BAB 1
USULAN GAGASAN

1.1 Deskripsi Umum Masalah

Gambar 1.1a adalah ilustrasi dari berbagai kasus *bullying* di Indonesia yang terjadi dalam berbagai bentuk mulai dari fisik, verbal, maupun melalui media sosial.

(a)

(b)

(c)

(d)

Gambar 1.1 (a) Kasus dugaan *pembullyan* di Sekolah Menengah Atas di Serpong. (b) Miss informasi terkait Siswi Sekolah Dasar menjadi korban perundungan. (c) Kampanye Pemilu 2024, ujaran kebencian terhadap kelompok minoritas meningkat. (d) Perkara kasus ujaran kebencian soal Papua TikToker.

Berbagai ilustrasi di atas menggambarkan beragam kasus *bullying* dan ujaran kebencian yang terjadi di Indonesia, menunjukkan isu mengenai *bullying* menjadi sorotan di tengah masyarakat. Mulai dari dugaan perundungan di Sekolah Menengah Atas di Serpong yang belum lama ini terjadi kasus *bullying* Gambar 1.1a yang melibatkan anak artis terkenal sebagai salah satu pelakunya [1]. Kasus tersebut bermula ketika korban ingin bergabung ke dalam geng di sekolah menengah atas di Tangerang [2]. Tindakan ini seharusnya tidak terjadi, terlebih di lingkungan pendidikan karena tindakan tersebut menyebabkan kerugian bagi korban baik secara psikologi maupun material. Disisi lain, kita harus selektif dalam menerima informasi dari berbagai sumber informasi karena Gambar 1.1b merupakan miss informasi tentang kasus *pembullyan* yang terjadi pada siswi sekolah dasar di salah satu Kabupaten di Jawa Timur [3]. Siswi yang dianggap sebagai korban merupakan seorang siswi berkebutuhan khusus yang mengalami gangguan komunikasi dan temperamen [4]. Saat merasa marah, ia sering membenturkan wajah dan kepalanya yang menyebabkan luka-luka lebam pada wajahnya. Foto-foto lebam siswi tersebut menjadi viral di media sosial, sehingga memicu keresahan di kalangan warganet yang mengira ia adalah korban perundungan [5]. Kejadian ini mengingatkan kita akan pentingnya bersikap kritis terhadap berbagai informasi yang diterima dari berbagai sumber informasi di media sosial [6].

Seiring berkembangnya teknologi komunikasi dan informasi, media sosial memiliki peranan yang penting di masyarakat umum untuk mencari informasi dan mengutarakan pendapat. Secara tidak sadar, hal ini berdapak pada tingkah laku pengguna media sosial yang sering kali menyampaikan ujaran kebencian kepada seorang tokoh maupun kelompok tertentu. Tanpa disadari tindakan tersebut merupakan salah bentuk dari *cyberbullying*, seperti contoh pada Gambar 1.2, yaitu beberapa bentuk contoh tindakan *cyberbullying* yang berupa ujaran kebencian (*hate comments*). *Hate comments* sering kali dijumpai pada media sosial yang meliputi rasisme, *body shaming*, *sexism*, *homophobia dan transphobia*, *xenophobia*, *religious intolerance*, dan komentar yang berisi ancaman, penghinaan serta pelecehan secara langsung, bahkan tanpa adanya alasan tertentu.

