BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terdahulu menjadi pedoman untuk melaksanakan penelitian tugas akhir. Penelitian terdahulu merupakan riset yang telah dilakukan pada bidan keilmuan yang sama dengan penelitian. Berikut penelitian terdahulu yang digunakan sebagai landasan penelitian tugas akhir.

Tabel 2.1 Penelitian Terdahulu

Penelitian Terdahulu	Penulis	Metode	Kelebihan	Kekurangan
Sentiment Analysis For Tiktok Shop's Closure In Indonesia Using Naive Bayes Models And NLP Journal of Theoretical and Applied Information Technology (2024) (Q4)	Henoch Juli Christanto, Steven Sondra Allen Widodo, Christine Dewi, Yerik Afrianto Singgalen, Dalianus Riantama, Andri Dayarana K. Silalahi.	Sentiment analysis terhadap penutupan Tiktok Shop menggunakan variasi Naïve Bayes dan vectorization. Kombinasi TextBlob dan CountVectorizer menghasilkan akurasi paling tinggi yaitu 86.60%. [28]	Menggunakan data Twitter untuk mengidentifikasi sentimen positif,negatif, dan netral. Menggunakan pembobotan TF-IDF dan Countvectorizer.	Analisis sentimen berfokus pada mengklasifikasikan sentimen positif, negatif, dan netral pada data twitter
Sentiment Analysis of the TikTok Tokopedia Seller Center Application Using Support	Faddilla Aulia Dara, Irfan Pratama	Sentiment analysis pada data ulasan aplikasi Tiktok Tokopedia Seller Center menggunakan	Menggunakan data ulasan aplikasi Tiktok Tokopedia Seller Center. Menggunakan Inset Lexicon	Menggunakan data imbalance dengan mayoritas kelas negatif mengakibatkan kelas positif memberikan

Vector Machine (SVM) and Naive Bayes Algorithms International Journal Software Engineering and Computer Science (IJSECS) (2025) (SINTA 4)		algoritma SVM dan Naïve Bayes pada sentimen negatif, positif, dan netral. Akurasi SVM lebih tinggi daripada Naïve Bayes. [45]	untuk labeling sentiment.	performa yang rendah
Sentiment Analysis of TikTok Shop Closure in Indonesia on Twitter Using Supervised Machine Learning Journal Of Electronics Electromedical Engineering And Medical Informatics (2024) (SINTA 2)	Noor Zalekha Al Habesyah, Rudy Herteno , Fatma Indriani , Irwan Budiman , Dwi Kartini	Sentiment analysis terhadap penutupan Tiktok Shop menggunakan SVM, Random Forest, Decision Tree, dan H2O serta SMOTE. Hasil menunjukan performa machine learning lebih baik dengan SMOTE. [46]	Menggunakan data Twitter untuk mengidentifikasi sentimen. Menyinggung peningkatan performa model klasifikasi dengan menggunakan split training dan testing data, oversampling dengan SMOTE, dan Cross Validation.	Terdapat model dengan performa klasifikasi yang lebih buruk dari menebak acak akibat jumlah data bersih yang digunakan kecil (3000 baris).
Sentiment Analysis of Public Acceptance of Covid-19 Vaccines Types in Indonesia using Naïve Bayes, Support	Dinar Ajeng Kristiyanti, Sri Hardani	Sentiment analysis penerimaan tipe vaksin COVID -19 pada masyarakat menggunakan algoritma Naive Bayes (NB), Support	Menggunakan data Twitter dengan keyword. Pembobotan data menggunakan TF-IDF. Labeling sentiment	Parameter *setting* mode klasifikasi SVM tidak ditentukan berdasarkan metode optimasi.

