BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Pada sub bab 2.1, disajikan penelitian-penelitian terdahulu yang menjadi acuan utama dalam penelitian ini. Tabel 2.1, memuat 10 penelitian sebelumnya yang memiliki relevansi dengan fokus penelitian yang akan dilakukan. Tujuan dibuat Tabel 2.1, adalah untuk memberikan gambaran komprehensif mengenai landasan teoritis yang mendukung kajian ini.

Tabel 2. 1 Penelitian Terdahulu

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
Analisis Sentimen Masyarakat Terhadap *Paylater* Menggunakan Metode *Naïve Bayes Classifier* [11].	ZONAsi Jurnal sistem Informasi	Alfandi Safira, Firman Noor Hasan (2023)	Menggunakan *framework CRISP-DM*, *Web Scrapping* menggunakan *library* snscrape, dan *labelling* data secara manual, *modelling* menggunakan *TextBlob* dan algoritma *Naïve Bayes Classifier* , dan *Evaluation* menggunakan *Confusion matrix*.	Evaluasi *model* dalam penelitian ini menggunakan *Confusion matrix* untuk mengukur tingkat akurasi masing-masing metode. Hasil evaluasi menunjukkan bahwa algoritma *Naïve Bayes* memiliki akurasi sebesar 91%, yang berarti *model* ini mampu mengklasifikasikan sentimen dengan tingkat ketepatan yang sangat tinggi.
Analisis Sentimen Pengguna Terhadap Aplikasi Indodana Di Google Play Store Menggunakan Metode *Naïve Bayes* [25].	*Journal of Informatics Management and Information Technology*	Rifqi Rizaldi, Riska Aryanti (2024)	Menggunakan *library google play scraper* untuk melakukan *scrapping*, *scraper inset lexicon* untuk *labelling*, dan *rating*, menggunakan *TF-IDF* untuk ekstraksi fitur, menggunakan *model Naïve Bayes,* dan untuk *Evaluation* menggunakan	Penelitian ini menganalisis sentimen pengguna terhadap aplikasi Indodana: *Paylater* & Pinjaman berdasarkan 500 ulasan dari Google Play

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
			Confusion matrix, dan *classification report*.	Store menggunakan algoritma *Multinomial Naïve Bayes* dengan dua metode pelabelan, yaitu *inset lexicon* dan *rating*. Hasil pelabelan *inset lexicon* menghasilkan 106 ulasan positif dan 367 ulasan negatif, sedangkan pelabelan *rating* menghasilkan 184 ulasan positif dan 278 ulasan negatif. Setelah melalui proses *preprocessing*, ekstraksi fitur menggunakan *TF-IDF*, dan pembagian data latih-uji (80%-20%), klasifikasi dengan *Naïve Bayes* menunjukkan bahwa metode *inset lexicon* mencapai akurasi 86% dan metode *rating* mencapai akurasi 87%. Meskipun akurasi *rating* sedikit lebih tinggi, nilai *precision*, *recall*, dan *f1-score* dari *inset lexicon* lebih baik.
Sentiment Analysis Of Online Loans	KOMPUTASI : JURNAL ILMIAH	Cita Suci Saputri, Arie Qur'ania,	Data dikumpulkan melalui *crawling* Twitter dengan kata	Distribusi sentimen menunjukkan

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
On Twitter Using Lexicon Based Methods And Support Vector Machine (SVM) [26].	ILMU KOMPUTER DAN MATEMATIKA, VOL. 21 (2) (2024)	Irma Anggraeni (2024)	kunci "pinjol" menggunakan *Tweepy* (*Python*). Tahap *preprocessing*meliputi *Case folding*, *tokenisasi*, *filtering*, dan *Stemming*. *Labeling* dilakukan secara otomatis menggunakan pendekatan *leksikon* berbasis *Inset Lexicon* untuk tiga kelas sentimen. *Klasifikasi* dilakukan dengan metode *Support Vector Machine (SVM)*, dan *evaluasi* dilakukan menggunakan *Confusion matrix*.	dominasi sentimen *negatif* (73,5%), diikuti *positif* (15,9%) dan *netral* (10,6%). *model SVM* mencapai *akurasi* 82,36%, dengan *presisi* 81,00%, *recall* 82,00%, dan *f1-score* 81,00%. Disimpulkan bahwa sentimen publik terhadap pinjol didominasi oleh keluhan terkait pinjol ilegal dan penipuan. Disarankan agar masyarakat lebih berhati-hati.
Analisis Sentimen Pengguna Sistem *Paylater* Menggunakan *Support Vector Machine* Metode Pembobotan *Lexicon* [17].	*Journal of Informatics and Computer Science*	Ferra Junian Wahidna, Paramitha Nerisafitra (2023)	*preprocessing*menggunakan *slang word replacement*, *Labeling* menggunakan metode pembobotan *lexicon*, ekstraksi fitur menggunakan *TF-IDF*, penyeimbangan data menggunakan *SMOTE*, *tuning hyperparameter* menggunakan *Grid search*.	Penelitian ini berhasil mengklasifikasikan sentimen pengguna layanan Shopee *Paylater* dan Go *Paylater* menggunakan algoritma *Support Vector Machine* (*SVM*) dengan pembobotan *Lexicon*. Setelah proses *crawling*, *preprocessing*, dan pelabelan, data dianalisis dan ditemukan bahwa mayoritas sentimen pengguna bersifat netral. *model* klasifikasi *SVM* yang dibangun menunjukkan

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
				performa terbaik pada rasio data latih dan uji 80:20, dengan akurasi sebesar 89,74% untuk Shopee *Paylater* dan 90,27% untuk Go *Paylater*. Hasil ini menunjukkan bahwa kombinasi metode *Lexicon*, *SMOTE* dan *SVM* efektif dalam mengolah dan mengklasifikasikan opini pengguna terkait layanan *Paylater* di media sosial.
Optimizing Sentiment Analysis of Electric Vehicles Through Oversampling Techniques on YouTube Comments [27].	Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI	Jessica Crisfin Lapendy, Andi Aulia Cahyana Resky, Andi Tenriola, Dewi Fatmarani Surianto, Udin Sidik Sidin (2025).	Penelitian ini dimulai dengan pengumpulan komentar YouTube menggunakan *Netlytic* dan pelabelan manual yang divalidasi dengan *Cohen's Kappa*. Data kemudian diproses melalui *cleansing*, normalisasi, *casefolding*, *stopword removal*, dan *stemming*. Augmentasi dilakukan dengan *Synonym Replacement* (EDA), sedangkan penyeimbangan data menggunakan *Random Oversampling*, *SMOTE*, dan *ADASYN*. Klasifikasi dilakukan menggunakan *Naive Bayes Multinomial* dan *SVM* dengan *RBF kernel*, serta divalidasi menggunakan confusion matrix dan *K-Fold Cross Validation* (k = 10).	Penelitian ini Hasil penelitian menunjukkan bahwa performa terbaik dicapai oleh kombinasi *SVM* dengan *Random Oversampling* dan *Synonym Replacement*, menghasilkan akurasi sebesar 97% serta nilai *F1-score* 97% untuk sentimen negatif dan 96% untuk sentimen positif. Analisis sentimen mengungkap isu utama yang sering muncul, seperti harga yang mahal, subsidi yang terbatas, dan kekhawatiran terhadap

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
				kualitas baterai. Perbandingan metode menunjukkan bahwa akurasi model *SVM* dengan kernel *Gaussian RBF* dengan augmentasi *Synonym Replacement* , dan teknik penyeimbangan data memiliki akurasi 96%
Analyzing sentiments on official online Lending platform in Indonesia With a Combination of Naïve Bayes and Lexicon Based Method [28].	*2022 International Conference on Science and Technology (ICOSTECH)*	Aldira Dwinusa Putra, Tachbir Hendro Pudjiantoro, Faiza Renaldi, Asep Id Hadiana (2022)	Data yang digunakan dalam penelitian ini adalah 4.059 komentar dari X mengenai aplikasi pinjaman *online* resmi di Indonesia, yang dibagi menjadi 70% data latih dan 30% data uji. Proses *Labeling* dilakukan menggunakan metode *lexicon-based* untuk menentukan polaritas sentimen positif atau negatif. Setelah *Labeling*, data diklasifikasikan menggunakan algoritma *Naïve Bayes*. Evaluasi *model* dilakukan menggunakan akurasi, *precision*, dan *recall*	Hasil penelitian ini menunjukkan bahwa analisis sentimen terhadap *platform* pinjaman *online* resmi di Indonesia menggunakan kombinasi metode *Naïve Bayes* dan *Lexicon-Based* berhasil mencapai tingkat akurasi 82,06%, dengan hasil dominan berupa komentar negatif dari pengguna X. Pendekatan gabungan ini efektif mengurangi kesalahan klasifikasi yang biasa terjadi jika hanya menggunakan satu metode, sehingga menghasilkan klasifikasi sentimen yang

