BAB I
PENDAHULUAN

1.1 Latar Belakang

Deteksi ujaran kebencian adalah suatu proses untuk memeriksa kumpulan data yang sangat tidak seimbang untuk menemukan kalimat kebencian, dimana ujaran kebencian hanya menyumbang persentase yang sangat kecil dari keseluruhan kumpulan data, sedangkan sebagian besar adalah non-kebencian tetapi menunjukkan karakteristik linguistik yang mirip dengan ujaran kebencian [1]. Semakin maraknya ujaran kebencian dalam politik pada media sosial dapat memicu tindakan kekerasan, prasangka negatif, dan bahkan perpecahan karena ujaran kebencian menyerang atau memprovokasi individu atau kelompok lain dalam berbagai aspek seperti ras, warna kulit, suku, jenis kelamin, disabilitas, orientasi seksual, kebangsaan, atau agama [2]. Salah satu kasus yang paling menonjol di Indonesia adalah kelompok Saracen, mereka mengunggah konten yang berisi ujaran kebencian yang ditujukan pada kelompok tertentu khususnya di bidang politik. Dalam kasus ini, polisi menetapkan empat pengurus Saracen sebagai tersangka [3]. Dengan adanya metode deteksi ujaran kebencian dapat membantu menganalisis suatu ujaran antara kebencian dan non-kebencian. Namun, pengembangan deteksi ujaran kebencian dibatasi oleh kualitas sumber *dataset*, penggunaan bahasa informal, definisi yang berbeda tentang apa yang dimaksud dengan ujaran kebencian [4].

Beberapa peneliti menggunakan *Machine Learning* (ML) untuk mendeteksi ujaran kebencian. Buntoro, G. A., dkk [5] melakukan penelitian analisis sentimen terhadap Pemilihan Presiden tahun 2019 dengan membangun model *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dengan menggabungkannya dengan 7 (tujuh) tokenisasi seperti, *Alphabetic Tokenizer, Character N-gram Tokenizer, Unigram, Bigram, Trigram, N-gram,* dan *Word Tokenizer* untuk mencapai akurasi tertinggi dengan mengklasifikasikan *tweet* dengan simbol tanda pagar. Penelitian ini menghasilkan nilai akurasi tertinggi 79.02% untuk algoritma *Support Vector Machine* (SVM) dengan tokenisasi abjad. Nilai akurasi terendah pada penelitian ini diperoleh untuk algoritma *Naïve Bayes Classifier* (NBC) dengan tokenisasi N-gram yang memiliki nilai akurasi sebesar 44.94%. Namun, model yang dibangun masih membuat sejumlah kesalahan ketika proses klasifikasi dataset dengan distribusi sentimen tidak seimbang serta diperlukan normalisasi lanjutan untuk menangani dataset yang berisi bahasa gaul.

Talita, A. S. dan Wiguna, A., [6] mengusulkan model *Word2Vec* dengan algoritma *Long Short-Term Memory* (LSTM) yang cukup baik dalam mendeteksi kalimat ujaran kebencian dengan nilai parameter *recall* mencapai 0.7021, yang berarti analisis ini dapat digunakan untuk mendeteksi kalimat ujaran kebencian karena *Long Short-Term Memory* (LSTM) memiliki memori yang besar dan tepat untuk menangkap ketergantungan jangka panjang antara kata-kata dalam teks pendek [7]. Kelemahan dari penelitian ini adalah nilai parameter akurasi dan *precision* cenderung rendah karena kalimat ujaran kebencian yang digunakan mencakup bahasa informal serta kualitas data latih maupun data uji yang perlu untuk ditingkatkan melalui pra-pemprosesan serta perlu dikombinasikan dengan metode penilaian kata.

Alfina, I., Fanany, M. I., dan Mulia, R., [8] melakukan penelitian untuk mendeteksi ujaran kebencian terhadap agama dan secara manual menganotasi *tweet* ke dalam dua kelas, yaitu *tweet* yang berisi ujaran kebencian dan yang tidak. Dengan menggunakan *Random Forest Decision Tree* (RFDT), tingkat *F-Measure* yang dihasilkan adalah 93.5%. Namun kelemahan dari penelitian ini adalah model BOW (*Bag of Words*) yang dipakai tidak memadai untuk mendeteksi ujaran kebencian, sehingga disarankan untuk menggunakan model TF-IDF yang mampu mengatasi kekurangan model BOW. TF menemukan frekuensi kata dalam dokumen tanpa membiaskan ukuran dokumen besar ataupun kecil dan IDF menemukan pentingnya sebuah kata dalam dokumen [9].

