BAB II

LANDASAN TEORI

2.1 Penilitian Terdahalu

Penelitian yang telah dilakukan sebelumnya penting sebagai landasan dalam melaksanakan penelitian yang akan dilakukan. Tabel 2.1 merupakan beberapa penelitian terdahulu.

Tabel 2. 1 Penelitian Terdahulu

1.	Nama Jurnal	Buletin Sistem Informasi dan Teknologi Islam, Vol 4, No 4, (2023) [7]
	Author	A. Anugrah Aqsaa, Irawatia, Lukman Syafieb,
	Metode	Metode penelitian yang digunakan adalah Metode *Naïve Bayes* dan *Support Vector Machine*
	Permasalahan	Permasalahan pada penilitian adanya dugaan transaksi mencurigakan di Kementerian Keuangan (Kemenkeu) yang menjadi sorotan di media sosial, terutama di Twitter
	Hasil dan Kesimpulan	Hasil dan kesimpulan pada penilitian Naïve Bayes mendapatkan nilai akurasi sebesar 71,7%, presisi sebesar 55,2%, recall sebesar 45,3%, dan f1-score sebesar 44,8%, sedangkan pada *SVM* mendapatkan nilai akurasi sebesar 74%, presisi sebesar 87,8%, recall sebesar 49,1%, dan f1-score sebesar 49,8%.
2.	Nama Jurnal	Jurnal Mahasiswa Teknik Informatika, Vol. 8, No. 4 (2024) [8]
	Author	Yulius Bambang Seran, Supatman
	Metode	Penilitian ini menggunakan metode *Support Vector Machine*
	Permasalahan	Permasalahan pada penilitian adanya pro dan kontra terhadap kinerja kerja Presiden Joko Widodo
	Hasil dan Kesimpulan	Hasil dan kesimpulan pada penilitian menunjukkan bahwa model *Support Vector Machine* mencapai

		tingkat akurasi sebesar 66%. Model ini menunjukkan performa bagusdalam identifikasikelaspositif, dengan nilai recall98% dan f1-score77%
3.	Nama Jurnal	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, *Vol. 7, No. 1, (2023)* [9]
	Author	Denny Manuel Yeremia Sinurat, Dian Eka Ratnawati, Dwija Wisnu Brata
	Metode	Penilitian ini menggunakan metode *Naïve Bayes Classifier* dan *SMOTE*
	Permasalahan	Permasalahan pada penilitian ini reaksi masyarakat terhadap kebijakan kenaikan cukai rokok sebesar 10% yang diberlakukan mulai tahun 2023
	Hasil dan Kesimpulan	Hasil penelitian ini menunjukkan akurasi tertinggi sebesar 74%, yang dicapai menggunakan algoritma *Naïve Bayes* dengan data seimbang hasil dari metode *SMOTE.*
4.	Nama Jurnal	Techno.COM, Vol. 21, No. 4, (2022) [15]
	Author	Fefbiansyah Hasibuan, Wowon Priatna, Tyastuti Sri Lestari
	Metode	Pada penelitian metode algoritma klasifikasi *Naive Bayes*
	Permasalahan	Permasalahan pada penilitan ini kelangkaan minyak goreng yang terjadi di Indonesia, yang menyebabkan berbagai opini dari masyarakat di media twitter terkait pera Kementerian Perdagangan Republik
	Hasil dan Kesimpulan	Hasil dan kesimpulan pada penelitian menggunakan algoritma Naïve Bayes menunjukkan nilai akurasi sebesar 84,24%.
5.	Nama Jurnal	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 6, No. 5, (2022) [16]
	Author	Jasico Da Comoro Aruan, Bayu Rahayudi, Achmad Ridok
	Metode	Pada penelitian ini menggunakan metode *Support Vector Machine*
	Permasalahan	Permasalahan pada penilitian pentingnya menilai sentimen masyarakat terhadap layanan RSUD untuk meningkatkan mutu pelayanan kesehatan dan mendukung proses akreditasi RSUD.