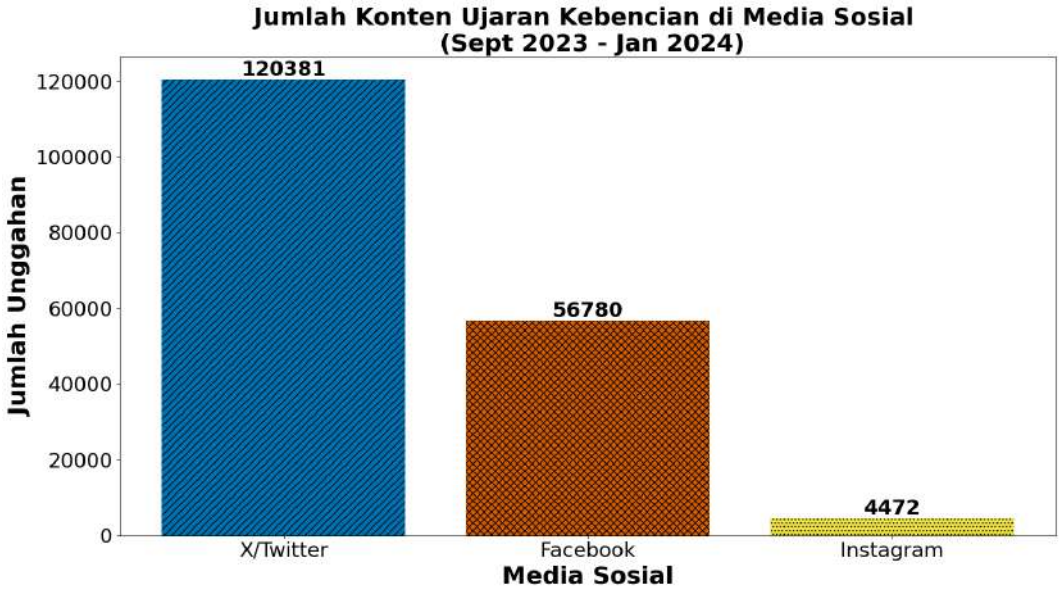
Selain itu, perkembangan teknologi informasi dan komunikasi telah membawa perubahan yang signifikan dalam cara manusia berinteraksi. Manusia menggunakan media sosial untuk berkomunikasi tanpa batas. Namun, kemudahan ini memunculkan tantangan baru bagi masyarakat, yaitu *cyberbullying* dan komentar kebencian. *Cyberbullying* merupakan tindakan intimidasi, pelecehan, atau penghinaan yang dilakukan oleh individu atau kelompok melalui media digital. Pelaku *cyberbullying* menyampaikan pesan negatif kepada korban dengan tujuan menyakiti atau

Gambar 1.2 Berbagai contoh tindakan *cyberbullying* di media sosial X.

merendahkan. Komentar kebencian adalah ungkapan negatif yang ditujukan kepada individu atau kelompok tertentu berdasarkan karakteristik seperti ras, agama atau orientasi seksual.

Tindakan tersebut secara nyata dapat dilihat pada pemilu 2024 Gambar 1.1c yang mengalami peningkatan terutama pada kelompok minoritas. Berdasarkan temuan riset sekretaris Aliansi Jurnalis Independent (AJI) Indonesia target korban berbagai macam kategori mulai dari personal, agama, hingga orietasi seksual. Seperti, ujaran kebencian yang ditujukan terhadap kelompok Yahudi terkait peristiwa serangan Israel di Gaza, perdebatan dengan meyerang personal tentang calon konsisten pemilu yang kecacatan hukum, dan lain sebagainya [7]. Selain itu, dalam dunia digital *cyberbullying* tidak dapat dihindarkan seperti kasus tiktoker Papua Gambar 1.1d. Pelaku dengan tidak bertanggung jawab mengunggah konten video yang dapat menimbulkan rasa kebencian terhadap aksi yang dilakukan oleh pendukung Eks Gubenur Papua pada saat pelaksanaan penjemputan dan pemakamannya di Papua [8]. Tindakan tersebut berujung pada tindak pidana terkait ujaran kebencian berdasarkan SARA yang dilakukan oleh pemilik, pengguna, dan yang menguasai akun media sosial TikTok. Sehingga, menyebabkan pemilik akun harus mempertanggung jawabkan tindakannya pada aparat kepolisian.

