

BAB II

TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Pada tahun 2024 Ariannor et al melakukan penelitian yang bertujuan untuk menganalisis sentimen pengguna terhadap aplikasi Info Pajak Kalsel (Kalimantan Selatan). Metode yang digunakan meliputi pengumpulan data melalui *web scraping*, pra-pemrosesan teks (termasuk konversi emoji menjadi teks), pembobotan, dan klasifikasi dengan *Naïve Bayes*. Hasil penelitian menunjukkan akurasi tertinggi sebesar 92% dengan pembagian data 10:90 (10% data uji, 90% data latih), serta *precision* 92%, *recall* 90%, dan *F1-score* 91%. Temuan utama penelitian ini adalah 59,94% ulasan bersifat negatif, dengan keluhan seperti error, kesulitan cek pajak, dan kinerja aplikasi yang buruk. Sementara itu, hanya 40,06% ulasan yang positif. Hasil ini menunjukkan perlunya perbaikan signifikan pada aplikasi untuk meningkatkan kepuasan pengguna[13].

Penelitian lain yang dilakukan oleh Ikhsan et al pada tahun 2025, bertujuan untuk menganalisis sentimen publik di X terkait kenaikan PPN menggunakan dua klasifikasi yaitu *Naïve Bayes* dan *Support Vector Machine* (SVM). Penelitian ini membandingkan tingkat akurasi dari kedua klasifikasi dengan menggunakan 2.852 dataset dalam format CSV. Data kemudian melalui proses *preprocessing*, implementasi algoritma, dan proses *balancing* data menggunakan teknik SMOTE. Hasil yang didapatkan adalah SVM memiliki akurasi 98% sebelum *balancing* dan 97% setelah *balancing*. Sementara *Naïve Bayes* mencapai akurasi 97% sebelum *balancing* dan 90% setelah *balancing*. Hasil ini menyimpulkan bahwa SVM menunjukkan kinerja yang lebih baik dan tidak terlalu terpengaruh secara signifikan oleh proses[14].

Pada tahun 2024, Ardiyansah & Parjito melakukan penelitian terhadap Luhut Binsar Pandjaitan di X menggunakan 4.008 data yang dikumpulkan melalui *web scraping*. Algoritma klasifikasi yang digunakan adalah *Naïve Bayes* dan *Support Vector Machine* (SVM) dengan tujuan untuk membandingkan kedua algoritma tersebut. Dilakukan serangkaian tahap pembersihan data hingga ekstraksi fitur menggunakan TF-IDF. Sebelum perbandingan, digunakan teknik SMOTE (*Synthetic*

Minority Over-sampling Technique) untuk menyelaraskan data agar kedua algoritma dapat belajar lebih seimbang pada setiap kategori sentimen. Hasil penelitian menunjukkan bahwa algoritma SVM unggul secara signifikan dibandingkan *Naïve Bayes*. SVM mencapai akurasi 99%, *precision* 99%, *recall* 100%, dan *F1-score* 99%. Sementara itu, *Naïve Bayes* hanya menghasilkan akurasi 95% dan menunjukkan performa yang lebih rendah pada metrik lainnya. Secara keseluruhan, penelitian ini menunjukkan bahwa SVM merupakan pilihan yang lebih baik dibandingkan *Naïve Bayes*[15].

Pada tahun 2024, Faisol melakukan penelitian terhadap M-Pajak. M-Pajak merupakan aplikasi digital yang dikembangkan oleh Direktorat Jenderal Pajak (DJP) untuk membantu masyarakat dalam mengelola perpajakan. Namun, aplikasi ini menerima banyak ulasan negatif di *Google Play Store*. Menggunakan tiga algoritma yaitu *Support Vector Machine* (SVM), *Naïve Bayes*, dan *K-Nearest Neighbors* (KNN). Dataset terdiri dari 3.033 ulasan yang diklasifikasikan sebagai positif atau negatif. Hasil penelitian menunjukkan bahwa SVM unggul dengan akurasi 86,41%, *precision* 85-86%, dan *recall* tinggi untuk sentimen negatif (98,12%). *Naïve Bayes* menempati posisi kedua (akurasi 66,95%), sementara KNN memiliki performa terendah (akurasi 45,01%). Penelitian ini menyimpulkan bahwa SVM merupakan metode yang paling efektif untuk analisis sentimen ulasan M-Pajak[16].

