BAB II

LANDASAN TEORI

2.1. Penelitian Terdahulu

Tabel 2.1 Penelitian Terkait

No	Author	Title	Journal	Method/Tools	Application/Result
1.	Xu, F., et al. (2023)	Comparative Analysis of SVM and Random Forest for Text Classification	*IEEE Access*	SVM, Random Forest, TF-IDF	Membandingkan efektivitas *Support Vector Machine* (SVM) dengan *Random Forest* dalam pengolahan data tekstual. Hasil menunjukkan bahwa Random Forest memiliki akurasi yang lebih tinggi dan lebih stabil terhadap *noise* dibandingkan SVM, meskipun waktu pelatihannya sedikit lebih lama.
2.	Al Amrani, Y., et al. (2022)	Random Forest and Support Vector Machine based Hybrid Approach to	*Procedia Computer Science*	SVM, Random Forest, Grid Search	Penelitian ini membuktikan bahwa penggunaan Grid Search untuk

		Sentiment Analysis			optimasi *hyperparameter* pada Random Forest dapat meningkatkan akurasi secara signifikan (dari 86% menjadi 91%) dibandingkan dengan parameter *default*.
3.	Singh, J., & Raghuvanshi, N. (2024)	Sentiment Analysis of Fintech Mobile App Reviews using Machine Learning	*International Journal of Information Management Data Insights*	Random Forest, Naive Bayes	Studi kasus pada aplikasi Fintech. Menemukan bahwa Random Forest memberikan *F1-Score* terbaik dalam mendeteksi sentimen negatif terkait masalah teknis dan penagihan, yang sangat relevan dengan kasus Adakami.
4.	Zhang, L. (2023)	Optimizing Hyperparameters in Text Mining Classifiers: A	*Journal of Big Data*	SVM, Random Forest, Grid Search)	Studi komprehensif yang menunjukkan bahwa tahap optimasi (tuning) adalah langkah krusial. SVM

		Comparative Study			sangat sensitif terhadap parameter *C* dan *Gamma*, sedangkan Random Forest sensitif terhadap *n_estimators*
5.	Pradha, S., et al. (2023)	Effective Sentiment Analysis of Social Media Data using Feature Selection and Ensemble Learning	*Applied Sciences*	Random Forest, TF-IDF	Studi ini membuktikan bahwa pendekatan Ensemble lebih efektif menangani ketidakseimbangan data daripada penggunaan pengklasifikasi tunggal dalam menangani data ulasan yang tidak seimbang (*imbalanced data*) dan memiliki dimensi fitur yang tinggi.
	Ramadani (2025)	Optimization of Random Forest Hyperparameters with Grid Search for Public Sentiment Classification	*IEEE International Conference*	Random Forest, Grid Search	Membuktikan bahwa penerapan Grid Search berhasil meningkatkan akurasi Random Forest secara signifikan dari 82.59% (default) menjadi 88.71%

					(optimized). Ini adalah bukti telak bahwa optimasi itu wajib.
	Rochmawati et al. (2025)	Comparison of Support Vector Machine (SVM) and Random Forest Algorithms in the Analysis of Social Media Sentiment	*IEEE International Conference*	SVM, Random Forest	Menemukan bahwa SVM (92%) sedikit mengungguli Random Forest (91%) pada data teks politik. Studi ini menegaskan bahwa SVM lebih superior dalam memisahkan hyperplane pada data yang bersih, namun RF tetap kompetitif.
	Mariska et al. (2025)	Comparison of SVM and Random Forest Algorithms in Sentiment Analysis of JMO Mobile Application	*IEEE International Conference*	SVM, Random Forest	Menunjukkan bahwa Random Forest (86.15%) sedikit lebih unggul daripada SVM (86.06%) karena kemampuannya menangani data ulasan aplikasi yang memiliki banyak variasi kata (high variability).