Berdasarkan penelitian yang dilakukan oleh Monash University bekerja sama dengan AJI Indonesia Gambar 1.3, ditemukan sebanyak 182.118 unggahan di media sosial yang mengandung ujaran kebencian selama masa kampanye Pemilu 2024. Penelitian ini mengungkapkan bahwa ujaran kebencian paling banyak ditemukan di platform X, dengan total 120.381 cuitan. Sementara itu, Facebook tercatat memiliki

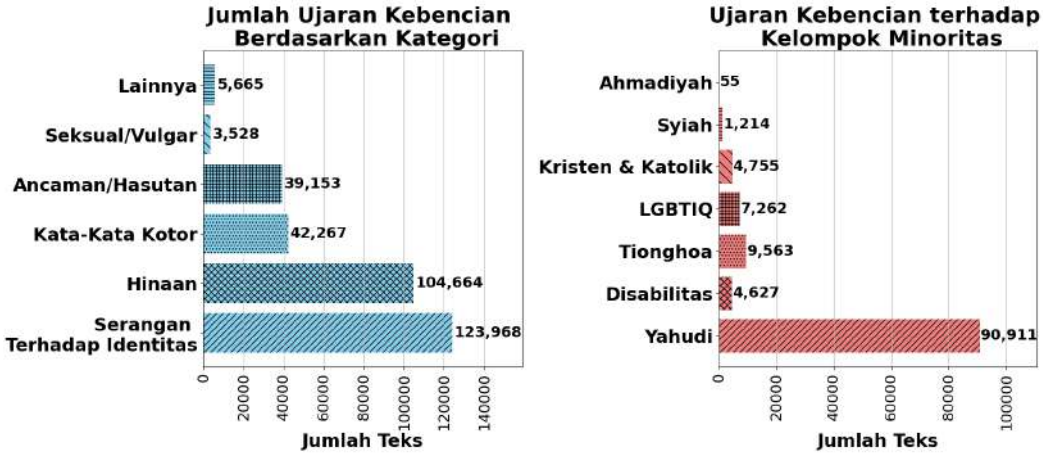


Gambar 1.3 Jumlah konten ujaran kebencian di media sosial.

56.780 unggahan dengan konten serupa, dan Instagram menyumbang 4.472 unggahan. Kami memilih platform X sebagai fokus pengumpulan dataset karena jumlah unggahan yang mengandung ujaran kebencian yang paling tinggi, yakni mencapai 120.381 data. Selain itu, eksploitasi data dari X dianggap lebih potensial mengingat Indonesia memiliki jumlah pengguna X yang signifikan, dengan sekitar 27,5 juta pengguna pada Oktober 2023, menjadikan Indonesia sebagai peringkat keempat terbesar pengguna X di dunia [9].

Berdasarkan pembahasan di atas, media sosial merupakan tempat yang kaya akan informasi. Namun penggunaannya harus hati-hati, terutama pada anak-anak. Perundungan tanpa memandang latar belakang dapat terjadi antara anak-anak yang seringkali menjadi korban atau pelaku. Hal ini menyoroti pentingnya menciptakan lingkungan yang aman dan bebas intimidasi, terutama di lembaga pendidikan. Kekerasan seperti ini tidak hanya terjadi di dunia nyata, namun juga menyebar di berbagai bidang, termasuk media sosial yang sering digunakan untuk melakukan *bullying* sehingga semakin memperparah dampak negatif bagi korbannya [10].

Berbagai fenomena *cyberbullying* yang terjadi di dunia maya berdampak negatif kepada korban, termasuk gangguan psikologis, penurunan kepercayaan diri, bahkan depresi. Menurut data dari Kementrian Komunikasi dan Informatika, kasus *cyberbullying* di Indonesia mengalami peningkatan setiap tahunnya [11]. Kondisi ini menunjukan bahwa urgensi untuk menangani permasalahan ini secara efektif semakin tinggi. Salah satu pendekatan yang dapat dilakukan adalah melalui deteksi otomatis menggunakan teknologi *Natural Language Processing* (NLP). Teknologi NLP memungkinkan sistem komputer memahami dan memproses bahasa manusia



Gambar 1.4 Jumlah ujaran kebencian berdasarkan kategori dan terhadap kelompok minoritas.

secara efektif. Dalam konteks bahasa Indonesia, tantangan utama dalam penerapan NLP adalah keterbatasan model dalam memahami bahasa.