Penelitian yang dilakukan oleh Eskiyaturrofikoh & Suryono terhadap aplikasi X pada tahun 2024 bertujuan untuk menganalisis sentimen pengguna terhadap aplikasi X. Metode klasifikasi yang digunakan adalah *Naïve Bayes* dan *Support Vector Machine* (SVM). Dataset yang digunakan terdiri dari 4.087 ulasan yang didapatkan melalui teknik *web scraping*, dengan 80% data untuk pelatihan dan 20% untuk pengujian. Temuan awal menunjukkan akurasi SVM sebesar 75,5% dan *Naïve Bayes* 75% sebelum SMOTE diterapkan. Setelah penggunaan SMOTE, akurasi SVM melonjak menjadi 81%, yang menunjukkan efektivitas SMOTE dalam menangani ketidakseimbangan kelas. Di sisi lain, akurasi *Naïve Bayes* tetap di 75,5%, mengindikasikan bahwa SMOTE tidak memberikan dampak signifikan pada model ini[17].

Tabel 2.1. Studi Sebelumnya

No	Penulis	Tahun	Metode	Hasil
1	Ariannor et al	2024	<i>Naïve Bayes</i>	Penelitian ini menemukan 59,94% ulasan bersifat negatif, dengan keluhan seperti error, kesulitan cek pajak, dan kinerja aplikasi yang buruk. Sementara itu, hanya terdapat 40,06% ulasan positif. Hasil ini menunjukkan perlunya perbaikan signifikan pada aplikasi untuk meningkatkan kepuasan pengguna.
2	Ikhsan et al	2025	<i>Naïve Bayes</i> dan <i>Support Vector Machine</i>	Hasil penelitian menunjukkan bahwa algoritma SVM memiliki akurasi 98% sebelum <i>balancing</i> dan 97% setelah <i>balancing</i> . Sementara itu, <i>Naïve Bayes</i> mencatat akurasi 97% sebelum <i>balancing</i> dan 90% setelah <i>balancing</i> . Secara keseluruhan, kedua algoritma menunjukkan performa yang baik dan seimbang dalam menganalisis sentimen publik.
3	Ardiyansah & Parjito	2024	<i>Naïve Bayes</i> dan <i>Support Vector Machine</i>	Penelitian ini menunjukkan bahwa penggunaan SMOTE (<i>Synthetic Minority Over-sampling Technique</i>) dapat menyeimbangkan data, sehingga

				didapatkan hasil <i>Naïve Bayes</i> mencapai akurasi 95%, sementara SVM menunjukkan performa lebih baik dengan akurasi 99%. Kesimpulannya SVM terbukti lebih akurat dibanding <i>Naïve Bayes</i>.
4	Faisol	2024	<i>Support Vector Machine</i> (SVM), <i>Naïve Bayes</i>, dan <i>K-Nearest Neighbors</i> (KNN)	Penelitian menunjukkan bahwa SVM unggul dengan akurasi 86,41%, <i>precision</i> 85-86%, dan <i>recall</i> tinggi untuk sentimen negatif (98,12%). <i>Naïve Bayes</i> menempati posisi kedua (akurasi 66,95%), sementara KNN memiliki performa terendah (akurasi 45,01%). Penelitian ini menyimpulkan bahwa SVM merupakan metode yang paling efektif untuk analisis sentimen ulasan M-Pajak.
5	Eskiyaturrofikoh & Suryono	2024	<i>Naïve Bayes</i> dan <i>Support Vector Machine</i> (SVM)	Hasil penelitian ini menemukan setelah SMOTE diterapkan, akurasi SVM melonjak ke 75% menjadi 81%. Sedangkan <i>Naïve Bayes</i> tidak menunjukkan peningkatan signifikan, tetap di 75,5%. menyiratkan bahwa SMOTE lebih menguntungkan bagi SVM .

