STOCK PRICE SENTIMENT
PADA SAHAM BBRI

TUGAS BESAR DATA MINING

Disusun oleh:

Muhammad Azka Nuril Islami (714220001)

Gaizka Wisnu Prawira (714220011)

Muhammad Fathir (714220021)

Salwa Mutfia Indah Putri (714220026)

Dosen Pengampu:

Nisa Hanum Harani, S.T., M.T., CDSP.,SFPC

NIK. 117.89.223

PROGRAM STUDI DIV TEKNIK INFORMATIKA

UNIVERSITAS LOGISTIK & BISNIS INTERNASIONAL

BANDUNG

2025

HALAMAN PERNYATAAN ORISINALITAS

Laporan tugas besar ini adalah hasil karya kami sendiri dan semua sumber, baik yang dikutip maupun dirujuk telah kami nyatakan dengan benar. Bilamana di kemudian hari ditemukan bahwa karya tulis ini menyalahi peraturan yang ada berkaitan dengan etika dan kaidah penulisan karya ilmiah yang berlaku, maka kami bersedia dituntut dan diproses sesuai dengan ketentuan yang berlaku.

Yang menyatakan,

Nama : Muhammad Azka Nuril Islami

NIM : 714220001

Tanda Tangan: .....

Tanggal: Kamis, 10 Juli 2025

Mengetahui,

Ketua :..... (.....tanda tangan. .....)

Dosen Pengampu Mata Kuliah : ..... (.....tanda tangan. .....)

KATA PENGANTAR

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga kami dapat menyelesaikan Laporan Tugas Besar Data Mining ini yang berjudul "Stock Price Sentiment pada Saham BBRI".

Laporan ini disusun untuk memenuhi tugas akhir mata kuliah Data Mining pada Program Studi D4 Teknik Informatika, Universitas Logistik dan Bisnis Internasional.

Kami mengucapkan terima kasih kepada:

- Dosen pengampu mata kuliah Data Mining atas bimbingan dan ilmunya selama perkuliahan berlangsung.
- Orang tua dan keluarga yang selalu memberikan dukungan moril dan semangat.
- Rekan satu kelompok atas kerja sama dan komitmen dalam menyelesaikan tugas ini bersama.

Kami menyadari bahwa laporan ini masih memiliki kekurangan. Oleh karena itu, kritik dan saran yang membangun sangat kami harapkan demi perbaikan di masa mendatang.

Bandung, 10 Juli 2025

Penyusun,

Muhammad Azka Nuril

Gaizka Wisnu Prawira

Muhammad Fathir

Salwa Mutfia Indah Putri

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Logistik dan Bisnis Internasional, saya yang bertanda tangan di bawah ini:

Nama : Muhammad Azka Nuril Islami

NIM : 714220001

Selaku ketua kelompok, menyatakan menyetujui untuk memberikan kepada Universitas Logistik dan Bisnis Internasional, hak bebas royalti noneksklusif (non-exclusive royalty free right) atas karya ilmiah kami yang berjudul, "STOCK PRICE SENTIMENT PADA SAHAM BBRI" beserta perangkat yang ada (jika diperlukan). Dengan hak ini, ULBI berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan tugas akhir kami selama tetap mencantumkan nama kami sebagai penulis/pemilik hak cipta.

Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Dibuat di : Bandung

Pada tanggal : 10 Juli 2025

Yang menyatakan,

Muhammad Azka Nuril Islami

Ketua Kelompok

ABSTRAK

ABSTRACT

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR RUMUS

DAFTAR NOTASI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pasar saham adalah komponen utama dari sistem keuangan suatu negara yang mencerminkan kesejahteraan ekonomi [1]. Fluktuasi harga saham biasanya dipengaruhi oleh berbagai faktor, seperti kondisi perusahaan, kebijakan ekonomi, dan faktor psikologis investor. Dalam praktiknya, keputusan pembelian dan penjualan saham tidak semata-mata didasarkan pada pertimbangan rasional, tetapi juga pada persepsi dan emosi para pelaku pasar yang dapat berubah sewaktu-waktu [2] . Keadaan ini berarti bahwa informasi dan opini publik dapat menjadi penyebab utama perubahan harga saham, terutama dalam jangka pendek [3].

Perkembangan teknologi informasi, khususnya media sosial seperti Twitter, telah mengubah cara individu dalam memberi dan menerima informasi. Twitter saat ini menjadi forum yang sangat aktif untuk memberikan komentar, berita, dan reaksi terhadap berbagai isu ekonomi dan perusahaan [4]. Kecepatan dan volume informasi yang disebarkan di Twitter berpotensi memberikan gambaran langsung tentang opini publik mengenai suatu perusahaan atau area bisnis[5]. Oleh karena itu, analisis sentimen tweet berpotensi menjadi sumber sekunder yang berharga untuk data dalam prediksi pasar [4][6].

Meskipun harga saham historis telah menjadi dasar model prediksi yang dibangun selama bertahun tahun, metode ini tidak lengkap. Model prediksi tradisional tidak dapat menangkap kekuatan psikologis dan pola pikir pasar yang berubah dengan cepat. Selama krisis atau pengumuman berita penting dalam situasi yang tidak stabil, data media sosial dapat memberikan peringatan lebih awal daripada indikator teknikal. Oleh karena itu, integrasi data sentimen dengan data historis sangat potensial untuk meningkatkan akurasi model prediksi saham.

Beberapa penelitian sebelumnya telah mengindikasikan bahwa sentimen media sosial berhubungan dengan pergerakan harga saham [7]. Hasil-hasil ini menunjukkan bahwa sentimen publik yang positif menyebabkan pergerakan harga saham naik, sementara sentimen negatif dapat menjadi pendahulu pergerakan harga turun. Namun, sebagian besar penelitian ini masih kurang dalam hal pengujian korelasi yang tepat tanpa mencoba meniru model prediksi yang lebih umum. Pendekatan yang lebih formal diperlukan untuk menggabungkan kedua jenis data di bawah kerangka model prediksi model prediksi yang terjamin.

Dengan potensi besar dari sentimen media sosial, penelitian ini bertujuan untuk membuat dan memvalidasi sebuah model untuk prediksi harga saham yang menggabungkan data masa lalu dan sentimen Twitter secara real-time [5][8]. Tujuannya adalah untuk mengidentifikasi apakah dengan menggabungkan kedua sumber tersebut dapat memberikan model yang lebih baik dan responsif terhadap perubahan pasar. Penelitian ini diharapkan dapat memberikan kontribusi yang berarti bagi perkembangan teknologi keuangan modern dan menginformasikan pengambilan keputusan investasi yang lebih berwawasan.

1.2 Rumusan Masalah

Berdasarkan latar belakang sebelumnya, dapat dirumuskan beberapa pertanyaan utama sebagaimana berikut:

A. Apakah analisis sentimen real time dari Twitter memiliki korelasi signifikan dengan pergerakan harga saham?
B. Bagaimana kinerja model prediksi harga saham yang hanya menggunakan data historis dibandingkan dengan model yang menggabungkan data historis dan sentimen Twitter?
C. Sejauh mana penambahan fitur sentimen Twitter dapat meningkatkan akurasi dan ketahanan model terhadap volatilitas pasar?

1.3 Tujuan penelitian

1.3.1. Tujuan Umum

Menganalisis pengaruh integrasi data sentimen Twitter terhadap akurasi model prediksi harga saham dibandingkan dengan model prediksi yang hanya menggunakan data historis.

1.3.2. Tujuan Khusus

A. Melakukan analisis sentimen pada data Twitter menggunakan model RoBERTa.
B. Menggabungkan data hasil analisis sentiment dengan data historis berdasarkan tanggal data.
C. Membangun dan mengevaluasi model klasifikasi sentimen dengan algoritma XGBoost, SVR, Random Forest, MLP, dan Logistic Regression.

1.4 Manfaat Penelitian

4.1.1 Manfaat Teoretis

A. Memberikan kontribusi pada pengembangan ilmu pengetahuan di bidang keuangan dan data science, khususnya terkait integrasi analisis sentimen media sosial dalam prediksi harga saham.
B. Menjadi referensi akademik mengenai penggunaan data sentimen Twitter untuk mendukung model prediksi harga saham yang lebih akurat.
C. Memperkaya literatur terkait pemanfaatan big data dan text mining dalam analisis pasar modal.

4.1.2 Manfaat Praktis

A. Membantu investor dalam mengambil keputusan investasi yang lebih tepat dengan mempertimbangkan informasi sentimen publik dari media sosial.
B. Memberikan wawasan bagi perusahaan sekuritas dan analis pasar untuk mengembangkan sistem prediksi harga saham yang lebih responsif terhadap dinamika pasar.

C. Menjadi dasar bagi pengembangan aplikasi atau sistem prediksi harga saham yang memanfaatkan integrasi data historis dan data sentimen secara real-time.

1.5 Ruang Lingkup

Penelitian ini memiliki ruang lingkup yang difokuskan pada pemanfaatan data historis harga saham dan data sentimen dari media sosial Twitter untuk membangun model prediksi harga saham. Adapun ruang lingkup penelitian ini dijabarkan sebagai berikut:
Penelitian ini dibatasi pada:

A. Penelitian menggunakan data historis saham dari perusahaan tertentu yang terdaftar di bursa, dalam periode waktu tertentu (misalnya satu tahun terakhir)
B. Informasi diambil dari tweet publik yang relevan dengan saham perusahaan tersebut menggunakan kata kunci atau tagar tertentu.
C. Sentimen akan diklasifikasikan ke dalam kategori positif, negatif, atau netral menggunakan metode pemrosesan bahasa alami (Natural Language Processing/NLP) seperti IndoBERT atau RoBERTa.
D. Model prediksi harga saham akan dikembangkan dengan pendekatan machine learning (XGBoost, SVR, Random Forest, MLP, dan Logistic Regression) dan dibandingkan antara model berbasis data historis saja dan model yang juga mengintegrasikan sentimen Twitter.

BAB II
TINJAUAN PUSTAKA

2.1 Kajian Teori dan Konsep Penting

2.1.1 Data Mining dan Machine Learning

A. Data mining adalah proses ekstraksi informasi berharga, pola, dan pengetahuan yang tersembunyi dari kumpulan data yang besar dan kompleks. Tujuan utamanya adalah untuk mengidentifikasi hubungan dan tren yang dapat digunakan untuk mendukung pengambilan keputusan strategis. Dalam esensinya, data mining merupakan teknik analisis yang menggunakan metode statistik, matematika, dan kecerdasan buatan untuk menggali pengetahuan yang belum diketahui secara otomatis dari data [9].

B. Machine learning merupakan bagian dari artificial intelligence/kecerdasan buatan yang membutuhkan data-data valid untuk proses belajarnya. Machine learning dapat membuat keputusan yang tepat dan cepat, serta dapat memberikan solusi terhadap berbagai permasalahan. Machine learning memiliki kemampuan untuk belajar sendiri dan memutuskan sesuatu tanpa harus diprogram berulang kali oleh manu-sia, hal ini dapat terjadi karena adanya pengalaman berbagai data yang dimiliki [10].