Dalam ranah klasifikasi teks, perdebatan mengenai efektivitas antara pendekatan linear dan ensemble terus menjadi topik riset yang krusial. Studi

komparatif oleh Xu et al. (2023) secara spesifik menyoroti bahwa meskipun Support Vector Machine (SVM) dikenal tangguh dan efisien dalam ruang berdimensi tinggi, algoritma ini cenderung sensitif terhadap outlier. Sebaliknya, Random Forest terbukti memiliki stabilitas yang lebih superior dan resistensi yang lebih baik terhadap gangguan data (noise) karena mekanisme voting dari banyak pohon keputusan, sebagaimana dikonfirmasi oleh Pradha et al. (2023) dalam penanganan data tidak seimbang (imbalanced data). Meskipun demikian, penelitian-penelitian tersebut masih memiliki keterbatasan, yaitu studi sebelumnya yang hanya membandingkan kedua model ini menggunakan parameter default. Padahal, Al Amrani et al. (2022) membuktikan bahwa akurasi model dapat melonjak signifikan (hingga 5-10%) jika dilakukan penyetelan parameter yang tepat. Oleh karena itu, penelitian ini tidak hanya sekadar membandingkan SVM dan Random Forest, melainkan menerapkan pendekatan Optimizing guna memastikan bahwa perbandingan dilakukan pada performa puncak (peak performance) masing-masing model, sehingga menghasilkan kesimpulan yang lebih valid dan objektif.

2.2. Teori yang berkaitan

2.2.1. Analisis Sentimen

Sebagai bagian integral dari NLP, analisis sentimen berfungsi untuk menambang opini subjektif yang terkandung dalam data teks. Proses ini bermuara pada klasifikasi polaritas pandangan, yakni memilah antara respon positif, negatif, maupun netral terhadap produk atau layanan tertentu . Bagi pelaku usaha, teknik ini menjadi instrumen vital dalam mendengarkan aspirasi pelanggan serta mengevaluasi reputasi perusahaan secara objektif guna perbaikan berkelanjutan.

2.2.2. Financial Technology Lending

Istilah *Financial Technology* (Fintech) merepresentasikan integrasi teknologi mutakhir ke dalam sektor keuangan yang bertujuan untuk mentransformasi atau memodernisasi layanan konvensional. Sejalan dengan

hal tersebut, Otoritas Jasa Keuangan (OJK) mendefinisikan fintech sebagai terobosan inovatif di industri finansial yang melahirkan berbagai model bisnis dan produk baru, dengan orientasi utama pada penciptaan efisiensi transaksi serta pemeliharaan stabilitas sistem keuangan.

Dalam ekosistem *fintech* nasional, layanan *Peer-to-Peer* (P2P) *Lending* mencatatkan pertumbuhan yang sangat progresif. Model bisnis ini berfungsi sebagai jembatan digital yang memfasilitasi interaksi langsung antara pemilik dana dan pencari pinjaman secara daring. Keunggulan utamanya terletak pada kemampuan mendisrupsi sistem perbankan konvensional melalui prosedur yang ringkas dan persyaratan fleksibel, sehingga mampu membuka akses keuangan bagi kelompok masyarakat yang selama ini belum terjangkau layanan perbankan (*unbanked*).

2.2.3. Adakami

Adakami merupakan salah satu platform layanan pinjam meminjam uang berbasis teknologi informasi (*fintech peer-to-peer lending*) yang beroperasi di Indonesia. Melalui pendekatannya yang mengutamakan kecepatan layanan dan kemudahan aksesibilitas syarat bagi pengguna., Adakami menyasar segmen masyarakat yang sebelumnya tidak terlayani oleh perbankan (*unbanked* dan *underbanked*).

Pemilihan platform ini sebagai fokus penelitian didasarkan pada posisinya yang strategis sebagai salah satu entitas dominan dalam ekosistem industri terkait. sehingga ulasan digital yang ditinggalkan oleh para nasabahnya menjadi sumber data yang kaya untuk dianalisis. Penelitian ini dirancang secara spesifik untuk memetakan sentimen pada berbagai aspek layanan dari Adakami guna menyajikan analisis holistik mengenai pengalaman nasabahnya. Analisis ini krusial untuk memahami persepsi nasabah terhadap area-area sensitif seperti proses penagihan dan transparansi biaya, yang telah diidentifikasi oleh penelitian sebelumnya sebagai sumber utama keluhan publik di industri pinjaman *online*[13].

2.2.4. Google Play Reviews

Kolom ulasan pada Google Play Store berfungsi sebagai wadah aspirasi digital di mana pengguna dapat menyalurkan penilaian subjektif mereka terhadap kinerja aplikasi yang telah dioperasikan. Ulasan ini merupakan sumber data yang sangat berharga karena berisi opini langsung, kritik, dan saran dari pengalaman nyata pengguna. Menganalisis data ini memungkinkan pengembang untuk memahami tingkat kepuasan, mengidentifikasi bug atau masalah, serta menemukan area yang perlu ditingkatkan untuk memperkaya pengalaman pengguna secara keseluruhan[3].

