BAB 2
TINJAUAN PUSTAKA

2.1 *Harassment*

Harassment mengacu pada tindakan yang bertujuan untuk merendahkan, mengancam, atau mempermalukan individu, baik secara fisik maupun verbal. Dalam konteks digital, *harassment* sering terjadi di media sosial, forum, dan aplikasi pesan singkat. Bentuk umum dari pelecehan ini meliputi penghinaan, ancaman, dan penyebaran fitnah, yang juga dikenal sebagai *cyberbullying*. Pelecehan online dapat berdampak signifikan terhadap kesehatan mental, seperti peningkatan stres dan kecemasan [24]. Oleh karena itu, deteksi otomatis *harassment* menjadi penting dalam mengatasi peningkatan insiden pelecehan di dunia maya [25].

2.2 Penghinaan

Penghinaan mengacu pada ungkapan yang ditujukan untuk merendahkan atau menyinggung individu melalui bahasa atau tindakan yang merendahkan. Dalam komunikasi daring, penghinaan sering kali terjadi di media sosial, forum, dan kolom komentar, yang sering memperburuk konflik dan berkontribusi pada lingkungan digital yang tidak bersahabat. Penelitian menunjukkan bahwa paparan terhadap penghinaan dapat menyebabkan dampak psikologis, termasuk penurunan harga diri dan peningkatan kecemasan, terutama di kalangan individu yang lebih muda [26, 27]. Sistem otomatis yang mampu mendeteksi penghinaan sangat penting untuk mendorong interaksi daring yang lebih sehat dan melindungi pengguna dari pelecehan verbal [28].

2.3 Pencemaran Nama Baik

Pencemaran nama baik melibatkan pembuatan pernyataan palsu tentang seseorang, yang dapat merusak reputasi atau menyebabkan tekanan psikologis. Dalam lanskap digital, pernyataan pencemaran nama baik sering diperkuat melalui platform media sosial dan forum, membuatnya lebih meluas dan merugikan. Studi menunjukkan bahwa pencemaran nama baik tidak hanya memengaruhi kesejahteraan mental korban, tetapi juga menimbulkan tantangan bagi platform

digital dalam memoderasi konten secara efektif [29, 30]. Implikasi hukum dan etika dari pencemaran nama baik di dunia maya semakin menyoroti perlunya metode deteksi otomatis yang akurat untuk mengidentifikasi dan mengurangi konten berbahaya semacam itu [31].

2.4 Algoritma dan Metode

2.4.1 *Natural Language Processing*

Natural Language Processing (NLP) merupakan cabang kecerdasan buatan yang berfokus pada memahami dan menghasilkan bahasa manusia, sehingga memungkinkan komputer untuk berinteraksi dengan manusia tanpa harus menggunakan bahasa yang diatur oleh mesin [32].

NLP memerlukan tahap pra-pemrosesan seperti tokenisasi, pemisahan, dan *stemming* yang dapat secara signifikan meningkatkan kualitas data kode sumber yang tidak terstruktur untuk teknik pengambilan informasi [33].

2.4.2 *Text Preprocessing*

Text Preprocessing adalah serangkaian langkah penting yang dilakukan sebelum analisis dimulai. Proses ini bertujuan untuk menyederhanakan data teks sehingga dapat dianalisis secara efektif, termasuk melalui identifikasi unit-unit seperti kata dan frasa, penghilangan elemen yang tidak relevan seperti karakter non-alfabet dan kata sambung, serta pengelolaan elemen semantis seperti konsep negasi [34].

Langkah-langkah *text preprocessing* yang dilakukan dalam penelitian ini terdiri dari beberapa tahapan utama. Tahap pertama adalah normalisasi *slang word*, diikuti dengan *text cleaning*, yang meliputi pengubahan semua teks menjadi huruf kecil (*lowercase*), penghilangan tanda baca (*punctuation*) dan angka (*numbers*), serta penghapusan spasi berlebih (*extra spaces*) untuk menyederhanakan format data. Selanjutnya, proses ini diikuti oleh *tokenization*, yaitu pemisahan teks menjadi unit-unit kecil seperti kata atau frasa, yang mempermudah analisis selanjutnya.

2.4.3 *Supervised Learning*

Supervised learning adalah salah satu pendekatan dalam *machine learning* di mana model dilatih menggunakan dataset yang sudah memiliki label atau output

yang benar. Dalam *supervised learning*, model belajar dari data berlabel untuk memprediksi output yang benar pada data baru yang belum dilihat sebelumnya. Metode ini sering digunakan dalam berbagai tugas klasifikasi dan regresi, di mana algoritma menerima data input dan output yang sesuai untuk membangun fungsi yang memetakan input ke output yang diharapkan [35].

Dalam penelitian ini, *supervised learning* digunakan untuk mendeteksi kasus *harassment* seperti hujatan dan pencemaran nama baik pada teks. Dataset yang digunakan sudah dilengkapi dengan label numerik, yaitu 0 untuk teks netral, 1 untuk teks yang mengandung penghinaan, dan 2 untuk teks yang mengandung pencemaran nama baik. Dengan dataset berlabel ini, model dilatih untuk mengenali pola-pola linguistik dalam teks yang mengindikasikan kategori tertentu. Pendekatan ini memungkinkan model untuk membangun fungsi klasifikasi yang mampu memetakan teks ke salah satu dari tiga kategori tersebut, sehingga dapat digunakan untuk mendeteksi bentuk *harassment* secara otomatis dan akurat.

