

## BAB II

### LANDASAN TEORI

#### 2.1 Penelitian terkait

Berikut adalah penelitian terdahulu yang mengkaji korelasi antara sentimen media sosial dan pergerakan harga saham PGAS.

Tabel 2. 1 Penelitian terkait

Judul Artikel, Penulis	Objek Penelitian	Metode	Sumber Data	Findings
<b>Stocks Price Forecasting Based On XGBoost and LSTM, Pham Hoang Vuong<sup>1</sup>, Trinh Tan Dat<sup>1</sup>, Tieu Khoi Mail, Pham Hoang Uyen<sup>2</sup> and Pham The Bao</b> [6]	Pengembangan dan evaluasi sistem peramalan harga saham (stock-price forecasting), khususnya membandingkan kinerja metode gabungan XGBoost + LSTM dengan model ARIMA sebagai baseline.	XGBoost, LSTM, & ARIMA	Data Forex (1 Januari 2008 – 19 Maret 2018)	<b>MSE:</b> $3.465 \times 10^{-7}$ (XGBoost+LSTM) vs $6.114 \times 10^{-7}$ (ARIMA) <b>MAE:</b> $3.825 \times 10^{-4}$ vs $4.149 \times 10^{-4}$ <b>RMSE:</b> $5.887 \times 10^{-4}$ vs $7.819 \times 10^{-4}$ Hasil ini mengindikasikan peningkatan akurasi peramalan harga saham dengan memanfaatkan seleksi fitur berbasis XGBoost diikuti LSTM deep learning
Investing Deep Stock Market Forecasting With Sentiment Analysis, Charalampos M. Liapis, Aikaterini Karanikola and Sotiris Kotsiantis[7]	Mengevaluasi dan membandingkan efektivitas berbagai metode <i>deep learning</i> dan state-of-the-art untuk peramalan pasar saham dengan menggabungkan data hasil <i>sentiment analysis</i> ke dalam ruang fitur model.	LSTM, GRU, & VADER	Data Saham (2 Januari 2018 – 24 Desember 2020), & Data Twitter.	Menambahkan fitur berbasis sentimen (TextBlob, VADER, FinBert) Bersama dengan harga penutupan yang dihaluskan (rolling mean -7) umumnya meningkatkan akurasi peramalan.
A Deep Learning-Based LSTM for Stock Price Prediction Using	Mengembangkan dan mengevaluasi model peramalan harga saham	LSTM & XGBoost	Kaggle, VADER,& TextBlob	Latih model LSTM (dua layer LSTM 50 unit + <i>Dense</i> layer) dan

Twitter Sentiment Analysis, Shimaa Ouf1, Mona El Hawary, Amal Aboutabl, Sherif Adel[8]	berbasis <i>deep learning</i> (LSTM) dan <i>ensemble learning</i> (XGBoost) pada tiga perusahaan (Apple, Google, Tesla), dengan fokus pada pengaruh integrasi data sentimen Twitter terhadap akurasi prediksi harga saham			XGBoost (dengan <i>grid search</i> hyperparameter) pada split data 90% latih (termasuk validasi 10%) dan 10% uji.
ID-SMSA: Indonesia stock market dataset for sentiment analysis, Jason Hartanto, Timothy Liundi, Rhio Sutoyo, Esther Widhi Andangsari.[9]	Penelitian ini memperkenalkan <b>ID-SMSA: Indonesian Stock Market Dataset for Sentiment Analysis</b> , yaitu kumpulan data tweet dalam Bahasa Indonesia (dengan terjemahan Bahasa Inggris) yang membahas 10 emiten berkapitalisasi pasar terbesar di Bursa Efek Indonesia per Maret 2023. Total terdapat <b>3.288 tweet</b> yang telah dilabeli sentimen positif, netral, atau negatif, menghasilkan dataset domain-spesifik pertama untuk analisis sentimen pasar saham Indonesia .	VADER	Crawldata X (Twitter), 12 Januari 2021 – 1 Maret 2024	ID-SMSA berisi 3.288 tweet tentang 10 emiten terbesar di BEI dengan distribusi sentimen 1.769 positif, 733 netral, 786 negatif dan tingkat kesepakatan anotator (Cohen's kappa) 0,779

Enhancing Sentiment analysis in stock market tweets through BERT-Based Knowledge Transfer, Emre Cicekyurt, Gokhan Bakal[10]	Pengaruh transfer pengetahuan berbasis BERT (FinBERT) untuk meningkatkan performa analisis sentimen tweet pasar saham .	BERT, LSTM, CNN,& XGBoost	Data Crawling Twitter (X), 9 Apr – 16 Jul 2020	Penambahan FinBERT-labeled data meningkatkan rata-rata F1-score tradisional sebesar 17% dan deep learning sebesar 20%, dengan model Random Forest unigram mencapai F1=69% (vs 54% awal), menegaskan korelasi positif kuat antara ukuran data latih dan performa klasifikasi
A Comparative Analysis of LSTM, ARIMA, XGBoost Algorithms in Predicting Stock Price Direction, Aiyegbeni Gifty , Dr. Yang Li[11]	Penelitian ini bertujuan mengumpulkan data histori harga penutupan saham Google selama lima tahun dan mengembangkan kerangka peramalan harga saham harian menggunakan algoritma XGBoost, LSTM, dan ARIMA .	XGBoost, ARIMA, & LSTM	Kaggle	XGBoost mencapai R <sup>2</sup> 99,35%, MAE 17,63, RMSE 30,24. LSTM mencapai R <sup>2</sup> 96,87%, MAE 49,35, RMSE 57,28. ARIMA mencapai R <sup>2</sup> 66,37%, MAE 140,12, RMSE 188,11. XGBoost ter-tuning meningkat menjadi R <sup>2</sup> 99,47%, MAE 15,98, RMSE 27,34. Gifty and Yang paper 20...Gifty and Yang paper 20... Ensemble tiga model mencetak R <sup>2</sup> 78,63%, MAE 149,94, RMSE 104,64. Gifty and Yang paper 20...Gifty and Yang paper 20... Ensemble XGBoost + LSTM mencetak R <sup>2</sup> 81,10%, MAE 107,20, RMSE 141,03.