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
				lebih akurat terhadap data opini masyarakat.
The Evaluation of Effects of Oversampling and Word Embedding on Analisis sentimen[29].	JURNAL INFOTEL Vol 17 No 1 (2025): februari 2025 17-37	Nur Heri Cahyana, Yuli Fauziah, Wisnalmawati, Agus Sasmito Aribowo, dan Shoffan Saifullah(2025)	Pelabelan sentimen secara manual untuk mengkategorikan komentar ke dalam sentimen positif dan negatif. tahapan pembersihan teks, normalisasi slang, *token*isasi, dan *Stemming*. Proses vektorisasi dilakukan menggunakan dua metode *word embedding*, yaitu *Word*2Vec dan *FastText*, untuk mengubah kalimat menjadi representasi vektor. Untuk mengatasi ketidakseimbangan kelas, data minoritas diperbanyak menggunakan tiga teknik *Oversampling* : *SMOTE*, Borderline-*SMOTE*, dan *ADASYN*. *model* klasifikasi kemudian dibangun menggunakan algoritma *Random Forest* dengan rasio data latih dan uji sebesar 80:20. Evaluasi performa *model* dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F-measure* untuk membandingkan efektivitas kombinasi berbagai teknik *embedding* dan *Oversampling* dalam meningkatkan kualitas analisis sentimen.	Penggunaan metode *ADASYN* dalam *model Random Forest* menunjukkan adanya peningkatan akurasi *model* pada seluruh *dataset* , dengan kenaikan rata-rata antara 4% hingga 9%. Peningkatan akurasi tertinggi tercatat pada *dataset* 3, di mana akurasi *model* berbasis *Word*2Vec mengalami peningkatan sebesar 8,8%, dari akurasi awal 74,2% menjadi 83,0% setelah menerapkan *ADASYN*. Hasil ini menunjukkan bahwa *ADASYN* cukup efektif dalam memperbaiki distribusi data tidak seimbang sehingga membantu meningkatkan kemampuan *model* dalam mengklasifikasikan sentimen secara lebih akurat.
Analisis Sentimen Twitter	Jurnal Ilmiah KOMPUTASI	Indira Mahayani, Dewi	Pengambilan data dalam penelitian dilakukan dengan	Hasil analisis sentimen dalam penelitian ini

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
terhadap Pembayaran Shopee*Paylater* Pada Aplikasi Belanja *Online* (Shopee) Menggunakan Metode *Lexicon Based* dan [5].		Agushinta R., Muhammad Edy Supriyadi (2020)	melakukan *crawling tweet* di *platform* X menggunakan kata kunci "Shopee*Paylater*" melalui API yang diakses menggunakan RStudio dengan menggunakan *library shiny*. Data yang diperoleh kemudian melalui proses *preprocessing*yang mencakup *Case folding*, *filtering*, normalisasi kata negasi, *Stopword Removal*, dan *Stemming*. Untuk tahap klasifikasi, digunakan pendekatan *lexicon-based* untuk menentukan polaritas sentimen, yang kemudian diuji akurasinya menggunakan algoritma *Naïve Bayes*. Evaluasi performa *model* dilakukan dengan menggunakan *Confusion matrix*, mencakup metrik akurasi, presisi, dan *recall*. mengukur kinerja *model* melalui akurasi, *recall*, dan *Confusion matrix*.	menunjukkan tingkat akurasi sebesar 82,52% dengan *error rate* sebesar 17,48%. Nilai presisi untuk sentimen positif mencapai 89,54%, sedangkan presisi untuk sentimen negatif hanya 47,06%, dengan nilai *recall* sebesar 89,54%. Distribusi sentimen dalam data menunjukkan dominasi sentimen positif sebanyak 290 *tweet* (84,3%) dan sentimen negatif sebanyak 54 *tweet* (15,7%). Berdasarkan hasil tersebut, dapat disimpulkan bahwa sentimen pengguna terhadap layanan Shopee*Paylater* cenderung positif, terutama terkait dengan kemudahan penggunaan, promo, dan ketersediaan *voucher*.
Semi-supervised Learning models for Analisis sentimen*on Marketplace Dataset* [30].	*International Journal of Artificial Intelligence & Robotics (IJAIR)*	Wisnalmawati, Agus Sasmito Aribowo, Yunie Herawati (2022)	Hasil penelitian ini menunjukkan bahwa *model semi-supervised learning* (SSL) yang dikembangkan untuk analisis sentimen pada data *marketplace* berhasil meningkatkan	Hasil penelitian ini menunjukkan bahwa *model semi-supervised learning* (SSL) yang dikembangkan

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
			kinerja klasifikasi, di mana *model Naïve Bayes* (NB) dan *Random Forest* (RF) dibandingkan dalam berbagai kondisi data; pada *dataset* tiga kelas (Market Data 1), *Random Forest* menunjukkan performa lebih baik dengan *f1-score* 0,65 dibandingkan NB 0,62, sedangkan pada *dataset* dua kelas (Market Data 2), *Naïve Bayes* unggul dengan *f1-score* 0,76 dibandingkan RF 0,71. secara keseluruhan, kinerja SSL sangat dipengaruhi oleh jumlah data latih dan kesesuaian pola fitur dalam dokumen terhadap *model Machine Learning*.	untuk analisis sentimen pada data *marketplace* berhasil meningkatkan kinerja klasifikasi, di mana *model Naïve Bayes* (NB) dan *Random Forest* (RF) dibandingkan dalam berbagai kondisi data; pada *dataset* tiga kelas (Market Data 1), *Random Forest* menunjukkan performa lebih baik dengan *f1-score* 0,65 dibandingkan NB 0,62, sedangkan pada *dataset* dua kelas (Market Data 2), *Naïve Bayes* unggul dengan *f1-score* 0,76 dibandingkan RF 0,71. secara keseluruhan, kinerja SSL sangat dipengaruhi oleh jumlah data latih dan kesesuaian pola fitur dalam dokumen terhadap *model Machine Learning*.
Sentiment Analysis of Indonesian People's Response Against the Paylater	*Vol.* 3 *No.* 4 (2023): Jurnal Bisnis dan Manajemen	Ani Nuraeni, Adhitia Erfina, Dede Sukmawan (2023).	*Sentiment analysis* menggunakan *Naïve Bayes Classifier* . Data dikumpulkan melalui *scraping* komentar Instagram (akun: Shopee, Gopay,	Berdasarkan hasil analisis sentimen, peringkat layanan *Paylater* dengan *sentimen* positif

Nama jurnal	Judul Jurnal	Penulis (Tahun)	Metode	Hasil
Payment Method Using the Naïve Bayes Algorithm [23].			Akulaku, Kredivo, Homecredit) dengan *library beautiful soup pada instagram*. *Preprocessing*meliputi *cleaning*, *Case folding*, *Tokenizing*, *normalization*, *stopwords removal*, dan *Stemming*. *Labeling* sentimen dilakukan secara otomatis dengan *TextBlob* (diterjemahkan via *Google Translate*). Evaluasi menggunakan *Confusion matrix* (akurasi, presisi, *recall*).	tertinggi adalah Akulaku (43,37%), diikuti oleh Kredivo (39,16%), Shopee *Paylater* (34,85%), Go*Paylater* (28,57%), dan Home Credit (25,31%). Dari segi akurasi klasifikasi sentimen, Home Credit mencatatkan nilai tertinggi sebesar 82,49%, disusul oleh Akulaku (75,47%), Go*Paylater* (74,79%), Shopee *Paylater* (66,88%), dan Kredivo (59,30%). Berdasarkan temuan ini, Akulaku direkomendasikan sebagai layanan *Paylater* dengan *sentimen* paling positif. Sementara itu, Home Credit tercatat sebagai layanan dengan *sentimen* negatif tertinggi (50,05%), yang sebagian besar disebabkan oleh keluhan terkait tingginya bunga dan masalah keamanan data.

Berdasarkan Tabel 2.1, ditampilkan beberapa referensi dari jurnal penelitian terdahulu yang telah dilakukan oleh peneliti sebelumnya dan menjadi rujukan dalam penelitian ini. Penelitian yang dilakukan oleh Safira & Hasan [11]. Menjadi fondasi penting dalam analisis sentimen layanan *Paylater* di Indonesia. Data dikumpulkan dari media sosial menggunakan teknik *web scraping*, kemudian dianalisis dengan menerapkan metode *Naïve Bayes* (NB), penelitian ini berhasil mencapai akurasi 91%. Hasilnya mengungkap dominai sentimen negatif terkait isu bunga tinggi dan denda keterlambatan, sekaligus membuktikan keunggulan NB dibandingkan *TextBlob* (61%) pada data berimbang.

Penelitian yang dilakukan oleh Rizaldi & Aryanti [25]. Memperdalam analisis dengan membandingkan dua metode pelabelan pada ulasan aplikasi Indodana (*Paylater* & pinjaman). Penelitian menggunakan *TF-IDF* dan *NB* untuk melakukan analisis sentimen. Hasilnya menunjukkan bahwa pelabelan berbasis *lexicon* (InSet) menghasilkan *precision* dan *recall* yang lebih baik (86%), meskipun akurasi pelabelan berdasarkan *rating* sedikit lebih tinggi (87%).

