung ujaran kebencian. Media sosial memungkinkan pengguna memproduksi dan menyebarkan konten dalam jumlah besar setiap detiknya. Sulit untuk menyaring jutaan unggahan secara *real-time* menggunakan metode manual [13]. Oleh karena itu, sistem otomatis dapat menjadi opsi untuk menangani volume konten yang sangat besar dalam waktu singkat.

Lembaga seperti Kementerian Komunikasi dan Digital Republik Indonesia (Kominfo) dan Kepolisian Negara Republik Indonesia dapat menggunakan sistem otomatis untuk melakukan pengawasan pada platform digital. Platform media sosial lokal juga dapat menggunakan sistem otomatis untuk memoderasi konten. Solusi teknis yang relevan terhadap tingginya intensitas konten bermuatan ujaran kebencian dengan keterbatasan kapasitas moderasi manual adalah pembuatan sistem pendeteksi ujaran kebencian [14]. Salah satu pendekatan yang dapat digunakan untuk mendeteksi ujaran kebencian secara lebih efektif dan memiliki skala yang lebih besar adalah Pembelajaran mesin *(machine learning)* [14].

Pembelajaran mesin kini telah berkembang menjadi alat yang penting untuk mengenali ujaran kebencian di media sosial [15]. Penelitian terbaru menyelidiki berbagai algoritma, metode rekayasa fitur, serta penyesuaian untuk bahasa tertentu dalam rangka meningkatkan ketepatan deteksi dan moderasi yang etis [16]. Beberapa model yang biasa digunakan adalah *K-Nearest Neighbors*, *Decision Tree*, *Naive Bayes*, *Logistic Regression*, *Support Vector Machines*, dan *Random Forest* [17, 18]. Seiring meningkatnya kebutuhan untuk memahami pola bahasa yang lebih rumit dan detail, metode yang menggunakan *deep neural networks* dan *transfer learning* mulai menunjukkan hasil yang lebih baik dibandingkan dengan model-model tradisional [19].

Teknik *deep neural* seperti *Convolutional Neural Network* (CNN) lebih efektif dalam mengenali pola bahasa yang rumit [20]. Mereka sering kali lebih baik daripada model-model tradisional, terutama dalam menangani ujaran kebencian yang lebih sulit atau halus [20, 21]. Contoh kalimat “Mereka itu suka makan gratisan, seperti biasa.” merupakan contoh ujaran kebencian yang halus dan tersirat, karena tidak menggunakan kata-kata kasar atau eksplisit. CNN merupakan

tipe algoritma *deep learning* yang dirancang khusus untuk menangani data yang memiliki struktur mirip kisi, seperti foto. CNN sering dipakai untuk berbagai tugas, seperti mengklasifikasikan gambar, mendeteksi objek, dan mengenali suara berkat kemampuannya dalam belajar dan mengeluarkan fitur penting dari data asal secara otomatis [22, 23].

Beberapa penelitian telah menyelidiki CNN baik sebagai model independen maupun dalam integrasi dengan arsitektur saraf lain untuk meningkatkan tingkat akurasi deteksi di berbagai bahasa dan media [24]. CNN telah menunjukkan performa yang baik dalam mengidentifikasi ujaran kebencian, dengan tingkat akurasi yang tinggi dan *F1-score* yang baik pada berbagai data dan bahasa. Sebagai contoh, model CNN berhasil mencapai akurasi 80,15% dan *F1-score* 80,35% dalam tugas mendeteksi ujaran kebencian secara umum [21]. Dalam konteks bahasa Arab, CNN berhasil meraih akurasi hingga 81% untuk tugas klasifikasi biner, dan mencapai *F1-score* 0,79 di platform Twitter di Arab Saudi [24, 25].