1.2 Analisis Masalah

Gambar 1.4 menggambarkan distribusi ujaran kebencian yang teridentifikasi berdasarkan kategori dan kelompok minoritas. Data ini memberikan wawasan tentang prevalensi dan pola penyebaran ujaran kebencian dalam berbagai kategori, termasuk ras, etnis, agama, orientasi seksual, dan kelompok lainnya yang rentan terhadap diskriminasi.

Berdasarkan data tersebut analisis mendalam terkait dampak yang ditimbulkan dari *cyberbullying* perlu dilakukan secara mendalam di berbagai aspek kehidupan. Fenomena ini tidak hanya berdampak pada individu, tetapi juga memiliki konsekuensi yang lebih luas bagi masyarakat, ekonomi, hukum, dan budaya. Oleh karena itu, perlu adanya pembahasan yang mendalam mengenai dampak *cyberrbullying* terhadap berbagai aspek kehidupan berrmasyarakat antara lain aspek ekonomi, sosial, hukum, psikologi, dan politik. Analisis yang mendalam terhadap berbagai aspek diharapkan dapat memberikan gambaran yang lebih baik mengenai pentingnya menangani kasus *cyberbullying* di media sosial.

Berdasarkan aspek ekonomi, *cyberbullying* dan komentar kebencian di media sosial dapat menyebabkan dampak yang signifikan. Salah satu dampaknya adalah biaya konseling yang mahal bagi korban perundungan yang memerlukan dukungan psikologis untuk pemulihan. Biaya ini menjadi beban tambahan yang membebani individu dan masyarakat secara umum. Selain itu, perusahaan yang menjadi sasaran komentar kebencian sering kali mengalami penurunan penjualan, karena rusaknya reputasi merek yang memengaruhi kepercayaan konsumen. Fenomena

disinformasi yang sering terjadi juga berkontribusi pada kerugian ekonomi, karena dapat menyebabkan konsumen membuat keputusan yang salah dalam memilih produk atau layanan, yang berujung pada penurunan permintaan dan pendapatan. Selain itu, *cyberbullying* dapat menyebabkan penurunan kesehatan mental bagi korban yang berakibat pada penurunan produktivitas di tempat kerja yang berdampak pada produktivitas perusahaan [12].

Berdasarkan aspek sosial, *cyberbullying* dan komentar kebencian di media sosial dapat menciptakan polarisasi di masyarakat, yaitu kelompok-kelompok dengan pandangan atau identitas yang berbeda saling terpecah dan memicu ketegangan. Ujaran kebencian yang disebarkan secara online sering kali memanfaatkan perbedaan ras, agama, politik, atau identitas sosial lainnya untuk menyerang atau mendiskreditkan pihak lain, yang pada akhirnya memperburuk hubungan antar kelompok. Polarisasi ini dapat memicu konflik horizontal, yaitu kelompok-kelompok dalam masyarakat saling bertentangan dan berpotensi terlibat dalam kekerasan atau pertikaian [13]. Selain itu, korban *cyberbullying* juga dapat menjadi terasing di masyarakat, karena mereka sering kali diisolasi, dijauhi, atau bahkan dipandang sebagai sasaran empuk dari kelompok-kelompok tertentu yang menganggap mereka sebagai ”musuh” [14]. Sebagai contoh dalam beberapa kasus, perdebatan politik di media sosial yang dipenuhi dengan komentar kebencian dapat menyebabkan demonstrasi besar yang berujung pada bentrokan antara kelompok pendukung yang berbeda. Hal ini menunjukkan bagaimana *cyberbullying* dan ujaran kebencian tidak hanya merusak individu secara langsung, tetapi juga dapat memecah belah masyarakat secara keseluruhan.