	Hasil dan Kesimpulan	Hasil dan kesimpulan pada penilitian ini metode *Support Vector Machine (SVM)* menghasilkan nilai akurasi sebesar 88%. Selain itu, nilai recall sebesar 87,5%, precision sebesar 90%, dan f1-score sebesar 87,5%
6.	Nama Jurnal	Jurnal Ilmiah Edutic, Vol.7, No.1, (2020) [10]
	Author	Dedi Darwis, Eka Shintya Pratiwi, A. Ferico Octaviansyah Pasaribu
	Metode	Pada penelitian ini menggunakan metode *Support Vector Machine*.
	Permasalahan	Permasalahan pada penilitan ini memahami opini masyarakat terhadap kinerja Komisi Pemberantasan Korupsi (KPK RI), terutama dalam konteks pemberantasan tindak pidana korupsi.
	Hasil dan Kesimpulan	Hasil dan kesimpulan penelitian ini menunjukkan nilai akurasi sebesar 82% dengan menggunakan algoritma *Support Vector Machine*, di mana sentimen dengan label negatif mendominasi sebesar 77%, diikuti oleh label netral sebesar 25%, dan label positif sebesar 8%.
7.	Nama Jurnal	Jurnal Teknologi Informasi dan Komunikasi, 6(4) (2022) [17]
	Author	Ali Ahmad, Windu Gata
	Metode	Pada penelitian ini menggunakan metode *Support Vector Machine*
	Permasalahan	Permasalahan pada penilitan ini memahami sentimen masyarakat Indonesia terhadap teknologi metaverse, terutama mengingat dampak signifikan yang ditimbulkan oleh perkembangan teknologi
	Hasil dan Kesimpulan	Hasil penelitian ini menunjukkan teknologi metaverse yang menunjukkan 66% bersikap netral, 17% negatif dan 16% positif, sedangkan dari hasil pengujian dengan algorithma *SVM* didapatkan hasil performansi *SVM* sebesar 87% .
8.	Nama Jurnal	Jurnal Informatika dan Teknik Elektro Terapan, Vol. 10 No. 1, (2022) [18]
	Author	Dianati Duei Putri, Gigih Forda Nama, Wahyu Eko Sulistiono
	Metode	Pada penelitian ini menggunakan metode *Naïve Bayes*

	Permasalahan	Permasalahan pada penelitian ini bagaimana masyarakat menyampaikan opini atau sentimen mereka terhadap kinerja Dewan Perwakilan Rakyat (DPR) melalui media social twitter
	Hasil dan Kesimpulan	Hasil penelitian ini menunjukkan bahwa algoritma *naive bayes* mendapatkan accuracy score sebesar 80%
9.	Nama Jurnal	Information Technology and Engineering (2023) [19]
	Author	Raymond Oetama, Yanfi Yanfi, Masagus M. Ikhsan Assiddiq
	Metode	Pada penelitian ini menggunakan algoritma *Support Vector Machine*
	Permasalahan	Permasalahan pada penelitian ini kasus penipuan yang melibatkan platform trading online seperti Binomo, yang sempat populer di Indonesia berkat promosi dari beberapa influencer.
	Hasil dan Kesimpulan	Hasil penelitian ini menunjukkan bahwa algoritma *SVM* memiliki akurasi sebesar 86% untuk data training dan 80% untuk data testing
10.	Nama Jurnal	*JOURNAL OF MULTIDISCIPLINARY ISSUES, 2(2) 1 - 21 (2022)* [20]
	Author	Vinson Phoan, Johan Setiawan
	Metode	Pada penelitian ini menggunakan algoritma *Support Vector Machine*
	Permasalahan	Permasalahan pada penelitian adanya fenomena pelecehan seksual, yang melibatkan tindakan verbal dan nonverbal dengan unsur pemaksaan terhadap korban.
	Hasil dan Kesimpulan	Hasil penelitian menggunkan *Support Vector Machine* menunjukkan bahwa akurasi 55,14% dan data pelecehan seksual yang dikumpulkan pada 16 Maret 2022, dengan 287 data yang diperoleh dari situs Twitter

