BAB II TINJAUAN PUSTAKA

Secara umum, ujaran kebencian adalah tindakan komunikatif yang dilakukan oleh individu atau kelompok yang memprovokasi, menghasut atau menyinggung individu atau kelompok lain berdasarkan berbagai aspek seperti ras, warna kulit, suku, jenis kelamin, disabilitas, orientasi seksual, kebangsaan, agama dan sebagainya[3]. Berdasarkan UU ITE Pasal 28 ayat 2, ujaran kebencian mencakup spektrum yang luas, mulai dari kata-kata kasar kepada orang lain, ujaran kebencian, hasutan hingga kebencian, dari ujaran yang sangat bias hingga hasutan hingga kebencian yang berujung pada kekerasan[4].

Hukum Indonesia juga memuat UU No. 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik. Menurut Pasal 28 jis pasal 45 ayat 2, siapa pun yang menyebarkan berita bohong, menyesatkan dan membangkitkan perasaan kebencian atau permusuhan dapat dihukum penjara hingga enam tahun[5]. Sebelumnya, pada 2017, mantan Gubernur DKI Jakarta Basuki Tjahaja Purnama pernah menjalani hukuman dua tahun penjara setelah dinyatakan bersalah melakukan penistaan agama. Saat itu, Ahok dalam pidatonya di Pulau Pramuka pada 27/09/2017 merujuk pada Surat QR Al-Maidah ayat 51[6]. Selain Ahok, Ahmad Dhani juga pernah mengalami hukuman Pidana atas ujaran kebencian yang dilakukannya, hingga dijatuhi hukuman 18 bulan penjara atau 1,5 tahun atas keputusan pengadilan Jakarta Selatan pada 28 Januari 2019[7].

Penggunaan sarana komunikasi saat ini telah berkembang begitu pesat dengan perkembangan teknologi komunikasi sehingga kita dihadapkan pada banyak cara penyampaian/penerimaan informasi melalui media tradisional seperti media cetak dan elektronik, dan yang paling maju adalah media sosial[8].

Ujaran kebencian merupakan hal yang sangat umum di media sosial, tidak terkecuali Twitter. Berbagai upaya telah dilakukan oleh pengembang, salah satunya adalah dengan menyewa moderator konten. Moderator bertanggung jawab untuk menyaring banyak hal yang dilaporkan oleh pengguna, mulai dari gambar yang menyinggung hingga ujaran kebencian.

Seiring berkembangnya jaman, semakin banyak pengguna baru yang akan menggunakan Twitter, semakin tinggi pula jumlah postingan yang akan dibuat oleh user, maka dari itu akan semakin tinggi pula jumlah cuitan yang mengandung ujaran kebencian yang akan. Jika sebuah sosial media memiliki banyak konten yang negatif, maka akan membutuhkan waktu lama pula untuk memverifikasi konten yang mengandung ujaran kebencian. Lamanya ujaran kebencian disahkan sangat merugikan dalam perselisihan yang ada karena hal tersebut dapat merugikan suatu pihak dalam waktu yang lama dan juga menciptakan suasana yang tidak nyaman bagi pengguna yang lain.

Dengan *Machine Learning* yang dapat mendeteksi ujaran kebencian, maka tidak dibutuhkan banyak moderator untuk menganalisis komentar dan konten Mengingat begitu banyak jenis ujaran kebencian, kemungkinan untuk tidak terdeteksi atau terlewat sangat tinggi. Selain itu, dengan adanya *Machine Learning* yang dapat mendeteksi ujaran kebencian, maka pekerjaan moderator konten lebih mudah lagi[9].