Berdasarkan aspek hukum, *cyberbullying* dan komentar kebencian di media sosial dapat melanggar ketentuan yang ada dalam Undang-Undang Informasi dan Transaksi Elektronik (UU ITE), yang mengatur penyalahgunaan teknologi untuk tujuan merugikan pihak lain. Pasal 27 dan Pasal 28 UU ITE mengatur larangan penyebaran konten yang dapat merugikan kehormatan, memfitnah, atau menebarkan kebencian berdasarkan suku, agama, ras, dan antar golongan (SARA) [15]. UU ITE menetapkan sanksi pidana bagi pelaku penyebaran ujaran kebencian, fitnah, atau pencemaran nama baik di dunia maya, yang dapat dihukum dengan pidana penjara dan denda. Selain itu, dalam beberapa kasus, *cyberbullying* dapat berujung pada tindak pidana pemerasan, terutama ketika pelaku mengancam korban untuk menyebarkan informasi pribadi atau merusak reputasi korban jika tuntutannya tidak dipenuhi. Oleh karena itu, tindakan *cyberbullying* tidak hanya berdampak pada kesejahteraan korban secara emosional, tetapi juga dapat mengarah pada sanksi hukum yang berat bagi pelaku, termasuk kriminalisasi dan pemerasan yang diatur

secara tegas dalam UU ITE.

Berdasarkan aspek politik, *cyberbullying* dan komentar kebencian di media sosial dapat memperburuk polarisasi politik yang sudah ada, seperti yang terlihat pada dua pemilihan presiden di Indonesia pada 2014 dan 2019. Selama periode tersebut, kelompok pendukung calon presiden saling menyerang dengan menggunakan istilah-istilah seperti ”cebong” dan ”kampret,” yang mengarah pada ujaran kebencian yang semakin intens. Serangan verbal ini tidak hanya mengarah pada perpecahan antar pendukung, tetapi juga menciptakan ketegangan politik yang meruncing. Ujaran kebencian yang beredar di media sosial memperburuk situasi, memicu demonstrasi besar-besaran yang pada akhirnya menyebabkan beberapa demonstran meninggal dunia dan bahkan peristiwa kekerasan lainnya, seperti terbakarnya halte bus di Jakarta [16]. Polarisasi politik yang dipicu oleh *cyberbullying* ini berpotensi memecah bangsa, mengancam stabilitas sosial dan politik negara, dan menciptakan suasana yang penuh kebencian dan kekerasan di kalangan masyarakat. Hal ini menunjukkan bahwa komentar kebencian dan perundungan online dapat memiliki dampak jauh lebih besar, mengarah pada disintegrasi sosial dan politik yang membahayakan keutuhan bangsa.

Berdasarkan aspek psikologis *cyberbullying* dan komentar kebencian, orang yang menjadi sasaran *cyberbullying* dan komentar kebencian mengalami serangan dan komentar yang berulang-ulang [17]. *Cyberbullying* di internet cenderung lebih berbahaya karena konten negatif disebarluaskan dan dapat berlangsung lama, sehingga meningkatkan tekanan psikologis korban. *Cyberbullying* dapat membuat korban merasa terisolasi dari lingkungan sosialnya, baik online maupun offline. Mereka mungkin menarik diri dari aktivitas sosial, menghindari bersosialisasi di dunia maya, dan kurang percaya diri dalam kehidupan sehari-hari. Hal ini dapat berdampak negatif pada kemampuan Anda untuk membentuk hubungan sosial yang sehat.

Berdasarkan analisis berbagai aspek seperti ekonomi, hukum, sosial, dan politik, dapat disimpulkan bahwa *cyberbullying* mempunyai dampak merugikan yang luas. Oleh karena itu, diperlukan cara yang cepat dan efisien untuk mengatasi fenomena tersebut. Pendekatan yang kami lakukan untuk mengatasi permasalahan *cyberbullying* adalah dengan menggunakan Artificial Intelligence (AI) yang didukung oleh teknologi NLP. Dengan menggunakan NLP, AI dapat secara otomatis mengidentifikasi, menganalisis, dan memfilter ujaran kebencian di media sosial secara *real time*, sehingga memungkinkan intervensi yang lebih cepat dan akurat. Teknologi ini membantu mengidentifikasi pola kebencian, mengurangi penyebaran informasi berbahaya, dan memberikan solusi proaktif untuk menciptakan lingkun-

gan digital yang lebih aman bagi masyarakat.