Pada tabel 2.1 berisi referensi dari studi-studi sebelumnya yang menjadi landasan atau dasar bagi penelitian yang sedang dilakukan. Dalam Analisis sentimen, terdapat beberapa algoritma machine learning yang populer, yaitu *Naive Bayes* dan *Support Vector Machine*. Kedua algoritma tersebut menjadi populer

karena menghasilkan nilai akurasi yang tinggi. Terdapat pada penelitian [15] menghasilkan nilai akurasi *Support Vector Machine* sebesar 88%. Pada penelitian [14] menghasilkan nilai akurasi *Naive Bayes* sebesar 84.24%. Penelitian [11] membahas tentang objek Kementerian keuangan yang menghasilkan nilai akurasi *Support Vector Machine* sebesar 74% dan nilai akurasi *Naive Bayes* sebesar 71.7% dengan pelabelan secara manual. Model *Naïve Bayes* yang menggunkan teknik *SMOTE* [13] menghasilkan nilai akurasi sebesar 74% dengan pelabelan data secara manual.

Fokus penelitian ini adalah pada analisis sentimen mengenai opini publik terhadap pemerintahan khususnya Kementerian Keuangan. Penelitian ini membandingkan dua algoritma yaitu *Support Vector Machine* dan *Naive Bayes*. Dataset yang digunakan mengalami ketidakseimbangan, teknik SMOTE diterapkan untuk menyimbangkan data dan meningkatkan performa model prediksi. Hasil penelitian akan divisualisasikan dalam bentuk *dashboard* pada sebuah *website* yang menampilkan sentimen analisis dari hasil opini dan persepsi masyarakat terhadap Kementerian Keuangan di Indonesia.

2.1 Tinjauan Teori

2.1.1 X

X merupakan *platform* media sosial ini memungkinkan pengguna untuk membagikan pesan singkat yang disebut *"tweet.* Setiap *tweet* dibatasi hingga 280 karakter dan dapat mencakup teks, gambar, video, atau tautan. X digunakan oleh jutaan orang di seluruh dunia untuk menyampaikan pemikiran, berita, opini, dan informasi lainnya secara *real-time*. Analisis sentimen *X* melibatkan mengumpulkan, memproses, dan menganalisis *tweet* untuk menilai apakah opini atau perasaan yang disampaikan dalam *tweet* tersebut termasuk dalam kategori positif, negatif, atau netral terhadap topik tertentu [21].

Opini publik mengacu pada pandangan, pendapat, atau sikap yang dimiliki oleh sekelompok orang dalam masyarakat terhadap suatu topik atau entitas. Dalam penggunaan *X* untuk memprediksi opini publik terhadap Kementerian Keuangan analisis sentimen dilakukan untuk menilai dan memahami respons

atau tanggapan publik yang diungkapkan melalui *tweet* terkait Kementerian Keuangan. Menganalisis *tweet* dapat memperoleh wawasan yang berguna untuk mengukur respons publik dan memprediksi tren opini yang berkaitan dengan kinerja atau kebijakan Kementerian Keuangan.

2.1.2 Analisis Sentimen

Analisis sentimen adalah menganalisis teks digital untuk menentukan apakah pesan tersebut memiliki emosional yang positif, negatif, atau netral Analisis sentimen bermanfaat untuk memahami pendapat yang terkandung dalam ulasan atau komentar yang digunakan oleh pengguna internet. Berbagai penelitian dalam bidang analisis sentimen telah dilakukan oleh sejumlah peneliti sebelumnya dengan tujuan untuk mendapatkan informasi dari suatu kumpulan data mengenai penilaian subjek yang diteliti. [22]. Proses analisis sentimen dengan pengumpulan data teks yang relevan dengan topik yang ingin dianalisis. Data tersebut kemudian diproses untuk menghilangkan noise, seperti tanda baca, kata-kata yang tidak relevan, dan emotikon. Setelah itu, dilakukan analisis sentimen untuk menentukan apakah sentimen dalam teks tersebut bersifat positif, negatif, atau netral.