Penelitian yang dilakukan Indira Mahayani et al [5]. melakukan analisis sentimen pengguna Twitter terhadap fitur pembayaran Shopee*Paylater* di aplikasi Shopee dengan menggabungkan metode *Lexicon Based* dan *Naïve Bayes*. Data diambil melalui *crawling* Twitter API 373 *tweets* mentah menggunakan *library shiny* menggunakan RStudio. Tahap *Preprocessing* meliputi, normalisasi kata negasi, *Stopword Removal*, dan *Stemming*. Hasil klasifikasi *Lexicon Based* menunjukkan dominasi sentimen positif (84.3%). Evaluasi dengan *Naïve Bayes* dan *Confusion matrix* menghasilkan akurasi 82.52% dengan presisi positif tinggi (89.54%). Penelitian menyimpulkan bahwa pengguna cenderung respons positif terhadap Shopee*Paylater*, serta merekomendasikan pengembangan sistem *realtime* dan penanganan *misspelling* untuk penelitian lanjutan.

Penelitian yang dilakukan oleh Wahidna & Nerisafitra [9], berfokus pada permasalahan ketidakseimbangan data dalam ulasan pengguna terhadap layanan *Paylater*. Dalam penelitian ini, mereka menggabungkan algoritma *Support Vector Machine* (*SVM*) dengan teknik *SMOTE* untuk menyeimbangkan jumlah data antar kelas sentimen, serta menggunakan *Grid search* untuk mencari kombinasi parameter terbaik pada *model*. Hasilnya, akurasi klasifikasi meningkat hingga 90,27% pada data ulasan dari Shopee dan Go *Paylater*, dan *model* juga mampu meningkatkan presisi dalam mendeteksi sentimen negatif yang sebelumnya sulit teridentifikasi karena jumlah datanya yang lebih sedikit.

Sementara itu, penelitian yang dilakukan oleh Jessica et al [27]. memperkuat temuan sebelumnya dengan menunjukkan bahwa *SMOTE* mampu meningkatkan akurasi *SVM* hingga 96% pada data tidak seimbang. Penelitian yang dilakukan oleh Cahyana et al [29]. juga mengonfirmasi efektivitas *ADASYN* dalam meningkatkan akurasi *model* berbasis *word embedding* sebesar 8,8%. Namun, kedua studi sepakat bahwa *TF-IDF* tetap lebih unggul karena kemampuannya mengekstrak kata kunci tanpa konteks panjang.

Penelitian yang dilakukan Cita Suci Saputri et al [26]. Melakukan analisis sentimen masyarakat terhadap pinjaman *online* (pinjol) di Twitter menggunakan kombinasi metode *Lexicon Based* dan *Support Vector Machine (SVM)*. Data diambil melalui *crawling* Twitter (7.631 *tweets*, menjadi 1.728 data setelah pembersihan) dengan kata kunci "pinjol". Tahap *preprocessing*meliputi *Case folding*, *token*isasi, penghapusan *noise*, dan *Stemming*. *Labeling* sentimen otomatis menggunakan *Lexicon Based* (*Inset lexicon*) mengklasifikasikan data menjadi positif, negatif, dan netral. Klasifikasi dengan *SVM* menghasilkan akurasi 82.36% dengan presisi 81.00% dan *recall* 82.00%. Hasil menunjukkan 73.5% sentimen negatif didominasi keluhan penipuan, bunga tinggi, dan pinjol ilegal, sementara hanya 15.9% positif terkait layanan resmi. Penelitian merekomendasikan peningkatan kewaspadaan masyarakat dalam memilih layanan pinjol.

penelitian yang menggunakan metode *hybrid* dilakukan oleh Aldira Dwinusa Putra et al [28]. yang menggabungkan algoritma *Naïve Bayes* dengan pendekatan *lexicon-based* pada 4.059 komentar Twitter mengenai layanan pinjaman *online*. Hasil penelitian tersebut menunjukkan akurasi sebesar 82,06%, yang menandakan bahwa kombinasi dua pendekatan ini cukup efektif dalam melakukan analisis sentimen. Namun, penelitian ini masih memiliki kekurangan, yaitu ketergantungan pada *lexicon* yang bersifat statis, artinya daftar kata dan makna yang digunakan tidak dapat beradaptasi secara otomatis dengan konteks baru, bahasa gaul, atau perkembangan kosakata di media sosial, sehingga dapat mengurangi akurasi dalam klasifikasi sentimen yang dinamis.

Penelitian yang dilakukan oleh Wisnalmawati et al [30]. membuka peluang penerapan *semi-supervised learning* (SSL) pada analisis sentimen *Paylater*. Meskipun menggunakan *Random Forest* dan NB, mekanisme *pseudo-Labeling* dalam studi ini berhasil mencapai *f1-score* 0,76 pada data dua kelas. Temuan ini menjadi dasar potensial untuk integrasi SSL dengan *SVM*, khususnya dalam memanfaatkan data tidak berlabel pada *platform* finansial.

Penelitian yang dilakukan oleh Ani Nuraeni et al [23]. menganalisis sentimen masyarakat Indonesia terhadap metode pembayaran *Paylater* di lima *marketplace* (Shopee, Gopay, Akulaku, Kredivo, Homecredit) menggunakan algoritma *Naïve Bayes*. Data diambil dari 3.603 komentar Instagram melalui teknik *scraping* dengan *Python* dengan menggunakan beautifulsoup. Hasil menunjukkan Akulaku memiliki sentimen positif tertinggi (43.37%), sementara Homecredit memiliki sentimen negatif dominan (50.05%) akibat bunga tinggi dan kebocoran data. Akurasi klasifikasi tertinggi dicapai pada data Homecredit (82.49%). Penelitian merekomendasikan Akulaku sebagai opsi *Paylater* terbaik berdasarkan sentimen positif, serta menyoroti perlunya peningkatan normalisasi otomatis dan eksplorasi algoritma lain untuk penelitian lanjutan.

Dalam penelitian ini, peneliti memutuskan untuk memberikan kontribusi penelitian dengan melakukan pembandingan algoritma *SVM* dan *Naïve Bayes* karena keduanya menunjukkan performa klasifikasi yang tinggi dan konsisten dalam analisis sentimen sehingga menciptakan urgensi dalam melakukan perbandingan untuk mencari algoritma yang menghasilkan *model* dengan akurasi terbaik [11][25][28][5][23][17][26]. Metode penyeimbangan data *SMOTE* dan *ADASYN* dipilih karena efektif meningkatkan akurasi pada data tidak seimbang. Peneliti memutuskan untuk membandingkan dua metode penyeimbangan data karena terinspirasi dari penelitian sebelumnya yang melakukan perbandingan teknik penyeimbangan data dan menunjukkan adanya potensi perbedaan kinerja yang signifikan antara metode yang digunakan [27]. Pendekatan *Semi-supervised* digunakan untuk meningkatkan performa *model* saat data berlabel terbatas, dengan memanfaatkan data tidak berlabel dalam proses pembelajaran [30].

2.2 Teori Penelitian

2.2.1 Financial Technology

Kemajuan teknologi informasi dan komunikasi telah mendorong terjadinya transformasi digital di berbagai sektor kehidupan, termasuk sektor keuangan. Salah satu inovasi yang muncul dari perkembangan ini adalah *financial technology* atau *fintech*, yang menawarkan pendekatan baru dalam penyediaan layanan keuangan. *Fintech* menggantikan sebagian peran lembaga keuangan konvensional dengan memanfaatkan teknologi digital. *Fintech* mampu meningkatkan efisiensi operasional, memperluas akses keuangan, serta memberikan kemudahan dan kenyamanan bagi pengguna [31].

Fintech merupakan integrasi antara teknologi dan sistem keuangan yang menghasilkan produk, layanan, dan *model* bisnis baru, yang memengaruhi stabilitas ekonomi, nilai tukar mata uang, hingga efektivitas keamanan dalam transaksi digital. Seiring dengan perkembangannya, layanan *fintech* semakin beragam dan dapat dikategorikan dalam beberapa jenis utama. Jenis-jenis layanan *fintech* meliputi *Peer-to-Peer* (P2P) *Lending*, yang memungkinkan

pinjam meminjam secara langsung antar individu tanpa melalui lembaga keuangan formal seperti layanan *Buy Now Paylater,* dan pinjaman *online*. Crowdfunding, yang berfungsi sebagai *platform* untuk penggalangan dana kolektif untuk mendukung proyek atau usaha tertentu. digital *payment*, mencakup sistem pembayaran elektronik seperti dompet digital dan transfer instan. Investment *platform* menyediakan layanan investasi daring yang fleksibel. Insurtech, yaitu penerapan teknologi dalam layanan asuransi. Robo-*advisory*, layanan pengelolaan serta konsultasi keuangan otomatis berbasis algoritma [31].

2.2.2 Paylater

Paylater merupakan salah satu inovasi penting dalam layanan *financial technology* (*fintech*) yang mengalami perkembangan pesat di Indonesia [3]. Layanan *Paylater* merupakan fasilitas keuangan yang memungkinkan pengguna melakukan pembayaran dengan skema cicilan tanpa harus bergantung pada kartu kredit. Tidak seperti sistem kredit konvensional, *Paylater* tidak memerlukan kartu fisik dan umumnya terintegrasi langsung ke dalam *platform E-commerce* maupun aplikasi digital lainnya, menjadikannya solusi yang praktis dan mudah diakses oleh berbagai lapisan masyarakat [25].