2.5 *Naive Bayes Classifier*

Naive Bayes adalah metode klasifikasi probabilistik sederhana yang menghitung probabilitas suatu kelas berdasarkan atribut yang ada, dengan asumsi bahwa atribut-atribut tersebut bersifat independen secara kondisional satu sama lain, diberikan kelas tertentu. Algoritma *Naive Bayes* termasuk dalam teknik *data mining* dan merupakan salah satu algoritma paling populer dalam bidang ini karena efektivitasnya dalam klasifikasi teks dan pengolahan data besar, baik dari segi presisi maupun efisiensi komputasi [36].

Metode ini didasarkan pada Teorema Bayes, yang digunakan untuk menghitung probabilitas suatu kelas dengan menggabungkan informasi dari berbagai atribut dalam data. Asumsi utama dalam *Naive Bayes* adalah bahwa atribut-atribut yang ada bersifat independen, artinya setiap atribut memberikan informasi terpisah untuk menentukan kelas target. Asumsi ini memungkinkan probabilitas gabungan $P(X_1, X_2, \ldots, X_n \mid Y)$ direpresentasikan sebagai hasil perkalian dari probabilitas individual:

$$P(Y \mid X_1, X_2, \ldots, X_n) \propto P(Y) \prod_{i=1}^{n} P(X_i \mid Y) \tag{2.1}$$

Dengan penyederhanaan ini, *Naive Bayes* menghitung probabilitas posterior untuk setiap kelas Y dan memprediksi kelas $\hat{Y}$ dengan probabilitas posterior

tertinggi:

$$\hat{Y} = \arg\max_{Y} \left(P(Y) \prod_{i=1}^{n} P(X_i \mid Y) \right) \tag{2.2}$$

Langkah ini menyelesaikan proses klasifikasi, karena algoritma mengevaluasi dan membandingkan probabilitas posterior di semua kelas yang mungkin.

Naive Bayes sangat efektif untuk klasifikasi teks karena data teks sering kali memiliki ruang fitur yang besar (misalnya, kata atau frasa), dan asumsi independensi menyederhanakan kompleksitas komputasi tanpa secara signifikan mengurangi akurasi dalam banyak aplikasi praktis. Sebagai contoh, dalam penelitian ini, penghinaan dan pencemaran nama baik dapat diklasifikasikan dengan mempertimbangkan frekuensi kata individu atau kemunculan istilah sebagai fitur. Klasifikator kemudian memprediksi kategori yang paling mungkin berdasarkan probabilitas posterior yang dihitung [36, 37].

2.5.1 *Multinomial Naive Bayes*

Salah satu varian *Naive Bayes* yang paling sering digunakan untuk tugas klasifikasi teks adalah *Multinomial Naive Bayes*. Metode ini sangat cocok ketika fitur merepresentasikan jumlah diskrit, seperti frekuensi kata atau istilah yang diekstraksi dari kumpulan data teks. Berbeda dengan klasifikator *Naive Bayes* standar yang mengasumsikan fitur kontinu, varian multinomial memodelkan probabilitas $P(X_i \mid Y)$ menggunakan frekuensi istilah di setiap kelas. Probabilitas tersebut dihitung sebagai berikut:

$$P(X_i \mid Y) = \frac{N_{i,Y} + \alpha}{N_Y + \alpha \cdot |V|} \tag{2.3}$$

Di mana:

- $N_{i,Y}$: Jumlah istilah X_i dalam dokumen yang termasuk kelas Y.
- N_Y: Jumlah total semua istilah dalam kelas Y.
- $|V|$: Ukuran kosakata.
- α: Parameter smoothing, biasanya diatur ke $\alpha = 1$ (*Laplace smoothing*) untuk menangani probabilitas nol.

Multinomial Naive Bayes sangat efektif untuk data teks dengan representasi *bag-of-words*. Dengan menggabungkan frekuensi kata, metode ini memberikan prediksi yang lebih akurat untuk tugas-tugas seperti penyaringan spam, analisis sentimen, dan kategorisasi jenis pelecehan (misalnya, penghinaan dan pencemaran nama baik). Pendekatan ini akan digunakan dalam penelitian ini untuk memodelkan dan mengklasifikasikan subkategori pelecehan.

Langkah-langkah implementasi *Naive Bayes Classifier* dalam penelitian ini adalah sebagai berikut:

1. **Menghitung Probabilitas Awal ($P(Y)$):** Frekuensi setiap kelas dalam dataset dihitung untuk menentukan probabilitas awal setiap kelas.

2. **Menghitung Likelihood ($P(X_i \mid Y)$):** Untuk setiap atribut X_i, probabilitasnya diberikan kelas Y dihitung. Dalam kasus *Multinomial Naive Bayes*, digunakan frekuensi istilah dan *Laplace smoothing*.

3. **Menghitung Probabilitas Posterior ($P(Y \mid X)$):** Menggabungkan prior dan likelihood menggunakan Teorema Bayes dan asumsi independensi.

4. **Memperkirakan Kelas:** Memilih kelas Y dengan probabilitas posterior tertinggi sebagai hasil prediksi.

Langkah-langkah ini memastikan bahwa *Naive Bayes Classifier* dan *Multinomial Naive Bayes* diterapkan secara efektif untuk tugas klasifikasi teks dalam penelitian ini, terutama untuk mendeteksi penghinaan dan pencemaran nama baik.