Vector Machine, and Long Short-Term Memory (LSTM) Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) (2023) (SINTA 2)		Vector Machine (SVM), dan Long Short-Term Memory (LSTM). Model klasifikasi SVM menghasilkan akurasi tertinggi pada 84.89%. [47]	menggunakan Vader. Menggunakan Wordcloud untuk menggambarkan vaksin berdasarkan sentimen positif dan negatif.	
Sentiment Analysis Of Comments On Sexual Harassment In Colleges On Four Popular Social Journal of Multidisciplinary Issues (2022) (GARUDA)	Vinson Phoan a, Johan Setiawan	Sentiment analysis terhadap respon masyarakat / netizen terhadap kasus kekerasan seksual di UMN menggunakan FastText dan SVM. Model SVM menghasilkan 55.14% akurasi.[48]	Menggunakan data empat media sosial Twitter, Instagram, Medium, dan Line Today. Menggunakan Word Embedding FastText. Labeling sentiment dilakukan secara manual oleh pelatih yang ditentukan.	Menggunakan satu model klasifikasi teks.
Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification (IJACSA) International Journal of Advanced	Slamet Riyanto, Imas Sukaesih Sitangga ng, Taufik Djatna, Tika Dewi Atikah	Menggunakan data *unbalanced multi-class text* untuk menemukan metrik evaluasi Precision (P), Recall (R), dan F1-score (F1) yang menghasilkan performa	Meyinggung *imbalanced class* dan menggunakan SMOTE. Data yang digunakan memiliki karakteristik undersampling, oversampling, dan synthetic.	Data yang digunakan bukan berupa data media sosial.

Computer Science and Applications (2023) (Q3)		lebih baik berdasarkan algoritma yang digunakan Multinomial Naive Bayes, K-Nearest Neigbors, Support Vector Machine, Random Forest, and Long Short-Term Memory. Hasil komparasi menunjukan metrik evaluasi yang paling cocok disesuaikan dengan kebutuhan.[49]	Menggunakan GridSearch untuk hyperparamters model klasifikasi.	
Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms Electronic Commerce Research (2025) (Q1)	Laleh Davoodi, József Mezei, Markku Heikkilä	Aspect-based sentiment classification menggunaka deep learning model pada data online review dan menghasilkan 14 topik serta RoBERTa menghasilkan akurasi di atas 90%.[33]	Menggunakan data ulasan di situs *ecommerce* Menggunakan aspect-based sentiment analysis yang terdiri dari ekstraksi aspek (AE) dan klasifikasi sentimen aspek (ASC).	Hasil model yang diperoleh belum tentu digunakan secara general pada set data ulasan pengguna
(3) Aspect Based Sentiment Analysis: Feature	Shakirah Mohd Sofi, Ali Selamat	Meningkatkan sentiment analysis terkait kasus corona virus	Menggunakan metode aspect based sentiment analysis dengan teknik *word*	

Extraction Using Latent Dirichlet Allocation (Lda) And Term Frequency - Inverse Document Frequency (Tf-Idf) In Machine Learning (Ml) Malaysian Journal of Information and Communication Technology (MyJICT) (2023)		menggunnakan LDA sebagai feature extraction dan SVM serta Naïve Bayes sebagai classifier. Akurasi pada SVM menunjukan akurasi tertinggi pada 85%.[32]	*embedding* TFIDF dan Count Vectorizer. Menerapkan fitur ekstrasi dengan *keyword list analysis* dengan mengidentifikasikan frekuensi kata yang paling sering muncul. Evaluasi topik menggunakan coherence score.	
Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm Telematika 2021	M. Didik R. Wahyudi, Agung Fatwanto, Usfita Kiftiyani, M. Galih Wonoseto	Pemodelan topik pada headlines berita COVID di detik.com menggunakan LDA. LDA menghasilkan 3 topik perbulan selama 1 tahun data penelitian dan menghasilkan akurasi klasifikasi 82.4% dengan model Naïve Bayes. [42]	Menggunakan pemodelan topik LDA. Menggunakan data berisi judul berita selama satu tahun. Evaluasi topik menggunakan Coherence score pada set data per bulan. Akurasi pembagian topik atau kelas dievaluasi lagi menggunakan Naïve Bayes.	Data yang digunakan adalah judul berita dari satu media/penerbit online.
(2) A Topic Modeling	Roman Egger	Pemodelan topik pada	Data yang digunakan	Metode penelitian dapat

Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers on sociology 2022	and Joanne Yu	saat pandemi tentang perjalanan saat pandemi menggunakan LDA, NMF, BERTopic, dan Top2Vec. Hasil menunjukan NMF dan BERTopic menghasilkan topik yang lebih koheren dan distribusi topik yang dihasilkan tidak tumpang tindih.[44]	adalah data media sosial Twitter. Pembobotan kata menggunakan TF-IDF. Evaluasi topik menggunakan coherence score cv dengan library Gensim. Label yang disematkan pada topik yang ditemukan adalah kata atau istilah yang paling besar pada topik tertentu.	digunakan pada data media sosial lainnya tetapi perlu diperhatikan karena data media sosial dipengaruhi oleh demografi pengguna dan kata-kata retoris.
(1) Does the Choice of Topic Modeling Technique Impact the Interpretation of Aviation Incident Reports? A Methodological Assessment MDPI Technologies 2025	Aziida Nanyonga, Keith Joiner, Ugur Turhan and Graham Wild	Pemodelan topik Latent Dirichlet Allocation (LDA), Bidirectional Encoder Representations from Transformers (BERT), Probabilistic Latent Semantic Analysis (pLSA), DAN Non-negative Matrik yang diterapkan pada laporan keselamatan	Data yang digunakan adalah data aviasi. Evaluasi topik NMF dan LDA menggunakan coherence score.	Data yang digunakan adalah data laporan aviasi.

		penerbangan. Hasil menunjukan NMF memberikan coherence score paling tinggi (0.7987) dan LDA memberikan generalization lebih baik dengan perplexity (-6.471). [50]		
Enhancing Aspect Based Sentiment Analysis with a Hybrid Model for Hindi Language Journal of Information Systems Engineering and Management (2025)	Vijay Kumar Soni, Dr. Smita Selot, Vibhoo Sharma, Senthil Athithan	Mengintegrasikan PoS ke dalam ekstraksi aspek dan klasfikasi aspek dengan model Hi-BERT. [34]	Meningkatkan ABSE dengan menerapkan PoS tagging pada saat ekstraksi aspek.	Data yang digunakan adalah data berbahasa hindi

Tabel 2.1 menunjukan penelitian dan studi tentang analisis sentimen dan *topic modeling* dalam menganalisis opini publik pada media sosial atau platform digital. Penelitian terdahulu [28] [46] [45] yang relevan pada objek aplikasi Tiktok Shop atau Tokopedia dan Tiktok Shop Seller Center menggunakan data ulasan aplikasi atau media sosial Twitter dengan variasi model klasifikasi sentimen. Fokus penelitian tersebut masih mengklasifikasikan sentimen positif, negatif, dan netral.

Penelitian lain seperti [47] [48] menunjukan penggunaan model klasifikasi sentimen pada isu sosial, ekonomi, dan kependudukan. Hasil dari penelitian terdahulu SVM mendominasi dengan memberikan performa dan akurasi yang palng

baik. Penelitian pada isu lain menunjukan metodelogi yang relevan untuk digunakan pada data media sosial.

Penelitian [40] memberikan informasi mengenai kebutuhan matriks evaluasi seperti F1-score, Precision, Recall, dan SMOTE untuk menangani data imbalanced. Cara *Cross Validation, Grid Search*, dan *split data train-test* juga merupakan cara untuk meningkatkan performa dan stabilitas model.

Penelitian [33] menggunakan *aspect-based sentiment analysis* yang terdiri tdari ekstraksi aspek (AE) dan klasifikasi sentimen aspek (ASC) untuk memperoleh tema atau isu dari umpan balik yang diberikan pelanggan secara online. Penelitian [32] menggunakan *topic modeling* LDA untuk ekstraksi serta teknik *word-embedding* TFIDF dan Count Vectorizer untuk membentuk corpus yang digunakan. Penelitian terdahulu [34]menggunakan PoS (*part-of-speech*) tagging untuk meningkatkan ABSA dengan mengintegrasikannya pada proses ekstraksi fitur.