2.1.2 Teknik yang Digunakan

A. Bi-LSTM dan RoBERTa

Bi-LSTM pengembangan dari model LSTM yang memproses data sekuensial dari dua arah, yaitu forward dan backward. Dengan arsitektur ini, Bi-LSTM mampu menangkap konteks kata sebelum dan sesudah secara lebih baik dibanding LSTM biasa. Hal ini membuat Bi-LSTM banyak digunakan dalam analisis sentimen karena dapat memahami dependensi kata dalam kalimat secara menyeluruh, sehingga meningkatkan akurasi klasifikasi sentimen teks.

RoBERTa (Robustly Optimized BERT Pretraining Approach) adalah model transformer yang merupakan pengembangan dari BERT dengan optimasi pada jumlah data pre-training, ukuran batch, dan strategi masking yang lebih dinamis. RoBERTa terbukti memiliki performa yang lebih tinggi dibanding BERT pada berbagai tugas NLP, termasuk analisis sentimen media sosial. Model ini dapat

memahami makna kata dalam konteks kalimat secara lebih mendalam dan kompleks, sehingga menghasilkan klasifikasi sentimen yang lebih akurat.

B. Algoritma Klasifikasi Sentimen

a. Support Vector Regression (SVR): Meskipun SVR umumnya digunakan untuk regresi, dalam penelitian analisis sentimen SVR dapat digunakan untuk memprediksi skor sentimen yang kemudian dipetakan menjadi kategori sentimen positif, negatif, atau netral.
b. Multilayer Perceptron (MLP): Merupakan model jaringan saraf tiruan (artificial neural network) yang terdiri dari beberapa lapisan tersembunyi (hidden layers) dan dapat menangkap pola non-linear kompleks dalam data teks.
c. Logistic Regression: Model regresi untuk klasifikasi biner atau multi-kelas yang banyak digunakan sebagai baseline pada analisis sentimen karena interpretasinya yang sederhana dan proses training yang cepat.
d. Extreme Gradient Boosting (XGBoost): Merupakan algoritma boosting berbasis decision tree yang memiliki performa tinggi dan efisien. XGBoost sering digunakan pada kompetisi data science karena akurasinya yang baik pada berbagai jenis dataset.
e. Random Forest: Algoritma ensemble learning yang menggabungkan banyak decision tree untuk meningkatkan akurasi dan mengurangi overfitting, efektif pada data dengan banyak fitur seperti representasi TF-IDF.

C. Model Prediksi Harga Saham

Untuk prediksi harga saham, model yang digunakan antara lain:

a. Long Short-Term Memory (LSTM): Jaringan saraf tiruan yang dirancang khusus untuk menangani data deret waktu dengan memperhatikan dependensi jangka panjang.
b. Random Forest dan SVR: Dapat digunakan untuk regresi harga saham atau klasifikasi arah pergerakan harga saham.

2.2 Studi Terkait (Penelitian Sejenis)

Beberapa penelitian sebelumnya telah dilakukan untuk mengintegrasikan data sentimen dengan data historis dalam prediksi harga saham. Penelitian oleh Maharani et al. melakukan post-training IndoBERT dengan korpus finansial Indonesia untuk meningkatkan akurasi analisis sentimen dan

topik di domain keuangan, menunjukkan potensi pengembangan model RoBERTa domain-spesifik untuk analisis sentimen finansial [11].

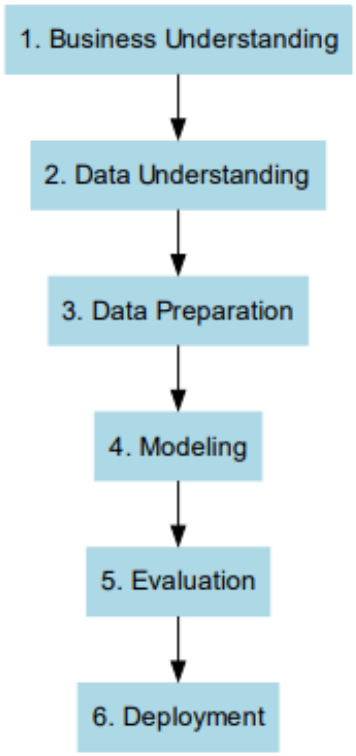
Penelitian lain menggabungkan data tweet dan berita dalam model prediksi harga saham menggunakan MLP dan LSTM, di mana hasilnya menunjukkan bahwa integrasi kedua sumber data tersebut dapat meningkatkan akurasi model dibandingkan hanya menggunakan data historis[12]. Selain itu, terdapat penelitian yang menggunakan analisis sentimen microblogging (Twitter) dan machine learning untuk prediksi harga saham, yang menunjukkan bahwa integrasi data sentimen dengan data historis mampu meningkatkan akurasi prediksi [13].

Penelitian lainnya juga menggabungkan data tweet dan berita dalam prediksi harga saham menggunakan metode machine learning seperti MLP dan LSTM, dan hasilnya menunjukkan bahwa kombinasi kedua sumber data tersebut dapat meningkatkan akurasi model prediksi dibandingkan hanya menggunakan data historis saham [14]. Studi lain yang menggunakan analisis sentimen microblogging (Twitter) dan machine learning untuk memprediksi pasar saham menunjukkan bahwa model yang mengintegrasikan data sentimen dari media sosial dengan data historis dapat memberikan hasil prediksi yang lebih akurat [15].

Selain itu, terdapat penelitian yang mengembangkan model prediksi harga saham dengan menggabungkan data historis dan sentimen Twitter menggunakan Bi-LSTM dan RoBERTa, dan hasil penelitian tersebut menunjukkan bahwa integrasi data sentimen dengan data historis secara signifikan meningkatkan akurasi prediksi dibandingkan model yang hanya menggunakan data historis [15]. Penelitian lainnya mengumpulkan lebih dari 12.000 komentar investor saham di China, dan menggunakan Bi-LSTM untuk prediksi harga saham dengan integrasi analisis sentimen. Hasilnya menunjukkan bahwa akurasi prediksi meningkat ketika sentimen dimasukkan ke dalam model [16]. Penelitian lain juga menggabungkan RoBERTa untuk analisis sentimen microblog (mirip Twitter) dan LSTM untuk prediksi harga saham, dan penelitian ini menunjukkan bahwa model integrasi mampu outperform model berbasis data historis saja [17].

2.3 Visualisasi (Diagram Alir Konsep)

Diagram alur berikut menggambarkan tahapan-tahapan utama yang dilakukan dalam penelitian ini, dimulai dari pemahaman permasalahan bisnis hingga tahap deployment. Penelitian ini mengadopsi model proses CRISP-DM (Cross Industry Standard Process for Data Mining) sebagai kerangka kerja utama karena model ini terbukti fleksibel dan banyak digunakan dalam proyek data mining. Setiap tahapan saling berkaitan secara iteratif, memungkinkan penyesuaian kembali terhadap proses sebelumnya bila ditemukan temuan baru dalam proses selanjutnya. Berikut adalah diagram alur dalam penelitian ini:



2.4 State of The Art

Perkembangan terkini dalam bidang prediksi harga saham menggunakan pendekatan data mining, khususnya yang menggabungkan data historis dan data sentimen media sosial. Dalam beberapa tahun terakhir, pemanfaatan media sosial seperti Twitter sebagai sumber data alternatif dalam prediksi pasar saham semakin berkembang.

Salah satu pendekatan mutakhir yang banyak digunakan adalah integrasi machine learning dengan data non-tradisional, seperti opini publik dari media sosial. Model seperti Long Short-Term Memory (LSTM) dan Random Forest telah digunakan secara luas karena kemampuannya dalam mengenali pola dari data waktu dan menangani ketidakpastian dalam data pasar yang fluktuatif.

Selain itu, teknologi Natural Language Processing (NLP) juga mengalami kemajuan pesat. Model analisis sentimen tidak lagi terbatas pada metode leksikal sederhana, tetapi mulai beralih ke model berbasis pembelajaran mendalam seperti BERT, XLNet, dan model pre-trained lainnya. Model ini dapat memahami konteks dan nuansa bahasa alami dengan lebih baik.

Secara umum, perkembangan terkini menunjukkan bahwa model prediksi yang menggabungkan data historis dan sentimen sosial dapat memberikan akurasi yang lebih tinggi dan respon yang lebih cepat terhadap perubahan pasar, terutama pada kondisi yang tidak stabil atau penuh ketidakpastian.

BAB III
METODOLOGI PENELITIAN

3.1 Tahapan penelitian

Dalam upaya memperoleh hasil analisis data yang terarah dan dapat dipertanggungjawabkan secara metodologis, digunakan pendekatan berbasis kerangka kerja yang telah teruji seperti CRISP-DM, sebuah model proses standar independen dari industri untuk data mining yang terdiri dari enam fase iterative [18].

Tabel 1, deskripsi proses CRISP-DM [2]

Fase	Deskripsi
Business Understanding	Memahami situasi bisnis, menentukan tujuan data mining, seperti klasifikasi (dalam laporan ini) serta kriteria keberhasilan, dan menyusun rencana proyek.
Data Understanding	Mengumpulkan data dari sumber yang relevan, mengeksplorasi, mendeskripsikan, dan memeriksa kualitas data menggunakan analisis statistik.
Data Preparation	Melakukan seleksi data dengan kriteria inklusi-eksklusi, membersihkan data yang berkualitas buruk, serta membangun atribut turunan sesuai model yang akan digunakan.
Modeling	Memilih teknik pemodelan, menyusun kasus uji, membangun model, menetapkan parameter, lalu mengevaluasi model sesuai kriteria yang telah ditentukan.
Evaluation	Memeriksa hasil model terhadap tujuan bisnis awal, menginterpretasi hasil, menentukan tindakan selanjutnya, dan melakukan review keseluruhan proses.
Deployment	Menerapkan hasil melalui laporan akhir atau komponen perangkat lunak, serta merencanakan pemantauan dan pemeliharaan implementasi model.

Namun, dalam laporan kali ini, proses metologi hanya akan dilaksanakan sampai pada tahap *Evaluation*, sehingga tahapan Deployment tidak menjadi fokus kajian.

3.2 Deskripsi Dataset

Tabel 2, Dataset

Tipe	Sumber	Ukuran	Attribut
sentimen	x/twitter	2232	15 (conversation_id_str, created_at, favorite_count, full_text, id_str, image_url, in_reply_to_screen_name, lang, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, username)
historis	website (investing.com)	37 / day	7 (Tanggal, Terakhir, Pembukaan, Tertinggi, Terendah, Vol., Perubahan%)

3.3 Algoritma / Data Mining Tools

Tabel 3, Algoritma

Algoritma	Alasan
Random Forest	Kuat pada data dengan interaksi antar fitur & tahan noise. Cocok untuk regresi harga saham berdasarkan banyak lag features.
XGBoost	Sering outperform model lain dalam kompetisi prediksi harga karena fokus pada residual dan regularisasi kuat.
Logistic Regression	Baseline linear sederhana untuk memeriksa apakah data cukup dijelaskan oleh relasi linear, sangat interpretatif.
SVR	Menangkap hubungan non-linear dengan kernel (misalnya RBF), memfokuskan prediksi pada pola utama dengan mengabaikan outlier kecil.
MLP	Neural network dasar untuk mempelajari representasi kompleks antar waktu tanpa harus memprogram interaksi secara manual.

