BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Kajian terdahulu disusun berdasarkan fungsi evidensinya. Rumpun pertama memetakan bukti mengenai skala, determinan, serta dampak *education-job mismatch* menggunakan data ketenagakerjaan dan kajian kebijakan. Rumpun kedua menelaah rancangan analisis sentimen, karakter data *Twitter*/X, *preprocessing*, TF-IDF, *class imbalance*, perbandingan SVM, *Naïve Bayes*, dan *Logistic Regression*, serta perkembangan model kontekstual. Setiap studi dinilai melalui kekuatan metodologis, keterbatasan, dan relevansinya terhadap desain penelitian ini. Pendekatan tersebut digunakan agar penelitian terdahulu menjadi dasar untuk menentukan gap objek, data, metode, dan evaluasi.

Tabel 2.1 Penelitian terdahulu terkait education-job mismatch

Ref.	Fokus	Metode	Temuan Utama	Kekuatan, Keterbatasan, & Relevansi
[1]	*Horizontal mismatch* lulusan pendidikan tinggi di Indonesia	Regresi logistik biner, Sakernas Februari 2022	33,50% pekerja lulusan pendidikan tinggi mengalami *horizontal mismatch,* status *mismatch* berkaitan dengan karakteristik individu, rumpun studi, pelatihan, dan pengalaman kerja.	Kekuatan: Sakernas nasional dan determinan terukur. Keterbatasan: tidak menangkap narasi pengalaman digital. Relevansi: dasar empiris *mismatch* lulusan pendidikan tinggi.
[8]	*Horizontal mismatch* lulusan SMK di Indonesia	Regresi logistik biner, Sakernas Agustus 2022	72,71% pekerja lulusan SMK mengalami *horizontal mismatch,* beberapa bidang keahlian memiliki *mismatch* sangat tinggi.	Kekuatan: sampel lulusan SMK besar dan rinci per bidang. Keterbatasan: satu periode dan berfokus pada status statistik. Relevansi: memperkuat masalah *link and match* vokasi.
[9]	Kesenjangan pendidikan dan kebutuhan pasar kerja	Multinomial logit, analisis vokasi,	Pendidikan vokasi berkaitan dengan *mismatch* dan *decent*	Kekuatan: menghubungkan vokasi, *mismatch*, dan

		mismatch, dan pendapatan layak	*income*, serta penting untuk kebijakan pembangunan SDM.	pendapatan layak. Keterbatasan: tidak menganalisis opini publik. Relevansi: memberi kerangka kebijakan pendidikan-tenaga kerja.
[10]	*Portrait of mismatch of education and work in Indonesia*	*Labor market brief* dan diskusi kebijakan	*Mismatch* berkaitan dengan produktivitas, efisiensi pasar kerja, dan kebutuhan penguatan keterampilan.	Kekuatan: ringkasan kebijakan yang mudah dibaca dan kontekstual. Keterbatasan: format brief membatasi detail metode. Relevansi: menjelaskan produktivitas, efisiensi, dan *reskilling*.
[11]	Determinan *education-job mismatch* pada pekerja terdidik	Multilevel multinomial logit, Sakernas 2022	Faktor individu dan regional memengaruhi peluang match, overeducation, dan undereducation.	Kekuatan: memasukkan faktor individu dan regional. Keterbatasan: indikator terstruktur tidak menangkap pengalaman subjektif. Relevansi: menunjukkan mismatch dipengaruhi struktur pasar kerja.
[12]	*Mismatch* dan *outcome* ekonomi pekerja	Analisis data ketenagakerjaan Indonesia	*Mismatch* memiliki kaitan dengan *outcome* ekonomi pekerja dan kualitas pemanfaatan kompetensi.	Kekuatan: menilai *outcome* ekonomi pekerja. Keterbatasan: fokus pada hubungan ekonometrika, bukan wacana digital. Relevansi: menguatkan dampak kesejahteraan *mismatch*.
[14]	*Vertical and horizontal mismatch* di Inggris	Pengukuran *objective mismatch* dan analisis lulusan	Sekitar sepertiga lulusan bekerja pada bidang tidak terkait, konsekuensi *mismatch* dapat berbeda menurut konteks.	Kekuatan: pembanding internasional dengan pengukuran objektif. Keterbatasan: konteks Inggris tidak langsung digeneralisasi ke Indonesia. Relevansi: menunjukkan *mismatch* bersifat lintas negara.
[13]	*Education-occupation mismatch and wage penalties in Indonesia*	Analisis ekonometrika berbasis Sakernas	*Overeducation* dan *mismatch* berkaitan dengan *wage penalties* serta mobilitas kerja.	Kekuatan: mengukur *wage penalty* dan mobilitas. Keterbatasan: tidak membahas sentimen atau strategi adaptasi.