Penelitian [50] [36] [42] menunjukan penggunaan *topic modeling* dan evaluasi hasil topik menggunakan *coherence score.* Pemodelan topik bergantung pada kumpulan kata pada kumpulan dokumen yang digunakan untuk mendapatkan topik yang koheren..

Penelitian terdahulu telah memberikan dasar metode untuk melakukan teknik analisis sentimen, pemodelan topik, penanganan data yang tidak seimbang, dan mengukur performa model. Pada penelitian tugas akhir ini, *Topic Modeling* LDA dan NMF menggunakan data ulasan pengguna aplikasi Tokopedia dan Tiktok Shop seller untuk menemukan topik sebagai aspek yang sering dibicarakan oleh pengguna dan sentimen pengguna terhadap topik atau aspek tersebut setelah integrasi layanan menggunakan SVM.

2.2 Teori tentang Topik Skripsi

2.2.1 Tokopedia dan Tik tok Shop Seller Center

Tokopedia dan TikTok Shop Seller Center adalah platform khusus yang dibuat untuk membantu penjual mengelola toko, produk, dan transaksi penjualan di TikTok Shop dan Tokopedia[51]. Aplikasi Tokopedia & TikTok Shop Seller adalah aplikasi pengelolaan toko online yang resmi beroperasi setelah kerja sama Tiktok Shop dan Tokopedia dibawah Pt

Tokopedia[16]. Aplikasi Seller Center membantu penjual untuk memantau penjualan, berinteraksi dengan pelanggan atau calon pelanggan, serta menerapkan strategi pemasaran dalam satu dashboard yang mudah diakses.

2.2.2 Text Classification

Klasifikasi teks merupakan salah satu metode *Natural Language Processing* (NLP) yang berguna untuk mengatur dan menganalisis teks dalam jumlah besar dan tidak terstruktur[52]. Model klasifikasi menerima *input* dalam bentuk numerik sehingga teknik klasifikasi teks melalui beberapa tahapan transformasi data sesuai kebutuhan klasifikasi. Data dalam bentuk teks dikonversi ke bentuk numerik menggunakan *feature exctraction* atau metode transformasi data ke dalam bentuk baru. *Feature extraction* pada *text mining* bertujuan memperoleh representasi numerik dari teks sebagai fitur model pembelajaran mesin [53]. *Feature extraction* seperti *word-embedding* menghasilkan representasi numerik yang diperoleh dari pembobotan kata berdasarkan frekuensi kemunculan kata pada kumpulan dokumen (*Bag-of-Words*) atau bobot kata diukur berdasarkan kemunculan bersama pada kondisi tertentu (Term Frequency-Inverse Document Frequency). Selain *word-embedding,* POS tagging (*part-of-speech tagging*) dapat digunakan sebagai fitur dalam klasifikasi teks[54]. POS membuat label kategori pada teks untuk menentukan label struktur tata bahasa seperti kata kerja, kata objek, dan kata subjek yang memberikan informasi mengenai struktur kalimat dan peran kata pada teks.

2.2.3 Topic Modeling

Pemodelan topik merupakan metode yang digunakan untuk menampilkan volume data yang besar dalam bentuk dimensi yang lebih sederhana. Tujuan pemodelan topik adalah untuk menyajikan tema tersembunyi, fitur utama, atau variabel yang tidak terlihat berdasarkan pola atau struktur pada data yang dipelajari. Pemodelan topik bekerja dengan mengidentifikasi kata dan frasa terpenting yang muncul bersamaan dalam

sekumpulan dokumen dan mengelompokkannya ke dalam topik-topik terpisah[55]. Topik yang ditemukan dapat digunakan untuk memahami subtema atau topik dari data teks dan untuk mengeksplorasi hubungan antar informasi yang berbeda.

