BAB 1
PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang semakin berkembang, kasus-kasus *harrassment* di dunia maya menjadi permasalahan yang semakin kompleks dan memerlukan penanganan yang cepat dan akurat [1]. Selama ini, proses analisis terhadap kasus-kasus tersebut melibatkan peran ahli bahasa yang bertugas menganalisis bahasa-bahasa tertentu, termasuk hujatan dan pencemaran nama baik, berdasarkan undang-undang pornografi dan UU ITE. Pihak berwenang seperti Bareskrim dan kepolisian biasanya mengundang ahli bahasa untuk menganalisis Berita Acara Pemeriksaan (BAP) guna menentukan apakah suatu kasus termasuk dalam kategori hujatan, pencemaran nama baik atau pelanggaran lainnya.

Pada wawancara pribadi dengan M. Niknik pada 18 September 2024, disampaikan bahwa seiring dengan perkembangan teknologi, sebuah *website* telah dikembangkan oleh pihak kepolisian yang memungkinkan masyarakat untuk melaporkan keluhan, termasuk kasus harassment. Meskipun *website* ini telah memfasilitasi proses pelaporan, sistem yang ada saat ini masih memiliki keterbatasan dalam hal kecepatan dan efisiensi penanganan laporan. Salah satu tantangan utama yang dihadapi adalah adanya batas waktu tiga jam sejak waktu pelaporan untuk memutuskan apakah suatu laporan dapat dilanjutkan atau tidak. Kondisi ini mengharuskan ahli bahasa untuk siap dihubungi kapan saja, bahkan pada tengah malam atau jam-jam yang tidak lazim. Hal ini berpotensi mempengaruhi akurasi analisis karena faktor kelelahan atau keterbatasan waktu [2].

Seiring dengan meningkatnya kasus terkait penggunaan bahasa dan perkembangan teknologi yang semakin pesat, Universitas Multimedia Nusantara berupaya responsif terhadap dinamika zaman dengan mengembangkan sistem penapisan kesalahan berbahasa Indonesia. Sistem ini bertujuan untuk mendukung otomatisasi dalam bidang jurnalistik [3, 4, 5]. Beberapa sistem penapisan bahasa yang telah selesai dikembangkan meliputi deteksi kesalahan ketik, penggunaan kata di-di, kata terikat, serta kata majemuk [6, 7, 8].

Namun demikian, seiring dengan berkembangnya sistem U-Tapis, identifikasi terhadap jenis teks tertentu, seperti yang mengandung unsur *harassment*, menjadi tantangan baru yang memerlukan perhatian khusus. Dalam

ranah linguistik forensik, berita atau teks memiliki potensi untuk dijadikan barang bukti yang konkret [9, 10, 11]. Oleh karena itu, pengembangan modul untuk mendeteksi teks yang mengandung *harassment* tidak hanya akan memperkuat posisi sistem U-Tapis sebagai pelopor otomatisasi di bidang jurnalistik, tetapi juga diharapkan memberikan kontribusi yang signifikan dalam upaya pencegahan dan penanggulangan kasus *harassment* di Indonesia.

Untuk menjawab tantangan ini, penting untuk mengidentifikasi kategori-kategori utama dalam *harassment*, seperti fitnah, hujatan, penghinaan, dan pencemaran nama baik. Menurut M. Niknik pada wawancaranya 23 September 2024, di antara kategori *harassment* yang ada, hujatan dan pencemaran nama baik merupakan salah satu kasus yang banyak ditemui di Indonesia. Fenomena ini menunjukkan bahwa deteksi hujatan dan pencemaran nama baik menjadi penting untuk dikembangkan dalam berbagai aplikasi berbasis teks [2].