				Relevansi: membuktikan konsekuensi ekonomi *mismatch*.
[15]	Ketidaksesuaian jurusan pada konteks pendidikan	Kajian pendidikan kejuruan	Ketidaksesuaian jurusan dapat berkaitan dengan konflik akademik, kepercayaan diri, dan tekanan psikologis.	Kekuatan: menyoroti dampak psikologis salah jurusan. Keterbatasan: konteks pendidikan, bukan outcome pekerjaan. Relevansi: menjelaskan faktor hulu sebelum transisi kerja.
[16]	Pengangguran lulusan SMK	Isu sepekan kebijakan publik	Pengangguran lulusan SMK menjadi perhatian kebijakan ketenagakerjaan.	Kekuatan: aktual untuk konteks kebijakan vokasi. Keterbatasan: bersifat ringkas dan bukan pengukuran langsung *horizontal mismatch*. Relevansi: memberi konteks tekanan transisi kerja.
[26]	Performa pasar kerja lulusan vokasi dan umum	Analisis *outcome* pasar kerja	Jalur pendidikan berhubungan dengan transisi kerja, stabilitas kerja, dan *outcome* pasar kerja.	Kekuatan: membandingkan *outcome* lulusan vokasi dan umum pada kondisi berbeda. Keterbatasan: tidak menganalisis sentimen. Relevansi: menghubungkan jalur pendidikan dan pengalaman pasar kerja.
[27]	*Mismatch* dan kesenjangan upah lulusan SMK	Heckman two-step, OLS, Blinder-Oaxaca	Dampak *mismatch* terhadap upah dapat bervariasi dan tidak selalu dipersepsikan seragam.	Kekuatan: menggunakan koreksi seleksi dan dekomposisi upah. Keterbatasan: fokus upah SMK, bukan opini digital. Relevansi: menjelaskan bahwa dampak *mismatch* dapat bervariasi.

Tabel 2.1 memperlihatkan bahwa penelitian mengenai *mismatch* selama ini kuat pada pengukuran statistik, determinan, dan dampak ekonomi. Sebagian besar penelitian menggunakan data resmi seperti Sakernas untuk menjelaskan siapa yang mengalami *mismatch* dan faktor apa yang memengaruhi nya [1], [8]. Penelitian lain

menambahkan bahwa faktor wilayah juga dapat memengaruhi peluang *mismatch* [11]. Kajian mengenai *wage penalties* menempatkan *mismatch* sebagai isu yang memiliki konsekuensi ekonomi [13]. Tabel tersebut juga menunjukkan celah penting, yaitu masih terbatasnya penelitian yang membaca bagaimana warganet membicarakan dan memaknai *mismatch* dalam ruang digital. Celah ini menjadi dasar bagi penelitian ini untuk menggunakan data media sosial *X* sebagai pelengkap perspektif statistik [17].

Tabel 2.2 Penelitian terdahulu terkait analisis sentimen dan algoritma machine learning

Ref.	Fokus	Metode	Temuan Utama	Kekuatan, Keterbatasan, dan Relevansi
[18]	Sentimen pembangunan IKN pada *X*	*Scraping*, *preprocessing*, TF-IDF, SVM, *LR*, *NB*	SVM memperoleh *accuracy* 80%, sedangkan *Logistic Regression* dan *Naïve Bayes* masing-masing 79% pada isu IKN.	Kekuatan: *platform* X, TF-IDF, dan tiga algoritma yang sama. Keterbatasan: topik, label, dan distribusi data berbeda. Relevansi: rujukan langsung desain komparasi.
[19]	Sentimen migrasi televisi digital pada *Twitter*	*Lexicon-based* dan komparasi *Multinomial NB*, SVM, *LR*	SVM memperoleh *accuracy* 94%, *Logistic Regression* 90%, dan *Multinomial Naïve Bayes* 88%.	Kekuatan: *dataset* dan metode label sama untuk tiga model. Keterbatasan: hanya 500 data dan dominan membahas *accuracy*. Relevansi: menunjukkan SVM dapat menjadi *baseline* kuat.
[20]	Sentimen pasca Pilpres 2024 di *Twitter*	Preprocessing dan komparasi *NB*, SVM, *LR*	Model klasik dapat dievaluasi secara terukur pada data *Twitter* Indonesia.	Kekuatan: isu politik Indonesia dan evaluasi multi-model. Keterbatasan: karakter data politik berbeda dari pengalaman karier. Relevansi: mendukung rancangan evaluasi *Twitter*.

[21]	Sentimen aplikasi *TikTok*	SVM, *Logistic Regression*, *Naïve Bayes*	*Logistic Regression* memperoleh *accuracy* 84% dan F1 83%; SVM 82% dan F1 81%; *Naïve Bayes* 79% dan F1 78%.	Kekuatan: 2.100 ulasan, tiga kelas, dan empat metrik. Keterbatasan: ulasan aplikasi tidak sama dengan unggahan spontan *X*. Relevansi: pembanding kuantitatif lintas domain.
[22]	Sentimen COVID-19 di Indonesia	*Bernoulli NB*, SVM, *LR*	*Logistic Regression* dan SVM memperoleh *accuracy* 99%, sedangkan *Bernoulli Naïve Bayes* 98% pada klasifikasi biner.	Kekuatan: hasil kuantitatif jelas untuk tiga model. Keterbatasan: klasifikasi biner dan nilai sangat tinggi tidak langsung sebanding. Relevansi: menegaskan pentingnya konteks *dataset*.
[23]	Sentimen ulasan *marketplace Tokopedia*	SVM, *Naïve Bayes*, *Logistic Regression*	Ketiga algoritma dapat digunakan sebagai *baseline* analisis sentimen bahasa Indonesia.	Kekuatan: teks Indonesia dan tiga *baseline*. Keterbatasan: domain *marketplace* serta sumber ulasan terstruktur. Relevansi: mendukung penggunaan model klasik pada bahasa Indonesia.
[24]	Perbandingan *Naïve Bayes* dan SVM	Komparasi algoritma klasifikasi sentimen	Performa model bergantung pada karakter data, fitur, dan *preprocessing*.	Kekuatan: menunjukkan performa bergantung fitur dan *preprocessing*. Keterbatasan: hanya membandingkan dua algoritma. Relevansi: mencegah asumsi model terbaik sejak awal.
[25]	Sentimen ulasan *LinkedIn*	SVM, *Naïve Bayes*, *Logistic Regression*	Perbandingan model dapat menghasilkan rekomendasi algoritma paling efektif pada domain tertentu.	Kekuatan: tiga model pada domain profesional dan *dataset* besar. Keterbatasan: data ulasan *LinkedIn*