Topik dari hasil pemodelan topik diukur berdasarkan kemudahan interpretasi atau hubungan semantik yang dapat ditangkap oleh manusia. Topik yang mudah interpretasi dan konsisten memiliki nilai koherensi atau *coherence score* yang tinggi. Nilai koherensi atau *coherence score* pada topik diperoleh dari menghitung nilai koheren pada masing-masing topik yang dihasilkan. Kata atau *term* dengan bobot tertinggi pada topik-topik diukur dan dicari kesamaan kata yang muncul. Perhitungan frekuensi kemunculan kata secara bersamaan pada dokumen tertentu menghasilkan nilai *coherence u_mass* sedangkan perhitungan nilai semantik menggunakan *sliding window* pada kata atau *term* menghasilkan nilai *coherence c_v* [56]. Nilai coherence Perhitungan nilai *coherence* bertujuan untuk mengurangi *human-readable* sebagai *input* evaluasi[57].

2.2.4 Analisis Sentimen

Analisis sentimen adalah salah satu teknik *Natural Language Processing* (NLP) yang mendeteksi sentimen positif atau negatif dalam data teks. Analisis sentimen sering digunakan untuk mendeteksi sentimen dalam data sosial atau memperoleh persepsi pelanggan[58]. Analisis sentimen berfokus pada polaritas teks atau mengklasifikasikan teks ke dalam kelas positif, negatif, dan netral serta mampu mendeteksi perasaan seperti emosi, atau ketertarikan. Data teks yang digunakan akan dilatih dan model akan menentukan klasifikasi teks menjadi positif, negatif, atau netral berdasarkan nilai yang diperoleh oleh model.

2.2.5 Aspect-based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) adalah teknik analisis sentimen yang mengidentifikasi fitur tertentu pada teks untuk mendapatkan analisis sentimen yang lebih spesifik[59]. ABSA terdiri dari identifikasi dengan

tema menggunakan pola kata tertentu pada dokumen berdasarkan kesamaan tema atau topik tertentu dan analisis sentiment yang diterapkan pada kelompok sehingga diperoleh distribusi kelompok sentimen terhadap kelompok aspek tertentu.

2.3 Teori tentang Framework/Algoritma yang digunakan

2.3.1 Machine Learning

Machine Learning memungkinkan komputer untuk belajar dari data serta membuat prediksi atau keputusan tanpa harus diprogram secara eksplisit[60]. Proses tersebut dilakukan dengan mengidentifikasi pola dalam data dan menggunakan pola tersebut untuk memproses data baru. *Machine learning* mampu mempelajari data yang yang membutuhkan *output* berdasarkan label pada data dan dapat menggunakan data yang tidak memiliki label dengan mempelajari pola atau struktur pada data tersebut. *Machine learning* dapat beradaptasi untuk meningkatkan tingkat akurasi dengan menerima lebih banyak data saat proses menjalankan tugas.

Metode *machine learning* membutuhkan evaluasi untuk mengetahui performa tugas yang dilakukan[49]. Metrik evaluasi adalah bentuk evaluasi yang memberikan ukuran terukur untuk menilai kinerja model pembelajaran mesin. Metrik yang paling umum digunakan untuk mengukur kinerja model meliputi *accuracy, precision, recall,* dan *f1-score. Precision* atau presisi adalah metrik untuk mengevaluasi model yang mengukur akurasi pengenalan model terhadap positif sejati dari total prediksi positif. *Recall* adalah metrik yang digunakan untuk mengevaluasi keakuratan model klasifikasi dalam mengidentifikasi positif yang benar dari jumlah total kelas positif dalam data aktual. *F1-score* akan menunjukan keseimbangan *precision* dan *recall.*

Pada metode klasifikasi teks[61], *micro accuracy* dan *macro accuray* dapat berguna dalam mengevaluasi model dengan mengukur akurasi model secara keseluruhan dengan memberikan bobot yang sama untuk semua kelas. Akurasi mikro dapat dihitung dengan menjumlahkan jumlah prediksi yang benar dari semua kelas dan membaginya dengan jumlah total prediksi

sedangkan akurasi makro ditentukan dengan menghitung akurasi untuk setiap kelas terlebih dahulu dan kemudian menghitung akurasi rata-rata semua kelas..

