BAB II

LANDASAN TEORI

2.1 Penelitian Terkait

Penelitian ini dilakukan dengan mempelajari penelitian – penelitian terdahulu yang memiliki keterkaitan dengan analisis sentimen menggunakan metode *Transformer*. Tabel 2.1 merupakan beberapa penelitian terlebih dahulu yang berkaitan dengan penelitian ini :

Tabel 2.1 Studi Literatur

No	Nama	Tahun	Judul
1.	Lazuardi& Asep Juarna	2023	ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI JOOX PADA ANDROID MENGGUNAKAN METODE *BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER (BERT)*
	Keterangan : Komputasi menggunakan 10.000 komentar, di mana 7.000 data dijadikan data latih, 2.010 sebagai data validasi, dan 990 data sebagai data uji. Skor dihitung dengan mengkombiinasikan akurasi baseline dengan skore recall yang memberikan akurasi F1-score. Hasil analisis sentimen adalah 41,92% *true* (sentimen) positif, 1,01% true netral, dan 35,95% true negatif, semuanya dari 990 data uji, dengan akurasi F1-score *BERT* turut-turut 86%, 51%, dan 76% sementara akurasi baseline adalah 83%, 79%, dan 75%, yang berarti ada peningkatan akurasi true positif sebesar 3,6% dan true negative sebesar 3,6%.		
No	Nama	Tahun	Judul

2.	Raden Mas Rizqi Wahyu Panca Kusuma Atmaja & Wiyli Yustanti	2022	Analisis Sentimen Customer Review Aplikasi Ruang Guru dengan Metode *BERT* (Bidirectional Encoder Representation from *Transformer*)
	Keterangan : Berdasarkan review komentar pengguna Ruang Guru di Google Playstore. Review pengguna mayoritas menyatakan positif. Dari 5437 *testing data*, 5254 komentar yang dinyatakan positif, 16 yang dinyatakan netral sentimen, dan 167 komentar dinyatakan negatif sentimen. Dapat disimpulkan bahwa metode *BERT* sangat efektif untuk di implementasi analisis sentimen.		
No	Nama	Tahun	Judul
3.	Ekka Pujo Ariesanto Akhmad	2023	Analisis Sentimen Ulasan Pengguna Aplikasi DLU Ferry Pada Google Playstore Menggunakan *BERT from Transformer*
	Keterangan : Hasil yang diambil dari penelitian ini adalah Model *BERT* berhasil menghasilkan akurasi sebesar 86% dengan pemilihan *hyperparameter*, yaitu *batch_size* 32, *Learning rate* 3e-6, dan *epoch* 5. Karena jumlah *dataset* yang Tida besar untuk pemrosesan kata, model menghadapi kesukaran untuk klasifikasi sentimen netral. Ada ketidakteraturan dalam proses labelisasi pada *dataset* yang bersifat mendua.		

2.2 Data Mining

Menurut (Nabila et al., 2021) Data Mining adalah tentang menemukan korelasi, pola, dan tren baru yang bermakna dengan mencari sejumlah besar data yang disimpan dalam arsip menggunakan pengenalan pola, teknik statistik dan

matematika. Penambangan data adalah pencarian pola dan informasi menarik dari sejumlah besar data.

Istilah Data Mining mencakup beberapa perspektif, seperti Data Mining atau pengenalan pola. Istilah "penambangan data" tepat karena tujuan utama penambangan data adalah untuk mengekstraksi informasi yang masih tersembunyi di dalam blok data.

2.3 *Deep Learning*

Menurut (Soebroto, 2019) *Deep Learning* adalah salah satu cabang dari *Machine Learning* yang terdiri dari algoritma pemodelan abstraksi tingkat tinggi pada data menggunakan sekumpulan fungsi transformasi non-linear yang ditata berlapis-lapis dan mendalam. *Deep Learning* sangat baik untuk diterapkan pada supervised *Learning*, unsupervised *Learning*, maupun reinforcement *Learning* dalam berbagai aplikasi seperti pengenalan citra, suara, klasifikasi teks, dan sebagainya.

Model pada *Deep Learning* pada dasarnya dibangun berdasarkan jaringan saraf tiruan (Neural Network), yang risetnya sudah berlangsung sejak era 80an namun baru-baru ini kembali bangkit dengan adanya komputer yang semakin cepat apalagi ditambah dengan adanya teknologi pada Big data.