2.3. Framework/Algoritma yang digunakan

2.3.1. Knowledge Discovery in Databases (KDD)

Konsep KDD merujuk pada pendekatan terstruktur yang dirancang khusus untuk menambang informasi berharga yang tersimpan di balik kompleksitas basis data yang besar. Tujuan utamanya bukan sekadar mengolah data, melainkan mengubah data mentah yang seringkali tidak berarti menjadi wawasan (*insight*) dan pengetahuan (*knowledge*) yang dapat ditindaklanjuti[16]. Proses ini memandang penemuan pengetahuan sebagai sebuah siklus yang terstruktur, memastikan bahwa pola yang ditemukan tidak hanya akurat secara statistik, tetapi juga valid, baru, bermanfaat, dan mudah dipahami oleh manusia.

Proses KDD dimulai dengan pemahaman mendalam bahwa data di dunia nyata jarang sekali dalam kondisi sempurna. Oleh karena itu, sebagian besar upaya dalam KDD difokuskan pada persiapan data secara cermat. Ini melibatkan serangkaian tugas penting seperti membersihkan data dari *noise* atau inkonsistensi, menangani data yang hilang, dan menyeragamkan format. Urgensi tahap persiapan data tidak dapat dikesampingkan karena kualitas luaran pengetahuan merupakan refleksi

langsung dari kualitas masukan data, sesuai dengan kaidah umum pemrosesan data yakni *Garbage In, Garbage Out*17].

Ketika data telah melalui tahap penyiapan, langkah selanjutnya adalah *data mining* yang menjadi motor penggerak KDD. Fase ini melibatkan aplikasi algoritma cerdas guna mendeteksi struktur pola atau model secara otomatis. Algoritma ini dapat melakukan berbagai tugas, seperti klasifikasi, regresi, pengelompokan (*clustering*), atau menemukan aturan asosiasi. Dalam konteks analisis sentimen[18], misalnya, *data mining* digunakan untuk melatih model yang mampu mengenali pola kata dan frasa yang mencirikan opini positif atau negatif, sebuah tugas yang mustahil dilakukan pada ribuan ulasan.

Namun, penemuan pola saja tidak cukup. KDD menekankan pentingnya evaluasi dan interpretasi guna memverifikasi bahwa pola-pola yang teridentifikasi memiliki validitas tinggi dan nilai signifikansi yang nyata. Pada fase ini, hasil dari algoritma *data mining* diuji validitasnya dan diukur kinerjanya menggunakan metrik-metrik statistik. Lebih dari itu, hasil tersebut diterjemahkan ke dalam konteks domain yang sedang diteliti. Tujuannya adalah untuk mengubah output teknis dari model menjadi narasi atau kesimpulan yang dapat dipahami dan digunakan untuk pengambilan keputusan.

Pada akhirnya, KDD berfungsi sebagai jembatan yang menghubungkan antara data teknis dan pengetahuan strategis. Kerangka kerja ini memastikan bahwa analisis data tidak berhenti pada laporan statistik, melainkan menghasilkan pemahaman mendalam yang dapat menjawab pertanyaan-pertanyaan penting dalam suatu masalah. Dengan pendekatannya yang sistematis dari pembersihan data hingga interpretasi hasil, KDD memberikan sebuah prosedur yang efektif untuk mendayagunakan potensi maksimal dari aset informasi yang dikelola.

2.3.2. *Support Vector Machine (SVM)*

Dalam kategori pembelajaran mesin terawasi (supervised learning), algoritma Support Vector Machine (SVM) sering menjadi pilihan utama peneliti karena keandalannya, terutama untuk tugas klasifikasi. Dikembangkan oleh Vladimir Vapnik dan rekan-rekannya, SVM bekerja berdasarkan prinsip statistika dan optimisasi untuk menemukan pemisah (klasifikator) terbaik antara dua atau lebih kelompok data. Karena kinerjanya yang terbukti kuat, terutama dalam ruang berdimensi tinggi, SVM menjadi salah satu metode pilihan utama untuk klasifikasi teks, termasuk analisis sentimen[19].