				berbeda dari percakapan *X*. Relevansi: pembanding domain kerja yang berdekatan.
[28]	*Stopword removal* dan *SMOTE* pada sentimen Indonesia	*LR*, *RF*, *Grid Search*, evaluasi *F1*	*Stopword removal* tidak selalu meningkatkan performa, *imbalance* perlu diperhatikan.	Kekuatan: menguji *preprocessing* dan penanganan *imbalance*. Keterbatasan: model serta sumber data berbeda. Relevansi: dasar *ablation* dan penggunaan F1.
[29]	Dataset sentimen *Twitter* Indonesia	Koleksi dan dokumentasi dataset	Dataset *Twitter* dapat digunakan untuk analisis opini publik dan *benchmarking*.	Kekuatan: dokumentasi *dataset* dan provenance. Keterbatasan: bukan komparasi model *mismatch*. Relevansi: rujukan audit serta transparansi *dataset*.
[30]	Dataset sentimen pemilu Indonesia	Pengumpulan data *Twitter API* dan publikasi dataset	Dokumentasi dataset mendukung *reproducibility* penelitian sentimen.	Kekuatan: dokumentasi pengumpulan dan publikasi *dataset*. Keterbatasan: domain pemilu dan kebijakan akses berbeda. Relevansi: memperkuat *reproducibility*.

Tabel 2.2 memperlihatkan bahwa penelitian analisis sentimen memiliki kekuatan pada alur komputasional yang dapat direplikasi, mulai dari pengumpulan data, *preprocessing*, representasi TF-IDF, pelatihan model, hingga evaluasi. Perbandingan beberapa algoritma juga menunjukkan bahwa tidak ada satu model yang selalu unggul pada semua domain, performa dipengaruhi oleh sumber data, jumlah kelas, distribusi label, aturan pelabelan, fitur, dan skema pembagian data. Keterbatasan yang sering muncul adalah penekanan pada *accuracy*, ukuran *dataset* yang terbatas, klasifikasi biner, serta penggunaan label otomatis tanpa validasi manusia. Penelitian ini menanggapi keterbatasan tersebut melalui evaluasi tiga

kelas, *macro F1*, *balanced accuracy*, MCC, *confusion matrix*, *cross-validation*, *quality gate* internal, dan *error analysis*.

Berdasarkan kedua rumpun literatur tersebut, posisi penelitian ini adalah menjembatani kajian *mismatch* yang kuat secara statistik dengan analisis sentimen yang mampu menangkap persepsi publik secara aktual. *Gap* penelitian ini terletak pada empat aspek. Pertama, gap objek material, karena penelitian *mismatch* umumnya menggunakan data resmi dan belum banyak menggunakan percakapan publik [1], [8]. Kedua, *gap* konteks, karena media sosial *X* dapat menangkap pengalaman dan opini yang bergerak cepat [17]. Ketiga, *gap* metode, karena analisis *mismatch* lebih banyak menggunakan pendekatan ekonometrika, sedangkan penelitian ini menggunakan *text mining* dan *machine learning* [18], [19]. Keempat, *gap* evaluasi komparatif, karena SVM, *Naïve Bayes*, dan *Logistic Regression* belum banyak dibandingkan secara spesifik pada topik ketidaksesuaian jurusan pendidikan dan pekerjaan berbasis data *X* berbahasa Indonesia [23], [25].

Sintesis kedua rumpun literatur menempatkan penelitian ini sebagai penghubung antara bukti statistik *mismatch* dan pembacaan opini publik berbasis teks. Gap objek terletak pada terbatasnya pemanfaatan percakapan publik untuk memahami pengalaman kerja tidak sesuai jurusan. Gap data terletak pada kebutuhan dokumentasi *batch*, deduplikasi, kurasi relevansi, kualitas bahasa, dan distribusi kelas secara transparan. Gap metode terletak pada belum banyaknya komparasi SVM, *Naïve Bayes*, dan *Logistic Regression* secara khusus untuk topik *education-job mismatch* berbahasa Indonesia. Gap evaluasi terletak pada kebutuhan metrik yang tidak hanya mengandalkan *accuracy* serta perlunya pembahasan generalisasi, kesalahan per kelas, dan keterbatasan label. Novelty penelitian terletak pada penerapan *pipeline* yang *auditable* untuk membaca isu *mismatch* dari data publik *X* dan menghubungkan hasil model dengan karakteristik data serta keterbatasannya.