Tabel 4, Tools

Tools	Fungsi
pandas	Mengelola data frame, generate lag features, moving average, RSI.
numpy	Operasi numerik cepat (vectorized), misalnya menghitung return.
matplotlib / seaborn	Visualisasi distribusi harga, heatmap korelasi, plot prediksi.

scikit-learn	Pipeline training: Random Forest, Logistic Regression, SVR, MLP; preprocessing (StandardScaler/MinMaxScaler), metrics (accuracy, confusion matrix).
xgboost	Model XGBoostClassifier, powerful untuk data tabular.
statsmodels	Uji stasionaritas (ADF Test) jika ingin mengecek pola.
yellowbrick	Visualisasi residual, ROC, class prediction error.
mlflow / wandb	Tracking experiment untuk tuning hyperparameter & logging metrics.

3.4 Evalusi Kinerja

Tabel 5, Evaluasi

Evaluasi	Penjelasan	Alasan
Accuracy	Proporsi prediksi arah benar (misalnya naik vs turun) dari seluruh prediksi.	Untuk baseline sederhana: seberapa sering model benar dalam memprediksi arah.
Precision	Dari semua yang diprediksi naik, berapa yang benar-benar naik.	Menghindari false signals, penting jika cost salah beli mahal.
Recall	Dari semua hari yang benar-benar naik, berapa yang terprediksi naik.	
F1-Score	Harmonik rata-rata precision dan recall.	Seimbang memerhatikan missed naik (false negative) dan false naik (false positive).
ROC AUC Score	Area under curve ROC yang membandingkan True Positive Rate vs False Positive Rate di semua threshold.	Untuk melihat kemampuan model membedakan naik vs turun terlepas threshold.
Confusion Matrix	Matriks jumlah prediksi naik/turun vs aktual naik/turun.	Memberi gambaran kesalahan model, bisa fokus memperbaiki misclassification.
Directional Accuracy	Persentase prediksi arah benar (mirip accuracy), tapi kadang dihitung dari return positif/negatif.	Sangat relevan dalam trading untuk memastikan prediksi arah lebih sering tepat.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Lingkungan Eksperimen

Seluruh proses eksperimen dilakukan menggunakan Google Colab dengan Python 3.10. Tools dan pustaka utama yang digunakan mencakup pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, dan transformers. Implementasi dilakukan dalam tiga tahap utama: preprocessing data, eksplorasi data (EDA), dan pembuatan serta evaluasi model prediksi.

4.2 Preprocessing Data

Proses *preprocessing* data merupakan tahap fundamental dalam analisis sentimen, terutama saat mengolah data mentah dari media sosial yang bersifat tidak terstruktur. Tahap ini bertujuan untuk menghasilkan data bersih, konsisten, dan siap digunakan dalam proses klasifikasi sentimen menggunakan model bahasa berbasis *transformer*. Berikut adalah tahapan yang dilakukan dalam preprocessing dataset sentimen:

4.2.1 Penggabungan Dataset

Empat dataset yang berasal dari sumber berbeda, yakni datasetX-Nuril2(2024-1).csv, datasetX-Nuril3(2024-2).csv, datasetX-Fathir(2025).csv, dan datasetX-wisnu(2025).csv digabungkan menggunakan fungsi pd.concat(). Tujuannya adalah untuk memperbesar ukuran data dan memperkaya variasi konteks kalimat yang digunakan dalam pelatihan dan evaluasi model.

```
df_x1 = pd.read_csv("../data/raw/datasetX-Nuril2(2024-1).csv")
df_x2 = pd.read_csv("../data/raw/datasetX-Nuril3(2024-2).csv")
df_x3 = pd.read_csv("../data/raw/datasetX-Fathir(2025).csv")
df_x4 = pd.read_csv("../data/raw/datasetX-wisnu(2025).csv")

df_combined_x = pd.concat([df_x1, df_x2, df_x3, df_x4],
ignore_index=True)
df_combined_x
```

4.2.2 Penghapusan Duplikasi

Langkah selanjutnya adalah menghapus entri duplikat berdasarkan kolom full_text untuk memastikan bahwa tidak ada redundansi data yang dapat mempengaruhi distribusi kelas sentimen.

```
df_combined_x = df_combined_x.drop_duplicates(subset='full_text')
```

4.2.3 Seleksi Kolom dan Tipe Data

Beberapa kolom yang dianggap tidak relevan terhadap proses analisis sentimen dihapus, seperti username, user_id_str, dan tweet_url. Selain itu, kolom created_at diubah menjadi format datetime untuk memudahkan proses sorting berdasarkan waktu.

```
df_combined_x['created_at'] =
pd.to_datetime(df_combined_x['created_at'])

df_combined_x = df_combined_x.drop(columns=['username', 'user_id_str',
'tweet_url', 'retweet_count', 'reply_count', 'quote_count', 'location',
'in_reply_to_screen_name', 'image_url', 'favorite_count',
'conversation_id_str'])
```

4.2.4 Seleksi Bahasa

Hanya data yang berbahasa Indonesia (kode 'in') yang dipertahankan. Hal ini penting untuk menjaga konsistensi bahasa dan agar sesuai dengan model klasifikasi sentimen yang digunakan, yaitu model pra-latih bahasa Indonesia.

```
df_combined_x = df_combined_x[df_combined_x['lang'] == 'in']
```

4.2.5 Klasifikasi Sentimen

Analisis sentimen dilakukan menggunakan transformers pipeline dari HuggingFace, dengan model "w11wo/indonesian-roberta-base-sentiment-classifier". Setiap entri teks diklasifikasikan ke dalam label sentimen (Positive, Neutral, atau Negative), dan disertai dengan skor kepercayaan dari model.

```
classifier = pipeline(
    "sentiment-analysis",
    model="w11wo/indonesian-roberta-base-sentiment-classifier",
)

df_combined_x['sentiment_result'] =
df_combined_x['full_text'].apply(lambda x: classifier(x)[0])
df_combined_x['sentiment'] =
df_combined_x['sentiment_result'].apply(lambda x: x['label'])
df_combined_x['score'] = df_combined_x['sentiment_result'].apply(lambda
x: x['score'])
```

4.2.6 Penyimpanan Dataset Bersih

Dataset yang telah diproses dan diklasifikasikan disimpan sebagai file CSV (dataset_sentiment.csv) untuk digunakan pada tahap analisis lanjutan atau pelatihan model pembelajaran mesin.

```
dataset_sentiment.to_csv('../data/processed/dataset_sentiment.csv', index=False)
```

4.2.7 Hasil Preprocessing Data

Sebelum dilakukan analisis lebih lanjut, data mentah hasil pengumpulan dari media sosial perlu melalui tahapan preprocessing untuk memastikan kualitas dan konsistensi data. Proses ini mencakup pembersihan data, penghapusan duplikasi, konversi format waktu, dan identifikasi sentimen terhadap setiap entri teks. Tahapan ini bertujuan untuk menghasilkan dataset yang valid, terstruktur, dan siap digunakan dalam proses analisis sentimen dan pemodelan lebih lanjut. Pada subbab ini disajikan hasil dari proses preprocessing yang dilakukan terhadap dataset yang telah dikumpulkan.

1. Struktur Data Sentimen

   Dataset hasil scraping tweet terdiri dari 1038 entri dengan empat atribut utama: created_at, full_text, id_str, dan lang. Seluruh entri terisi lengkap tanpa adanya nilai kosong, sebagaimana ditunjukkan oleh kode berikut:

   ```
   df_combined_x.info()
   ```

   Hasil:

   ```
   <class 'pandas.core.frame.DataFrame'>
   Index: 1038 entries, 0 to 2231
   Data columns (total 4 columns):
    #   Column      Non-Null Count  Dtype
   ---  ------      --------------  -----
    0   created_at  1038 non-null   datetime64[ns, UTC]
    1   full_text   1038 non-null   object
    2   id_str      1038 non-null   int64
    3   lang        1038 non-null   object
   dtypes: datetime64[ns, UTC](1), int64(1), object(2)
   memory usage: 40.5+ KB
   ```

2. Distribusi Kelas Sentimen

   Setelah klasifikasi sentimen dilakukan, distribusi label menunjukkan bahwa mayoritas tweet tergolong dalam kategori netral (736 entri), diikuti oleh negatif (159 entri) dan positif (141 entri). Hal ini menunjukkan potensi ketidakseimbangan kelas yang perlu diperhatikan dalam proses modeling.

   ```
   dataset_sentiment['sentiment'].value_counts()
   ```

Hasil:
```
sentiment
neutral     736
negative    159
positive    141
Name: count, dtype: int64
```

3. Struktur Dataset Historis Saham

   Dataset historis saham terdiri dari 39 entri yang mencerminkan data pergerakan harga saham dalam rentang waktu tertentu. Terdapat tujuh atribut utama, yaitu Tanggal, Terakhir, Pembukaan, Tertinggi, Terendah, Vol., dan Perubahan%. Semua entri pada dataset ini terisi penuh, yang menunjukkan tidak adanya nilai null.

```
dataset_historis.info()
```

Hasil:
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39 entries, 0 to 38
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Tanggal     39 non-null     object
 1   Terakhir    39 non-null     float64
 2   Pembukaan   39 non-null     float64
 3   Tertinggi   39 non-null     float64
 4   Terendah    39 non-null     float64
 5   Vol.        39 non-null     object
 6   Perubahan%  39 non-null     object
dtypes: float64(4), object(3)
memory usage: 2.3+ KB
```

4. Struktur Dataset Model

   Dataset model merupakan hasil integrasi antara data historis saham dan hasil agregasi dari analisis sentimen harian. Dataset ini juga terdiri dari 39 entri, dengan total 13 atribut yang mencakup informasi harga, volume, perubahan persentase, serta fitur-fitur sentimen seperti avg_signed_sentiment, count_positive, count_negative, count_neutral, dan total_tweets. Hampir seluruh kolom terisi penuh, kecuali lima fitur sentimen yang memiliki satu entri kosong (NaN). Hal ini masih dapat ditangani dengan teknik interpolasi atau penghapusan baris, tergantung pendekatan analisis.

```
dataset_model.info()
```

Hasil:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39 entries, 0 to 38
Data columns (total 13 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Tanggal              39 non-null     datetime64[ns]
 1   Terakhir             39 non-null     float64
 2   Pembukaan            39 non-null     float64
 3   Tertinggi            39 non-null     float64
 4   Terendah             39 non-null     float64
 5   Vol.                 39 non-null     object
 6   Perubahan%           39 non-null     object
 7   date                 39 non-null     object
 8   avg_signed_sentiment  38 non-null     float64
 9   count_positive       38 non-null     float64
 10  count_negative       38 non-null     float64
 11  count_neutral        38 non-null     float64
 12  total_tweets         38 non-null     float64
dtypes: datetime64[ns](1), float64(9), object(3)
memory usage: 4.1+ KB
```

4.3 Pembentukan Dataset Model

Tahapan pembentukan dataset model dilakukan dengan cara menggabungkan data historis peristiwa dengan hasil analisis sentimen dari data media sosial. Langkah ini bertujuan untuk memperoleh satu data frame terpadu yang siap digunakan dalam proses pemodelan dan analisis lanjutan.