Penelitian terkait deteksi ujaran kebencian (*hate speech*), komentar kasar (*abusive comments*), dan fenomena serupa seperti hinaan (*insult*) serta pencemaran nama baik (*defamation*) telah banyak dilakukan menggunakan metode seperti *Naive Bayes*, *Support Vector Machine* (SVM), dan *deep learning*. Dari beberapa penelitian berbasis algoritma *Naive Bayes* yang ditinjau, Sari (2019) mencatat akurasi tertinggi sebesar 83% dalam deteksi hinaan menggunakan Naive Bayes Classifier (NBC) dengan pendekatan *Term Frequency-Inverse Document Frequency* (TF-IDF) berbasis *unigram* dan metode *K-Fold Cross Validation* (k=10) [12, 13, 14, 15, 16, 17, 18]. Penelitian Wibowo (2023) menyusul dengan akurasi 80,73% dalam mendeteksi *cyberbullying* (pencemaran nama baik) pada platform Twitter [17]. Selain itu, penelitian Handayani et al. (2023) memberikan kontribusi penting dalam multiklasifikasi *hate speech*, dengan mendeteksi subkategori seperti ujaran kebencian terhadap individu, kelompok, agama, ras, fisik, dan gender. Namun, performa deteksi pada beberapa kelas masih menunjukkan kekurangan, terutama pada kelas minoritas seperti *Hate Speech Religion*, *Hate Speech Race*, dan *Hate Speech Group*, yang memiliki *F1-score* rendah untuk kelas positif, masing-masing 0,19, 0,16, dan 0,24. Ketidakseimbangan ini disebabkan oleh distribusi data yang tidak merata serta *support* yang buruk pada class tertentu, sehingga penelitian lanjutan perlu difokuskan pada pendekatan multiklasifikasi, untuk meningkatkan akurasi model pada klasifikasi multiklas [19].

Meskipun pendekatan multiklasifikasi telah dikembangkan, masih diperlukan sistem yang lebih general untuk mendeteksi kategori yang lebih luas seperti *harassment*. Penelitian ini mengambil langkah lebih lanjut dengan berfokus

pada subkategori yang lebih spesifik dari *harassment*, yakni penghinaan (*insult*) dan pencemaran nama baik (*defamation*).

Pendekatan ini menjadi semakin relevan, mengingat pentingnya aspek linguistik forensik dalam mendukung pengembangan teknologi kecerdasan buatan. Dalam ranah ilmu komputer, penelitian ini memiliki urgensi tinggi karena menyatukan aspek linguistik forensik dengan otomatisasi berbasis teknologi kecerdasan buatan. Pendekatan yang digunakan dalam penelitian ini tidak hanya menjawab kebutuhan dunia jurnalistik untuk mendeteksi subkategori *harassment*, tetapi juga mendukung pengembangan algoritma yang dapat memproses data teks secara efektif dan efisien.

Dengan demikian, *Naive Bayes Classifier* (NBC) dipilih karena kesederhanaan dan efisiensinya dalam menangani data teks, terutama pada konteks yang relevan dengan penelitian ini [20]. NBC telah terbukti efektif dalam mendeteksi pola pada data teks seperti ujaran kebencian dan teks bernuansa negatif, yang relevan dengan permasalahan penelitian ini [21]. Tantangan utama dalam penelitian ini adalah mengembangkan algoritma yang dapat mengakomodasi kompleksitas linguistik, termasuk ragam bahasa informal yang sering ditemukan dalam teks *online*, sekaligus meningkatkan akurasi deteksi terhadap bentuk *harassment* tertentu seperti penghinaan dan pencemaran nama baik, yang memiliki karakteristik linguistik berbeda [22, 23].

Selain itu, penelitian ini tidak hanya bertujuan untuk mengembangkan algoritma yang lebih adaptif, tetapi juga memberikan solusi praktis bagi sistem pelaporan *online*. Dengan integrasi pendekatan berbasis *Natural Language Processing* (NLP), penelitian ini menjawab kebutuhan praktis seperti percepatan sistem pelaporan *online*, serta memberikan kontribusi pada pengembangan teknologi analisis teks yang lebih adaptif dalam skala besar.

1.2 Rumusan Masalah

1. Bagaimana cara mengembangkan modul U-Tapis untuk deteksi kasus *harassment* penghinaan dan pencemaran nama baik dengan *Naive Bayes Classifier*?