2.2 Teori yang berkaitan

2.2.1 Konsep ketidaksesuaian jurusan pendidikan dan pekerjaan

Ketidaksesuaian jurusan pendidikan dan pekerjaan merupakan kondisi ketika bidang pendidikan, keterampilan, atau kompetensi yang dimiliki individu

tidak sepenuhnya selaras dengan pekerjaan yang dijalani. Dalam literatur ketenagakerjaan, kondisi ini sering dibahas sebagai *education-job mismatch* atau *skills mismatch*. *Education-job mismatch* dapat dilihat dari dua dimensi utama, yaitu *vertical mismatch* dan *horizontal mismatch*. *Vertical mismatch* terjadi ketika tingkat pendidikan pekerja lebih tinggi atau lebih rendah daripada tingkat pendidikan yang dibutuhkan oleh pekerjaan, sedangkan *horizontal mismatch* terjadi ketika bidang pendidikan atau keahlian pekerja tidak sesuai dengan bidang pekerjaan yang dilakukan [1], [2], [7].

Pada konteks penelitian ini, fokus utama adalah *horizontal mismatch* karena topik penelitian membahas ketidaksesuaian antara jurusan pendidikan dengan pekerjaan. Konsep *vertical mismatch* tetap dijelaskan karena keduanya sering muncul bersamaan dalam diskusi transisi pendidikan ke kerja. BPS menyoroti bahwa *mismatch* pada pemuda dapat berhubungan dengan *wage penalties* dan inefisiensi ekonomi [7], sementara studi-studi berbasis Sakernas menunjukkan bahwa *horizontal mismatch* terjadi pada lulusan pendidikan tinggi maupun lulusan SMK [1], [8]. Dengan demikian, *mismatch* dipahami sebagai persoalan struktural pasar kerja.

2.2.2 Skills Mismatch, Produktivitas, dan Link and Match

Skills mismatch adalah kondisi ketika keterampilan yang dimiliki pekerja tidak sesuai dengan keterampilan yang dibutuhkan oleh pekerjaan. Ketidaksesuaian ini dapat menyebabkan *underutilization of skills*, kebutuhan pelatihan tambahan, produktivitas yang tidak optimal, dan risiko ketidakpuasan kerja [3], [4], [5]. Dalam konteks pendidikan dan ketenagakerjaan, konsep *link and match* digunakan untuk menggambarkan upaya penyelarasan antara kurikulum, pelatihan, sertifikasi, dan kebutuhan industri [9], [10].

Apabila *link and match* tidak berjalan optimal, lulusan dapat memasuki pekerjaan yang tidak sesuai dengan bidang pendidikannya. Dari sisi individu, *mismatch* dapat berkaitan dengan rumpun studi, pelatihan, dan pengalaman kerja sebelum lulus [1]. Dari sisi kelembagaan dan pasar kerja, masalah ini juga berkaitan dengan penyelarasan pendidikan, pelatihan, dan kebutuhan industri [9], [10]. Oleh karena itu, teori *skills mismatch* dan *link and match* menjadi

dasar untuk memahami mengapa percakapan publik mengenai “salah jurusan” atau “kerja tidak sesuai jurusan” dapat muncul di media sosial.

2.2.3 Media Sosial X sebagai Sumber Data Opini Publik

Media sosial *X* merupakan platform mikroblog yang memungkinkan pengguna menyampaikan opini, pengalaman, keluhan, komentar, dan respons terhadap isu sosial secara cepat. DataReportal mencatat besarnya ekosistem digital yang mendukung penggunaan media sosial sebagai sumber data penelitian, termasuk jumlah pengguna internet, pengguna media sosial, dan jangkauan audiens *X* [17]. Walaupun data *X* tidak dapat disamakan dengan survei representatif nasional, platform ini relevan untuk menangkap narasi publik yang bersifat aktual dan spontan.

Pada penelitian ini, *X* digunakan untuk menangkap opini warganet mengenai ketidaksesuaian jurusan pendidikan dan pekerjaan. Narasi seperti “salah jurusan”, “kerja tidak sesuai jurusan”, “jurusan tidak kepakai”, atau “kuliah berbeda dengan kerja” mencerminkan pengalaman sosial yang tidak selalu tertangkap oleh data statistik resmi. Dengan demikian, data *X* berfungsi sebagai pelengkap data ketenagakerjaan karena dapat memperlihatkan bagaimana warganet membicarakan, menilai, dan memaknai *mismatch* dalam kehidupan sehari-hari [17]. Dataset berbasis *Twitter*/*X* juga telah digunakan dalam penelitian opini publik sehingga prosedur dokumentasi data tetap perlu dijelaskan secara transparan [29], [30].

2.2.4 Analisis Sentimen

Analisis sentimen adalah proses mengidentifikasi dan mengklasifikasikan opini dalam teks ke dalam kategori tertentu, seperti positif, negatif, dan netral [31]. Sentimen positif menunjukkan pandangan yang menerima, membenarkan, atau melihat sisi manfaat dari kerja tidak sesuai jurusan. Sentimen negatif menunjukkan keluhan, kekecewaan, kekhawatiran, atau dampak buruk akibat *mismatch*. Sentimen netral menunjukkan informasi, pertanyaan, atau pernyataan deskriptif tanpa kecenderungan emosional yang kuat [32].