4.3.1 Membaca dan Membersihkan Data Historis

Langkah awal yaitu memuat data historis dari direktori ../data/processed/dataset_historis.csv dan memastikan kolom tanggal dalam format datetime.

```
dataset_historis =
pd.read_csv("../data/processed/dataset_historis.csv")

dataset_historis['Tanggal'] =
pd.to_datetime(dataset_historis['Tanggal'], dayfirst=True,
errors='coerce')


dataset_sentiment['date'] = dataset_sentiment['created_at'].dt.date
dataset_historis['date'] =
pd.to_datetime(dataset_historis['Tanggal']).dt.date
```

Tahap ini memastikan bahwa data historis memiliki format tanggal yang sesuai dan dapat digunakan untuk penggabungan data selanjutnya.

4.3.2 Konversi dan Penandaan Sentimen

Data sentimen dari media sosial sebelumnya telah dianalisis dan diklasifikasikan ke dalam tiga kategori: positive, neutral, dan negative. Kategori ini kemudian dikonversikan ke dalam nilai numerik menggunakan dictionary sentiment_sign.

```
sentiment_sign = {'positive': 1, 'neutral': 0, 'negative': -1}
dataset_sentiment['sentiment_sign'] =
dataset_sentiment['sentiment'].map(sentiment_sign)
```

Selanjutnya, skor sentimen yang telah diberikan dari hasil analisis kemudian dikalikan dengan nilai tanda (sign) untuk mendapatkan skor bertanda (signed score).

```
dataset_sentiment['signed_score'] = dataset_sentiment['sentiment_sign']
* dataset_sentiment['score']
```

4.3.3 Agregasi Sentimen Harian

Data sentimen kemudian dikelompokkan berdasarkan tanggal dengan melakukan agregasi terhadap beberapa metrik penting:

- avg_signed_sentiment: rerata skor sentimen bertanda per hari.
- count_positive, count_negative, count_neutral: jumlah masing-masing jenis sentimen dalam satu hari.
- total_tweets: total jumlah tweet yang dianalisis pada tanggal tersebut.

Kode berikut digunakan untuk menghasilkan agregasi ini:

```
dataset_grouped = dataset_sentiment.groupby('date').agg(
    avg_signed_sentiment=('signed_score', 'mean'),
    count_positive=('sentiment', lambda x: (x=='positive').sum()),
    count_negative=('sentiment', lambda x: (x=='negative').sum()),
    count_neutral=('sentiment', lambda x: (x=='neutral').sum()),
    total_tweets=('sentiment', 'count')
).reset_index()
```

Hasil dari agregasi ini menghasilkan data sentimen harian yang siap digabungkan dengan data historis.

4.3.4 Penggabungan Data Historis dan Sentimen

Setelah data historis dan data sentimen sama-sama memiliki kolom date, kedua data tersebut digabungkan menggunakan metode merge dengan jenis *left join*. Hal ini memastikan semua baris pada data historis tetap dipertahankan meskipun tidak semua tanggal memiliki data sentimen.

```
dataset_model = pd.merge(dataset_historis, dataset_grouped, on='date',
how='left')
```

Langkah ini menghasilkan satu data frame baru bernama dataset_model yang menggabungkan informasi historis dan sentimen sosial secara lengkap.

4.3.5 Ekspor Dataset Model

Dataset model yang telah terbentuk kemudian disimpan ke dalam file .csv untuk keperluan analisis lanjutan.

```
dataset_model.to_csv('../data/processed/dataset_model.csv',
index=False)
```

Dengan demikian, proses pembentukan dataset model telah selesai. Dataset ini berisi data historis peristiwa beserta dimensi sosial yang diwakili oleh analisis sentimen dari media sosial pada hari yang sama. Ini memungkinkan pendekatan pemodelan yang lebih komprehensif dengan mempertimbangkan faktor sosial dan temporal sekaligus.

4.4 Eksplorasi Dataset Saham dan Sentimen

Sebelum membangun model prediktif, penting untuk memahami karakteristik dasar dari data yang digunakan. Tahapan ini dikenal sebagai Exploratory Data Analysis (EDA). Melalui EDA, kita dapat mengidentifikasi pola, distribusi, outlier, nilai hilang, serta hubungan antar variabel yang dapat berpengaruh terhadap model akhir. Tahap ini juga berfungsi sebagai landasan dalam pengambilan keputusan pada proses pra-pemrosesan dan pemilihan fitur. Langkah pertama dalam EDA adalah memuat dataset dan memastikan bahwa format serta struktur data telah sesuai.

4.4.1 Memuat dan Menyiapkan Dataset

Langkah pertama dalam eksplorasi data adalah memuat dataset dari file yang telah diproses sebelumnya, yaitu dataset_model.csv. Dataset ini telah melewati proses pra-pemrosesan awal seperti pembersihan dan penggabungan sumber data. Setelah data dimuat, dilakukan pengecekan struktur data dan penghapusan nilai kosong (missing values) untuk memastikan kualitas data yang akan digunakan dalam analisis selanjutnya.

```
import pandas as pd

from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("../data/processed/dataset_model.csv")

data.info()

data.dropna(inplace=True)
```

4.4.2 Pemeriksaan Autokorelasi Harga Saham

Untuk memahami pola waktu (time dependency) pada harga saham, digunakan analisis autokorelasi. Plot ACF (Autocorrelation Function) digunakan untuk melihat apakah nilai harga saham pada suatu waktu dipengaruhi oleh nilai pada waktu sebelumnya. Hal ini penting untuk mengidentifikasi kemungkinan adanya komponen musiman atau tren jangka panjang.

```
plot_acf(data['Terakhir'].dropna(), lags=30)
plt.title("Autocorrelation of Closing Price")
plt.show()
```

Gambar 4. 1 Autocorrelation of Closing Price

Hasil plot menunjukkan bahwa nilai autokorelasi menurun secara bertahap dan tetap signifikan hingga sekitar lag ke-8, yang menunjukkan adanya autokorelasi positif kuat jangka pendek. Setelah itu, nilai autokorelasi berubah menjadi negatif dan tetap signifikan secara moderat hingga sekitar lag ke-20 sebelum akhirnya mendekati nol.

4.4.3 Hubungan Harga Saham dan Rata-rata Sentimen

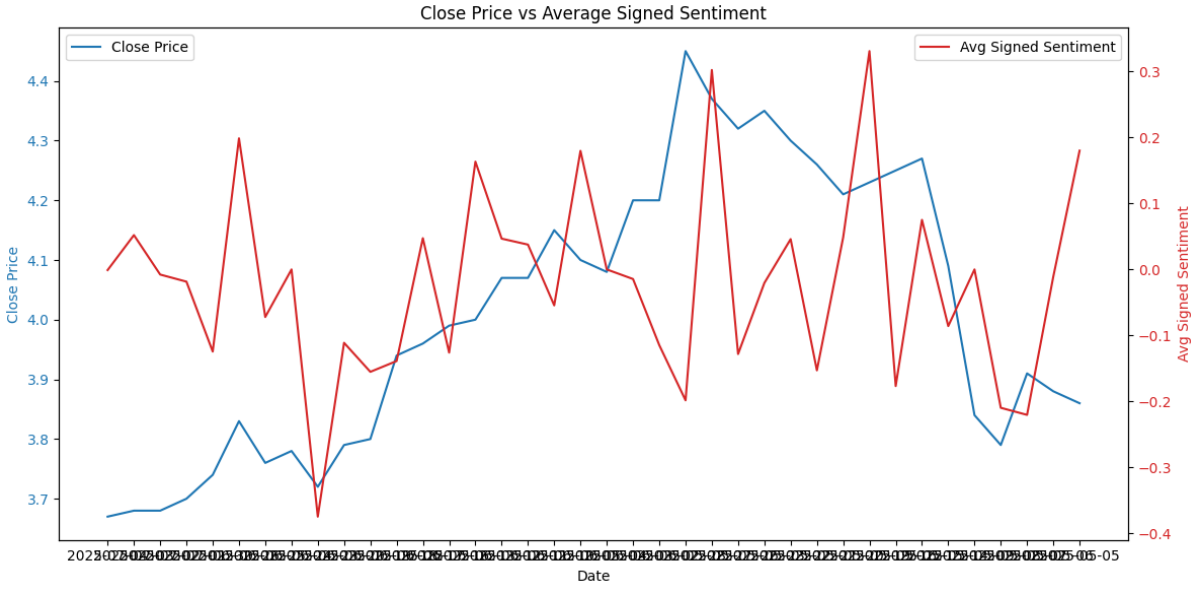
Visualisasi ini bertujuan untuk melihat keterkaitan antara pergerakan harga saham dengan sentimen publik yang diukur melalui nilai rata-rata sentimen bertanda (avg_signed_sentiment). Dengan menggunakan grafik dua sumbu, ditampilkan tren harga saham dan sentimen dalam rentang waktu yang sama untuk mengamati apakah perubahan sentimen dapat mempengaruhi harga saham.

```
fig, ax1 = plt.subplots(figsize=(12,6))

color = 'tab:blue'
ax1.set_xlabel('Date')
ax1.set_ylabel('Close Price', color=color)
ax1.plot(data['date'], data['Terakhir'], color=color, label='Close
Price')
ax1.tick_params(axis='y', labelcolor=color)
ax1.legend(loc='upper left')

ax2 = ax1.twinx()
color = 'tab:red'
ax2.set_ylabel('Avg Signed Sentiment', color=color)
ax2.plot(data['date'], data['avg_signed_sentiment'], color=color,
label='Avg Signed Sentiment')
ax2.tick_params(axis='y', labelcolor=color)
ax2.legend(loc='upper right')

plt.title('Close Price vs Average Signed Sentiment')
plt.tight_layout()
plt.show()
```



Gambar 4. 2 Close Price vs Average Signed Sentiment

Berdasarkan visualisasi tersebut, terlihat bahwa pergerakan sentimen tidak secara langsung selaras dengan tren harga saham. Terdapat beberapa periode di mana lonjakan atau penurunan sentimen tidak diikuti oleh perubahan signifikan dalam harga saham, dan sebaliknya. Hal ini menandakan bahwa meskipun sentimen publik memiliki peran dalam membentuk ekspektasi pasar, hubungannya terhadap harga saham bersifat lemah atau tidak linier dalam jangka pendek.

4.4.4 Konversi Volume Transaksi

Kolom volume transaksi (Vol.) pada dataset masih dalam format string yang mengandung satuan seperti 'K' (ribu) dan 'M' (juta). Agar dapat digunakan dalam analisis numerik dan korelasi, kolom ini perlu dikonversi menjadi nilai numerik murni. Fungsi parse_volume digunakan untuk melakukan parsing nilai string menjadi angka desimal yang sesuai.