Pertama kali diperkenalkan di Indonesia pada tahun 2018, *Paylater* hadir sebagai bentuk inovasi sistem pembayaran yang memberikan fleksibilitas kepada pengguna untuk melakukan transaksi terlebih dahulu dan menyelesaikan pembayaran di kemudian hari. Hal ini memberikan kemudahan dan kenyamanan, terutama dalam memenuhi kebutuhan harian atau keperluan mendesak, tanpa harus melalui prosedur yang rumit sebagaimana dalam pengajuan kredit formal [4]. Perkembangan layanan ini sangat pesat, didukung oleh kolaborasi antara perusahaan *fintech* dan *E-commerce*. Traveloka menjadi pelopor dalam mengimplementasikan *Paylater* di Indonesia melalui kerja sama dengan PT Dana Pasar Pinjaman. Seiring berjalannya waktu, banyak *platform* digital lainnya seperti Tokopedia, Shopee, dan Bukalapak juga mulai

mengadopsi sistem serupa, sehingga meningkatkan popularitas *Paylater* secara signifikan di kalangan pengguna digital [6].

Antusiasme masyarakat terhadap *Paylater* tercermin dari pertumbuhan pengguna yang sangat signifikan. Berdasarkan data Otoritas Jasa Keuangan (OJK), jumlah kontrak pembiayaan *Paylater* di Indonesia mencapai 79,92 juta pada tahun 2023, naik drastis dari hanya 4,63 juta kontrak pada tahun 2019, dengan rata-rata pertumbuhan tahunan mencapai 144,35%. Hingga Maret 2024, outstanding piutang pembiayaan *Paylater* tercatat sebesar Rp6,13 triliun, meningkat 23,9% secara tahunan (*year-on-year*), yang mengindikasikan tingginya minat masyarakat terhadap sistem pembayaran ini [6]. Melihat tren tersebut, Kepala Eksekutif Pengawasan Lembaga Pembiayaan OJK, Agusman, menyatakan bahwa kinerja *Paylater* diperkirakan akan terus meningkat seiring dengan perkembangan teknologi yang kian menyederhanakan proses transaksi digital. *Paylater* menjadi bagian komponen dari ekosistem *fintech* yang tidak hanya menawarkan kenyamanan, tetapi juga memperluas inklusi keuangan di Indonesia [7].

2.2.3 Analisis sentimen

Analisis sentimen merupakan metode untuk mengidentifikasi serta mengkategorikan opini atau ekspresi emosional dalam teks menjadi dua jenis polaritas, yakni positif, dan negatif. Teknik ini banyak dimanfaatkan untuk menelusuri opini masyarakat terhadap berbagai topik seperti produk, layanan, atau isu tertentu, terutama yang bersumber dari *platform* digital seperti media sosial, forum daring, dan ulasan pengguna. Informasi yang dihasilkan dari analisis ini dapat digunakan sebagai dasar pengambilan keputusan strategis oleh perusahaan maupun pembuat kebijakan [32]. Tujuan utama dari analisis sentimen adalah untuk memperoleh pemahaman otomatis mengenai sikap publik yang terekam dalam bentuk teks, sehingga dapat digunakan sebagai dasar dalam pengambilan keputusan berbasis data. Contohnya, dalam dunia bisnis, perusahaan dapat menganalisis tanggapan konsumen terhadap produk atau layanan yang mereka tawarkan dan merumuskan strategi yang lebih tepat

sasaran [15]. Secara umum, terdapat tiga pendekatan utama dalam penerapan analisis sentimen**:**

1. Pendekatan Berbasis kamus *(Lexicon-Based Approach)*
   Mengandalkan daftar kata yang telah diberi label sentimen untuk mengukur polaritas teks [33].
2. Pendekatan Berbasis Pembelajaran Mesin (*Machine Learning Approach*)
   Menggunakan algoritma klasifikasi seperti *Support Vector Machine (SVM)*, *Naïve Bayes*, atau *Random Forest* untuk mempelajari pola dari data berlabel dan memprediksi sentimen pada data baru [15].
3. Pendekatan Hibrida (*Hybrid Approach*)
   Merupakan gabungan dari dua pendekatan sebelumnya untuk memperoleh hasil yang lebih akurat dan adaptif [28].

2.2.4 Machine Learning

Machine Learning merupakan salah satu cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan sistem komputer untuk belajar dan beradaptasi dari data yang tersedia tanpa harus diprogram secara langsung. Di dalam *Machine Learning*, algoritma dan *model* statistik dimanfaatkan untuk mengevaluasi data serta mengidentifikasi pola atau keterkaitan penting yang dapat digunakan dalam proses pengambilan keputusan maupun prediksi. Pendekatan ini berperan dalam meningkatkan performa sistem prediktif serta akurasi hasil yang dihasilkan [31].

Machine Learning dibagi menjadi tiga jenis utama, yaitu *supervised learning*, *unsupervised learning*, dan *semi-supervised learning*. *Supervised learning* menggunakan data yang telah diberi label untuk melatih *model* agar mampu melakukan klasifikasi atau prediksi dengan akurat [28]. Sementara itu, *unsupervised learning* bekerja dengan data yang tidak memiliki label, memungkinkan *model* untuk mengeksplorasi struktur tersembunyi dalam data yang kompleks. Metode ini sering diterapkan pada data yang beragam seperti teks, gambar, audio, maupun video [34]. *semi-supervised learning* merupakan pendekatan gabungan antara *unsupervised learning,* dan *supervised learning*

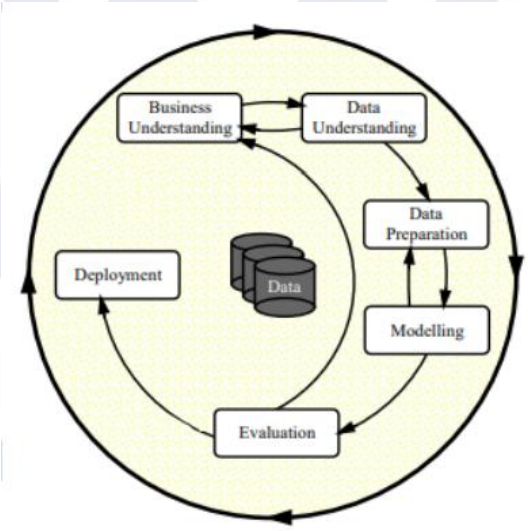
yang memanfaatkan sebagian data berlabel dan sebagian lainnya tidak berlabel dalam proses pelatihan *model* [30].

2.2.5 Twitter (X)

X adalah *platform* media sosial yang berfungsi sebagai layanan *microblogging*. X Memungkinkan penggunanya untuk membagikan pemikiran, informasi, maupun opini secara singkat dan cepat. Melalui X, pengguna dapat mengunggah *tweet*s dalam berbagai format, seperti teks, gambar, video, atau tautan (*URL*), dengan batas maksimum 280 karakter. *Platform* ini juga menyediakan fitur tambahan seperti *retweet, reply*, dan *like* yang memungkinkan interaksi antar pengguna serta memperluas jangkauan penyebaran informasi. Twitter sering dimanfaatkan dalam berbagai bidang, mulai dari jurnalisme, pemasaran, analisis opini publik, hingga penelitian akademik [31].

2.2.6 Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM merupakan sebuah kerangka kerja yang memetakan proses dalam kegiatan *Data Mining*. Diformulasikan sebagai respons terhadap kebutuhan penyelesaian masalah yang sering muncul dalam konteks bisnis dan penelitian [35]. Berdasarkan Gambar 2.1, Terdapat enam tahapan utama yang terdapat dalam struktur *CRISP-DM* [35]:



Gambar 2. 1 Tahapan Crisp-Dm

Sumber : [35]

1. *Business Understanding*

Fokus dari tahap *Business Understanding* adalah untuk memperoleh pemahaman yang komprehensif terkait tujuan dan kebutuhan dari perspektif bisnis. Langkah pertama adalah mengekstraksi informasi yang terkait untuk merumuskan permasalahan yang akan dipecahkan. Langkah kedua merencanakan strategi yang akan mengarah pada pencapaian tujuan yang telah ditetapkan [35].

2. *Data Understanding*

Data Understanding dimulai dengan proses pengambilan data serta penerapan pendekatan untuk memahami karakteristik dan sifat data yang telah di kumpulkan. Tahap ini juga bertujuan untuk mengidentifikasi permasalahan yang terkait dengan kualitas data, mengeksplorasi wawasan yang terkandung dalam data, serta menggali informasi tersembunyi yang terdapat dalam *dataset* tersebut. Pemahaman yang baik terhadap data pada tahap ini menjadi dasar penting untuk menentukan strategi analisis dan pemodelan selanjutnya [35].

3. *Data Preparation*

Data Preparation adalah proses data mentah diolah dan dibersihkan hingga mencapai bentuk akhir atau data yang akan digunakan pada tahap *modelling*. Pada tahap ini, dilakukan berbagai langkah seperti pembersihan data dari duplikasi, penanganan data hilang (*missing values*), Pembersihan teks dari simbol, dan transformasi data. Tujuannya adalah memastikan bahwa data berada dalam format yang konsisten, terstruktur, dan optimal untuk menghasilkan model yang akurat dan andal [35].