Secara umum, *Deep Learning* dapat digunakan untuk :

1. Supervised *Learning* (Pembelajaran Terarah): Dalam pembelajaran terarah, algoritma dilatih menggunakan data yang sudah diketahui labelnya (data terlabel). Tujuannya adalah untuk mempelajari hubungan antara fitur-fitur (features) pada data input dengan label yang sesuai.

Setelah melalui tahap pelatihan, algoritma dapat memprediksi label dari data baru yang belum pernah dilihat sebelumnya.

2. Unsupervised *Learning* (Pembelajaran Tak Terarah): Dalam pembelajaran tak terarah, algoritma berusaha mengidentifikasi pola atau struktur yang terdapat dalam data input yang tidak memiliki label (data tak terlabel). Algoritma ini dapat digunakan untuk tugas seperti pengelompokan (clustering), reduksi dimensi (dimensionality reduction), dan pencarian asosiasi (association discovery).
3. Reinforcement *Learning* (Pembelajaran Penguatan): Dalam pembelajaran penguatan, algoritma belajar melalui interaksi dengan lingkungan. Algoritma ini mengambil tindakan (action) dalam lingkungan tertentu dan menerima umpan balik (feedback) berupa reward atau hukuman (punishment) sebagai tanggapan atas tindakan tersebut. Tujuan utama dari pembelajaran penguatan adalah untuk mengoptimalkan keputusan yang diambil oleh agen untuk mencapai tujuan tertentu.

2.4 Natural Language Processing

Menurut (Prasetyo et al., 2021) Natural Language Processing (NLP) adalah kombinasi ilmu komputer dan kecerdasan buatan yang berkaitan dengan linguistik. NLP berurusan dengan bagaimana mesin memahami bahasa manusia untuk berinteraksi satu sama lain. NLP memungkinkan komputer untuk mempelajari dan memahami bahasa manusia, memungkinkan komputer untuk berkomunikasi dengan manusia. Bahasa manusia itu unik karena diciptakan khusus untuk menyampaikan makna. Membuat komputer memahami bahasa manusia adalah tugas yang sulit karena bahasa manusia memiliki struktur yang kompleks. Juga,

setiap bahasa itu unik dan dapat memiliki banyak arti. Contohnya dapat dilihat pada kalimat berikut, “Look at the dog with one eye”, di mana kalimat tersebut dapat memiliki arti “melihat anjing dengan satu mata” atau “melihat anjing yang mempunyai mata satu”.

Dua teknik terpenting untuk memahami NLP adalah analisis sintaksis dan analisis semantik. Kedua teknik tersebut digunakan untuk memeriksa struktur bahasa. Analisis sintaksis mengacu pada tata bahasa sedangkan analisis semantik mengacu pada interpretasi kalimat.

Analisis sintaksis adalah teknik yang digunakan untuk menyusun kalimat agar memiliki tata bahasa yang benar. Analisis sintaksis adalah tentang menentukan struktur kalimat seperti subjek, predikat, kata benda, kata kerja, kata ganti, dll. Sistem mampu membaca frase, yang dibagi menjadi kata-kata dan akhirnya membuat deskripsi terstruktur. Teknik ini memungkinkan Anda untuk menyederhanakan kalimat agar lebih mudah menemukan informasi. Selain itu, analisis sintaksis juga dapat membantu mengidentifikasi kata atau frasa baru atau tidak biasa.

Suatu kalimat dapat disebut kalimat jika sekurang-kurangnya terdiri dari subjek dan predikat, misalnya kalimat “Andi makan”. Dengan menggunakan teknik analisis sintaksis, komputer dapat membedakan antara subjek ("Andi") dan predikat ("makan"). Kalimat yang dibentuk belum tentu masuk akal, karena analisis sintaksis hanya memeriksa struktur kalimat.