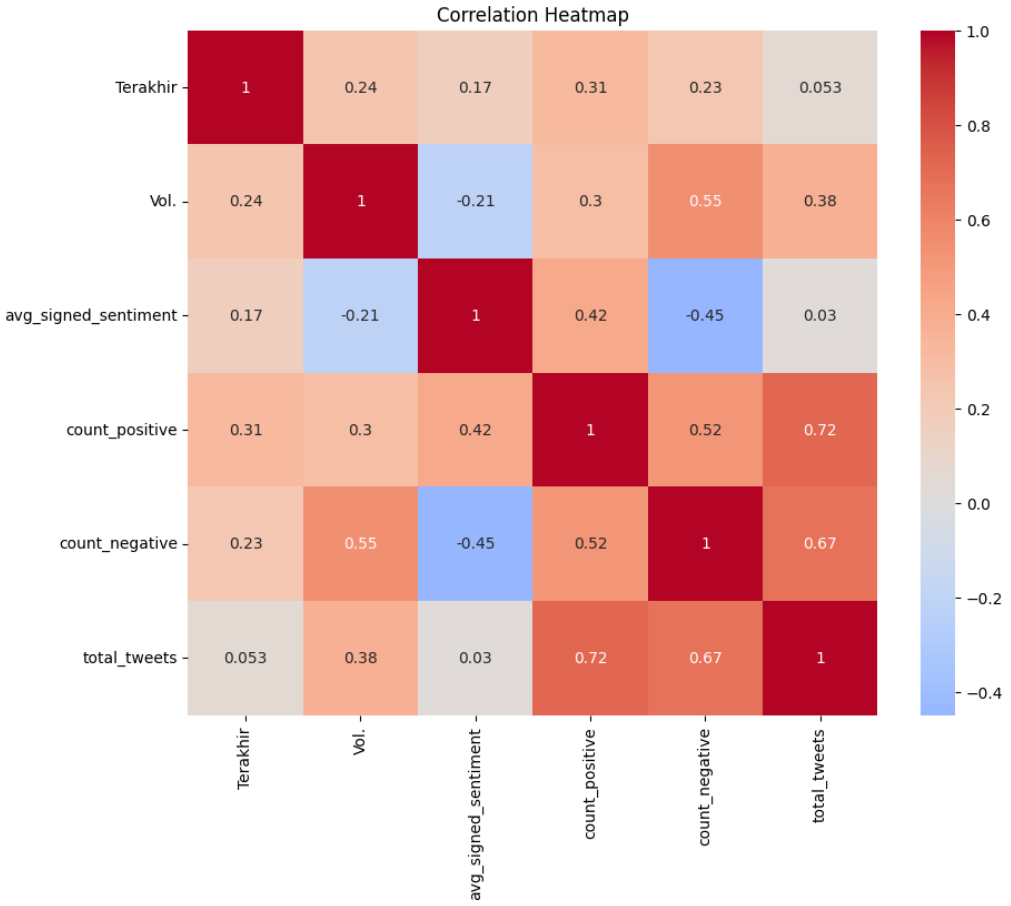
```
def parse_volume(vol_str):
    if isinstance(vol_str, str):
        vol_str = vol_str.replace(',', '.')  # ubah koma jadi titik
        if vol_str.endswith('M'):
            return float(vol_str[:-1]) * 1_000_000
        elif vol_str.endswith('K'):
            return float(vol_str[:-1]) * 1_000
        else:
            return float(vol_str)
    return vol_str  # jika sudah float atau NaN

data['Vol.'] = data['Vol.'].apply(parse_volume)
```

4.4.5 Korelasi Antar Variabel

Analisis korelasi membantu mengidentifikasi sejauh mana hubungan antar variabel numerik seperti harga saham, volume transaksi, sentimen, dan jumlah tweet. Heatmap korelasi memberikan visualisasi yang informatif mengenai kekuatan dan arah hubungan antar variabel, yang dapat digunakan sebagai dasar dalam pemilihan fitur untuk model prediktif.

```
plt.figure(figsize=(10,8))
corr = data[['Terakhir', 'Vol.', 'avg_signed_sentiment',
'count_positive', 'count_negative', 'total_tweets']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Heatmap')
plt.show()
```



Gambar 4. 3 Corelation Heatmap

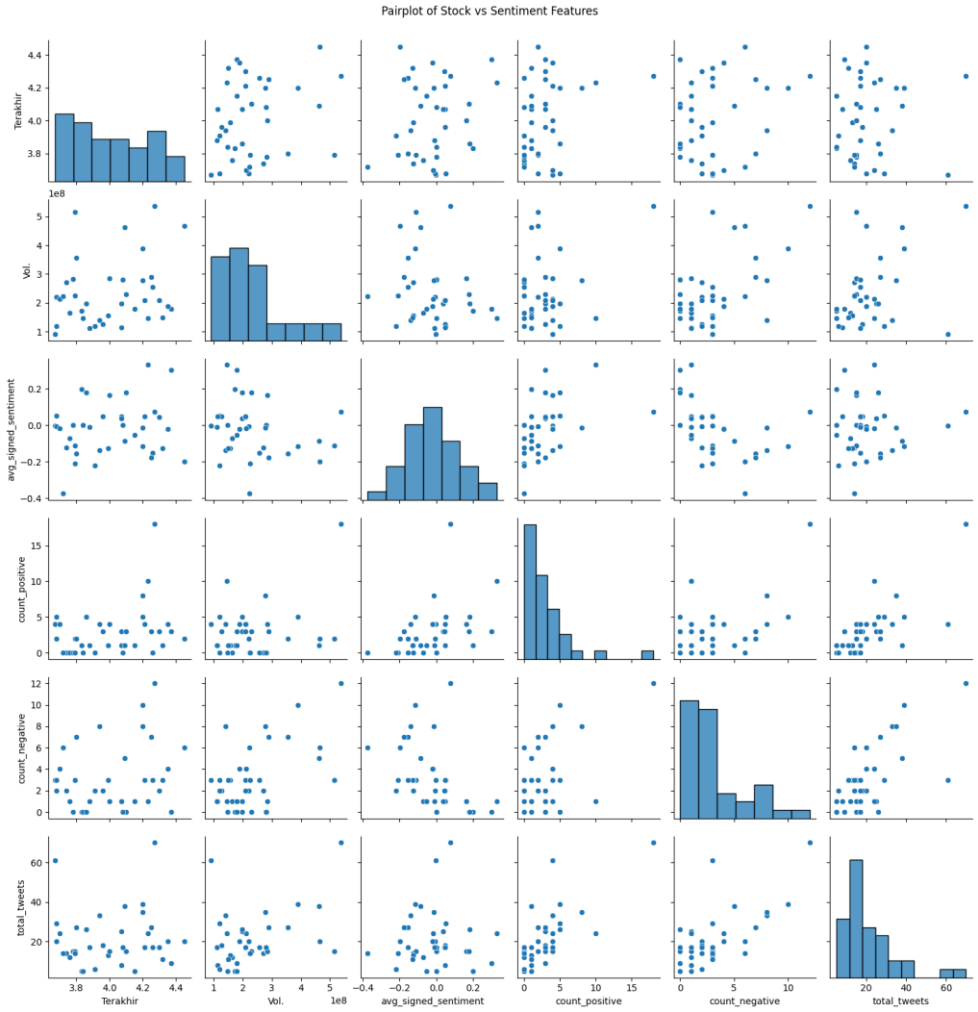
Berdasarkan heatmap di atas, ditemukan beberapa poin penting:

- Volume transaksi (Vol.) menunjukkan korelasi positif moderat terhadap harga penutupan saham (Terakhir) sebesar 0.24, yang mengindikasikan bahwa lonjakan aktivitas pasar cenderung berasosiasi dengan perubahan harga.
- Rata-rata sentimen (avg_signed_sentiment) hanya memiliki korelasi lemah terhadap harga saham (0.17). Ini menyiratkan bahwa meskipun sentimen memiliki nilai informasi, pengaruhnya terhadap harga bersifat tidak dominan dan kemungkinan berperan sebagai faktor pelengkap.
- Jumlah tweet positif dan negatif memiliki korelasi yang kuat terhadap total tweet, masing-masing sebesar 0.72 dan 0.67, yang logis mengingat total tweet merupakan akumulasi dari semua kategori sentimen.
- Jumlah tweet positif (count_positive) menunjukkan korelasi lebih kuat terhadap rata-rata sentimen (avg_signed_sentiment) (0.42) dibandingkan jumlah tweet negatif (-0.45), yang mengindikasikan bahwa sentimen rata-rata lebih sensitif terhadap sentimen positif.

4.4.6 Distribusi dan Hubungan Antar Fitur

Visualisasi distribusi dan hubungan antar variabel dapat dilakukan dengan menggunakan pairplot. Visualisasi ini memudahkan dalam melihat pola sebaran data dan kemungkinan hubungan linear/non-linear antar variabel seperti harga saham, volume, dan sentimen. Ini juga berguna untuk mendeteksi outlier secara visual.

```
sns.pairplot(data[['Terakhir', 'Vol.', 'avg_signed_sentiment',
'count_positive', 'count_negative', 'total_tweets']])
plt.suptitle("Pairplot of Stock vs Sentiment Features", y=1.02)
plt.show()
```



Gambar 4. 4 Pairplot of Stock vs Sentiment Features

Dari visualisasi ini, terlihat bahwa sebagian besar fitur seperti count_positive, count_negative, dan total_tweets menunjukkan pola distribusi miring ke kanan (right-skewed), mengindikasikan bahwa sebagian besar nilai berada pada rentang rendah dengan beberapa nilai ekstrem yang tinggi. Distribusi volume (Vol.) juga tampak sangat bervariasi, dengan nilai yang tersebar luas dan beberapa titik ekstrem.