Pendekatan analisis sentimen dapat dilakukan dengan metode berbasis kamus atau *lexicon-based* dan metode berbasis pembelajaran mesin atau *machine learning-based* [32]. Pada penelitian ini, pelabelan sentimen dilakukan menggunakan lexicon-based labeling berbasis domain [33]. SVM, *Naïve Bayes*, dan *Logistic Regression* kemudian digunakan sebagai model klasifikasi untuk mempelajari pola label tersebut. Kombinasi *preprocessing*, representasi TF-IDF, dan evaluasi model klasifikasi juga digunakan dalam penelitian sentimen berbasis media sosial [18], [19].

Dalam konteks linguistik sentimen Bahasa Indonesia, polaritas ditentukan oleh hubungan antarunsur bahasa dalam kalimat. Kata negasi seperti "tidak", "bukan", "belum", "ga", "gak", dan "nggak" dapat membalik atau mengubah orientasi sentimen. Kata penguat seperti "sangat", "banget", "sekali", "terlalu", dan "parah" dapat meningkatkan intensitas emosi. Selain itu, media sosial banyak memuat *slang*, singkatan, campuran bahasa, *emoji*, sarkasme, serta frasa multi-kata. Oleh karena itu, analisis sentimen Bahasa Indonesia perlu memperhatikan normalisasi kata tidak baku, keberadaan negasi, dan konteks frasa, bukan hanya menghitung frekuensi kata tunggal.

Penelitian tentang *IndoBERTweet* menunjukkan bahwa teks *Twitter* Indonesia memiliki *vocabulary mismatch* dan karakter domain yang berbeda dari teks formal, sehingga pemrosesan teks media sosial membutuhkan perhatian khusus terhadap bentuk informal dan domain percakapan [34]. Temuan ini mendukung keputusan penelitian untuk mengevaluasi *preprocessing* dan tidak menghapus negasi secara agresif.

2.2.5 Lexicon-Based Labeling

Lexicon-based labeling adalah pendekatan pelabelan sentimen yang menggunakan kamus kata atau frasa dengan polaritas tertentu untuk menentukan kecenderungan positif, negatif, atau netral pada teks [33], [35]. Pendekatan ini relevan ketika penelitian membutuhkan proses pelabelan awal yang konsisten, transparan, dan dapat direplikasi. Dalam penelitian ini, *lexicon* disusun secara *domain-specific* agar sesuai dengan konteks ketidaksesuaian jurusan pendidikan dan pekerjaan. Kata atau frasa negatif mencakup indikator

seperti “salah jurusan”, “tidak sesuai jurusan”, “tidak terpakai”, “susah kerja”, atau “menyesal”. Kata atau frasa positif mencakup indikator seperti “bisa belajar”, “ada kesempatan”, “skill lebih penting”, “tidak masalah”, atau “pengalaman baru”. Untuk konteks bahasa Indonesia, *InSet Lexicon* juga relevan sebagai rujukan daftar kata sentimen pada microblogs berbahasa Indonesia [36].

Pelabelan *lexicon* pada penelitian ini mempertimbangkan konteks sederhana seperti negasi, *intensifier*, *diminisher*, *neutral cues*, dan frasa multi-kata. Negasi seperti “tidak”, “bukan”, dan “belum” perlu dipertahankan karena dapat mengubah polaritas sentimen, sedangkan *intensifier* seperti “banget”, “sangat”, atau “terlalu” dapat memperkuat nilai sentimen [33]. Dengan demikian, frasa seperti “tidak sesuai jurusan” tidak boleh diperlakukan sama dengan kata “sesuai” yang berdiri sendiri. Literatur analisis sentimen media sosial juga menekankan tantangan slang, sarkasme, negasi, dan konteks pendek, sehingga aturan *lexicon* perlu disusun secara hati-hati dan tetap diakui keterbatasannya [35], [36].

IndoBERT tidak digunakan sebagai metode utama karena penelitian ini difokuskan pada perbandingan algoritma *machine learning* klasik berbasis TF-IDF, yaitu SVM, *Naïve Bayes*, dan *Logistic Regression*. Pemilihan ini mempertimbangkan keterbacaan proses, efisiensi komputasi, kemudahan replikasi, serta kebutuhan interpretasi fitur dan frasa dominan. Selain itu, label penelitian diperoleh melalui *lexicon-based labeling*, sehingga model klasik lebih sesuai untuk mengevaluasi hubungan antara preprocessing, representasi TF-IDF, *class_weight*, dan performa klasifikasi.

IndoBERT dan variannya tetap relevan sebagai pembanding lanjutan karena model berbasis transformer terbukti kuat untuk tugas NLP Bahasa Indonesia. *IndoBERTweet* dikembangkan khusus untuk data *Twitter* Indonesia [34], Indo LEGO-ABSA menunjukkan potensi model generatif untuk *aspect-based sentiment analysis* Bahasa Indonesia [37], dan *IndoBERT-Sentiment* menekankan pentingnya konteks topik dalam klasifikasi sentimen Bahasa Indonesia [38]. Namun, *fine-tuning IndoBERT* idealnya memerlukan label manual yang kuat dan tervalidasi agar model tidak hanya mempelajari *noise*

dari label otomatis. Oleh karena itu, *IndoBERT* ditempatkan sebagai rekomendasi penelitian lanjutan setelah validasi manual label diperkuat.