Teknik yang digunakan bisa berupa analisis kata kunci, analisis statistik, atau menggunakan model *Machine Learning*. Hasil dari analisis sentimen kemudian diinterpretasikan untuk memahami kesimpulan atau tren umum yang terkait dengan topik atau entitas yang dianalisis. Analisis sentimen juga membantu dalam mendeteksi isu-isu yang penting bagi masyarakat, mengukur tingkat kepercayaan publik, serta mengevaluasi efektivitas komunikasi dan strategi pemerintah [23]. Pemahaman yang lebih baik tentang sentimen masyarakat, lembaga atau kementerian dapat merespons secara lebih efektif dan memperbaiki kinerja untuk kepentingan publik.

2.1.3 Text Preprocessing

Preprocessing teks adalah langkah pertama dalam menghilangkan gangguan atau *noise* pada data teks agar dapat diolah lebih lanjut dengan lebih efisien [24]. Proses *preprocessing* teks meliputi serangkaian rutinitas dan langkah untuk menyiapkan data agar dapat digunakan dalam fungsi pengambilan data dari sistem penambangan teks. Tujuan dari langkah-langkah ini adalah untuk melakukan pembersihan dan persiapan data teks agar dapat dijalankan dengan lebih efisien dalam proses analisis atau penambangan informasi, di antaranya [25]:

1. *Case Folding*

   Proses mengubah teks dalam kalimat menjadi huruf kecil dilakukan meskipun terdapat nama kota dan entitas lainnya. Hal ini disebabkan karena ulasan yang diperoleh dari pengumpulan data tidak memiliki format yang seragam seperti ulasan lainnya.

2. *Cleaning*

   Proses menghilangkan teks yang mengandung awalan tertentu, tag *HTML*, tanda baca, serta angka menggunakan ekspresi reguler (regex).

3. *Stop Word Removal*

   Proses penghapusan kata yang tidak relevan dalam kalimat dilakukan berdasarkan daftar *Stop Word* dalam bahasa Inggris atau bahasa Indonesia.

4. *Tokenization*

   Proses membagi teks menjadi unit-unit kecil yang disebut token, seperti kata-kata, frasa, atau kalimat, untuk memudahkan analisis dan pemrosesan data.

5. *Stemming*

   Proses mengubah kata-kata ke dalam bentuk dasarnya untuk mengenali kata-kata yang memiliki arti yang sama.

2.1.4 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) merupakan salah satu teknik yang digunakan dalam pemrosesan bahasa alami dan analisis teks

untuk mengevaluasi pentingnya sebuah kata dalam suatu dokumen atau dalam keseluruhan korpus teks [26]. Metode ini menghitung skor untuk kata-kata berdasarkan dua faktor utama, yaitu seberapa sering kata tersebut muncul dalam dokumen (TF) dan seberapa umum kata tersebut muncul dalam seluruh korpus teks (IDF). Dengan menggabungkan nilai TF dan IDF, skor TF-IDF memberikan informasi yang lebih tepat mengenai pentingnya suatu kata dalam konteks teks tertentu.

2.1.5 Wordcloud

Wordcloud merupakan cara visual untuk menampilkan kata-kata yang sering muncul dalam sebuah teks. Semakin sering kata tersebut muncul, semakin besar ukurannya dalam wordcloud. Visualisasi ini memudahkan melihat kata-kata kunci atau tema utama dari teks secara cepat. Wordcloud sering digunakan dalam analisis teks untuk memberikan gambaran cepat mengenai kata-kata yang dominan atau penting dalam sebuah dokumen, artikel, atau kumpulan data seperti tweet. Alat ini membantu mengidentifikasi tema atau topik utama dari teks secara visual dan mudah dipahami [27].