4. *Data Modelling*

Fase *modelling* merupakan tahap di mana berbagai teknik pemodelan diterapkan dan diuji untuk menentukan metode terbaik dalam menganalisis data. Proses ini melibatkan eksplorasi dan evaluasi terhadap beberapa algoritma, termasuk penyesuaian *hyperparameter* untuk mengoptimalkan performa model. Tujuannya adalah memperoleh model yang paling efektif

dan efisien dalam menyelesaikan permasalahan yang telah diidentifikasi sebelumnya [35].

5. *Evaluation*

Evaluation merupakan tahap yang melakukan analisis secara mendalam kepada *model* yang digunakan serta peninjauan menyeluruh terkait kemampuan *model*. Berfungsi untuk mencapai tujuan penelitian yang telah ditetapkan. Proses ini mencakup penilaian yang komprehensif terhadap kinerja *model*, memastikan bahwa *model* tersebut sesuai dengan tujuan penelitian yang telah ditetapkan [35].

6. *Deployment*

Deployment merupakan proses ketika *model* yang telah dilatih diterapkan dalam lingkungan produksi. Tujuannya adalah agar *model* tersebut dapat digunakan secara nyata oleh pengguna. Umumnya, *deployment* dilakukan melalui pembangunan aplikasi atau situs web yang mudah diakses dan dioperasikan oleh pengguna. [35].

2.3 *Framework* Dan *algoritma* penelitian

2.3.2 Web Scrapping

Web Scraping adalah teknik otomatis yang digunakan untuk mengekstraksi data dari situs web secara sistematis. Proses ini dilakukan oleh program yang disebut *web scraper*, yang dirancang untuk mengumpulkan informasi dari berbagai halaman web. Teknik ini sering dimanfaatkan untuk memperoleh data dari situs *E-commerce*, *platform* berita, atau sumber informasi publik lainnya. Data yang berhasil dikumpulkan kemudian dapat dianalisis untuk berbagai tujuan, seperti riset pasar, pengembangan produk, atau analisis tren [36].

2.3.3 TweetHarvest

Tweet Harvest adalah sebuah *library Python* yang dirancang khusus untuk melakukan *web scraping* pada *platform* Twitter. *Library* ini memungkinkan pengguna untuk secara efisien mengumpulkan *tweet* berdasarkan kata kunci tertentu, bahasa, dan rentang tahun yang diinginkan.

Tweet Harvest memiliki kemampuan untuk menyesuaikan pencarian, pengguna dapat menargetkan *tweet* yang relevan untuk analisis lebih mendalam. Data yang dikumpulkan kemudian dapat disimpan dalam format CSV, yang memudahkan pengolahan dan analisis lebih lanjut, seperti analisis sentimen atau pemetaan tren yang sedang berkembang di Twitter. Format CSV yang dihasilkan mencakup informasi penting seperti teks *tweet*, tanggal posting, serta metadata lainnya yang dapat digunakan untuk penelitian lebih lanjut [37].

Untuk dapat menggunakan *Tweet Harvest*, pengguna perlu terlebih dahulu memperoleh *authorization token* dari Twitter. *Token* ini memberikan izin akses untuk mengumpulkan data *tweet* secara langsung sesuai dengan kebijakan dan regulasi yang berlaku di API Twitter. Untuk mendapatkannya, pengguna harus melakukan *login* ke akun Twitter mereka melalui browser dan kemudian mengekstrak *token* dari sesi *login* yang aktif. Proses ini penting karena Twitter membatasi akses data dengan menggunakan *token* untuk memastikan bahwa hanya pengguna yang sah yang dapat mengakses informasi tersebut. Penggunaan *token* ini menyebabkan *Tweet Harvest* dapat bekerja sesuai dengan ketentuan API Twitter, yang memungkinkan pengambilan data dengan cara yang legal dan sesuai dengan kebijakan *platform* [37].

2.3.4 Text preprocessing

Text preprocessing merupakan sebuah proses yang sangat penting dalam pembuatan *model* untuk sentimen analisis. Data yang diambil sering kali tidak konsisten dan mengandung *noise* seperti simbol, angka, akhiran kata, dan kata-kata yang tidak memiliki relevansi dengan sentimen. *Noise* dan data yang tidak konsisten dapat memengaruhi performa *model* saat digunakan dalam proses klasifikasi. Tujuan dari tahap ini adalah untuk membersihkan data dari *noise* dan ketidaksesuaian, mengubah data ke dalam format yang lebih cocok, dan mengurangi data yang memiliki duplikasi, gangguan, dan ketidaksesuaian. Teknik ini dapat diterapkan dalam berbagai tahap, termasuk klasifikasi dokumen, pengelompokan data, ekstraksi informasi, analisis sentimen, dan pengambilan informasi. Penggunaan teknik ini secara tepat dapat meningkatkan

akurasi sistem dalam memahami dan mengolah data teks secara lebih efektif [21].

1. *Case folding*

Case folding adalah suatu proses dalam pengolahan data teks yang bertujuan mengubah seluruh huruf kapital menjadi huruf kecil. Langkah ini penting untuk memastikan konsistensi dalam penulisan kata, sehingga kata yang sama tidak dianggap berbeda hanya karena perbedaan kapitalisasi. Proses ini membantu mengurangi redundansi dan meningkatkan akurasi analisis teks [21].

2. *Cleaning data*

Proses Cleaning Data merupakan proses penghapusan simbol, karakter selain huruf alfabet, dan emoticon dari *tweet*. Langkah pembersihan dilaksanakan untuk mengurangi keberadaan karakter atau simbol yang tidak diinginkan atau tidak memiliki makna. Proses ini membantu mengurangi *noise* dan *outlier* pada teks [21].

3. *Tokenization*

Tokenization adalah tahap awal dalam pemrosesan teks, di mana sebuah teks dibagi-bagi menjadi unit-unit kata yang disebut *token*. Pemisahan ini umumnya didasarkan pada karakter spasi sebagai batas antar kata. Selain itu, pada tahap ini, karakter-karakter khusus seperti simbol, angka, dan tanda baca juga akan dieliminasi untuk menyederhanakan data dan memastikan bahwa hanya informasi tekstual yang relevan yang diproses lebih lanjut [21].

4. *Stopword Removal*

Proses *Stopword Removal* pada teks adalah langkah untuk menghapus kata-kata yang kurang bermakna dalam kalimat teks. Contoh dari *stopword* dalam bahasa Indonesia meliputi kata-kata seperti "yang", "atau", dan "di". Penghapusan kata-kata ini bertujuan untuk memfokuskan analisis pada kata-kata yang lebih informatif dan relevan terhadap konteks data [21].

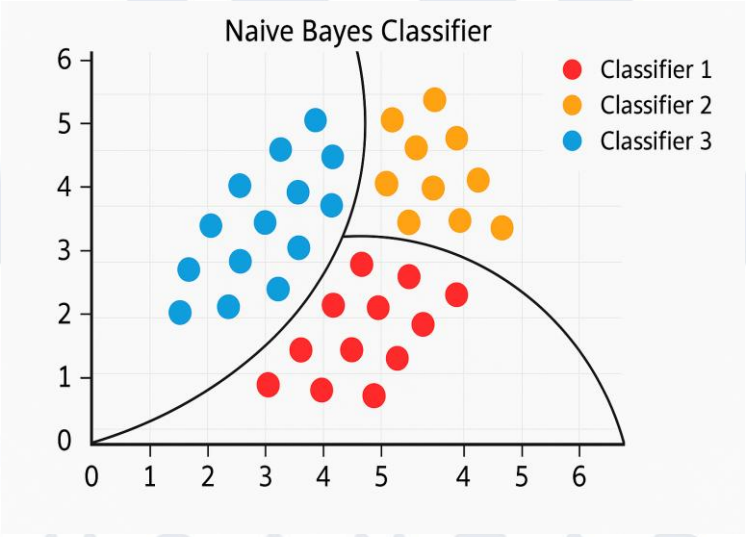
5. *Stemming*

Stemming merupakan proses penting dalam pengolahan bahasa alami yang bertujuan untuk menemukan bentuk dasar atau akar kata dari setiap kata dalam suatu kalimat. Proses ini dilakukan dengan cara menghilangkan berbagai

imbuhan, baik yang berada di awal (awalan), di tengah (sisipan), maupun di akhir (akhiran) dari sebuah kata. Tujuannya adalah untuk menyederhanakan variasi morfologis kata agar menjadi bentuk yang seragam, sehingga dapat memudahkan dalam proses analisis teks, seperti klasifikasi atau pencarian kata kunci. Sebagai contoh, kata "mengambil" akan diubah menjadi "ambil" karena bentuk dasar atau lemma dari kata tersebut adalah "ambil" tanpa imbuhan "meng" yang menandakan aspek verbal dalam bahasa Indonesia [21].