Prinsip fundamental SVM terletak pada konstruksi *hyperplane* terbaik yang berfungsi sebagai batas pemisah antar kelas data. Jika diproyeksikan ke dalam ruang dua dimensi, pembatas ini bermanifestasi sebagai garis lurus. Dalam ruang tiga dimensi, representasi pemisah ini berkembang menjadi bidang datar atau sub-ruang seiring bertambahnya dimensi data. Fokus utama SVM bukanlah menemukan pemisah rata-rata, melainkan menentukan satu-satunya garis pemisah yang menawarkan margin pemisahan maksimal.

Hyperplane yang dianggap "terbaik" adalah yang memiliki **margin** atau jarak terbesar antara dirinya dengan titik data terdekat dari setiap kelas. Titik-titik data terdekat yang "menopang" atau menentukan posisi *hyperplane* inilah yang disebut sebagai **support vectors**. Dengan memaksimalkan margin ini, SVM bertujuan untuk menciptakan model yang memiliki kemampuan generalisasi yang baik, Ini bermakna bahwa model mampu menghasilkan prediksi yang valid secara konsisten, baik terhadap data latih maupun terhadap data baru yang belum pernah diproses oleh sistem.

2.3.3. Random Forest

Random Forest dikenal sebagai algoritma yang mampu menghasilkan prediksi lebih stabil dengan cara menggabungkan *output* dari berbagai *decision trees*. Dalam fase pelatihan, algoritma ini membangun struktur pohon yang majemuk, kemudian menetapkan hasil akhir berdasarkan modus dari kelas yang dihasilkan oleh masing-masing pohon individu . Pendekatan ini terbukti efektif untuk memproses data dalam volume besar sekaligus meminimalisir risiko *overfitting* yang umum terjadi pada penggunaan *decision tree* konvensional [21].

2.3.4. Confusion Matrix

Guna memvalidasi performa model SVM yang telah terbentuk, diperlukan metode pengukuran yang akurat. *Confusion Matrix* diadopsi sebagai alat evaluasi utama, di mana hasil prediksi model terhadap data uji diringkas dalam bentuk matriks yang mempertemukan label prediksi dengan label yang sebenarnya.

Tabel ini memberikan visualisasi yang jelas mengenai di mana model berhasil melakukan prediksi dengan benar dan di mana ia melakukan kesalahan[22]. Dalam konteks penelitian analisis sentimen tiga kelas (Positif, Negatif, dan Netral), *Confusion Matrix* akan berbentuk tabel 3x3.

Untuk menyederhanakan penjelasan, kita akan menggunakan konteks biner (misalnya, hanya kelas Positif dan Negatif). Komponen utama dari *Confusion Matrix* adalah sebagai berikut:

1. **TP (True Positive):** Kondisi di mana prediksi model selaras dengan data aktual, yaitu data positif diklasifikasikan sebagai positif .
2. **TN (True Negative):** Keadaan di mana model memprediksi data negatif dengan benar sesuai dengan label aslinya .

3. **FP (False Positive):** Kesalahan klasifikasi di mana data negatif dianggap positif oleh model, fenomena ini juga disebut sebagai *Type I Error* .
4. **FN (False Negative):** Kegagalan model dalam mengenali data positif sehingga salah melabelinya sebagai negatif, atau disebut *Type II Error* .

Berpijak pada empat elemen fundamental *Confusion Matrix*, dilakukan perhitungan terhadap sejumlah metrik evaluasi guna mendapatkan gambaran numerik mengenai reliabilitas model. Indikator kinerja yang menjadi fokus dalam studi ini terdiri dari:

1. **Accuracy** Tingkat keberhasilan model dalam memprediksi data secara tepat digambarkan oleh metrik Akurasi. Nilai ini diperoleh dengan menjumlahkan seluruh prediksi benar (TP + TN) lalu membaginya dengan total populasi data dalam himpunan pengujian.

   Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$ x 100%

2. **Precision** Tingkat ketepatan model dalam mengklasifikasikan label positif dinilai melalui metrik Presisi. Indikator ini menjawab pertanyaan krusial mengenai persentase kebenaran dari total prediksi positif yang dihasilkan, yang sangat menentukan kredibilitas model dalam menghindari kesalahan deteksi positif.

   Precision = $\frac{TP}{TP + FP}$ x 100%

3. **Recall** Kapabilitas model dalam mendeteksi kembali seluruh data positif yang eksis diukur melalui metrik *Recall*. Indikator ini menghitung proporsi data positif yang sukses dikenali dari total data aktual, sehingga sangat esensial untuk meminimalisir risiko terlewatnya informasi positif.