2. Berapa tingkat akurasi, *F1 Score*, *precission*, dan *recall* deteksi *harrasment* hinaan dan pencemaran nama baik dalam bahasa Indonesia yang dapat dicapai oleh algoritma *Naive Bayes Classifier*?

1.3 Tujuan Penelitian

1. Mengembangkan modul U-Tapis untuk mendeteksi kasus *harassment* dalam bentuk penghinaan dan pencemaran baik secara otomatis menggunakan algoritma *Naive Bayes Classifier*, dengan fokus pada kalimat-kalimat dalam bahasa Indonesia.

2. Mengukur tingkat akurasi, *F1 Score*, *precission*, dan *recall* dari algoritma *Naive Bayes Classifier* dalam mendeteksi kasus *harassment* berbasis penghinaan dan pencemaran nama baik, sehingga dapat diketahui sejauh mana model ini mampu mengenali pola-pola penghinaan dan pencemaran nama baik dalam bahasa Indonesia secara efektif.

1.4 Urgensi Penelitian

Dalam era digital yang semakin berkembang, kasus hujatan di dunia maya menjadi masalah yang mendesak dan memerlukan penanganan cepat serta akurat. Saat ini, penanganan laporan hujatan masih sangat bergantung pada analisis manual oleh ahli bahasa, yang sering kali terhambat oleh keterbatasan waktu dan tenaga. Kondisi ini berpotensi memperlambat proses pengambilan keputusan dalam sistem pelaporan daring, terutama ketika batas waktu yang ketat harus dipenuhi. Oleh karena itu, penelitian ini menjadi sangat penting untuk mengembangkan solusi berbasis teknologi yang dapat mengotomatisasi deteksi hinaan dan pencemaran nama baik. Dengan menerapkan algoritma Naive Bayes Classifier pada modul U-Tapis, diharapkan proses klasifikasi teks hujatan dapat dipercepat dan efisiensi sistem pelaporan meningkat. Penelitian ini juga diharapkan dapat mengurangi ketergantungan terhadap ahli bahasa dan memberikan kontribusi nyata dalam penanganan kasus harassment di dunia maya, khususnya dalam konteks hukum di Indonesia.

1.5 Luaran Penelitian

Target utama dari penelitian ini adalah menghasilkan modul U-Tapis yang dapat mendeteksi hujatan dalam teks berbahasa Indonesia dengan akurasi tinggi menggunakan algoritma Naive Bayes Classifier. Selain pengembangan teknologi yang diharapkan dapat diterapkan dalam sistem pelaporan daring yang ada, penelitian ini juga bertujuan untuk menghasilkan luaran akademis berupa publikasi

ilmiah. Targetnya adalah menembus jurnal publikasi nasional yang terakreditasi, khususnya yang berfokus pada bidang Natural Language Processing (NLP), keamanan siber, atau pemrosesan bahasa alami. Dengan demikian, penelitian ini tidak hanya berdampak pada implementasi praktis, tetapi juga berkontribusi pada perkembangan ilmu pengetahuan di tingkat global.

1.6 Manfaat Penelitian

1. Meningkatkan efektifitas alur pemrosesan kasus *harassment* di Indonesia. Dengan pendeteksian otomatis kasus *harassment*, modul U-Tapis dapat mempercepat proses pengambilan keputusan tanpa harus selalu bergantung pada ketersediaan ahli bahasa.

2. Mengetahui efektivitas model terhadap efisiensi waktu dan tenaga ahli. Dengan menghitung F1 *Score* pada hasil pengujian menggunakan algoritma *Naive Bayes Classifier*, penelitian ini dapat memberikan wawasan tentang efisiensi tenaga dan waktu dalam proses deteksi kasus *harassment* di Indonesia. Sistem ini dapat meningkatkan efisiensi dan produktivitas alur kerja kepolisian, sehingga pengalokasian tenaga dan waktu dapat lebih baik dilaksanakan.