Hubungan antar variabel tampak lemah secara visual, terutama antara harga saham (Terakhir) dengan variabel sentimen. Meskipun terdapat beberapa pola hubungan positif moderat, seperti

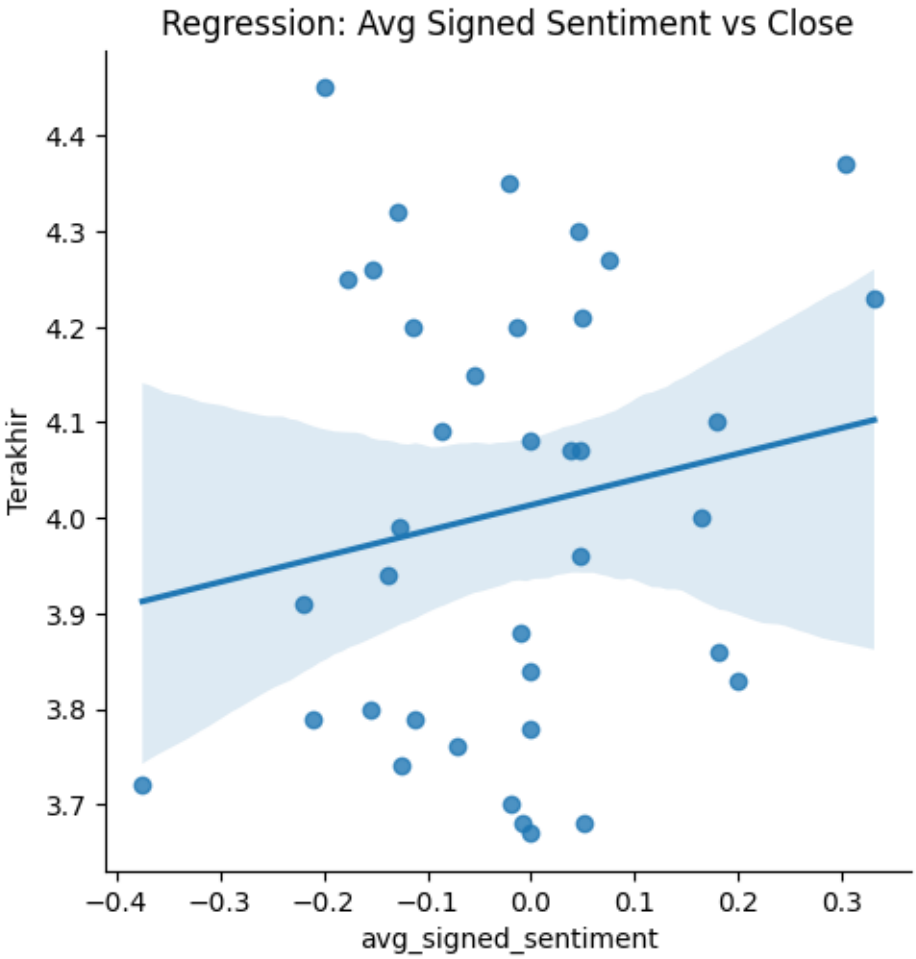
antara count_positive dan total_tweets, secara umum tidak ditemukan hubungan linear yang kuat antar sebagian besar fitur. Hal ini menguatkan hasil dari analisis korelasi sebelumnya bahwa faktor-faktor sentimen sosial media mungkin bersifat pelengkap daripada penentu utama dalam fluktuasi harga saham.

Secara keseluruhan, pairplot memberikan gambaran menyeluruh terhadap struktur data dan memberikan wawasan awal yang penting sebelum proses pemodelan lebih lanjut dilakukan.

4.4.7 Regresi Harga Saham terhadap Sentimen

Untuk mengevaluasi apakah sentimen publik berpengaruh secara linier terhadap harga saham, dilakukan analisis regresi linier sederhana antara avg_signed_sentiment dan Terakhir. Plot regresi memberikan gambaran tentang arah dan kekuatan hubungan antara kedua variabel tersebut, serta potensi signifikansinya secara statistik.

```
sns.lmplot(x='avg_signed_sentiment', y='Terakhir', data=data)
plt.title('Regression: Avg Signed Sentiment vs Close')
plt.show()
```



Gambar 4. 5 Regression: Avg Signed Sentiment vs Close

Untuk menguji potensi hubungan linier antara sentimen publik dan harga saham, dilakukan analisis regresi linier sederhana antara nilai rata-rata sentimen bertanda (avg_signed_sentiment)

dan harga penutupan saham (Terakhir). Visualisasi ini bertujuan untuk mengevaluasi sejauh mana perubahan dalam sentimen publik dapat menjelaskan variasi harga saham.

Dari grafik regresi, terlihat adanya tren linier positif antara kedua variabel, yang menunjukkan bahwa peningkatan nilai sentimen cenderung diikuti oleh peningkatan harga saham. Namun demikian, sebaran data yang cukup menyebar di sekitar garis regresi mencerminkan adanya variabilitas yang tinggi. Hal ini mengindikasikan bahwa meskipun terdapat kecenderungan arah hubungan yang positif, sentimen bukanlah satu-satunya faktor yang memengaruhi harga saham. Dengan demikian, hasil regresi ini memperkuat pemahaman bahwa faktor sentimen dapat menjadi salah satu indikator dalam model prediksi harga saham, tetapi perlu dikombinasikan dengan variabel-variabel fundamental atau teknikal lainnya untuk meningkatkan akurasi prediksi.

4.5 Tabel Hasil Eksperimen / Model

Model	Acc. Mean	Prec. Mean	Recall Mean	F1 Mean	ROC AUC	Dir. Acc	Acc. Std	Prec. Std	Recall Std	F1 Std	ROC AUC Std	Dir. Acc Std
RandomForest_3_fold	0.6111	0.4167	0.6667	0.5079	0.8958	0.5000	0.2079	0.3118	0.4714	0.3675	0.1062	0.1361
RandomForest_5_fold	0.6111	0.4167	0.6667	0.5079	0.8333	0.5000	0.2079	0.3118	0.4714	0.3675	0.1179	0.1361
RandomForest_10_fold	0.6111	0.4167	0.6667	0.5079	0.8958	0.5000	0.2079	0.3118	0.4714	0.3675	0.1062	0.1361
XGBoost_3_fold	0.5556	0.3833	0.6667	0.4762	0.8333	0.5000	0.2079	0.3064	0.4714	0.3563	0.2357	0.1361
XGBoost_5_fold	0.6667	0.4722	0.6667	0.5524	0.6759	0.6111	0.2357	0.3356	0.4714	0.3913	0.1249	0.0786
XGBoost_10_fold	0.5556	0.3833	0.6667	0.4762	0.8333	0.5000	0.2079	0.3064	0.4714	0.3563	0.2357	0.1361
LogReg_3_fold	0.5556	0.2500	0.3333	0.2857	0.8009	0.5556	0.2079	0.3536	0.4714	0.4041	0.1540	0.0786
LogReg_5_fold	0.6111	0.4167	0.5000	0.4524	0.7963	0.5556	0.2079	0.3118	0.4082	0.3515	0.2144	0.0786
LogReg_10_fold	0.5556	0.2500	0.3333	0.2857	0.8009	0.5556	0.2079	0.3536	0.4714	0.4041	0.1540	0.0786
SVC_3_fold	0.6667	0.5000	0.6667	0.5556	0.5833	0.6111	0.2722	0.4082	0.4714	0.4157	0.4249	0.2833
SVC_5_fold	0.6667	0.5000	0.6667	0.5556	0.5833	0.6111	0.2722	0.4082	0.4714	0.4157	0.4249	0.2833
SVC_10_fold	0.6667	0.5000	0.6667	0.5556	0.5833	0.6111	0.2722	0.4082	0.4714	0.4157	0.4249	0.2833
MLP_3_fold	0.6111	0.4167	0.5000	0.4524	0.7917	0.5556	0.2079	0.3118	0.4082	0.3515	0.1559	0.0786
MLP_5_fold	0.6111	0.4444	0.5000	0.4667	0.7917	0.6111	0.2833	0.4157	0.4082	0.4110	0.2946	0.2833
MLP_10_fold	0.6111	0.4444	0.5000	0.4667	0.7917	0.6111	0.2833	0.4157	0.4082	0.4110	0.2946	0.2833

4.6 Interpretasi Hasil Evaluasi Model

Metrik	Model Terbaik	Nilai	Catatan
Accuracy	SVC dan XGBoost_5_fold	0.6667	Konsisten pada skema 3, 5, dan 10 fold
Precision	SVC (semua fold)	0.5000	Presisi tinggi, mengurangi false positive
Recall	Random Forest (semua fold), XG Boost (semua fold), SVC (semua fold)	0.6667	Model cukup baik mendeteksi kelas positif
F1-score	SVC dan XGBoost_5_fold	0.5556/0.5524	Menunjukkan keseimbangan antara presisi dan recall
ROC AUC	RandomForest (3&10fold)	0.8958	Kemampuan terbaik dalam membedakan kelas
Stabilitas	LogReg_5_fold, MLP_3_fold	std rendah	Variasi antar fold kecil, artinya performa relatif konsisten

4.7 Analisis Keunggulan dan Keterbatasan Model

1. Random Forest

Keunggulan:

- Memiliki ROC AUC tertinggi (0.8958), artinya sangat baik dalam membedakan antara kelas positif dan negatif.
- Cocok untuk data yang kompleks dan memiliki fitur non-linier.
- Tidak mudah overfitting karena menggunakan banyak pohon.

Keterbatasan:

- F1-score dan precision rendah, yang artinya banyak false positives atau false negatives.

- Performanya tidak meningkat signifikan meskipun jumlah fold ditambah (3, 5, dan 10 fold memiliki hasil serupa).
- Cenderung kurang presisi pada kelas minoritas.

2. XGBoost

Keunggulan:

- Mencapai kombinasi metrik yang seimbang terutama pada skema 5-fold (akurat dan presisi bagus).
- Cenderung memiliki generalisasi lebih baik karena optimisasi gradien dan regularisasi internal.
- F1-score tinggi (0.5524) menunjukkan keseimbangan presisi dan recall.

Keterbatasan:

- Versi 3-fold dan 10-fold menunjukkan penurunan performa, mengindikasikan sensitif terhadap jumlah data latih/validasi.
- Implementasi dan tuning lebih kompleks dibanding model lain.

3. SVC (Support Vector Classifier)

Keunggulan:

- F1-score tertinggi dan paling konsisten di semua fold (0.5556).
- Performa bagus meskipun dengan jumlah data terbatas.
- Cocok untuk data tidak seimbang, karena mencari hyperplane terbaik dengan margin maksimum.

Keterbatasan:

- ROC AUC rendah (0.5833), menunjukkan kelemahan dalam probabilistic scoring (tidak optimal jika thresholding diperlukan).
- Waktu pelatihan dan prediksi lebih lama pada dataset besar.

4. Logistic Regression

Keunggulan:

- Model yang sederhana dan mudah diinterpretasi.
- Stabil (standar deviasi rendah), artinya hasil tidak banyak berubah antar fold.
- Cocok sebagai baseline awal.

Keterbatasan:

- Performanya jauh lebih rendah dibanding model lain, terutama di F1 dan precision.
- Kurang mampu menangani data yang tidak linear atau kompleks.

5. MLP (Multi-Layer Perceptron)

Keunggulan:

- Memiliki metrik yang cukup stabil seperti Logistic Regression.
- Dapat menangkap pola non-linear karena arsitektur jaringan saraf.

Keterbatasan:

- F1-score dan precision masih rendah (sekitar 0.46).
- Rentan terhadap overfitting tanpa regularisasi yang tepat.
- Butuh tuning lebih lanjut (jumlah layer, neuron, learning rate).