   Recall = $\frac{TP}{TP + FN}$ x 100%

4. **F1-Score** Sebagai representasi rata-rata harmonik antara Presisi dan *Recall*, *F1-Score* berfungsi sebagai metrik tunggal yang merefleksikan keseimbangan kedua aspek tersebut. Indikator ini menjadi sangat vital dalam skenario pengolahan *dataset* dengan distribusi kelas yang tidak seimbang (*imbalanced*), di mana nilai *F1-Score* yang tinggi menandakan bahwa model mampu menjaga harmoni antara ketepatan prediksi dan jangkauan deteksi.

   $$\text{F1-Score} = 2 \text{ x } \frac{Precision \; x \; Recall}{Precision + Recall}$$

Dalam penelitian ini, *Confusion Matrix* akan dihasilkan dari pengujian model SVM terhadap data ulasan Adakami yang telah dilabeli. Hasil dari metrik-metrik di atas akan dianalisis untuk menarik kesimpulan tentang performa model. Sebagai contoh, *Precision* yang tinggi untuk kelas Negatif menandakan bahwa ketika model memprediksi sebuah ulasan sebagai negatif, prediksi tersebut kemungkinan besar benar. Sementara itu, *Recall* yang tinggi untuk kelas Negatif berarti model mampu mengidentifikasi sebagian besar ulasan yang memang benar-benar negatif.

Evaluasi ini akan menjadi dasar untuk menyatakan kelayakan model klasifikasi sentimen yang dibangun sebelum digunakan untuk menganalisis keseluruhan dataset ulasan.

2.4. Tools/software yang digunakan

2.4.1. Python

Salah satu bahasa pemrograman yang sangat dianggap unggul karena berbasis objek, bersifat open source, serta memiliki beragam aplikasi yang meliputi pembuatan situs web dan analisis data. Bahasa pemrograman ini dikenal dengan keunggulannya dalam fleksibilitas, memudahkan pengembangan perangkat lunak, serta kemampuannya dalam otomatisasi tugas-tugas kompleks.

Python memainkan peran penting dalam melakukan sentimen analisis pada opini publik terkait layanan pinjaman online Adakami. Berikut beberapa keunggulan dan cara Python digunakan:

Keunggulan Python:

a. Mudah dipelajari dan digunakan: Python memiliki sintaksis yang sederhana dan banyak dokumentasi yang tersedia, sehingga mudah dipelajari dan digunakan oleh siapa saja, termasuk pemula.
b. Beragam library dan framework: Python memiliki banyak library dan framework yang telah dikembangkan khusus untuk analisis data dan sentimen, seperti NLTK, TextBlob, SpaCy, dan VADER.
c. Open source dan gratis: Python adalah bahasa open source dan gratis, sehingga siapa saja dapat menggunakannya tanpa harus membeli lisensi.
d. Multiplatform: Python dapat berjalan diberbagai platform, termasuk Windows, Linux, dan macOS.

2.4.2. Library Pengolahan Data dan Teks

Beberapa *library* Python akan menjadi komponen inti dalam alur kerja penelitian:

1. **Pandas:** *Library* Untuk pengelolaan struktur data ulasan Adakami, penelitian ini memanfaatkan pustaka Pandas, yang memungkinkan manipulasi format DataFrame secara efisien sebelum masuk ke tahap analisis.
2. **NLTK (Natural Language Toolkit):** Merupakan *library* fundamental untuk tugas-tugas NLP. Secara teknis, penelitian ini mendayagunakan modul NLTK untuk melakukan segmentasi teks menjadi token (*tokenizing*) dan menyaring kosa kata yang memiliki frekuensi tinggi namun minim makna (*stopword removal*).
3. **Sastrawi:** *Library* ini khusus dirancang untuk NLP Bahasa Indonesia. Operasi inti yang dijalankan meliputi *stemming*, sebuah teknik pemetaan

kata berimbuhan menuju kata dasar (misal: transformasi 'penagihan' ke 'tagih'), yang menjadi prasyarat krusial bagi efektivitas analisis berbasis leksikon.

4. **Scikit-learn (Opsional):** Jika diperlukan untuk perbandingan atau validasi model, *library* ini menyediakan implementasi berbagai algoritma *machine learning* yang siap pakai.