Penelitian yang dirujuk adalah studi yang dilakukan oleh Arroyo dengan menggunakan kumpulan data TRAC-1. Dataset tersebut berisi 12.041 item data yang diklasifikasikan ke dalam beberapa kategori menurut derajat ujaran kebencian. Beberapa algoritma seperti Naive Bayes, Perceptron, SVM dan Passive Aggressive digunakan dalam penelitian ini. Penelitian ini menggunakan TF-IDF, Bag of Words, WISSE, dan N-gram. Dalam penelitian ini, perceptron dianggap sangat tidak stabil. Naive Bayes dan SVM memiliki kinerja yang lebih baik dan lebih stabil. Menggabungkan TF-IDF dan n-gram menunjukkan kinerja yang stabil dengan peningkatan akurasi sekitar 10% dibandingkan dengan TF-IDF saja. TF-IDF tidak digunakan di SVM dan Naive Bayes berkinerja lebih baik dibandingkan dengan Bag of Words. Secara keseluruhan, SVM dan Naive Bayes mencetak kinerja terbaik dan akurasi tertinggi dibandingkan dengan Perceptron dan Passive Aggressive[10].

Penelitian selanjutnya adalah penelitian yang dilakukan oleh Samghabadi. Penelitian ini mirip dengan penelitian sebelumnya karena menggunakan dataset

TRAC 2018. Kumpulan data ini berasal dari Facebook, yang terdiri dari 12.000 pidato kebencian dalam bahasa Inggris dan 12.000 India. Logistic Regression dan algoritma SVM digunakan dalam penelitian ini. Penyelidikan ini dilanjutkan dengan menggunakan algoritma Logistic Regression karena pada akhirnya ditentukan untuk mengungguli SVM. TF-IDF, N-Grams, Sentient, Word2Vec, LIWC dan Gender Probability juga digunakan dalam penelitian ini. Hasil eksperimen pada dataset bahasa Inggris menunjukkan bahwa TF-IDF memiliki akurasi tertinggi, diikuti oleh n-gram dan Word2Vec, mengungguli sentimen, LIWC, dan probabilitas gender lebih dari 20%. Hasil ketika TF-IDF, Word2Vec dan n-gram digabungkan adalah 58,75%, 0,71% lebih tinggi daripada ketika TF-IDF digunakan sendiri[11].

Penelitian yang terakhir adalah Penelitian dari Del Vigna. Penelitian ini menggunakan dataset Facebook yang berisi 17.567 komentar dalam bahasa Italia. Penelitian ini menggunakan algoritma SVM dan LSTM. Word2vec juga digunakan untuk melengkapi SVM dan LSTM. Hasil survei menunjukkan bahwa SVM 4% lebih akurat daripada LSTM[12].

Penelitian yang terakhir adalah Penelitian dari Hakiem M, Dalam penelitian ini, digunakan metode Naïve Bayes dengan memanfaatkan fitur N-gram dan seleksi fitur Information Gain. N-gram yang digunakan meliputi Unigram, Bigram, dan kombinasi Unigram-bigram. Data penelitian terdiri dari 250 data ujaran kebencian dan 250 data bukan ujaran kebencian, dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Hasil terbaik diperoleh menggunakan fitur Unigram tanpa seleksi fitur Information Gain, dengan akurasi 84%, presisi 92%, recall 79,31%, dan f-measure 85,18%. Kesimpulannya, metode Naïve Bayes dengan fitur Unigram tanpa seleksi fitur Information Gain adalah yang terbaik untuk klasifikasi ujaran kebencian di Twitter[13].

Tabel 2. 1 Tabel Pembanding Penelitian

Unsur Pembanding	**Arroyo**	**Samghabadi**	**Del Vigna**	**Hakiem M**	**Penulis**
Sumber Dataset	Instagram	Facebook	Facebook	Twitter	Twitter
Algoritma Klasifikasi	Naive Bayes, Perceptron, SVM dan Passive Aggressive	Logistic Regression, SVM	SVM, LTSM	SVM	SVM, Naïve Bayes, Logistic Regression
Feature extraction	TF-IDF, Bag of Words, WISSE, dan N-gram	TF-IDF, N-Grams, Sentient, Word2Vec, LIWC dan Gender Probability	Word2Vec	N-Gram	N-Gram, Word2vec, TF-IDF

Pada penelitian ini akan dilakukan eksperimen untuk mengklasifikasikan ujaran kebencian dengan menggunakan berbagai macam algoritma yang dinilai cocok. Tidak semua algoritma memberikan hasil yang baik dikarenakan sangat bergantung pada dataset