Berdasarkan pada uraian dan kajian literatur yang telah dijelaskan di atas, maka diusulkan sebuah model menggunakan *Long Short-Term Memory* (LSTM) dengan menggunakan metode penilaian kata *Term Frequency - Inverse Document Frequency* (TF-IDF) untuk mendeteksi kalimat ujaran kebencian yang efektik. Selain menggunakan metode penilaian kata TF-IDF, untuk mengatasi ketergantungan pada kuantitas dan kualitas sumber *dataset* juga dapat diatasi dengan mengolah sumber *dataset* melalui beberapa pra-pemrosesan, seperti *Data Cleaning, Case Folding, Tokenizing, Filtering*, dan *Stemming* [2]. Berdasarkan latar belakang tersebut, maka dilakukan penelitian dengan mengambil judul **"Deteksi Ujaran Kebencian Dalam Domain Politik Pada Media Sosial Dengan Algoritma Long Short-Term Memory (LSTM)"**.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan di atas, rumusan masalah dalam penelitian ini adalah bagaimana menghasilkan model untuk dapat melakukan deteksi ujaran kebencian dengan akurasi yang tinggi tanpa ketergantungan pada kuantitas dan kualitas

sumber *dataset*. Pengontrolan *noise* yang tepat pada tahapan pra-pemrosesan serta pemilihan metode penilaian kata yang tepat akan mempengaruhi hasil deteksi menjadi lebih baik.

1.3 Tujuan

Tujuan yang ingin dicapai dari pelaksanaan penelitian ini adalah untuk menerapkan pra-pemrosesan yang baik agar dapat mempengaruhi hasil deteksi dan akurat serta untuk menerapkan model *Long Short-Term Memory* (LSTM) yang dikombinasikan dengan metode penilaian kata *Term Frequency - Inverse Document Frequency* (TF-IDF) untuk menghasilkan akurasi yang tinggi dalam mendeteksi ujaran kebencian.

1.4 Manfaat

Adapun manfaat dari penelitian ini antara lain.

1. Model ini dapat digunakan dan diterapkan pada media sosial yang lain untuk mendeteksi teks ujaran kebencian.
2. Model ini dapat digunakan oleh penegak hokum untuk mendeteksi ujaran kebencian secara otomatis.
3. Model ini dapat menjadi pertimbangan untuk menambah sarana ilmu pengetahuan serta sebagai referensi tambahan dalam penelitian lebih lanjut.

1.5 Ruang Lingkup

Ruang lingkup dalam penelitian ini antara lain.

1. Penelitian ini akan berfokus pada media sosial *Twitter*.
2. Penelitian ini menggunakan dataset Deteksi Ujaran Kebencian dari Alfina, I., Fanany, M. I., dan Mulia, R. [8]. Dalam dataset ini terdapat 453 *tweet* yang bukan ujaran kebencian dan 260 *tweet* yang merupakan ujaran kebencian. Dataset ini tersedia secara umum dan dapat diakses melalui link https://github.com/ialfina/id-hatespeech-detection. Kemudian dataset juga diambil dari https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection. sebanyak 13.170 dataset yang telah diklasifikasikan sebagai ujaran kebencian dan bukan ujaran kebencian. Selain itu, dataset juga diambil secara *real-time* dari media sosial *Twitter* sebanyak 5.000 data.
3. Dataset pada penelitian ini akan melewati beberapa pra-pemrosesan, seperti *Data Cleaning, Case Folding, Tokenizing, Filtering*, dan *Stemming*.
4. Penelitian ini akan menggunakan metode penilaian kata *Term Frequency - Inverse Document Frequency* (TF-IDF).

5. Penelitian ini menyajikan model yang berfokus pada pendeteksian teks berbahasa Indonesia.

1.6 Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi Literatur

   Pada tahapan ini dilakukan pengumpulan bahan referensi yang berhubungan dengan penelitian, seperti: *Neural Network,* ujaran kebencian, metode *Long Short-Term Memory* (LSTM), metode penilaian kata *Term Frequency - Inverse Document Frequency* (TF-IDF), dan bahan referensi lain terkait dengan penelitian mengenai deteksi ujaran kebencian dari beberapa penelitian sebelumnya.

2. Tahap Analisis

   Pada tahapan ini dilakukan proses untuk mengidentifikasi data yang dibutuhkan, masalah dan tantangan yang harus diselesaikan dan menjelaskan solusi yang diusulkan untuk menyelesaikan masalah dan tantangan yang ada. Proses dalam *machine learning* akan digambarkan dalam bentuk *flowchart*.

3. Perancangan Model

   Pada tahapan ini dilakukan perancangan model yang dimulai dari mengumpulkan set data teks *tweet*, menerapkan pra-pemrosesan data, dan menerapkan *Long Short-Term Memory* (LSTM) untuk memprediksi ujaran kebencian.

4. Pengujian

   Pada tahapan ini dilakukan beberapa pengujian, seperti:

   a. Membandingkan metode *Long Short-Term Memory* (LSTM) dengan dan tanpa pra-pemrosesan yang baik untuk menguji hasil deteksi ujaran kebencian.
   b. Membandingkan metode *Long Short-Term Memory* (LSTM) dengan metode penilaian kata TF-IDF dan tanpa TF-IDF untuk menguji hasil deteksi ujaran kebencian.
   c. Melakukan pengujian terhadap model yang dihasilkan untuk mendeteksi ujaran kebencian dengan beragam jenis dan karakteristik *tweet.*
   d. Membandingkan metode *Long Short-Term Memory* (LSTM) dengan beberapa metode lain, seperti *Naïve Bayes Classifier* (NBC), *Support Vector Machine* (SVM), dan *Random Forest Decision Tree* (RFDT).

5. Menarik kesimpulan dari hasil pengujian.
6. Menyusun laporan Tesis