2.2. Dasar Teori

2.2.1 Analisis Sentimen

Analisis sentimen merupakan metode yang melibatkan teknik komputasi untuk mengkategorikan data teks berdasarkan emosi yang terkandung di dalamnya. Ketersediaan data teks yang besar dan pendapat seseorang terhadap suatu topik menjadi pendorong meningkatnya popularitas analisis sentimen. Analisis sentimen sangat bergantung pada dataset yang digunakan. Dataset dengan kalimat panjang memerlukan metode analisis yang berbeda dibandingkan dengan dataset kalimat pendek[18]. Analisis sentimen bertujuan untuk mengidentifikasi emosi dari teks yang dibagikan secara online, misalnya pada platform X[19].

Bing Liu mendefinisikan analisis sentimen sebagai bidang ilmu yang mempelajari opini, emosi, evaluasi, penilaian, sikap, dan perasaan terhadap kualitas suatu hal, mulai dari produk hingga isu atau atributnya. Lebih lanjut, Bing Liu mengklasifikasikan analisis sentimen sebagai masalah *Supervised Learning*. Proses analisis sentimen dapat disederhanakan menjadi tiga langkah: pengklasifikasian data, evaluasi sentimen, dan visualisasi hasilnya[20]. Terdapat beberapa tipe analisis sentimen yang berbeda, di antaranya adalah[21]:

1. *Fine grained sentiment* adalah jenis analisis yang memberikan penilaian sentimen dengan skala yang lebih spesifik, mulai dari sangat positif hingga sangat negatif. Penilaian bintang pada *e-commerce* menunjukkan tingkat kepuasan pelanggan adalah contoh dari analisis.
2. *Emotion detection* adalah jenis analisis yang tidak hanya menilai positif dan negatif tetapi juga berfokus pada identifikasi emosi pelanggan.
3. *Intent-based analysis* adalah jenis analisis yang berfokus pada penggalian maksud sebenarnya dari ekspresi pelanggan dalam bentuk tulisan.
4. *Aspect-based sentiment analysis* adalah jenis analisis yang digunakan saat mengidentifikasi komponen-komponen tertentu yang menjadi sumber sentimen positif maupun negatif dalam sebuah opini.

Pentingnya analisis sentimen pada data media sosial didasari oleh kecenderungan manusia untuk semakin terbuka dalam mengekspresikan pikiran dan perasaan mereka secara daring. Analisis sentimen mampu mengategorikan kalimat opini menjadi positif, negatif, dan netral. Hasil klasifikasi ini memberikan informasi yang berharga bagi berbagai entitas, baik perusahaan maupun instansi[22].

2.2.2 Media Sosial

Media sosial merupakan platform daring yang memfasilitasi pengguna untuk berinteraksi, menyebarkan informasi, dan membuat konten seperti blog, media pertemanan, wiki, forum diskusi, serta lingkungan virtual[23]. Media sosial memiliki kemampuan untuk meniadakan batasan waktu, lokasi, dan dimensi, sehingga memungkinkan manusia mempercepat penyebaran informasi. Media sosial telah terintegrasi dalam kehidupan sehari-hari manusia karena kemampuannya memenuhi kebutuhan manusia sebagai makhluk sosial dan keinginan untuk mengekspresikan diri secara berbeda[24]. Berdasarkan *We Are Social*, jumlah pengguna aktif media sosial di Indonesia per Januari 2024 mencapai 139 juta (49,9% dari total populasi)[5]. Media sosial dapat dikategorikan menjadi tiga jenis yaitu[25]:

1. Aplikasi pesan singkat, yang banyak digunakan antara lain *Whatsapp*, *Facebook Messenger*, *Line*, dan *Telegram*.
2. Jejaring Sosial, *Facebook* dan *Twitter* merupakan contoh yang paling populer di kalangan pengguna.
3. Platform berbagi, contoh yang populer adalah *YouTube* dan *Instagram*.