2.3.2 Non-negative Matrix Factorization (NMF)

NMF adalah algoritma dekomposisi non-probabilistik yang menggunakan teknik faktorisasi matriks[62]. NMF merupakan algoritma *unsupervised learning* sehingga data yang dipelajari tidak memerlukan informasi yang diberi label sebelumnya. NMF dapat menggunakan representasi data yang telah diubah menjadi bobot kata dengan kata-kata penting memiliki bobot yang lebih tinggi seperti TF-IDF (*Term Frequency - Inverse Document Frequency*). NMF bekerja dengan memecah matriks kata-dokumen (V) menjadi dua matriks yang memiliki dengan dimensi yang lebih rendah yaitu matriks *term-topic* (W) dan matriks *topic-document* (H) sehingga NMF digunakan untuk mengurangi dimensionalitas dengan mengubah dimensi data menjadi ruang berdimensi lebih rendah. NMF menunjukan menunjukkan kontribusi/bobot setiap faktor (topik) dalam merekonstruksi dokumen.

$$V = WH$$

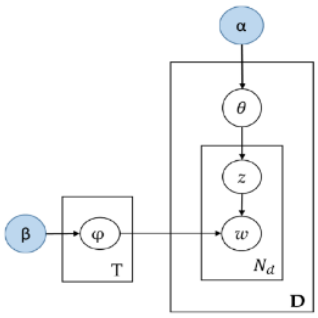
Gambar 2.1 NMF matrix[63]

Dari gambar 2.1 dapat diketahui tujuan NMF adalah untuk menemukan perkiraan V menggunakan hasil dua matriks W dan H.

2.3.3 Latent Dirichlet Allocation (LDA)

LDA adalah metode *topic modeling* berbasis probabilistik yang mengasumsikan setiap dokumen terdiri dari campuran beberapa topik dan setiap topik terdiri dari sejumlah kata dengan probabilitas tertentu[31]. LDA merupakan metode pemodelan topik yang mudah diterapkan karena tidak memerlukan pengetahuan sebelumnya dalam mengidentifikasi kosakata dan dapat mendeteksi hubungan yang bermakna antara dokumen dalam kumpulan data. LDA bertujuan untuk mengidentifikasi topik-topik tersembunyi dengan menentukan distribusi topik dalam setiap dokumen dan

distribusi kata-kata dalam setiap topik. LDA memerlukan *corpus* atau kumpulan dokumen serta perkiraan jumlah topik untuk menghasilkan sekumpulan topik yang berisi distribusi kata-kata sebagai perwakilan dan distribusi topik pada setiap dokumen. LDA menggunakan representasi dokumen sebagai sekumpulan kata dan frekuensi kemunculan (*Bag-of-Words*). Berdasarkan proses generatif pada gambar 2.2, kata-kata dalam dokumen hanya merupakan variabel teramati sedangkan (ϕ dan θ) merupakan variabel laten dan (α dan β) menunjukan hiperparameter.



Gambar 2.2 Latent Dirichlet Allocation Model[59]

α adalah parameter untuk distribusi topik per dokumen
β adalah parameter untuk distribusi kata per topik
D menunjukan jumlah dokumen
N jumlah kata dalam dokumen tertentu
Φ adalah distribusi kata pada topik tertentu
Θ adalah distribusi topik pada dokumen tertentu

Model LDA menghasilkan *coherence_score* yang dapat dianalisis untuk menentukan kualitas topik dan nilai *coherence* yang tinggi menunjukan model yang lebih baik[32].