2.1.6 Scraping

Scraping merupakan proses mengumpulkan data dari sebuah website secara otomatis menggunakan program atau *script*. Tujuannya adalah Untuk mengambil informasi yang ditampilkan di halaman web, seperti teks, gambar, tanpa harus menyalin secara manual. Data ini kemudian bisa digunakan untuk berbagai keperluan, seperti analisis, penelitian, atau pembuatan aplikasi. Scraping biasanya dilakukan dengan bantuan alat atau bahasa pemrograman, seperti Python menggunakan library seperti BeautifulSoup atau Scrapy [28].

2.1.7 SMOTE

SMOTE singkatan dari (*Synthetic Minority Over-sampling Technique*) teknik oversampling dalam *Machine Learning* untuk menangani ketidakseimbangan data. Teknik ini membuat data sintetis baru dari kelas

minoritas dengan memilih titik data minoritas dan menambah titik sintetis di antara tetangga terdekatnya. Hal ini berguna untuk menambah jumlah data pada kelas minoritas serta mencegah saat melatih model pada dataset yang tidak seimbang [29]. *SMOTE* dapat meningkatkan jumlah sampel pada kelas minoritas tanpa melakukan duplikasi data.

2.1.8 Confusion Matrix

Confusion matrix merupakan tabel yang digunakan untuk melihat seberapa baik model klasifikasi dalam memprediksi sesuatu. Tabel ini menunjukkan empat kemungkinan hasil. Pertama, *True Positive (TP)* adalah ketika model benar memprediksi sesuatu sebagai positif. Kedua, *True Negative (TN),* ketika model benar memprediksi sesuatu sebagai negatif. Ketiga, False Positive (FP), saat model salah memprediksi positif padahal sebenarnya negatif. Terakhir, False Negative (FN), ketika model salah memprediksi negatif padahal sebenarnya positif. Confusion matrix bisa memahami kesalahan yang dibuat oleh model dan menghitung seberapa akurat model [30] .

2.1.9 Accuracy

Akurasi merupakan ukuran kinerja model klasifikasi yang menunjukkan seberapa sering model memberikan prediksi yang benar. Akurasi dihitung dengan cara membagi jumlah prediksi yang benar, baik *True Positive* maupun *True Negative,* dengan total prediksi yang dibuat. [31].

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

Rumus 2. 1 Rumus Akurasi

2.1.10 Precision

Presisi adalah metrik yang digunakan dalam evaluasi model klasifikasi untuk menunjukkan seberapa tepat model saat memprediksi kelas positif. Presisi mengukur berapa banyak prediksi positif yang benar dibandingkan dengan semua prediksi positif yang dibuat oleh model [32].

$$Precision = \frac{TP}{TP + FP}$$

Rumus 2. 2 Rumus Presisi

2.1.11 Recall

Recall merupakan metrik evaluasi untuk model klasifikasi yang menunjukkan seberapa efektif model dalam mengidentifikasi semua contoh positif yang sebenarnya. Recall mengukur berapa banyak kasus positif yang sebenarnya (*True Positive*) berhasil diprediksi dengan benar oleh model dibandingkan dengan seluruh jumlah kasus positif yang ada [33], [34]

$$Recall = \frac{TP}{TP + FN}$$

Rumus 2. 3 Rumus Recall

2.1.12 F-measure

F-measure dikenal sebagai F1-score, adalah metrik yang menggabungkan presisi dan recall menjadi satu angka untuk memberikan gambaran yang seimbang tentang kinerja model klasifikasi. F1-score digunakan Ketika ingin menyeimbangkan antara presisi dan *recall* terutama jika ada ketidakseimbangan antara jumlah kelas positif dan negative [34].