2.2.3 X

X adalah platform media sosial yang berfungsi sebagai sarana informasi, menghubungkan pengguna dengan berbagai informasi yang relevan dengan minat mereka. X hadir setelah popularitas *Facebook*, menawarkan format baru berupa *microblogging* dengan batasan karakter pada setiap unggahan (*tweet*). Awalnya dibatasi 140 karakter, kemudian ditingkatkan menjadi 280

karakter karena dianggap kurang memadai. Fitur ini memudahkan pengguna untuk saling berbagi informasi[26]. Berikut ini beberapa fitur yang ada pada X[27]:

1. *Tweet* adalah pesan singkat berisi informasi beragam topik yang memungkinkan terjadinya interaksi antara pengguna dengan pengikut maupun non-pengikutnya.
2. *Trending* adalah fitur yang menampilkan kata, frasa, atau tagar yang sedang menjadi topik perbincangan populer di kalangan pengguna.
3. *Retweet* adalah fitur untuk membagikan postingan pengguna lain.

3.2.4 Pajak

Berdasarkan Undang-Undang No 28 Tahun 2007 Perubahan Ketiga atas Undang-Undang Nomor 6 Tahun 1983 tentang Ketentuan Umum dan Tata Cara Perpajakan, mendefinisikan pajak sebagai kontribusi wajib kepada negara yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-Undang, dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat. Menurut Prof. DR. Rachmat Sumitro, pajak adalah iuran rakyat kepada kas negara, yang merupakan transfer kekayaan dari masyarakat ke pemerintah. Kontribusi ini didasarkan pada undang-undang dan bertujuan untuk membiayai pengeluaran rutin negara. Kelebihan dana yang terkumpul digunakan sebagai tabungan publik, yang menjadi sumber utama pendanaan investasi publik[28].

Membayar pajak adalah bentuk nyata dari tanggung jawab sebagai warga negara dan partisipasi aktif setiap Wajib Pajak dalam memenuhi kewajiban perpajakannya. Tindakan ini secara langsung dan kolektif berkontribusi pada pendanaan negara dan pelaksanaan pembangunan nasional[29]. Pajak memegang peranan penting dalam kehidupan bernegara karena menjadi sumber utama pendanaan pembangunan di Indonesia. Oleh karena itu, pajak memiliki beberapa fungsi sebagai berikut[30]:

1. Fungsi Anggaran (*Budgetair*), yaitu sebagai sumber pendapatan negara, pajak berfungsi untuk membiayai pengeluaran-pengeluaran negara. Untuk menjalankan tugas-tugas rutin negara dan melaksanakan pembangunan, negara membutuhkan biaya. Biaya ini dapat diperoleh dari penerimaan pajak.
2. Fungsi Mengatur (*Regulerend*), yaitu pajak dapat dimanfaatkan sebagai instrumen untuk mencapai tujuan tertentu. Misalnya, untuk menarik investasi, baik domestik maupun asing, pemerintah memberikan berbagai insentif berupa pengurangan pajak.
3. Fungsi Stabilitas, yaitu pemerintah memiliki sumber daya finansial untuk mengimplementasikan kebijakan yang bertujuan menjaga stabilitas harga dan mengendalikan inflasi. Upaya ini dapat dilakukan dengan mengatur jumlah uang yang beredar di masyarakat, memungut pajak secara efektif, serta menggunakan dana pajak secara efisien.
4. Redistribusi Pendapatan, yaitu dana pajak yang terkumpul oleh negara dialokasikan untuk membiayai berbagai kepentingan publik, termasuk pembangunan infrastruktur dan sektor lainnya.