1.3 Analisis Solusi yang Sudah Ada

Penelitian yang telah ada tentang deteksi kalimat *abusive* pada teks bahasa Indonesia menggunakan arsitektur IndoBERT menunjukkan kemampuan mengklasifikasikan teks *abusive*, terutama pada dataset kedua, dengan rata-rata *F1-Score* mencapai 76,32% setelah penambahan data pada kelas minoritas untuk mengatasi ketidakseimbangan data. IndoBERT masih mengalami keterbatasan dalam memprediksi kelas minoritas saat dataset tidak seimbang, yang dapat mempengaruhi akurasi klasifikasi pada semua kelas [18].

Penelitian tentang analisis sentimen terhadap *cyberbullying* di X menggunakan algoritma BiLSTM dengan akurasi 82,29% memiliki beberapa kekurangan yang perlu diperhatikan. Meskipun model ini mampu mencapai akurasi yang baik, hasilnya cenderung lebih efektif dalam mendeteksi tweet yang tidak mengandung *cyberbullying*, yang menunjukkan adanya ketidakseimbangan dalam pengklasifikasian antara kelas mayoritas dan minoritas. Selain itu, penggunaan hanya satu metode pembobotan kata, yaitu Word2Vec, dapat membatasi pemahaman model terhadap konteks dan variasi bahasa dalam tweet berbahasa Indonesia. X merupakan *platform* yang sangat dinamis dengan variasi bahasa yang khas, penting untuk menggunakan model yang memiliki kemampuan yang lebih spesifik dalam memahami bahasa Indonesia. Penggunaan model yang tidak sepenuhnya dioptimalkan untuk bahasa Indonesia bisa menyebabkan kesalahan dalam interpretasi sentimen yang ditulis dalam bahasa yang tidak standar atau mengandung ungkapan dan singkatan. Ketersediaan data dengan distribusi kategori yang tidak seimbang juga dapat mempengaruhi akurasi model dalam mendeteksi *cyberbullying* [19].

Penelitian tentang analisis sentimen terhadap *cyberbullying* di X menggunakan algoritma Naïve Bayes memperoleh akurasi sebesar 86,62%. Namun, terdapat kekurangan pada penggunaan dataset yang memiliki jumlah variabel terbatas, yang dapat membatasi kemampuan model dalam mengklasifikasikan data dengan variasi lebih luas. Penelitian ini hanya mengklasifikasikan sentimen positif dan negatif tanpa mempertimbangkan sentimen netral, yang seharusnya dapat memberikan gambaran yang lebih lengkap mengenai respons terhadap *cyberbullying*. Selain itu, algoritma Naïve Bayes mungkin tidak optimal dalam menangani bahasa tidak baku, ungkapan, atau variasi bahasa daerah yang sering muncul di media sosial [20].

Penelitian tentang optimasi analisis sentimen terhadap *cyberbullying* di Instagram menggunakan metode *Particle Swarm Optimization* (PSO) dengan algoritma

Support Vector Machine (SVM) dan Naïve Bayes memperoleh hasil akurasi 78,60% untuk kombinasi PSO-SVM dan 78,00% untuk PSO-Naïve Bayes, dengan masing-masing dukungan class precision 100% dan 99,74%. Penting untuk dicatat bahwa dalam konteks analisis sentimen *cyberbullying* di media sosial Indonesia, model yang digunakan harus mampu memahami bahasa Indonesia dengan baik, mengingat kompleksitas dan nuansa yang terkandung dalam bahasa sehari-hari yang digunakan oleh pengguna media sosial di Indonesia [21].