2.3.5 Naïve Bayes

Algoritma *Naïve Bayes* merupakan salah satu metode dalam *Machine Learning* yang menggunakan prinsip dasar dari *Bayes Theorem* [28]. Algoritma ini bekerja dengan mengasumsikan bahwa setiap *feature* atau atribut bersifat independen satu sama lain. Keunggulan dari algoritma ini terletak pada kemampuannya dalam melakukan perhitungan yang cepat dan efisien, serta performanya yang cukup baik dalam menangani masalah *text classification*. Cara kerja algoritma ini ditunjukkan pada Gambar 2.2, di mana *feature space* dibagi menjadi beberapa wilayah klasifikasi berdasarkan nilai *posterior probability* dari setiap kelas. *model* akan menghitung kemungkinan suatu data masuk ke dalam kelas tertentu dengan melihat distribusi fitur dalam masing-masing kelas.



Gambar 2. 2 Visualisasi Naïve Bayes
Sumber : [38]

Equation 2.1, menampilkan rumus dasar dari *Bayes' Theorem* yang digunakan sebagai landasan dalam proses perhitungan probabilitas tersebut [39].

$$P(c \mid x) = \frac{P(x|c).P(c)}{P(x)} \quad (2.1)$$

$P(c|x)$ = sebuah probabilitas kata x muncul pada kelas c.

$P(c)$ = probabilitas kata pada kelas c.

$P(x) =$ probabilitas kemunculan pada kata x.

Tabel 2.2, menunjukkan *pseudocode* dari algoritma *SVM* yang menggambarkan langkah-langkah utama dalam proses pelatihan dan klasifikasi [26].

Tabel 2. 2 Pseudocode Naïve Bayes

Algoritma 2 Naïve Bayes (NB)
Input:
D = {(x₁, y₁), ..., (xₙ, yₙ)} // Data pelatihan
xᵢ = [xᵢ₁, ..., xᵢ_d] ∈ ℕ⁺ // Vektor hitung dari d fitur (misal: jumlah kata)
yᵢ ∈ {c₁, ..., c_K} // Label kelas
Hiperparameter:
α ≥ 0 // Parameter smoothing Laplace (misal α = 1)
Output:
Probabilitas prior kelas P(c),
Probabilitas likelihood fitur P(xⱼ | c)
Fase Pelatihan:
1: Untuk setiap kelas c ∈ {c₁, ..., c_K} lakukan:
2: N_c ← jumlah sampel dengan label yᵢ = c
3: P(c) ← N_c / n // Probabilitas prior kelas
4: T_c ← 0 // Total jumlah token untuk kelas c
5: Untuk setiap fitur j = 1 sampai d lakukan:
6: T_{cⱼ} ← jumlah nilai xᵢⱼ dari semua xᵢ dengan yᵢ = c
7: T_c ← T_c + T_{cⱼ}

8: Untuk setiap fitur j = 1 sampai d lakukan:

9: $P(x_j \mid c) \leftarrow (T_{c_j} + \alpha) / (T_c + \alpha \cdot d)$ // Likelihood dengan smoothing

Fase Prediksi:

10: Fungsi Predict(x_baru):

11: Untuk setiap kelas $c \in \{c_1, ..., c_K\}$ lakukan:

12: log_prob[c] ← log P(c)

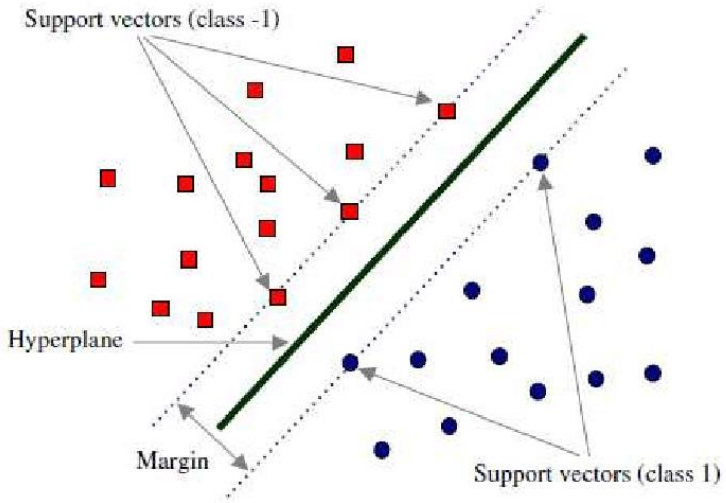
13: Untuk j = 1 sampai d lakukan:

14: log_prob[c] ← log_prob[c] + x_baru[j] × log $P(x_j \mid c)$

15: Kembalikan kelas c* yang memiliki log_prob[c] terbesar

2.3.6 Support Vector Machine(SVM)

Algoritma *SVM* merupakan sebuah teknik pembelajaran terawasi (*supervised learning*) dalam klasifikasi yang diperkenalkan pada tahun 1992. Proses implementasinya melibatkan tahap pelatihan untuk memahami serta mengenali objek yang akan dianalisis, dilanjutkan dengan tahap pengujian [31]. *SVM* bekerja dengan mencari garis atau hiperbidang (*hyperplane*) yang optimal untuk memisahkan dua kelas data yang berbeda, misalnya, kelas Positif (+1) dan kelas Negatif (-1). Ilustrasi dalam Gambar 2.3, menggambarkan batas pemisahan dengan *margin* maksimum, dimana *hyperplane* yang memiliki *margin* terbaik diharapkan dapat memberikan hasil klasifikasi yang lebih akurat [40].



Gambar 2. 3 Hyperplane Svm
Sumber : [40]

SVM merupakan algoritma yang dipakai untuk menyelesaikan masalah klasifikasi dengan menggunakan persamaan atau pertidaksamaan yang bersifat *linear*. Penerapan teknik *kernel* pada *SVM*, membuat *model* dapat memetakan data ke dalam ruang vektor berdimensi lebih tinggi. Sehingga memungkinkan pemisahan data dengan *hyperplane* yang optimal, dengan demikian masalah yang awalnya tidak dapat diselesaikan secara *linear* dapat diatasi.

Tabel 2.3, menampilkan persamaan *kernel* yang digunakan:

Tabel 2. 3 Rumus Pengujian Uat
Sumber : [40]

Jenis *Kernel*	*model*
Linear	$K\ (x.x') = x.x$
Polynomial	$K\ (x.x') = (x.x' + \mathrm{c})'$
RBF Gaussian	$K\ (x.x') = \exp\ (-y\|\|x - x'\|\|^2)$
Siqmoid	$K\ (x.x) = \tanh(\alpha x.x + \beta)$

Tabel 2.4, menunjukkan *pseudocode* dari algoritma *SVM* yang menggambarkan langkah-langkah utama dalam proses pelatihan dan klasifikasi [26].

Tabel 2. 4 Pseudocode Svm

Algoritma 1 Pseudocode Algoritma SVM

Input:

D = {(x_1, y_1), ..., (x_n, y_n)}, dengan $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$

Tipe kernel K(x_i, x_j), misal:

- Linear: $K(x_i, x_j) = x_i^T x_j$

- Polinomial: $K(x_i, x_j) = (\gamma\, x_i^T x_j + coef0)^{degree}$

- RBF: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

Hyperparameter: C > 0, γ > 0 (untuk RBF), degree (untuk polinomial), coef0

Output:

Vektor support x_i, koefisien α_i, dan bias b

1: Hitung matriks Gram $K \in \mathbb{R}^{n \times n}$:

Untuk i = 1 sampai n:

Untuk j = 1 sampai n:

$K_{ij} \leftarrow K(x_i, x_j)$

2: Selesaikan program kuadratik berikut:

Maksimalkan: $L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K_{ij}$

Dengan syarat: $0 \leq \alpha_i \leq C$, untuk semua i

$$\sum_i \alpha_i y_i = 0$$

3: Pilih vektor support:

SupportVectors $\leftarrow \{x_i \mid \alpha_i > 0\}$

4: Hitung bias b:

Pilih salah satu vektor support x_s dengan $0 < \alpha_s < C$

$b \leftarrow y_s - \sum_i \alpha_i y_i K(x_i, x_s)$

5: Kembalikan model:

$$f(x) = \sum_i \alpha_i y_i K(x_i, x) + b$$

2.3.7 Term Frequency – Inverse Document Frequency (TF-IDF)

Term Frequency - Inverse Document Frequency (TF-IDF) merupakan sebuah teknik ekstraksi fitur yang digunakan untuk memberikan bobot kepada setiap kata dalam teks. *TF-IDF* menghitung seberapa sering sebuah kata muncul dalam suatu dokumen. Untuk menghitung bobot *TF-IDF*, dapat menggunakan rumus yang diberikan dalam referensi [21].

1. Term Frequency (TF)

Term Frequency adalah metode pembobotan kata yang sangat sederhana, perhitungan IDF dapat dilihat pada rumus 2.2 [21].

$$tf(t,d) = \frac{n_i j}{\sum_k n_i, j} \quad (2.2)$$

Keterangan :

$tf(t,d)$ = Frekuensi *term.*

$n_i j$ = Total kemunculan seluruh kata pada dokumen.

$\sum_k n_i, j$ = total seluruh kata pada dokumen.