2.2.6 Kata Kunci dan Frasa Domain Mismatch

Kata kunci dalam penelitian ini memiliki dua fungsi utama. Pertama, kata kunci digunakan pada tahap pengumpulan data untuk menemukan unggahan yang relevan, misalnya "salah jurusan", "kerja tidak sesuai jurusan", "jurusan tidak kepakai", "kerja di luar jurusan", "banting setir karier", "career switch", dan "lowongan semua jurusan". Kedua, kata kunci dan frasa domain digunakan pada tahap interpretasi hasil untuk mengetahui isu dominan pada masing-masing sentimen. Oleh karena itu, kata kunci berfungsi sebagai dasar konseptual dalam membaca narasi publik tentang *mismatch* jurusan-pekerjaan. Praktik penggunaan data *X* untuk membaca opini publik juga digunakan pada penelitian dataset sosial sebelumnya [29], [30].

2.2.7 Text Mining dan Natural Language Processing

Text mining adalah proses mengekstraksi informasi dari data teks tidak terstruktur agar dapat dianalisis secara sistematis [39]. Dalam penelitian media sosial, *text mining* perlu didukung oleh *Natural Language Processing* (NLP) karena teks *X* sering mengandung singkatan, salah ketik, *slang*, *emoji*, *mention*, *hashtag*, *URL*, dan campuran bahasa. Oleh karena itu, tahap *preprocessing* menjadi penting untuk mengubah teks mentah menjadi data yang lebih bersih dan dapat diproses oleh algoritma.

Tahapan *preprocessing* yang digunakan dalam penelitian ini meliputi *case folding*, *cleaning*, normalisasi kata tidak baku, tokenisasi, *stopword removal*, dan *stemming*. Beberapa tahapan seperti *stopword removal* dan *stemming* perlu diterapkan secara hati-hati karena dapat menghilangkan makna penting pada teks pendek. Temuan Putra Negara menunjukkan bahwa *stopword removal* tidak selalu meningkatkan performa model, sehingga tahap *preprocessing* perlu dirancang berdasarkan kebutuhan data dan dievaluasi dampaknya terhadap performa klasifikasi [28].

2.2.8 Ekstraksi Fitur TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan metode representasi teks yang mengubah kata dalam dokumen menjadi bobot numerik. *Term Frequency* menggambarkan seberapa sering suatu kata muncul dalam sebuah dokumen, sedangkan *Inverse Document Frequency* menurunkan bobot kata yang terlalu sering muncul di banyak dokumen. Dengan demikian, kata yang khas pada dokumen tertentu akan memiliki bobot lebih tinggi dibanding kata umum [40].

Rumus dasar TF-IDF yang digunakan dalam representasi fitur teks adalah sebagai berikut:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad \textbf{(2.1)}$$

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad \textbf{(2.2)}$$

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \mathit{I}\mathrm{DF(t)} \quad \textbf{(2.3)}$$

Keterangan: t merupakan *term*/kata, d merupakan dokumen, f(t,d) merupakan frekuensi kemunculan term t pada dokumen d, N merupakan jumlah seluruh dokumen, dan df(t) merupakan jumlah dokumen yang mengandung term t. Rumus ini menunjukkan bahwa kata yang sering muncul pada suatu dokumen, tetapi tidak terlalu umum pada seluruh korpus, akan memperoleh bobot yang lebih tinggi [40].

Dalam penelitian ini, TF-IDF digunakan karena ketiga algoritma yang dibandingkan, yaitu SVM, *Naïve Bayes*, dan *Logistic Regression*, memerlukan input numerik. Penggunaan TF-IDF juga sesuai dengan karakter data teks pendek seperti unggahan *X* karena mampu menonjolkan kata atau frasa yang membedakan sentimen positif, negatif, dan netral. Pengujian dapat menggunakan unigram dan bigram agar frasa penting seperti "tidak sesuai", "salah jurusan", atau "kerja beda" tetap dapat tertangkap sebagai fitur [40], [41].

2.2.9 Analisis N-gram dan Frasa Dominan

Analisis *n-gram* digunakan untuk menangkap kata atau frasa yang sering muncul sebagai satuan makna [39]. *Unigram* merepresentasikan kata tunggal, sedangkan bigram dan trigram merepresentasikan kombinasi dua atau tiga kata. Pada konteks penelitian ini, n-gram penting karena makna *mismatch* sering muncul dalam bentuk frasa, misalnya "tidak sesuai", "salah jurusan", "banting setir", "pindah jalur", dan "kerja apa". Setelah data diberi label, rata-rata bobot TF-IDF pada frasa dominan dapat digunakan untuk menjelaskan isu utama yang membentuk sentimen negatif, netral, dan positif.

2.2.10 Evaluasi Model Klasifikasi

Evaluasi model klasifikasi dilakukan untuk mengukur kemampuan algoritma dalam memprediksi kelas sentimen. Metrik yang digunakan dalam penelitian ini adalah *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* menunjukkan proporsi prediksi benar dari seluruh data uji. *Precision* menunjukkan ketepatan prediksi pada suatu kelas, *recall* menunjukkan kemampuan model menangkap data yang benar-benar termasuk kelas tersebut, sedangkan *F1-score* merupakan rata-rata harmonis antara *precision* dan *recall* [18], [22].