Penelitian ini menggunakan metode SVM untuk menganalisis sentimen terkait cyberbullying pada platform X (sebelumnya Twitter). Data yang digunakan berjumlah 1000 tweet dengan keyword pencarian ”burik”, ”jelek”, dan ”gendut”. Hasil analisis menunjukkan akurasi sebesar 87%, presisi 88%, *recall* 99%, dan *f1-score* 93%. Meskipun demikian, penting untuk memastikan bahwa model yang digunakan memiliki kemampuan untuk memahami bahasa Indonesia dengan baik, mengingat karakteristik bahasa yang digunakan di media sosial, seperti kata-kata ungkapan dan ekspresi yang sering muncul [22].

Penelitian ini menggunakan model *Bidirectional Encoder Representations from Transformers* (BERT) yang telah di *fine-tuning* untuk menganalisis sentimen ulasan film *Dirty Vote*, dengan fokus utama pada akurasi. Hasil penelitian menunjukkan bahwa model BERT mencapai akurasi lebih dari 0.8 pada dataset validasi. Namun, penting untuk ditekankan bahwa untuk mencapai hasil yang optimal, model harus memahami bahasa Indonesia dengan baik, mengingat ulasan yang dianalisis menggunakan bahasa Indonesia. Keberhasilan model dalam mengatasi kompleksitas bahasa Indonesia akan sangat mempengaruhi keakuratan hasil analisis sentimen [23].

Penelitian ini membandingkan akurasi BERT dan DistilBERT dalam analisis sentimen ulasan kursus. Meskipun keduanya memiliki nilai *F1 Score* yang mirip, model belum mampu menangkap kelas minoritas dengan baik. *Transfer learning* menggunakan model yang telah dilatih terbukti efektif, meskipun masih ada kesulitan dalam menangani data terbatas. Hal ini menunjukkan bahwa pemahaman bahasa Indonesia yang baik sangat penting agar model dapat lebih akurat dalam menganalisis sentimen [24].

Penelitian ini menggunakan algoritma SVM yang dioptimasi dengan PSO untuk menganalisis sentimen masyarakat terhadap anggota Kelompok Penyelenggara Pemungutan Suara (KPPS) di media sosial X menjelang Pemilu 2024 di Indonesia. Dari 702 data opini yang dikumpulkan, setelah melalui tahap *preprocessing*, tersisa 688 data dengan kata ”meninggal” yang paling banyak disebutkan. Hasil analisis menunjukkan akurasi 70%, dengan 99% opini bersentimen positif dan 1% opini

bersentimen negatif. Meskipun demikian, hasil ini menunjukkan bahwa model masih kesulitan dalam menangkap sentimen minoritas dengan lebih baik [25].

Penelitian ini menguji kemampuan model dalam memahami sentimen publik terkait kedatangan pengungsi rohingya di Indonesia menggunakan dua metode, yaitu SVM dan Naïve Bayes, dengan fokus pada analisis sentimen dalam bahasa Indonesia. Dataset yang digunakan berisi 3350 tweet yang telah dibersihkan dan dibagi menjadi data pelatihan dan pengujian dengan rasio 70:30. Hasil penelitian menunjukkan bahwa model SVM memiliki akurasi yang lebih tinggi (76%) dibandingkan dengan Naïve Bayes yang hanya mencapai 70%. Setelah dilakukan optimasi dengan metode SMOTE untuk menangani ketidakseimbangan data, kedua model menunjukkan peningkatan performa. Model SVM mencatatkan *precision* 0.74, *recall* 0.73, dan *f1-score* 0.74, yang menunjukkan kemampuannya dalam lebih memahami dan menganalisis sentimen yang disampaikan dalam bahasa Indonesia. Performa model SVM yang lebih baik mencerminkan kemampuan yang lebih tinggi dalam mengidentifikasi sentimen yang terkandung dalam tweet berbahasa Indonesia, baik dalam bentuk opini positif maupun negatif [26].