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

Rumus 2. 4 Rumus F-Measure

2.2 Algoritma dan *Framework*

2.2.1 *Machine Learning*

Machine Learning merupakan bidang yang terus berkembang, berfokus pada pengenalan pola dan pembelajaran berbasis komputer dalam kecerdasan buatan. *Machine Learning* digunakan berbagai algoritma pembelajaran, baik yang bersifat diawasi (*supervised*) maupun tidak diawasi (*unsupervised*) untuk memprediksi dan mendukung pengambilan

keputusan otomatis berdasarkan sekumpulan data [35]. Tujuan utama *ML* adalah memberikan kemampuan kepada sistem komputer untuk belajar dari pengalaman, memahami pola data, dan meningkatkan kinerjanya seiring waktu tanpa pemrograman eksplisit. *Machine Learning* tidak hanya diinstruksikan untuk mengeksekusi tugas tertentu, tetapi juga diberikan kemampuan untuk belajar dari pengalaman. Proses implementasi *Machine Learning* dapat disusun ke dalam tiga tahap, yaitu persiapan, pemodelan data, dan evaluasi.

1. Persiapan

   Tahap persiapan dimulai dengan pemilahan atribut atau parameter yang akan digunakan dalam pemodelan *Machine Learning*. Proses ini melibatkan pemilihan variabel atau fitur yang signifikan untuk memastikan kontribusi maksimal dalam pembentukan model yang akurat.
2. Pemodelan data

   Tahap pemodelan data adalah representasi matematis dari hubungan antara fitur dan label. Selama tahap pelatihan, model mempelajari parameter dan karakteristik dari data latihan.
3. Evaluasi.

   Tahap evaluasi menggunakan data yang tidak terlihat sebelumnya untuk memastikan bahwa model dapat memberikan prediksi yang akurat dan dapat diandalkan pada situasi dunia nyata.

2.2.2 Naive Bayes

Naive Bayes merupakan metode klasifikasi sederhana yang dapat mengestimasi probabilitas dengan menggabungkan variasi dan frekuensi nilai dari dataset yang tersedia. Algoritma ini menggunakan teorema Bayes untuk memperkirakan probabilitas atribut yang independen satu sama lain, diberikan nilai pada variabel kelas. *Naive Bayes* berdasarkan asumsi sederhana bahwa nilai atribut secara bersyarat saling bebas jika nilai *output* telah diketahui [36]. Nilai *output* sudah diketahui, probabilitas mengamati

bersama-sama adalah hasil kali dari probabilitas individu. Keunggulan *Naive Bayes* terletak pada kebutuhan jumlah data pelatihan yang relatif kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi [37]. Metode ini sering memberikan kinerja yang baik dalam kebanyakan situasi dunia nyata yang kompleks dibandingkan dengan ekspektasinya. Prediksi *Bayes* didasarkan pada teorema Bayes sebagai berikut :

$$\mathrm{P}(X) = \frac{P(H)\ X\ P(H)}{P(x)}$$

Rumus 2. 5 Rumus Naive Bayes

Ket:

- X : Data dengan class yang belum diketahui
- H : Hipotesis data merupakan suatu class spesifik
- P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)
- P(H) : Probabilitas hipotesis H (prior probabilitas)
- P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
- P(X) : Probabilitas X

2.2.3 Support Vector Machine (SVM)

SVM (*Support Vector Machine*) adalah suatu algoritma yang dikenal baik karena mampu menghasilkan solusi yang optimal dalam melakukan klasifikasi [38]. Algoritma ini diperkenalkan oleh Vapnik sebagai model *Machine Learning* berbasis kernel yang dapat digunakan untuk klasifikasi dan regresi. *SVM* bekerja dengan membangun *hyperplane* optimal atau batas keputusan dalam ruang fitur, yang memisahkan berbagai kelas data. Pendekatan ini memungkinkan *SVM* untuk memberikan solusi klasifikasi yang baik, terutama dalam konteks data yang tidak linier dan memiliki dimensi tinggi.