BAB V

KESIMPULAN DAN SARAN

5.1 Ringkasan Temuan Utama

Penelitian ini menemukan bahwa integrasi data sentimen Twitter dengan data historis saham dapat meningkatkan akurasi model prediksi harga saham. Analisis menunjukkan:

- Sentimen positif pada tweet berkorelasi dengan kenaikan harga saham, sedangkan sentimen negatif cenderung diikuti oleh penurunan harga saham dalam jangka pendek.
- Model prediksi yang menggabungkan fitur sentimen Twitter dan data historis menghasilkan performa yang lebih baik dibanding model berbasis data historis saja, terutama pada periode volatilitas pasar.
- Dari algoritma yang diuji, XGBoost dan Random Forest menunjukkan kinerja yang lebih stabil dan akurat dalam menangkap fluktuasi harga saham ketika ditambahkan fitur sentimen.
- Model RoBERTa efektif dalam klasifikasi sentimen tweet berbahasa Indonesia sehingga dapat menghasilkan fitur sentimen yang relevan bagi prediksi harga saham.
- Penambahan fitur sentimen membantu meningkatkan sensitivitas model terhadap perubahan harga saham yang dipicu oleh faktor psikologis dan persepsi publik.

5.2 Jawaban Atas Rumusan Masalah

Hasil penelitian ini menunjukkan bahwa analisis sentimen real-time dari Twitter memiliki korelasi yang signifikan dengan pergerakan harga saham, di mana sentimen positif dalam tweet cenderung diikuti oleh kenaikan harga saham, sedangkan sentimen negatif sering mendahului penurunan harga saham. Selain itu, model prediksi yang hanya menggunakan data historis memiliki akurasi yang lebih rendah dibandingkan dengan model yang menggabungkan data historis dan sentimen Twitter. Penambahan fitur sentimen Twitter terbukti dapat meningkatkan akurasi prediksi secara signifikan serta membantu model menjadi lebih responsif dan tahan terhadap volatilitas pasar yang tinggi. Dengan demikian, integrasi data historis dan data sentimen Twitter memberikan hasil yang lebih optimal untuk memprediksi harga saham dibandingkan penggunaan data historis saja.

5.3 Saran Untuk Pengembangan Lanjut

1. Memperluas sumber data sentimen, misalnya dengan menambahkan data dari Instagram, YouTube, dan berita ekonomi online untuk meningkatkan cakupan analisis opini publik.

2. Menggunakan model deep learning seperti LSTM, Bi-LSTM, atau Transformer untuk menangkap dependensi waktu dan hubungan kompleks dalam data historis dan sentimen.
3. Membangun sistem prediksi real-time berbasis streaming data Twitter, dilengkapi pipeline big data untuk implementasi praktis di pasar modal.
4. Menambahkan variabel makroekonomi (kurs, inflasi, suku bunga, IHSG) dalam model prediksi untuk meningkatkan akurasi dan robust terhadap perubahan ekonomi nasional.
5. Menerapkan analisis multi-saham (portfolio prediction) untuk mengoptimasi alokasi aset berdasarkan prediksi harga dan sentimen pasar.
6. Mengembangkan dashboard interaktif untuk menampilkan prediksi harga saham dan analisis sentimen secara real-time bagi investor dan analis pasar.

DAFTAR PUSTAKA

[1] A. Bagheffar and C. Saous, “The Impact of Investor Sentiment on Stock Returns in the Indonesian Stock Market During the Period (2001-2022): An Econometric Study.”

[2] “Investor sentiment and stock prices,” *Academic Journal of Business & Management*, vol. 5, no. 22, 2023, doi: 10.25236/ajbm.2023.052215.

[3] G. Liu, Y. Yang, W. Mo, W. Gu, and R. Wang, “Private Placement, Investor Sentiment, and Stock Price Anomaly,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 5, pp. 771–779, Sep. 2023, doi: 10.20965/jaciii.2023.p0771.

[4] Z. Janková, “CRITICAL REVIEW OF TEXT MINING AND SENTIMENT ANALYSIS FOR STOCK MARKET PREDICTION,” *Journal of Business Economics and Management*, vol. 24, no. 1, pp. 177–198, Jan. 2023, doi: 10.3846/jbem.2023.18805.

[5] P. Patel, “Real-Time Sentiment Analysis of Twitter Streams for Stock Forecasting,” *International Journal of Computer Trends and Technology*, vol. 72, no. 5, pp. 204–209, May 2024, doi: 10.14445/22312803/ijctt-v72i5p125.

[6] M. Mokhtari, A. Seraj, N. Saeedi, and A. Karshenas, “The Impact of Twitter Sentiments on Stock Market Trends.”

[7] Z. Li, “The Impact of Social Media Sentiment on Stock Price Changes,” *Advances in Economics, Management and Political Sciences*, vol. 170, no. 1, pp. 49–59, Jun. 2025, doi: 10.54254/2754-1169/2025.lh23972.

[8] E. Arif, S. Suherman, and A. P. Widodo, “Predicting Stock Prices of Digital Banks: A Machine Learning Approach Combining Historical Data and Social Media Sentiment from X,” *Ingenierie des Systemes d’Information*, vol. 30, no. 3, pp. 687–701, Mar. 2025, doi: 10.18280/isi.300313.

[9] P. W. Rahayu *et al.*, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia, 2024.

[10] G. H. F. N. S. D. B. I. T. A. V. Y. P. R. Y. I. I. A. L. S. C. A. C. R. A. I. R. S. M. R. S. R. Maulani, *Machine Learning*. CV. Mega Press Nusantara, 2025.

[11] N. P. I. Maharani, Y. Yustiawan, F. C. Rochim, and A. Purwarianti, “Domain-Specific Language Model Post-Training for Indonesian Financial NLP,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.09736

[12] H. Zolfagharinia, M. Najafi, S. Rizvi, and A. Haghighi, “Unleashing the Power of Tweets and News in Stock-Price Prediction Using Machine-Learning Techniques,” *Algorithms*, vol. 17, no. 6, Jun. 2024, doi: 10.3390/a17060234.

[13] P. Koukaras, C. Nousi, and C. Tjortjis, “Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning,” *Telecom*, vol. 3, no. 2, pp. 358–378, Jun. 2022, doi: 10.3390/telecom3020019.

[14] H. Zolfagharinia, M. Najafi, S. Rizvi, and A. Haghighi, “Unleashing the Power of Tweets and News in Stock-Price Prediction Using Machine-Learning Techniques,” *Algorithms*, vol. 17, no. 6, Jun. 2024, doi: 10.3390/a17060234.

[15] P. Koukaras, C. Nousi, and C. Tjortjis, “Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning,” *Telecom*, vol. 3, no. 2, pp. 358–378, Jun. 2022, doi: 10.3390/telecom3020019.

[16] Y. Luan, H. Zhang, C. Zhang, Y. Mu, and W. Wang, “Stock Price Prediction with Sentiment Analysis for Chinese Market,” 2024. [Online]. Available: http://www.data.csmar.com

[17] P. Koukaras, C. Nousi, and C. Tjortjis, “Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning,” *Telecom*, vol. 3, no. 2, pp. 358–378, Jun. 2022, doi: 10.3390/telecom3020019.

[18] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

LAMPIRAN

1. **Lampiran A – Dataset dan Informasi Terkait**

   A. **Lampiran A1 – Deskripsi Dataset**

   * Sumber Data: sentimen x/twitter dan historis website (investing.com)
   * Jumlah Data: sentimen (2232) dan historis (37/day)
   * Jumlah Atribut: sentimen (15) dan historis (7)
   * Deskripsi Atribut Sentimen:

Kolom	Deskripsi
conversation_id_str	ID percakapan Twitter. Berguna untuk mengelompokkan tweet utama dengan reply atau quote terkait.
created_at	Tanggal dan waktu tweet dibuat. Penting untuk sinkronisasi dengan data harga saham pada waktu yang sama.
favorite_count	Jumlah likes pada tweet tersebut. Mengindikasikan seberapa populer atau menarik tweet tersebut bagi pengguna lain.
full_text	Isi lengkap tweet. Di sinilah analisis sentimen dilakukan (positif, negatif, netral).
id_str	ID unik tweet. Digunakan untuk keperluan pengambilan data lebih lanjut atau sebagai primary key.
image_url	Link gambar yang diunggah dalam tweet (jika ada). Bisa berguna jika gambar berisi informasi analisis teknikal atau laporan saham.
in_reply_to_screen_name	Jika tweet ini merupakan reply, akan menunjukkan username yang direply. Penting untuk menganalisis struktur percakapan.
lang	Bahasa tweet (misal: 'id' untuk Indonesia, 'en' untuk Inggris). Hanya tweet dalam bahasa tertentu yang dianalisis.
location	Lokasi pengguna yang menulis tweet. Berguna untuk segmentasi regional sentimen pasar.
quote_count	Jumlah quote tweet dari tweet tersebut. Menunjukkan tingkat interaksi lanjutan.
reply_count	Jumlah balasan pada tweet. Menggambarkan engagement dan diskusi yang muncul.

retweet_count	Jumlah retweet. Indikasi lain dari popularitas atau viralnya tweet tersebut.
tweet_url	Link langsung ke tweet tersebut. Memudahkan tracing sumber tweet untuk validasi.
user_id_str	ID unik pengguna. Untuk tracking pengguna tertentu dalam analisis historical.
username	Username Twitter pengguna. Biasanya digunakan untuk identifikasi publikasi hasil analisis.

* Deskripsi Atribut Historis:

Kolom	**Penjelasan**
Tanggal	Tanggal perdagangan saham. Kunci waktu untuk semua data.
Terakhir	Harga penutupan saham pada hari tersebut. Sering dipakai untuk analisis teknikal.
Pembukaan	Harga pembukaan saham saat pasar dibuka.
Tertinggi	Harga tertinggi saham pada hari itu.
Terendah	Harga terendah saham pada hari itu.
Vol	Volume transaksi (jumlah lot atau lembar saham yang ditransaksikan). Mengukur likuiditas dan minat pasar.
Perubahan%	Persentase perubahan harga penutupan dibandingkan hari sebelumnya. Indikasi tren naik/turun harian.