Penelitian ini membandingkan dua algoritma dalam analisis sentimen terhadap opini masyarakat Indonesia mengenai perkembangan teknologi metaverse di media sosial X, yaitu Naïve Bayes dan *logistic regression*. Dalam eksperimen ini, digunakan dataset sebanyak 6728 komentar yang dianalisis menggunakan pendekatan text mining. Hasil awal menunjukkan bahwa *logistic regression* memiliki akurasi sedikit lebih tinggi (91%) dibandingkan Naïve Bayes (90%), meskipun kedua model menunjukkan hasil yang kurang optimal pada *precision*, *recall*, dan *F1-Score*, terutama karena ketidakseimbangan data yang dominan pada sentimen positif (5799 data) dibandingkan sentimen negatif (795 data). Sebagai alternatif solusi untuk mengatasi masalah ini, diterapkan optimasi SMOTE untuk menyeimbangkan jumlah data pada kedua sentimen. Setelah optimasi SMOTE, model *logistic regression* menunjukkan peningkatan yang signifikan dengan akurasi 95%, *precision* 94%, *recall* 93%, dan *F1-Score* 95%, sementara Naïve Bayes juga mengalami perbaikan dengan akurasi 91%, namun nilai *precision*, *recall*, dan *F1-Score* lebih rendah dibandingkan *logistic regression*. Penelitian ini menegaskan pentingnya optimasi SMOTE dalam menangani ketidakseimbangan data dan menunjukkan bahwa *logistic regression* lebih unggul dalam memprediksi sentimen terkait metaverse. Ke depan, disarankan untuk membandingkan algoritma klasifikasi teks lain untuk mendapatkan performa yang lebih baik [27].

Cybersentinel adalah aplikasi deteksi *cyberbullying* yang menggunakan pendekatan *machine learning* dan analisis sentimen dengan *VADER Lexicon* un-

tuk menganalisis komentar berbahasa Indonesia di media sosial. Aplikasi ini mengandalkan algoritma SVM yang dioptimalkan dengan *GridSearchCV* untuk meningkatkan akurasi dalam klasifikasi komentar sebagai *bullying* atau tidak. Namun, meskipun model ini dioptimalkan, aplikasi ini masih menghadapi tantangan dalam memahami konteks bahasa Indonesia secara penuh, karena bahasa Indonesia memiliki banyak variasi dan nuansa yang sulit dipahami oleh model yang lebih umum. Dengan hasil akurasi yang mencapai 98,83%, aplikasi ini menunjukkan kemampuan yang cukup baik, meskipun tidak sepenuhnya efisien dalam menangani kompleksitas bahasa Indonesia [28]–[31].

1.4 Kesimpulan dan Ringkasan Bab 1

Cyberbullying dan ujaran kebencian di media sosial di Indonesia merupakan permasalahan serius yang berdampak luas, baik bagi kesehatan mental individu maupun stabilitas sosial. Pengguna mengalami gangguan psikologis seperti stres, depresi, bahkan keinginan untuk mengakhiri hidup. Selain dampak pada individu, ujaran kebencian juga memperkeruh suasana politik, memicu konflik sosial, dan meningkatkan polarisasi di masyarakat. Di sisi lain, *platform* media sosial sering kali kesulitan menanggulangi kasus ini karena tingginya volume data serta variasi bahasa dan slang lokal. Oleh karena itu, penerapan teknologi seperti NLP menjadi sangat penting untuk mendeteksi dan mengurangi penyebaran konten negatif secara otomatis. Berbagai penelitian telah dilakukan menggunakan model AI, termasuk IndoBERT, BiLSTM, dan model *deep learning* lainnya. Hasilnya cukup menjanjikan, namun masih ada tantangan dalam meningkatkan akurasi deteksi, terutama karena kompleksitas bahasa Indonesia dan keterbatasan dataset yang tersedia. Upaya kolaboratif dari berbagai pihak, termasuk peneliti, perusahaan teknologi, dan pemerintah, diperlukan untuk menciptakan sistem deteksi yang lebih efektif. Penggunaan model AI yang canggih diharapkan dapat menekan angka penyebaran ujaran kebencian dan *cyberbullying*, sehingga menciptakan lingkungan digital yang lebih aman dan kondusif bagi semua pengguna.