2.2.5 Coretax

Coretax, atau Sistem Administrasi Coretax, adalah sebuah sistem yang memanfaatkan teknologi dengan tujuan memodernisasi pengelolaan administrasi perpajakan melalui implementasi teknologi informasi yang lebih mutakhir. Sistem Coretax diharapkan dapat memadukan berbagai aspek administrasi perpajakan ke dalam satu platform yang lebih efisien dan terpadu[9]. Lebih lanjut, Coretax dirancang untuk mengidentifikasi wajib pajak yang lalai dalam pembayaran, sehingga Direktorat Jenderal Pajak (DJP) dapat memberikan sanksi denda atau melakukan sosialisasi untuk meningkatkan kesadaran dan kepatuhan pajak. Menteri Keuangan Sri Mulyani menyatakan bahwa Indonesia akan mengimplementasikan Coretax sebagai teknologi perpajakan termutakhir yang pernah ada di Indonesia[31].

Dalam rangka modernisasi administrasi perpajakan Indonesia, DJP mengambil langkah yang didasari oleh Peraturan Menteri Keuangan Nomor 81 Tahun 2024 tentang Ketentuan Perpajakan dalam Rangka Pelaksanaan Sistem Inti Administrasi Perpajakan. Peraturan ini mengharuskan integrasi semua aplikasi pajak dalam Coretax, sehingga menciptakan sistem yang terpusat dan kohesif untuk mengelola kewajiban perpajakan dan memastikan pengawasan yang konsisten. Implementasi Coretax tidak hanya mempermudah proses administrasi bagi wajib pajak, tetapi juga memperkuat otoritas pajak melalui pengumpulan dan analisis data secara *real-time*[32] .

2.2.6 *Naïve Bayes Classifier* (NBC)

Naïve Bayes Classifier (NBC) merupakan metode klasifikasi statistik yang memprediksi probabilitas suatu entitas menjadi anggota kelas tertentu. *Naïve Bayes* berakar pada teorema Bayes, mengasumsikan independensi antar atribut. Keunggulan utama *Naïve Bayes* terletak pada kesederhanaannya yang dikombinasikan dengan akurasi klasifikasi yang tinggi, klasifikasi data dilaksanakan dalam dua fase. Fase awal mencakup pembangunan model melalui proses pelatihan (*model training*), sedangkan fase berikutnya berfokus pada penentuan kategori untuk data yang belum terklasifikasi. Salah satu keunggulan *Naïve Bayes* ada pada persyaratan data pelatihan yang minimal untuk memperoleh estimasi parameter yang memadai guna keperluan klasifikasi[33].

Dikemukakan oleh Thomas Bayes, seorang ilmuwan asal Inggris, *Naïve Bayes* adalah algoritma kategorisasi yang memanfaatkan prinsip probabilitas dan statistik untuk memprediksi kemungkinan peristiwa mendatang berdasarkan data historis. Teorema tersebut dikombinasikan dengan asumsi Naïve, yang menegaskan independensi antar atribut dalam model. Dalam implementasinya, algoritma *Naïve Bayes* mengasumsikan independensi antar atribut, artinya setiap atribut dianggap tidak memiliki hubungan atau pengaruh terhadap atribut lainnya, terlepas dari potensi korelasinya. Salah

satu keterbatasan *Naïve Bayes* adalah asumsinya bahwa semua fitur tidak saling terkait, yang bisa mengurangi keakuratannya pada data kompleks[34].