Secara umum, metrik evaluasi yang digunakan dalam penelitian ini dapat dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(2.4)}$$

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(2.5)}$$

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(2.6)}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \textbf{(2.7)}$$

Keterangan: TP adalah *True Positive*, TN adalah *True Negative*, FP adalah False *Positive*, dan FN adalah *False Negative*. Karena penelitian ini menggunakan tiga kelas sentimen, yaitu positif, negatif, dan netral, nilai *precision*, *recall*, dan *F1-score* dapat dihitung per kelas kemudian dirata-ratakan menggunakan *macro average* atau *weighted average* [41].

Untuk memperkuat evaluasi pada dataset yang tidak seimbang, penelitian ini juga menambahkan *balanced accuracy*, *Matthews Correlation Coefficient*, *Cohen's Kappa*, dan *class weight*. *Balanced accuracy* menghitung rata-rata *recall* tiap kelas sehingga lebih adil ketika jumlah data antar kelas tidak seimbang. MCC mengukur korelasi antara label aktual dan prediksi, sedangkan *Cohen's Kappa* mengukur tingkat kesepakatan prediksi terhadap label aktual setelah memperhitungkan kemungkinan kesepakatan secara acak.

Pada data sentimen, distribusi kelas sering tidak seimbang. Misalnya, sentimen negatif dapat lebih banyak daripada positif atau netral. Oleh karena itu, evaluasi tidak boleh hanya mengandalkan *accuracy*. Penelitian ini menggunakan *F1-score*, terutama *macro average* atau *weighted average*, agar performa setiap kelas tetap diperhatikan. *Confusion matrix* juga digunakan untuk melihat kelas mana yang paling sering tertukar, misalnya netral yang diklasifikasikan sebagai negatif karena mengandung kata keluhan [28], [41].

2.2.11 Data Splitting, Stratifikasi, dan Class Weight

Data splitting dilakukan untuk memisahkan data latih, validasi, dan uji agar proses pemilihan model tidak bias. Data latih digunakan untuk membangun model, data validasi digunakan untuk membandingkan parameter dan memilih konfigurasi terbaik, sedangkan data uji digunakan hanya untuk evaluasi akhir. Stratified split diperlukan agar proporsi kelas positif, negatif, dan netral tetap relatif seimbang pada setiap subset. Pada model tertentu, *class_weight* dapat digunakan untuk membantu model memperhatikan kelas minoritas [41]**.**

2.2.12 Etika Penelitian Data Media Sosial

Penelitian berbasis data media sosial perlu memperhatikan etika pengumpulan dan penyajian data. Walaupun unggahan *X* bersifat publik,

peneliti tetap perlu melakukan anonimisasi identitas pengguna, tidak menampilkan *username* atau ID akun, tidak menyebarkan ulang konten sensitif secara verbatim, serta menyimpan data secara aman. Kutipan unggahan, apabila digunakan dalam pembahasan, perlu disamarkan atau diparafrasekan agar tidak merugikan pengguna. Prinsip ini penting karena penelitian menganalisis pengalaman karier dan pendidikan yang dapat bersifat personal [42].

2.3 Framework/Algoritma yang digunakan

1) CRISP-DM

Framework yang digunakan dalam penelitian ini adalah adaptasi CRISP-DM (*Cross Industry Standard Process for Data Mining*). CRISP-DM dipilih karena penelitian ini berbasis data mining dan membutuhkan alur yang sistematis dari pemahaman masalah hingga evaluasi hasil. Tahapan yang digunakan meliputi *business/research understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *reporting*. Dalam konteks penelitian ini, *business understanding* berfokus pada masalah *mismatch* jurusan-pekerjaan, *data understanding* berfokus pada karakter data *X*, *data preparation* mencakup pembersihan dan pelabelan teks, modeling mencakup penerapan *SVM*, *Naïve Bayes*, dan *Logistic Regression*, sedangkan *evaluation* mencakup perbandingan performa model [43].

2) Support Vector Machine

Support Vector Machine adalah algoritma klasifikasi yang mencari *hyperplane* terbaik untuk memisahkan data ke dalam kelas tertentu. Pada data teks, fitur TF-IDF biasanya menghasilkan dimensi yang tinggi, sehingga Linear SVM sering digunakan karena efektif untuk *data sparse*. SVM berusaha memaksimalkan margin antar kelas agar model mampu melakukan generalisasi pada data baru [44]**.** Dalam penelitian sentimen *Twitter*/*X*, SVM sering menghasilkan performa kompetitif dibandingkan algoritma klasik lain [18], [19], [45].

Secara sederhana, fungsi keputusan pada SVM dapat dituliskan sebagai berikut:

$$f(x) = w \cdot x + b \quad \textbf{(2.8)}$$

$$y_i(w \cdot x_i + b) \geq 1 \quad \textbf{(2.9)}$$

Keterangan: w adalah vektor bobot, x adalah vektor fitur TF-IDF, b adalah bias, dan y_i adalah label kelas. Prinsip utama SVM adalah mencari *hyperplane* yang dapat memisahkan kelas dengan margin paling besar sehingga model mampu mengklasifikasikan data baru secara lebih stabil [45].

3) Naïve Bayes

Naïve Bayes adalah algoritma probabilistik yang menggunakan Teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen satu sama lain. Meskipun asumsi independensi ini sederhana, *Naïve Bayes* banyak digunakan pada klasifikasi teks karena cepat, ringan, dan cukup efektif untuk data berbasis frekuensi kata. Pada penelitian ini, varian Naïve Bayes yang dapat digunakan adalah *Multinomial Naïve Bayes* atau *Bernoulli Naïve Bayes*, menyesuaikan representasi fitur yang digunakan [46].