2. Inverse Document Frequencey (IDF)

Inverse Document Frequency (IDF) digunakan untuk melihat kemunculan setiap kata pada kumpulan kelas, perhitungan IDF dapat dilihat pada rumus 2.3 [21].

$$idf = \log \frac{N}{df_j} \quad (2.3)$$

Keterangan :

N = Total kelas.

df_j = Total kelas j yang memiliki isi kata i.

3. Menghitung *TF-IDF* (*Term Frequency Inverse Document*)

perhitungan *TF-IDF* dapat dilihat pada rumus 2.3 [21].

$$W_{ij} = tf_{ij} \times idf \quad (2.4)$$

Keterangan :

W_{ij} = Bobot kata I pada kelas j

tf_{ij} = Total kemunculan kata I pada kelas j.

df_j = Total Kelas j yang berisi kata i.

2.3.8 SMOTE

SMOTE (*Synthetic Minority Oversampling Technique*) adalah sebuah teknik *Oversampling* yang telah dirancang untuk mengatasi permasalahan ketidakseimbangan antar kelas dalam *dataset*. Pendekatan ini bekerja dengan cara menciptakan sampel-sampel tambahan dari kelas minoritas dengan cara menghasilkan instansi-instansi baru berdasarkan kombinasi yang cermat dari tetangga-tetangga terdekatnya. Tujuan utamanya adalah untuk mencapai

keseimbangan dalam *dataset* dengan meningkatkan jumlah sampel pada kelas minoritas yang biasanya memiliki representasi yang lebih rendah [21].

2.3.9 ADASYN (Adaptive Synthetic Sampling)

ADASYN (*Adaptive Synthetic Sampling*) merupakan salah satu metode *oversampling* berbasis adaptif yang dikembangkan untuk secara efisien menangani masalah ketidakseimbangan kelas (*class imbalance*) dalam data pelatihan. Sebuah tantangan yang umum dijumpai dalam pengembangan *model* pembelajaran mesin. Ketidakseimbangan ini terjadi ketika jumlah sampel antar kelas dalam sebuah dataset tidak seimbang secara signifikan, di mana kelas minoritas memiliki jumlah data yang jauh lebih sedikit dibandingkan dengan kelas mayoritas. Situasi ini dapat menyebabkan algoritma pembelajaran cenderung memfokuskan proses pelatihan pada kelas mayoritas, sehingga menurunkan akurasi dan sensitivitas terhadap kelas minoritas [41].

Teknik ini bekerja dengan menghitung distribusi kepadatan (*density distribution*) dari *instance* kelas minoritas untuk mengidentifikasi area-area dalam ruang fitur yang dianggap sulit dipelajari oleh model, seperti area dengan kepadatan rendah atau yang berada di dekat batas klasifikasi. Melalui informasi tersebut, teknik ini kemudian melakukan proses interpolasi linier antara *instance* minoritas tersebut dengan *k-nearest neighbors* terdekatnya, yaitu tetangga-tetangga terdekat dalam ruang vektor fitur. Tujuan dari interpolasi ini adalah untuk menghasilkan *sampel sintetik* baru yang menyerupai pola distribusi alami kelas minoritas, dengan mempertimbangkan struktur lokal data. Sampel-sampel tersebut dihasilkan secara proporsional terhadap tingkat kesulitan wilayah tersebut, sehingga area yang lebih sulit atau rentan terhadap kesalahan klasifikasi akan memperoleh lebih banyak *sintetik sample* yang pada akhirnya diharapkan dapat meningkatkan performa model [41].

2.3.10 Grid search

Grid search merupakan teknik untuk melakukan penyetelan *hyperparameter* (*hyperparameter tuning*) yang bertujuan meningkatkan performa *model*. Proses ini difasilitasi oleh fitur *Grid searchCV* dalam pustaka

scikit-learn, yang menjalankan *Cross Validation* secara terstruktur. Pada praktiknya, *Grid search* bekerja dengan melakukan evaluasi pada seluruh kombinasi *hyperparameter* yang telah ditentukan. Setelah seluruh kombinasi tersebut diuji, teknik ini akan memilih *best parameter*, yakni konfigurasi *hyperparameter* dengan hasil performa terbaik dari proses pengujian yang dilakukan [20].

2.3.11 Confusion matrix

Confusion matrix merupakan salah satu alat evaluasi untuk melakukan analisa kinerja model klasifikasi yang telah dilakukan *training*. *Confusion matrix* pada dasarnya adalah sebuah data yang memungkinkan kita untuk membandingkan nilai yang sebenarnya atau aktual dengan nilai yang diprediksi oleh *model*, yang nantinya dapat digunakan untuk menyusun berbagai matriks penilaian [23].

1. Accuracy

Accuracy adalah perbandingan antara jumlah prediksi yang akurat atau benar, bedasarkan ketepatan memprediksi label positif maupun label negatif, dengan jumlah keseluruhan data yang tersedia. Perhitungannya dapat dilakukan melalui persamaan (2.5) [23].

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{Tp+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (2.5)$$

2. *Precision*

Precision adalah parameter pengukuran yang digunakan untuk mengevaluasi kemampuan *model* dalam melakukan prediksi yang akurat terhadap kelas positif, dengan mempertimbangkan jumlah keseluruhan data yang teridentifikasi sebagai kelas positif. Nilai *precision* yang tinggi menunjukkan bahwa sebagian besar prediksi positif memang benar-benar relevan atau sesuai. Perhitungan presisi dilakukan dengan membagi jumlah prediksi positif yang tepat oleh total jumlah prediksi positif, Perhitungannya dapat dilakukan melalui persamaan (2.6) [23].

$$\boldsymbol{Precision} = \frac{\boldsymbol{Tp+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (2.6)$$

3. *Recall*

Recall adalah sebuah metrik yang mencerminkan kemampuan *model* dalam mengenali atau mengidentifikasi data yang sebenarnya terklasifikasikan sebagai positif. Dalam perhitungannya, *Recall* mengukur sejauh mana *model* mampu mengidentifikasi dengan benar jumlah data positif dari keseluruhan sampel yang sebenarnya memiliki label positif, Perhitungannya dapat dilakukan melalui persamaan (2.7) [23].

$$\boldsymbol{Recall} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FN}) \quad (2.7)$$

5. F1-Score

f1-score adalah sebuah parameter pengukuran yang mencerminkan hubungan antara presisi (*precision*) dan *recall* yang sebelumnya telah dihitung. Parameter ini memberikan gambaran tentang bagaimana *model* mencapai keseimbangan antara kemampuan dalam mengklasifikasikan dengan benar kelas positif *(Precision)* dan dalam mengidentifikasi secara komprehensif semua entitas yang sesungguhnya positif (*recall*), Perhitungannya dapat dilakukan melalui persamaan (2.8) [23].

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \quad (2.8)$$

***6.* Matrik Evaluasi Utama Dalam Klasifikasi**

True Positive (TP) mengacu pada kondisi positif yang berhasil diprediksi secara akurat oleh *model*, sebagaimana ditunjukkan dalam persamaan sebelumnya. Sementara itu, *True Negative* (TN) merepresentasikan kondisi negatif yang juga diprediksi dengan tepat. Sebaliknya, *False Positive* (FP) terjadi ketika *model* secara keliru memprediksi kasus negatif sebagai positif. Adapun *False Negative* (FN) merupakan kesalahan prediksi di mana kasus positif diklasifikasikan secara salah sebagai negatif.

2.3.12 Cross Validation

Cross Validation adalah teknik *resampling* data yang sering digunakan untuk memperkirakan performa model prediksi serta meningkatkan kinerja model melalui penyesuaian parameter [42]. Penggunaan Cross Validation

memungkinkan model untuk melakukan generalisasi terhadap kemampuan prediksi, sehingga hasil yang diperoleh tidak hanya baik pada data latih, tetapi juga pada data yang belum pernah dilihat sebelumnya. Penerapan Cross Validation juga dapat mengurangi risiko terjadinya *overfitting*.

Terdapat berbagai variasi dalam penerapan Cross Validation, namun *K*-Fold Cross Validation dianggap sebagai salah satu metode yang paling optimal karena memanfaatkan seluruh data untuk pelatihan dan validasi [43]. Metode ini menghasilkan evaluasi performa model yang lebih representatif, serta dapat mengurangi bias dan kesalahan dalam pengukuran. Dalam *K*-Fold Cross Validation, data dibagi menjadi *K* sub-sampel yang berukuran kurang lebih sama. Model kemudian dievaluasi sebanyak *K* kali, di mana pada setiap iterasi, satu sub-sampel digunakan sebagai data uji dan sisanya sebagai data latih. Hasil evaluasi dari setiap iterasi kemudian dirata-ratakan untuk memperoleh estimasi performa model secara keseluruhan [43]. Dalam penelitian ini, digunakan metode 10-Fold Cross Validation, yaitu data dibagi menjadi sepuluh bagian. Proses pelatihan dan pengujian dilakukan sebanyak sepuluh kali, sehingga seluruh data mendapat kesempatan untuk menjadi data uji.