Algoritma *Support Vector Machine (SVM)* digunakan untuk menemukan *hyperplane* terbaik dalam ruang N-dimensi yang dengan jelas

memisahkan titik data. *Hyperplane* merupakan suatu fungsi yang berperan sebagai pemisah antara kelas. Fungsi utama dari *Support Vector Machine (SVM)* adalah untuk memisahkan data menjadi dua kelas yang berbeda dengan menggunakan *hyperplane*. *Hyperplane* ini bisa berupa garis atau bidang yang memiliki margin maksimum, yang memisahkan kedua kelas. tersebut secara optimal. Rumus *SVM* untuk klasifikasi sebagai berikut:

$$Largef(x) = sign\left(sum_{i=1}^{n} yi\ alpha_{\ i}\ K(x_{i,}x\right) + b)$$

Rumus 2. 6 Rumus Support Vector Machine

Ket:

- (x): fungsi prediksi
- x: vektor fitur input
- y: label kelas (+1 atau -1)
- α: vektor bobot
- K(xi,x): fungsi kernel yang menghitung jarak antara dua vektor fitur
- b: bias

2.2.4 *CRISP-DM*

CRISP-DM atau *Cross Industry Standard Process* for *Data mining*, adalah suatu standar dalam pemrosesan *data mining* yang telah dikembangkan. Dalam standar ini, data melewati serangkaian fase yang terstruktur dan jelas, mengikuti metodologi yang efisien . Metodologi ini digagas oleh *CRISP-DM* Consortium, sebuah kelompok yang menghasilkan standar industri yang diterima secara luas dalam bidang *data mining* dan analisis data. Metodologi ini terdiri dari enam tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment*. Metodologi ini melibatkan enam tahapan yang dapat diuraikan sebagai berikut [39]:

1. *Business Understanding*

Pada tahap ini, beberapa kegiatan melibatkan pemahaman kebutuhan dan tujuan dari perspektif bisnis. Selanjutnya, pengetahuan diartikan ke dalam bentuk pendefinisian masalah dalam *data mining*, dan kemudian merumuskan rencana serta strategi untuk mencapai tujuan *data mining*.

2. *Data Understanding*

   Tahap ini dimulai dengan pengumpulan data, dilanjutkan dengan deskripsi data, dan mengevaluasi kualitas data. Data dijelajahi untuk memahami karakteristiknya, pola yang mungkin ada, serta masalah atau kekurangan yang perlu diatasi.

3. *Data Preparation*

   Pada tahap ini, langkah-langkah melibatkan pembentukan dataset akhir dari data mentah. Beberapa kegiatan yang dilakukan mencakup pembersihan data, pemilihan data (*Data Selection*) untuk rekaman dan atribut, serta transformasi data (*Data Transformation*). Semua langkah ini bertujuan untuk menyediakan input yang bersih dan relevan untuk proses pemodelan berikutnya.

4. *Modelling*

   Tahap ini melibatkan pemilihan dan penerapan model *data mining* yang sesuai dengan tujuan proyek. Model ini dapat mencakup teknik seperti regresi, klasifikasi, clustering, atau yang lainnya.

5. *Evaluation*

   Tahap ini melibatkan evaluasi kinerja pola yang dihasilkan oleh algoritma. Parameter untuk evaluasi komparatif algoritma mencakup *Confusion Matrix*, yang mengacu pada nilai akurasi, presisi, dan *recall*.

6. *Deployment*

   Setelah model berhasil dievaluasi, langkah terakhir melibatkan implementasi model ke dalam lingkungan produksi. Pada tahap ini, dilakukan pembuatan laporan dan artikel jurnal dengan menggunakan model yang telah dihasilkan.

2.2.5 *SEMMA*

SEMMA merupakan singkatan dari *Sample, Emplore, Modify, Model*, dan *Assess*, sebuah metodologi yang dikembangkan oleh SAS Institute. Metode ini dirancang untuk membantu pengguna dalam melakukan proses *data mining* dengan lebih efisien. Dengan lima tahapan yang jelas yaitu *Sample, Explore, Modify, Model*, dan *Assess*. *SEMMA* memberikan kerangka kerja yang mudah dipahami dan digunakan untuk memprediksi variabel-variabel yang relevan dalam proyek *data mining*. Tahapan-tahapan ini masing-masing memiliki peran uniknya sendiri dalam keseluruhan proses dan memberikan manfaat yang signifikan dalam mengelola proyek *data mining* dengan efektif [40].