Terdapat tiga jenis model *Naïve Bayes*, yaitu *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, *Multinomial Naïve Bayes*[35]:

1. *Gaussian Naïve Bayes* merupakan varian *Naïve Bayes Classifier* yang mengintegrasikan distribusi normal dalam perhitungannya. Algoritma ini memungkinkan klasifikasi data numerik yang mengikuti distribusi *Gaussian*, serta dapat menangani data kategorikal.
2. *Bernoulli Naïve Bayes* adalah varian dari *Naïve Bayes* yang dirancang untuk data di mana fitur-fiturnya bersifat biner, seperti representasi kehadiran atau ketiadaan kata dalam dokumen. Meskipun *primarily* digunakan untuk kasus tersebut, Bernoulli juga cocok untuk data kategorikal asalkan atribut-atributnya dapat direpresentasikan sebagai variabel biner.
3. *Multinomial Naïve Bayes* adalah algoritma klasifikasi teks yang berfungsi memprediksi kategori dokumen berdasarkan frekuensi kemunculan kata-kata di dalamnya.

2.2.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan algoritma *supervised learning* yang efektif untuk klasifikasi dan regresi. Boser, Guyon, dan Vapnik adalah pihak yang memperkenalkan *Support Vector Machine*, sebuah metode machine learning pada *Annual Workshop on Computational Learning Theory* tahun 1992[36]. Prinsip kerja SVM adalah mencari garis pemisah paling optimal untuk memisahkan kategori data dalam ruang multidimensi. Meskipun awalnya dirancang untuk klasifikasi biner, SVM kini mampu menangani kasus multi-kelas serta masalah non-linear dengan memproyeksikan data ke ruang yang lebih besar menggunakan trik kernel[15]. Kernel diartikan sebagai fungsi yang memproses fitur data, mengubahnya dari representasi dasar menjadi fitur baru dengan dimensi yang diperbesar[37].

Support Vector Machine (SVM) unggul dalam model klasifikasi dengan konsep matematis yang kokoh, sehingga menghasilkan akurasi yang baik. Algoritma ini bekerja dengan mengidentifikasi *hyperplane* optimal, yaitu fungsi yang bertugas memisahkan kelas-kelas, dengan memaksimalkan jarak di antara mereka[38]. *Support vector* adalah titik data yang paling dekat dengan *hyperplane*. *Margin* kemudian diartikan sebagai dua kali jarak *support vector* ke *hyperplane*. Sebuah *hyperplane* dikatakan baik jika *margin* yang dihasilkannya adalah yang terbesar. SVM adalah metode yang efektif untuk memecahkan masalah dengan data berdimensi tinggi dan jumlah sampel pelatihan yang kecil. Metode ini didasarkan pada prinsip *Structural Risk Minimization* (SRM), yang memaksimalkan *margin* dan meminimalkan batas atas risiko[39].

2.2.8 Confusion Matrix

Confusion Matrix adalah tabel yang merangkum seberapa akurat sebuah model mengklasifikasikan data uji dengan menunjukkan jumlah hasil prediksi yang benar dan salah. Hasil ini adalah *True Positive* (TP) model benar memprediksi kelas positif, *True Negative* (TNg) model benar memprediksi kelas negatif, *True Neutral* (TNT) model benar memprediksi kelas netral, *False Positive* (FP) model salah memprediksi kelas positif, dan *False Negative* (FNg) model salah memprediksi kelas negatif, *False Neutral* (FNt) model salah memprediksi kelas netral yang masing-masing merepresentasikan berbagai jenis akurasi dan kesalahan klasifikasi. Berikut ini merupakan contoh Tabel 2.2 *confusion matrix*[40].

Tabel 2. 2. Confusion Matrix

Kelas Aktual	Kelas Prediksi		
	<i>Negative</i>	<i>Neutral</i>	<i>Positive</i>
<i>Negative</i>	<i>True Negative</i>	<i>False Negative</i>	<i>False Negative</i>
<i>Neutral</i>	<i>False Neutral</i>	<i>True Neutral</i>	<i>False Neutral</i>
<i>Positive</i>	<i>False Positive</i>	<i>False Positive</i>	<i>True Positive</i>