2.3.13 SahabatAI

Gemma2 9B CPT Sahabat-AI adalah sebuah *model* bahasa besar (*Large Language model/LLM*) yang dikembangkan dan dioptimalkan untuk bahasa Indonesia serta berbagai dialeknya. *model* ini dikembangkan oleh PT GoTo Gojek Tokopedia Tbk dan AI Singapore, dan merupakan hasil pelatihan lanjutan dari *model* dasar Gemma2 9B CPT SEA-Lionv3. Proses pelatihan ini melibatkan penggunaan sekitar 50 miliar *token* untuk meningkatkan kemampuannya dalam pemahaman bahasa Indonesia [44]. Sahabat-AI merupakan ekosistem yang diinisiasi oleh perusahaan teknologi dan telekomunikasi Indonesia, yaitu GoTo Group dan Indosat Ooredoo Hutchison. *model* ini mendukung beberapa bahasa, termasuk Bahasa Indonesia, Bahasa Inggris, Bahasa Jawa, dan Bahasa Sunda. Gamma2 9B CPT Sahabat-AI v1 merupakan *model* bertipe *decoder* yang menggunakan *tokenizer* default yang juga digunakan pada Gamma-2-9B,

dengan panjang konteks mencapai 8192 *token*. Fitur ini memungkinkan *model* untuk memproses teks dalam jumlah yang lebih besar secara efisien, sehingga mendukung aplikasi beragam dalam analisis bahasa dan pemrosesan teks skala besar [44]. Fungsi penggunaan *LLM* Gamma2 9B CPT Sahabat-AI v1 pada penelitian ini untuk tahap *Deployment* digunakan untuk mendukung proses analisis dengan menyediakan ringkasan otomatis serta memberikan *insight* mengenai dampak dari setiap kalimat dan kata yang telah diklasifikasi.

2.3.14 Zero-shot

Zero-shot adalah pendekatan dalam pemrosesan bahasa alami (*Natural Language Processing* atau NLP) yang memanfaatkan kemampuan *model* bahasa besar (*Large Language model* atau LLM) untuk menyelesaikan suatu tugas hanya berdasarkan instruksi yang diberikan dalam bentuk teks (prompt), tanpa memerlukan pelatihan khusus atau contoh sebelumnya. Pendekatan ini memungkinkan *model* memahami dan menanggapi berbagai jenis permintaan secara langsung, berdasarkan pengetahuan luas yang telah diperoleh selama proses pelatihan[19].

2.3.15 User Acceptance Testing

User Acceptance Testing (*UAT*) adalah tahap evaluasi yang dilakukan untuk menilai apakah aplikasi atau sistem yang dikembangkan telah sesuai dengan kebutuhan serta ekspektasi pengguna. Proses ini biasanya dilaksanakan oleh pengguna akhir, dengan tujuan memastikan bahwa sistem benar-benar siap digunakan sebelum dilakukan perilisan. Hasil dari pengujian ini memberikan informasi penting bagi tim pengembang mengenai sejauh mana sistem berhasil memenuhi harapan pengguna [45]. *UAT* digunakan pada fase *Deployment* pada penelitian ini, di mana sistem yang telah dibangun yaitu website analisis sentimen berbasis *Streamlit* diperiksa dan diuji langsung oleh pengguna. Tujuannya adalah untuk memastikan bahwa seluruh fitur yang dikembangkan, seperti *Data Preparation*, *sentiment analysis*, *visualisasi*, dan *summary insight*, dapat berjalan dengan baik serta sesuai dengan kebutuhan analisis dari pengguna akhir.

2.3.16 Skala Likert

Skala pengukuran *Likert* merupakan hasil penciptaan dari Rensis *Likert*. Skala ini digunakan dalam penyusunan kuesioner dan sering dimanfaatkan dalam penelitian survei. Fungsinya adalah untuk mengukur pendapat dan sikap dari para responden. Responden diminta untuk mengisi kuesioner dengan menunjukkan tingkat persetujuan terhadap sistem yang telah dirancang [46]. *Skala Likert* dapat disusun dalam bentuk pilihan ganda atau *checklist*. Responden diminta memilih tingkat persetujuan terhadap suatu pernyataan dengan memilih salah satu opsi yang tersedia. Biasanya, disediakan lima tingkat pilihan dalam skala tersebut, masing-masing dengan skor tertentu sebagaimana ditunjukkan pada tabel 2.5 [46]:

Tabel 2. 5 Skor Skala Likert
Sumber : [46]

Sangat Tidak Setuju (STS)	Tidak Setuju(TS)	Netral (N)	Setuju(S)	Sangat Setuju(SS)
1	2	3	4	5

Perhitungannya hasil sekala *Likert* dapat dilakukan melalui persamaan (2.9) [46]:

$$PresentaseSkor = \frac{(SS*5)+(S*4)+(N*3)+(TS*2)+(STS*1)}{5*Jumlah\ Responden} * 100\% \quad (2.9)$$

Hasil skor persentase dapat digunakan untuk mengukur skala kepuasan responden dengan menggunakan berdasarkan tabel 2.6 [46]:

Tabel 2. 6 Persentase Skor Skala Likert
Sumber : [46]

Persentase skor	Jawaban
0% - 20%	Sangat Tidak Setuju
21% - 40%	Tidak Setuju
41% - 60%	Netral
61% – 80%	Setuju
81% - 100%	Sangat Setuju

2.4 *Tools* Dan *Software* Penelitian

2.4.1 Streamlit

Streamlit adalah *framework* berbasis *open-source* yang digunakan untuk membangun aplikasi web secara interaktif menggunakan *Python*. *Framework* ini dirancang agar mudah digunakan, terutama oleh pengembang yang tidak memiliki latar belakang dalam pemrograman web seperti *HTML* atau *CSS* [40]. Pendekatan *scripting* yang sederhana dan sintaks yang menyerupai penulisan kode *Python* pada umumnya, *Streamlit* memungkinkan proses pengembangan aplikasi menjadi lebih cepat dan efisien. Selain itu, *Streamlit* menyediakan berbagai elemen *user interface (UI)* seperti tombol, kolom *Input* teks, grafik visualisasi, dan tampilan tabel, yang mendukung pembuatan aplikasi interaktif secara dinamis tanpa konfigurasi yang kompleks [40].

2.4.2 Python

Python adalah bahasa pemrograman dapat diakses secara bebas dan dikembangkan oleh siapa saja, karena bersifat *open source*. Bahasa ini memiliki aplikasi luas dalam berbagai bidang, mulai dari pengembangan situs web, pengolahan data, hingga pembuatan permainan [47]. Keunggulan *Python* juga terletak pada ketersediaan beragam pustaka sumber terbuka yang komprehensif dan jelas. Berkat kelebihan ini, *Python* diakui memiliki kemampuan yang handal dalam menangani proyek-proyek seperti *Big Data, Data Mining, Data Science, Deep Learning*, dan yang sedang populer saat ini, yaitu *Machine Learning* [47].

2.4.3 Google Colaboratory

Google Collaboratory (*Google Colab*) adalah sebuah *platform* berbasis cloud yang digunakan untuk analisis data dan pengembangan *model Machine Learning*. *Platform* ini memungkinkan pengguna menulis dan mengeksekusi kode *Python* secara langsung, serta menggabungkannya dengan elemen rich text seperti grafik, gambar, HTML, LaTeX, dan lainnya dalam satu dokumen yang disimpan di Google Drive [48]. Salah satu keunggulan utama *Google Colab* adalah kemudahan dalam menjalankan kode tanpa perlu melakukan instalasi

perangkat lunak atau pengaturan lingkungan tambahan. Adapun beberapa keunggulan yang ditawarkan *Google Colab* antara lain:

1. Ketersediaan Beragam Library

Google Colab telah menyediakan berbagai pustaka *Machine Learning* dan analisis data yang telah diinstal sebelumnya, sehingga mempercepat proses pengembangan.

2. Penyimpanan Berbasis Cloud

Melalui integrasi dengan Google Drive, pengguna dapat menyimpan, mengakses, dan mengelola *file* dari berbagai perangkat secara fleksibel.

3. Fitur Kolaborasi

Google Colab mendukung kolaborasi antar pengguna, memungkinkan banyak orang untuk mengedit dan menjalankan kode secara bersamaan dalam satu dokumen.

2.4.4 Visual studio Code

Visual studio Code (VS Code) merupakan editor kode sumber yang bersifat open-source dan dapat digunakan pada berbagai *platform*, yang dikembangkan oleh Microsoft. Editor ini dirancang untuk memenuhi kebutuhan para pengembang perangkat lunak di era modern, dengan antarmuka yang sederhana, ringan, dan sangat dapat disesuaikan. VS Code mendukung berbagai bahasa pemrograman, dilengkapi dengan ekstensi yang beragam, serta integrasi dengan Git, yang secara keseluruhan dapat meningkatkan produktivitas pengembang dalam menulis kode. Beberapa fitur utama yang dimiliki oleh VS Code antara lain IntelliSense yang cerdas, kemampuan debugging secara langsung dari editor, serta ekosistem ekstensi yang kaya, menjadikannya pilihan yang banyak digunakan dalam berbagai jenis proyek, mulai dari pengembangan web, aplikasi cloud, hingga pengembangan lintas *platform* [34].