Analisis semantik adalah teknik yang digunakan untuk memahami makna dan interpretasi struktur bahasa. Bahasa orang lain dapat dipahami berdasarkan intuisi

dan keterampilan berbahasa. Komputer tidak memiliki intuisi dan pengetahuan seperti itu, sehingga membutuhkan metode lain, yaitu semantik. Semantik merupakan proses yang penting karena output semantik yang diharapkan adalah makna yang terkandung dalam input tersebut. Analisis semantik memproses teks untuk mengidentifikasi dan memahami topik yang dibahas. Semantik juga mempelajari hubungan antara berbagai konsep dalam sebuah teks. Misalnya, jika teks tersebut memuat kata "money" dan "accounting", maka topik pembahasannya terkait dengan "economy".

2.5 Analisis Sentimen

Menurut (Wahyudi et al., 2021) Analisis sentimen adalah teknik komputer untuk mempelajari subjektivitas opini, perasaan, dan teks. Tugas dasar analisis sentimen adalah mengklasifikasikan polaritas teks yang terjadi pada dokumen, kalimat, atau opini. Polaritas berarti apakah teks suatu dokumen, kalimat atau pendapat memiliki sisi positif atau negatif .

Adapun tahapan dalam melakukan analisis sentiment antara lain sebagai berikut:

1. Pemilihan Data: Tahap pertama adalah memilih sumber data yang akan dianalisis. Data ini bisa berupa teks dari ulasan pelanggan, media sosial, survey, atau sumber lainnya yang mengandung opini atau sentimen.
2. Pra-Pemrosesan Data: Langkah ini melibatkan pra-pemrosesan pada data teks sebelum dilakukan analisis sentimen. Pra-pemrosesan dapat mencakup case folding (mengubah teks menjadi huruf kecil semua), tokenisasi (mengubah teks menjadi unit-unit kecil seperti kata-kata), penghapusan stopwords (kata-kata umum yang tidak memberikan makna

khusus), dan pembersihan teks (misalnya, menghapus tanda baca atau karakter khusus).

3. Ekstraksi Fitur: Pada tahap ini, fitur-fitur yang relevan diekstraksi dari data teks. Fitur-fitur ini dapat berupa kata-kata kunci, frasa, atau aspek-aspek tertentu yang berkaitan dengan sentimen yang ingin dianalisis. Misalnya, jika analisis sentimen dilakukan terhadap ulasan produk, fitur-fitur bisa termasuk kata-kata yang berkaitan dengan kualitas, harga, atau layanan.
4. Klasifikasi Sentimen: Tahap ini melibatkan penerapan algoritma klasifikasi untuk menentukan polaritas sentimen dari setiap data teks. Algoritma klasifikasi dapat berbasis aturan (seperti leksikon sentimen yang memetakan kata-kata dengan label sentimen), pendekatan hibrida (kombinasi leksikon dengan teknik *Machine Learning*), atau algoritma *Machine Learning* murni (seperti Naive Bayes, Support Vector *Machine*s, atau Decision Trees).
5. Evaluasi dan Interpretasi: Setelah sentimen diklasifikasikan, tahap ini melibatkan evaluasi hasilnya dan interpretasi hasilnya dalam konteks yang relevan. Hasil analisis sentimen dapat dievaluasi menggunakan metrik seperti akurasi, presisi, recall, atau F1-score. Interpretasi hasilnya dapat dilakukan dengan menganalisis distribusi sentimen, tren, atau membandingkan hasil antara kelompok data yang berbeda.
6. Pelaporan dan Visualisasi: Langkah terakhir adalah menyusun laporan hasil analisis sentimen dan mengkomunikasikan temuan kepada pihak yang berkepentingan. Visualisasi data seperti grafik, diagram, atau word

cloud juga bisa digunakan untuk memperjelas dan memvisualisasikan hasil analisis sentimen.

2.6 Labeling BERT

Menurut (Nayla et al., 2023) BERT adalah sebuah algoritma *Deep Learning* keluaran Google yang masih berkaitan dengan NLP (Natural Language Processing). Algoritma ini adalah invasi dari model *Transformer* dimana model tersebut memproses sebuah kata pada kalimat berdasarkan ada atau tidaknya kaitan antara kata tersebut dengan kalimat secara keseluruhan. Algoritma BERT memproses konteks penuh dengan cara melihat pola yang muncul pada sebelum atau sesudah kata.

2.7 Bahasa Pemrograman

Menurut (Premana et al., 2022) Bahasa pemrograman atau biasa disebut bahasa komputer atau bahasa pemrograman komputer adalah instruksi standar untuk mengelola komputer. Bahasa pemrograman ini adalah seperangkat aturan sintaksis dan semantik yang digunakan untuk mendefinisikan program komputer.