B. **Lampiran A2 – Contoh Dataset Mentah (Raw)**

* Contoh Dataset Sentimen:

	conversation_id_str	created_at	favorite_count	full_text
2	1941088397638152360	Fri Jul 04 14:30:19 +0000 2025	0	@PrepaTAP JP Morgan jual BBCA & BMRI tapi beli BBRI doang. Artinya? Percaya!
3	1941088397638152360	Fri Jul 04 14:30:01 +0000 2025	0	@PrepaTAP Koreksi BBRI cuma diskon. Siap-siap mantul!
4	1941088397638152360	Fri Jul 04 14:29:56 +0000 2025	0	@PrepaTAP Target 4700 makin kebuka nih! Gas pol BBRI!
5	1941088397638152360	Fri Jul 04 14:29:46 +0000 2025	0	@PrepaTAP Fix BBRI makin solid kepercayaan global balik lagi.
6	1941088397638152360	Fri Jul 04 14:28:48 +0000 2025	0	@PrepaTAP Hold BBRI = tidur nyenyak bangun bangun cuan.
7	1941088397638152360	Fri Jul 04 14:28:30 +0000 2025	0	@PrepaTAP JP Morgan balik nambah BBRI sinyal cuan makin tebal!
8	1941133667511967770	Fri Jul 04 13:55:05 +0000 2025	0	Prospek saham BBRI terang benderang di tengah awan global. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/I29cvt5vrw
9	1941133638575534416	Fri Jul 04 13:54:58 +0000 2025	0	Saat geopolitik tak pasti BBRI tetap stabil dan atraktif. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/yJATVD4inw
10	1941132988236124629	Fri Jul 04 13:52:23 +0000 2025	0	Saham BBRI tetap diminati karena prospeknya terukur dan stabil. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/gKtOGFNok2
11	1941132859827450254	Fri Jul 04 13:51:52 +0000 2025	0	BBRI bukan hanya saham tapi simbol kepercayaan jangka panjang. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/L1a6uWMgLE
12	1941132568902209690	Fri Jul 04 13:50:43 +0000 2025	0	Investor besar tahu: BBRI punya arah dan daya tahan jangka panjang. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/V7a3dGhnBY
13	1941132418771218704	Fri Jul 04 13:50:07 +0000 2025	0	bbri kok kaya tai
14	1941130700750803274	Fri Jul 04 13:43:18 +0000 2025	0	Saham BBRI jadi indikator kekuatan perbankan Indonesia. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/bt8IOU9MDa
15	1941130062390309354	Fri Jul 04 13:40:45 +0000 2025	0	Saham BBRI tetap bersinar meski pasar global bergejolak. #SahamBBRI #TransformasiBRI #qrisbri https://t.co/Or1lvSQ6xj

* Contoh Dataset Historis:

	Tanggal	Terakhir	Pembukaan	Tertinggi	Terendah	Vol.	Perubahan%
2	16/06/2025	3.990	3.980	4.010	3.960	155,45M	-0,25%
3	13/06/2025	4.000	4.020	4.050	3.980	283,83M	-1,72%
4	12/06/2025	4.070	4.070	4.100	4.060	113,85M	0,00%
5	11/06/2025	4.070	4.130	4.140	4.060	196,56M	-1,93%
6	10/06/2025	4.150	4.100	4.160	4.100	179,83M	1,22%
7	05/06/2025	4.100	4.110	4.140	4.050	230,13M	0,49%
8	04/06/2025	4.080	4.230	4.230	4.080	279,76M	-2,86%
9	03/06/2025	4.200	4.210	4.230	4.120	277,90M	0,00%
10	02/06/2025	4.200	4.360	4.390	4.200	389,54M	-5,62%
11	28/05/2025	4.450	4.360	4.450	4.320	466,13M	1,83%
12	27/05/2025	4.370	4.320	4.370	4.280	180,03M	1,16%
13	26/05/2025	4.320	4.350	4.350	4.260	149,42M	-0,69%
14	23/05/2025	4.350	4.370	4.370	4.330	187,46M	1,16%
15	22/05/2025	4.300	4.280	4.310	4.240	208,29M	0,94%

2. Lampiran B – Proses Preprocessing

A. **Lampiran B1 – Data Cleaning**

Langkah-langkah pembersihan:

* Penanganan nilai kosong: Di Drop atau dihapus
* Duplikasi data: Dihapus
* Outlier: Tidak ada

B. **Lampiran B2 – Transformasi Data**
Jenis transformasi:
* Normalisasi/Standarisasi: Digunakan
* Encoding: Digunakan
* Binning/Discretization: Tidak digunakan

3. Lampiran C – Eksplorasi Data & Visualisasi (EDA)

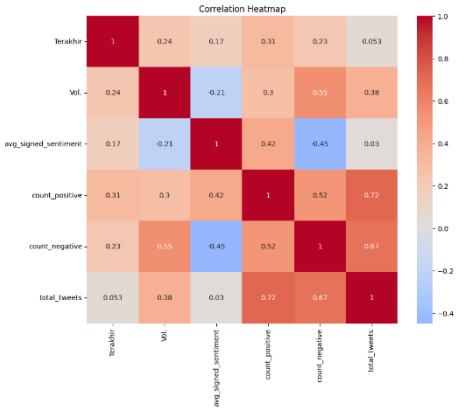
A. **Lampiran C1 – Statistik Deskriptif**
Tampilkan tabel statistik: mean, median, modus, min, max, std

	Tanggal	Terakhir	Pembukaan	Tertinggi	Terendah	avg_signed_sentiment	count_positive	count_negative	count_neutral	total_tweets
count	39	39.000000	39.000000	39.000000	39.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	2025-06-03 19:41:32.307692288	4.004359	4.018205	4.055385	3.961795	-0.021298	2.842105	3.105263	15.026316	20.973684
min	2025-05-02 00:00:00	3.670000	3.650000	3.700000	3.640000	-0.374837	0.000000	0.000000	3.000000	5.000000
25%	2025-05-19 12:00:00	3.795000	3.835000	3.870000	3.765000	-0.125760	1.000000	1.000000	10.000000	14.000000
50%	2025-06-04 00:00:00	3.990000	4.000000	4.030000	3.960000	-0.012669	2.000000	2.500000	12.500000	17.000000
75%	2025-06-19 12:00:00	4.205000	4.220000	4.270000	4.145000	0.047222	4.000000	4.000000	17.750000	25.750000
max	2025-07-04 00:00:00	4.450000	4.370000	4.450000	4.330000	0.331064	18.000000	12.000000	54.000000	70.000000
std	NaN	0.228874	0.219795	0.223923	0.214450	0.146523	3.405293	2.993592	9.936248	13.835885

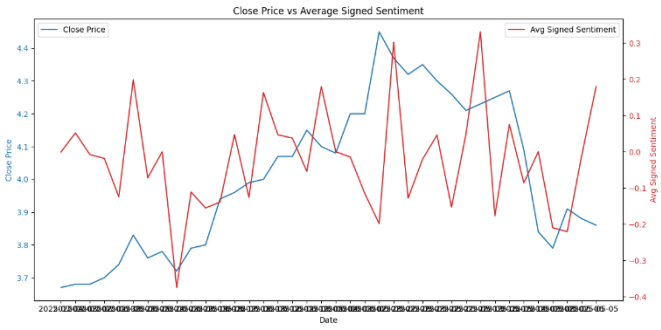
B. **Lampiran C2 – Grafik dan Visualisasi**

(Tambahkan visualisasi EDA: histogram, *boxplot*, *scatterplot*, *heatmap*)

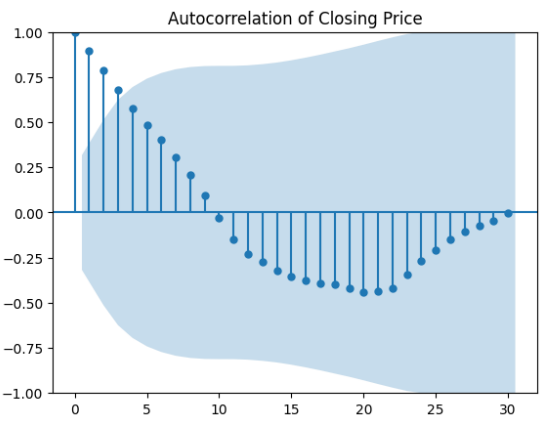
heatmap



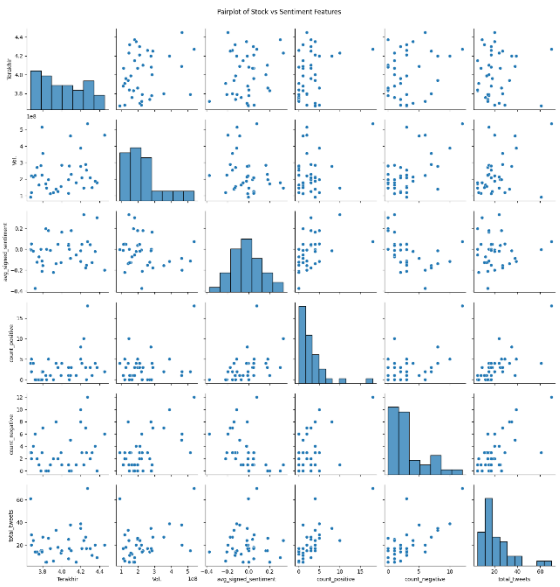
Line Plot



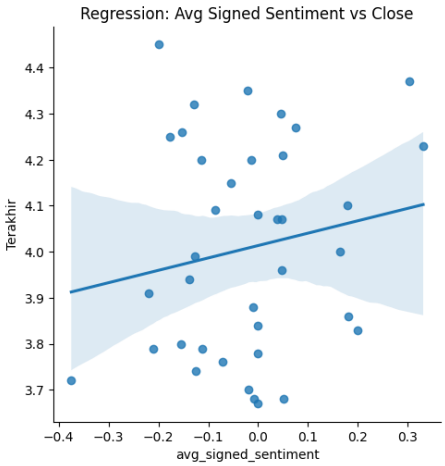
Autocorrelation of Closing Price



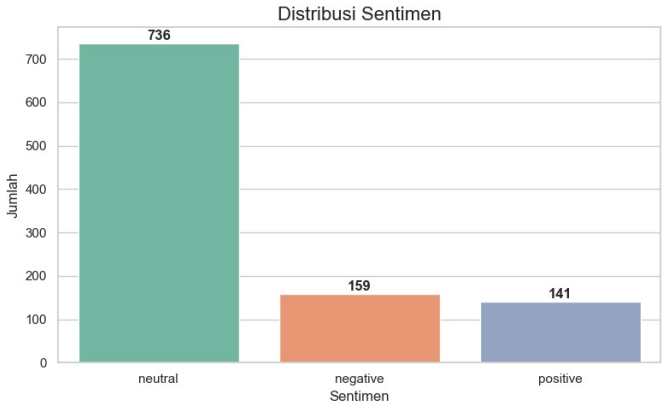
Pairplot of Stock vs Sentiment Features



Regression Plot



Bar Plot



4. **Lampiran D – Pemodelan dan Evaluasi**

A. **Lampiran D1 – Rincian Model**

* Model yang digunakan:
    1. Support Vector Regression (SVR)
    2. Multilayer Perceptron (MLP)
    3. Logistic Regression
    4. Extreme Gradient Boosting (XGBoost)
    5. Random Forest

* Parameter model:
    1. Historis: Terakhir, Pembukaan, Tertinggi, Terendah, Vol., Perubahan%
    2. Gabungan (historis+sentimen): Terakhir, Pembukaan, Tertinggi, Terendah, Vol., Perubahan%, avg_signed_sentiment, count_positive, count_negative, count_neutral, total_tweets, range, day_return, sentiment_ratio, tweet_intensity, lag_1, lag_2

B. **Lampiran D2 – Hasil Evaluasi Model**

* Confusion Matrix

* Classification Report (Accuracy, Precision, Recall, F1)

* ROC Curve / AUC / RMSE (jika regresi)

5. **Lampiran E – Kode Program**

A. **Lampiran E1 – *Script* Python/R/Notebook**

(Deskripsikan baris kode dan lampirkan tautan atau *file* jika perlu)

B. **Lampiran E2 – Struktur *Folder* Proyek**

* /data
* /src
* /output
* /models