Penelitian ini dilakukan dengan tujuan merancang model CNN untuk mendeteksi ujaran kebencian berbahasa Indonesia. CNN dipilih untuk identifikasi pola kebencian secara efektif karena kemampuan untuk mengekstraksi fitur lokal yang relevan dari teks [26]. CNN dapat dilatih *from scratch* dengan jumlah data yang relatif kecil, selama distribusi datanya representatif. Kombinasi CNN dengan embedding berbasis konteks seperti Word2Vec dapat lebih menguatkan performa model secara keseluruhan. Cara ini diharapkan dapat mengembangkann sistem klasifikasi yang tidak hanya akurat tetapi juga sesuai dengan karakteristik linguistik bahasa Indonesia yang berkaitan dengan ujaran kebencian.

1.2 Rumusan Masalah

1. Bagaimana cara merancang model *CNN* dan menerapkan *embedding* Word2Vec untuk mendeteksi ujaran kebencian berbahasa Indonesia?
2. Bagaimana performa model dalam mendeteksi ujaran kebencian berbahasa Indonesia berdasarkan *accuracy*, *precision*, *recall*, *F1-score*, dan *conffusion matrix*.

1.3 Batasan Permasalahan

1. Data yang digunakan terbatas pada *dataset* komentar twitter yang diambil dari github Muhammad Okky Ibrohim [27]. *Dataset* yang dihasilkan beserta

panduan anotasinya dipublikasikan secara terbuka. *Dataset* ini telah melalui tahapan kurasi dan anotasi yang ketat, mencakup verifikasi oleh beberapa anotator dengan pedoman anotasi yang jelas serta validasi berlapis untuk memastikan konsistensi dan reliabilitas label. Selain itu *dataset* ini sudah digunakan dalam beberapa penelitian yang memanfaatkan *dataset* yang sama dalam konteks deteksi ujaran kebencian

2. Klasifikasi pada *dataset* dibagi menjadi positif dan negatif.
3. Penelitian berfokus pada perancangan dan hasil model untuk deteksi ujaran kebencian.

1.4 Tujuan Penelitian

1. Merancang model untuk deteksi ujaran kebencian berbahasa Indonesia menggunakan *CNN* dengan menerapkan *embedding* Word2Vec.
2. Mengevaluasi performa model dalam mendeteksi ujaran kebencian berbahasa Indonesia berdasarkan *accuracy*, *precision*, *recall*, *F1-score*, dan *conffusion matrix*.

1.5 Manfaat Penelitian

Penelitian ini memiliki beberapa manfaat yang dapat dijabarkan sebagai berikut:

1. Menerapkan dan memperkuat pemahaman teoritis dan praktis mengenai konsep machine learning dan *Natural Language Processing* (NLP), melalui proses perancangan serta pengujian model *Convolutional Neural Network* (CNN) untuk tugas klasifikasi ujaran kebencian.
2. Menerapkan pendekatan CNN diharapkan mampu meningkatkan ketepatan dalam mengidentifikasi konten bermuatan negatif secara otomatis.
3. Hasil dari penelitian ini dapat digunakan sebagai referensi bagi peneliti lain yang ingin meneliti penerapan machine learning, khususnya CNN, untuk deteksi ujaran kebencian.

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 berisi latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
- Bab 2 LANDASAN TEORI
  Berisi teori-teori yang relevan dengan penelitian ini, termasuk penjelasan tentang ujaran kebencian, platform X, *deep learning*, Word2Vec, dan penggunaan algoritma *Convolutional Neural Network* (CNN) untuk klasifikasi.
- Bab 3 METODOLOGI PENELITIAN
  Menjelaskan metodologi penelitian yang mencakup studi literatur, pengumpulan data, pemrosesan data, pembagian data, perancangan model, dan evaluasi model.
- Bab 4 HASIL DAN DISKUSI
  Menjelaskan hasil dari penelitian yang menggunakan algoritma *Convolutional Neural Network* (CNN) untuk deteksi ujaran kebencian.
- Bab 5 KESIMPULAN DAN SARAN
  Bab ini berisi kesimpulan dari penelitian yang telah dilakukan serta rekomendasi untuk penelitian lanjutan.