Dalam skripsi ini akan menggunakan beberapa Bahasa pemrograman diantara lain adalah :

2.7.1 Python

Menurut (Mohammad Daniel Gumilar, 2021) Python adalah Bahasa pemrograman tingkat tinggi (high-level programming language) yang berjalan dalam sistem yang diinterpretasikan dan dapat digunakan untuk berbagai tujuan (general purpose). Python pertama kali dikembangkan pada awal 1990-an oleh Guido van Rossum di Scichting Mathematik Centrum (CWI) di Belanda.

2.7.2 Javascript

Menurut (Reza & Putra, 2021) Javascript adalah bahasa yang mewakili kumpulan skrip yang akan dieksekusi dalam dokumen HTML. Bahasa pemrograman Javascript adalah bahasa pemrograman komputer yang sangat mampu menambahkan lebih banyak fungsionalitas ke bahasa HTML dengan memungkinkan perintah dijalankan di sisi pengguna, yaitu. di sisi browser dan bukan di server web.

2.8 Streamlit

Menurut (Putranto et al., 2023) Streamlit adalah framework berbasis Python yang bersifat open source. Framework ini dibuat untuk memudahkan developer membuat program data science dan *Machine Learning* interaktif berbasis web. Salah satu keunggulan Streamlit adalah pengembang tidak perlu mengubah tata letak situs web menggunakan CSS, HTML, dan Javascript karena kerangka Streamlit telah menyediakan ini dengan fitur yang disertakan dalam kerangka.

2.9 SQL Lite

Menurut (ANGGA SETIYADI, 2021) SQLite adalah suatu library yang menerapkan mesin database self-contained, serverless, zero-configuration, dan transactional. Self-contained berati SQLite membutuhkan sedikit sekali dukungan dari library eksternal atau dari sistem operasi. Serverless berati SQLite dalam mengakses database baik itu read atau write dapat secara langsung dari file database tanpa melalui proses server dan tidak mendukung pengaksesan secara remote (artinya database SQLite bisa dikendalikan dari jarak jauh dengan adanya jaringan komputer (“Computer Network”), baik melalui jaringan lokal (intranet) atau

internet), dimana kebanyakan mesin SQL database diterapkan sebagai proses server yang terpisah.

2.10 *Transformer*

Menurut (Yessy Asri, Dwina Kuswardani, Listra Firgia Missianes Horhoruw & Siti Aisyah Ramadhana, 2024) *Transformer* adalah model dalam dunia *Deep Learning* yang diperkenalkan oleh Vaswani dkk. Pada tahun 2017 model ini menjadi terkenal karena kemampuanya dalam menangani tugas-tugas pemrosesan bahasa alami (Natural Language Processing, NLP) dengan sangat baik. Sebelumnya model-model seperti LSTM dan GRU sering digunakan, tetapi *Transformer* memberikan kinerja yang lebih baik dan paralelisasi yang lebih efisien. Model transduksi urutan saraf yang bersaing umumnya mengadopsi struktur encoder-decoder. Dalam sistem ini, bagian yang disebut "encoder" BERT tugas mengubah urutan simbol-simbol input menjadi bentuk representasi yang terus-menerus disebut sebagai z. Setelah mendapatkan representasi z, decoder kemudian menghasilkan urutan output simbol-simbol satu per satu.

2.11 Cleaning

Menurut (Akhmad, 2023) Cleaning adalah proses yang digunakan untuk menghilangkan angka, beberapa simbol, url, username, hastag, spasi berlebih, tanda baca, emoji, dan pengulangan karakter yang ada pada kalimat. Cleaning menggunakan regular expression untuk menemukan karakter yang akan dihapus.