2.2.6 *KDD*

Knowledge Discovery in Database Process (KDD) merupakan metode yang digunakan dalam *data mining* untuk menemukan informasi berharga dan pola yang tersembunyi dalam data. Definisi KDD oleh Fayyed et al. (1996) menggambarkan *KDD* sebagai proses menggunakan teknik *data mining* untuk mengidentifikasi pola yang signifikan, melibatkan algoritma untuk mengekstrak pola dari data. Dunham (2003) merangkum proses *KDD* dalam beberapa tahapan, yaitu seleksi data, pra-proses data, transformasi data, *data mining*, dan interpretasi dan evaluasi. Seleksi data melibatkan pemilihan data yang relevan, pra-proses data adalah pembersihan dan integrasi data, transformasi data mengubah format data, *data mining* adalah inti proses *KDD* dengan penggunaan algoritma, dan tahap terakhir adalah interpretasi dan evaluasi hasil untuk memahami temuan dan mengevaluasi [41].

2.3 Software dan *Tools* yang digunakan

2.3.1 *Google Colab*

Colaboratory atau *Colab* merupakan produk yang dikembangkan oleh *Google Research*. *Colab* memungkinkan pengguna untuk menulis dan menjalankan kode Python secara bebas melalui browser, dan sangat cocok untuk keperluan *Machine Learning*, analisis data, serta pembelajaran. Secara teknis, *Colab* adalah layanan notebook *Jupyter* yang dihosting dan dapat digunakan tanpa memerlukan konfigurasi tambahan. Selain itu, *Colab* juga memberikan akses gratis ke sumber daya komputasi, termasuk GPU. [42].

Google *Colab* menyediakan fungsi kolaborasi yang memungkinkan beberapa pengguna untuk bekerja bersama dalam satu notebook secara real-time, memudahkan kerja tim pada proyek analisis sentimen. Integrasi dengan Google Drive juga menyederhanakan manajemen proyek, sementara tersedianya berbagai library *Python* seperti *Pandas, NumPy, Matplotlib,* dan *Tweepy* mempermudah pengguna dalam analisis dan visualisasi data. Dokumentasi yang komprehensif dan tutorial online juga membantu pengguna memahami dan memanfaatkan fitur-fitur Google *Colab* dengan lebih baik, menjadikannya opsi populer untuk analisis sentimen dan pembelajaran mesin lainnya.

2.3.2 *Python*

Python adalah bahasa pemrograman tingkat tinggi yang banyak digunakan di berbagai bidang, seperti pengembangan *web*, analisis data, kecerdasan buatan, dan pengembangan perangkat lunak. Bahasa ini dikenal karena sintaksisnya yang mudah dimengerti dan dipelajari. *Python* menjadi pilihan yang populer bagi para pengembang karena kemudahannya dalam mengelola kode [43]. Dalam konteks analisis sentimen, *Python* sangat berguna karena kemampuannya dalam pengolahan data, analisis statistik, dan pemodelan. Beberapa library yang sering digunakan seperti Pandas untuk manipulasi data, *NLTK* untuk

pemrosesan bahasa alami, dan Scikit-learn untuk pemodelan *Machine Learning*, semuanya dapat diintegrasikan dengan *Python* untuk melakukan analisis sentimen dengan efektif. Dengan menggunakan kombinasi library dan tools ini, para analis dapat membersihkan data, menganalisis sentimen teks, membangun model klasifikasi, dan menghasilkan visualisasi untuk mendukung interpretasi hasil analisis. *Python* memberikan fleksibilitas dan kekuatan yang dibutuhkan untuk mengelola dan menganalisis data teks dengan efisien dalam konteks analisis sentiment.