Rumus dasar Teorema Bayes yang menjadi dasar algoritma *Naïve Bayes* adalah sebagai berikut:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \quad \textbf{(2.10)}$$

Untuk klasifikasi teks, bentuk sederhananya dapat dituliskan sebagai berikut:

$$P(C \mid d) \propto P(C) \times \prod_i P(w_i \mid C) \quad \textbf{(2.11)}$$

Keterangan: C adalah kelas sentimen, X atau d adalah dokumen teks, dan w_i adalah kata ke-i dalam dokumen. Meskipun menggunakan asumsi independensi antarfitur, *Naïve Bayes* sering digunakan pada klasifikasi teks karena sederhana, cepat, dan mampu menjadi pembanding yang kuat untuk model linear lainnya.

4) Logistic Regression

Logistic Regression adalah algoritma klasifikasi linear yang memodelkan probabilitas suatu data masuk ke kelas tertentu. Untuk klasifikasi lebih dari dua kelas, *Logistic Regression* dapat menggunakan pendekatan multinomial atau *one-vs-rest*. Algoritma ini sering digunakan pada analisis sentimen karena relatif stabil, mudah diinterpretasikan, dan bekerja baik dengan fitur TF-IDF. Dalam beberapa penelitian sentimen Indonesia, *Logistic Regression* menunjukkan performa yang kompetitif dibandingkan SVM dan *Naïve Bayes* [18], [21], [41].

Pada klasifikasi biner, *Logistic Regression* menggunakan fungsi sigmoid untuk menghasilkan probabilitas kelas, dengan rumus sebagai berikut:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-z}} \quad \textbf{(2.12)}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad \textbf{(2.13)}$$

Untuk klasifikasi multikelas seperti positif, negatif, dan netral, *Logistic Regression* dapat menggunakan pendekatan multinomial atau *softmax*:

$$P(y = k \mid x) = \frac{exp(z_k)}{\sum_j exp(z_j)} \quad \textbf{(2.14)}$$

Keterangan: x adalah vektor fitur TF-IDF, β adalah parameter model, z adalah skor linear, dan k adalah kelas sentimen. Rumus tersebut menunjukkan bahwa *Logistic Regression* mengubah kombinasi linear fitur menjadi probabilitas kelas sehingga kelas dengan probabilitas tertinggi dipilih sebagai hasil prediksi.

5) Rancangan Perbandingan Algoritma

Perbandingan algoritma dalam penelitian ini dilakukan dengan menggunakan dataset, *preprocessing*, skema pembagian data, dan metrik

evaluasi yang sama. Hal ini penting agar hasil perbandingan tidak bias. SVM, *Naïve Bayes*, dan *Logistic Regression* dilatih menggunakan fitur TF-IDF [40]. Model terbaik dipilih berdasarkan kombinasi *accuracy*, *precision*, *recall*, *F1-score*, dan interpretasi *confusion matrix* [18], [22]. Dengan rancangan ini, penelitian dapat menjawab rumusan masalah kedua secara objektif.

2.4 Tools/Software yang Digunakan

Tabel 2.3 Tools/software yang digunakan dalam penelitian

No	Tools/Software	Fungsi	Output
1	*Python*	Bahasa pemrograman utama untuk pengolahan data, *preprocessing*, pemodelan, dan evaluasi.	*Script*/*notebook* eksperimen.
2	*Jupyter Notebook*/*Google Colab*	Lingkungan eksperimen agar proses analisis terdokumentasi secara bertahap.	Notebook penelitian.
3	Akses resmi/API atau *tweet-harvest* untuk *scraping X*	Mengambil data unggahan publik berdasarkan kata kunci, periode tertentu, dan variasi query domain mismatch.	Dataset mentah dalam format *CSV*/*JSON*.
4	*pandas* dan *numpy*	Membersihkan, menggabungkan, memfilter, dan mengolah dataset.	Dataset siap *preprocessing* dan analisis.
5	*Library NLP* Bahasa Indonesia	Normalisasi, tokenisasi, *stopword removal*, dan *stemming* teks Indonesia.	Teks yang lebih terstandar.
6	*scikit-learn*	TF-IDF, pembagian data, pemodelan SVM/NB/LR, dan evaluasi metrik.	Model, *confusion matrix*, *accuracy*, *precision*, *recall*, *F1-score*.
7	*Matplotlib*	Visualisasi distribusi sentimen, distribusi waktu, perbandingan performa model, *confusion matrix*, dan frasa dominan.	Grafik hasil analisis.
8	*Mendeley*	Manajemen sitasi dan daftar pustaka.	Daftar pustaka sesuai format.

Tabel 2.3 merangkum perangkat yang digunakan dalam penelitian. Pemilihan *tools* tersebut mengikuti kebutuhan alur penelitian, mulai dari pengumpulan data *X*, pembersihan data, pemodelan, evaluasi, hingga penulisan laporan. *Python* dan *scikit-learn* menjadi perangkat utama karena mendukung implementasi TF-IDF serta algoritma SVM, *Naïve Bayes*, dan *Logistic Regression*. Sementara itu, *Mendeley* digunakan untuk menjaga konsistensi sitasi karena penelitian ini menggunakan banyak sumber ilmiah dan data resmi.