2.12 Tokocrypto

Tokocrypto adalah exchange cryptocurrency yang memungkinkan pengguna untuk membeli, menjual, dan memperdagangkan berbagai aset digital, seperti

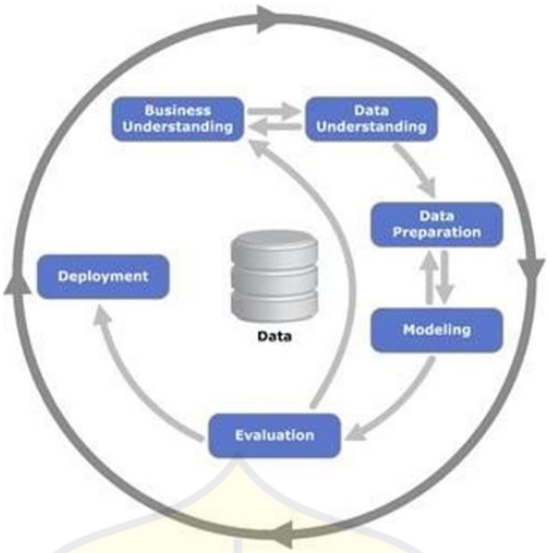
Bitcoin, Ethereum, dan berbagai altcoin lainya. Platform ini BERTujuan untuk memberikan akses yang mudah dan aman bagi pengguna di Indonesia untuk berpartisipasi dalam pasar cryptocurrency.

Tokocrypto dirancang dengan antarmuka yang sederhana dan mudah digunakan, cocok untuk pemula maupun trader berpengalaman. Tokocrypto menggunakan berbagai langkah keamanan, termasuk penyimpanan aset yang aman dan otentikasi dua faktor (2FA), untuk melindungi akun pengguna. Tokocrypto juga memiliki aplikasi mobile yang memungkinkan pengguna untuk trading kapan saja dan di mana saja.

Tokocrypto BERTujuan untuk mempermudah akses ke dunia cryptocurrency bagi masyarakat Indonesia. Dengan fitur yang ramah pengguna, fokus pada keamanan, dan dukungan edukasi, Tokocrypto mejadi salah satu pilihan populer bagi trader dan investor cryptocurrency di Indonesia.

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

Menurut (Purnomo et al., 2023) CRISP-DM (Cross Industry Standard Process for Data Mining) adalah standar pemrosesan penambangan data yang dirancang untuk mengambil data yang ada melalui setiap langkah terstruktur, terdefinisi dengan baik, dan efisien. CRISP-DM bukan satu-satunya standar untuk penambangan data, tetapi ini adalah yang paling populer saat ini. Berdasarkan hasil survei data science pm, CRISP-DM digunakan dua hingga tiga kali lebih sering daripada empat standar yang paling umum digunakan. Berikut pada Gambar 2.1 merupakan proses dari metodologi CRIPS-DM :



Gambar 2.1 Proses Metodologi CRIPS-DM

1. Business Understanding

Tahap ini adalah untuk mengumpulkan data melalui metode web scraping pada aplikasi Tokocrypto. Analisis baru dapat dilakukan setelah data berhasil dikumpulkan melalui metode web scrapping pada aplikasi Tokocrypto.

2. Data Understanding

Data akan dianalisis apakah sudah cukup dan layak untuk dilakukan pengolahan data atau perlu melakukan pengumpulan data Kembali.

3. Data Preperation

Setelah mendapatkan data yang diperlukan, maka selanjutnya untuk dapat digunakan modeling perlu dikakukan pre-processing untuk menghilangkan kerusakan (kurang baku, singkatan) atau kata yang tidak terdapat pada KBBI (Kamus Besar Bahasa Indonesia), kemudian melakukan tokenize atau pemisahan kata, melakukan labeling pada ulasan dengan membuat diksi kosa kata yang berkonotasi positif dan negatif, melakukan pembobotan nilai kata

dengan EMBEDDING, serta melakukan split data dengan membuat data training dan data testing.

4. Modeling

Pada tahap ini membuat analisis deskriptif pada sentimen analisis dengan menerapkan metode *Transformer*. Jika perlu penyesuaian bisa Kembali ke tahap Data Preparation.

5. Evaluation

Pada tahap ini melakukan pengujian menggunakan Confusion Matrix, dilakukan untuk mendapatkan hasil akurasi dengan mempertimbangkan True Positive, True Negative, False Positive dan False Negative pada label actual dan predict dari algoritma yang digunakan dalam melakukan klasifikasi.

6. Deployment

Setelah mendapatkan hasil berupa nilai model pada tahap modelling dan Evaluation, selanjutnya akan dilakukan deploy aplikasi.