2.3.4 Support Vector Machine (SVM)

Support vector machine (SVM) adalah jenis algoritma *machine learning* yang dapat digunakan untuk klasifikasi biner dan *multi-class* data. Teknik klasifikasi dapat diterapkan untuk pemrosesan data teks. SVM digunakan untuk mengurutkan dua kelompok data dengan klasifikasi yang sama. Algoritma akan menarik garis *(hyperplanes*) untuk memisahkan kelompok menurut pola. Support Vector Machine (SVM) adalah algoritma *supervised*

learning yang dapat digunakan untuk menyelesaikan permasalahan klasifikasi atau dengan menggunakan kernel untuk mendeteksi *support vector*. SVM digunakan untuk mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. SVM menggunakan trik kernel untuk memetakan data ke ruang fitur berdimensi tinggi tanpa melakukan transformasi langsung sehingga data berdimensi tinggi dapat dikategorikan pada dimensi yang tak terhingga (*infinite*). SVM bekerja dengan baik pada data berdimensi tinggi termasuk data teks yang direpresentasikan dalam bentuk vektor[64].

2.3.5 CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah metodologi yang menjadi standar industri dalam penelitian data mining sehingga proses analisis data berjalan terstruktur dan sistematis[65]. *Framework* CRISP-DM dapat diterapkan di berbagai bidang penelitian termasuk analisis sentimen dan *topic modeling*. CRISP-DM terdiri dari enam tahapan utama yang saling berkaitan dan dapat dilakukan secara iteratif:

1. Business Understanding

Tahap *business understanding* berfokus pada memahami tujuan dan kebutuhan proyek bisnis atau penelitian. Pada tahap ini, penelitian menetapkan objektif dari penelitian *data mining* dan memilih ala atau teknologi yang akan digunakan.

2. Data Understanding

Tahap *data understanding* berisi pengumpulan data yang dibutuhkan pada penelitian data mining, mengidentifikasi data yang diperoleh, eksplorasi data yang diperoleh untuk menemukan hubungan dan informasi, serta mengecek kualitas data yang tersedia.

3. Data Preparation

Tahap ini merupakan tahap transformasi data mentah menjadi format yang siap digunakan untuk pemodelan. Proses data preparation berisi

pembersihan teks seperti menghilangkan tanda baca, stopwords, tokenisasi, serta pembobotan kata menggunakan TF-IDF.

4. Modeling

Tahap *modeling* meliputi pemilihan model algoritma yang akan digunakan, *splitting* set data yang digunakan, membangun model yang akan digunakan, serta melakukan komparasi performa model agar hasil yang diperoleh lebih akurat.

5. Evaluation

Tahap evaluasi dilakukan untuk menilai kualitas dan kinerja model. Evaluasi *topic modeling* dilakukan menggunakan coherence score sedangkan evaluasi klasifikasi sentimen menggunakan metrik akurasi, presisi, recall, dan F1-score.

6. Deployment

Tahap ini bertujuan untuk menyajikan hasil analisis kepada pengguna akhir atau pihak yang berkepentingan. Tahap *deployment* meliputi merancang rencana *deploy* model, mengawasi dan melakukan evaluasi berkala, serta membuat laporan seperti dashboard yang berisi visualisasi hasil penelitian *data mining*.

2.4 Teori tentang tools/software yang digunakan

2.4.1 Google Colab

Google Colab adalah layanan Google yang menyediakan layanan Jupyter Notebook berbasis *cloud* serta lingkungan atau *environment* untuk *data mining* dan proyek analisis data[66]. Google Colab menunjang aktivitas *machine learning* atau *deep learning* dengan menyediakan akses GPU dan TPU gratis. Google Colab mampu menangani kode pemrograman python, visualisasi data, dan narasi dalam satu skrip dokumen.Google Colab mendukung enviroment Jupyter Notebook untuk pengembangan dan pelatihan model *data mining* seperti model klasifikasi, pengelompokan, dan regresi[67].

2.4.2 Python

Python adalah bahasa pemrograman popular yang dapat digunakan untuk pembelajaran mesin atau *machine learning*, pengembangan web dan aplikasi desktop, serta dapat digunakan untuk analisis data[68]. Python memiliki sintaks yang sederhana dan mudah digunakan sehingga Python bahasa yang bagus untuk dipelajari bagi pemula. Python memungkinkan pemrograman berorientasi objek dan prosedural. Python memiliki library yang menunjang teknik penambangan dan analisis data.