## 2.2 Teori Penelitian

### 2.2.1 Saham

Saham merupakan bukti kepemilikan atas sebuah perusahaan. Seorang pemilik saham memiliki hak atas aset dan pendapatan dari perusahaan yang bersangkutan. Dua alasan utama seseorang membeli saham dari suatu perusahaan yaitu untuk mendapatkan bagian keuntungan yang diperoleh perusahaan atau disebut juga dividen dan menjual kembali saham dengan harga yang lebih tinggi. Saham yang dapat dijual dan dibeli masyarakat umum diperdagangkan di pasar modal. Bursa Efek Indonesia (BEI) merupakan pasar modal Indonesia dimana saham-saham perusahaan yang sudah terdaftar diperdagangkan. Sebuah perusahaan memperdagangkan sahamnya untuk mendapatkan modal tambahan. Agar sahamnya dapat diperdagangkan di pasar saham, sebuah perusahaan terlebih dahulu harus melalui proses *Initial Public Offering* (IPO). [12]

Tingkatan dukungan sosial dari lingkungan sekitar sangat memengaruhi kesiapan seseorang dalam mengambil keputusan. Saat individu merasakan adanya dorongan, pengakuan, atau dorongan moral dari keluarga, teman, atau rekan kerja, ia cenderung lebih percaya diri dan termotivasi untuk bertindak. Dukungan ini tidak hanya berupa pujian atau semangat verbal, tetapi juga bantuan konkret-misalnya, pemberian informasi, sumber daya, atau contoh perilaku positif-yang memperkuat keyakinan bahwa tindakan yang diambil akan mendapat hasil yang konstruktif. Tanpa adanya dukungan tersebut, niat baik sekalipun sering kali berhenti di tahap pemikiran karena rasa ragu atau takut akan kegagalan.

Selain dukungan sosial, niat seseorang untuk bertindak merupakan kompas internal yang mengarahkan keputusan. Niat ini terbentuk dari gabungan keyakinan atas manfaat tindakan, norma pribadi, dan ekspektasi terhadap hasil. Semakin kuat persepsi bahwa tindakan akan membawa keuntungan-baik secara materi, emosional, maupun status-semakin besar energi psikologis yang dicurahkan untuk merencanakan dan melaksanakan langkah konkret. Dalam konteks pengambilan keputusan, niat juga dipengaruhi oleh self-efficacy sejauh mana individu merasa

memiliki kemampuan dan kontrol atas proses tersebut. Tanpa niat yang jelas dan kuat, seseorang mudah terombang-ambing oleh keraguan meski kondisi eksternal mendukung.

Namun, niat dan dukungan sosial saja tidak cukup jika situasi yang memungkinkan tidak hadir. Keberadaan yang tepat-misalnya peluang yang nyata, waktu yang sesuai, atau kondisi lingkungan yang kondusif-akan menentukan apakah tindakan benar-benar direalisasikan. Situasi ini mencakup aspek fisik (akses ke sarana, kondisi cuaca, lokasi), kontekstual (aturan, kebijakan, budaya organisasi), dan temporal (deadline, ketersediaan waktu). Misalnya, meski seseorang sangat berniat memulai usaha baru dan didukung penuh, tanpa modal awal atau izin usaha yang memadai, niat itu hanya akan mengendap. Dengan demikian, teori Snehendu B. Kar menegaskan bahwa interaksi sinergis antara dukungan sosial, niat internal, dan kondisi situasional menjadi kunci utama agar individu benar-benar bertindak sesuai keputusan yang diambil[11]. Membahas akan permasalahan sebelumnya, maka dapat diambil point penting bahwa terdapat situasi yang memungkinkan untuk para investor tersebut mengambil suatu tindakan dalam menentukan permasalahan saham. Suatu tindakan tersebut bisa berasal dari hasil analisa yang digunakan untuk menentukan akan permasalahan dari saham tersebut. Mengetahui akan itu, maka saham bisa menjadi point bagi para investor tersebut untuk mengambil tindakan apa yang harus dilakukan dari hasil analisa menggunakan metode tertentu. Namun, dalam berinvestasi saham, tentu saja ada risiko yang perlu diperhatikan. Harga saham dapat berubah-ubah dengan cepat dan seringkali sulit diprediksi. Mengingat para investor kerap kali bertujuan untuk mendapatkan keuntungan sebesar - besarnya, maka analisa akan bagaimana saham itu bekerja di masa depan menjadi pekerjaan mereka. Untuk mendapatkan keuntungan yang besar, para investor dapat melakukan analisis dalam memprediksi harga saham [12].

### **2.2.2 Investasi**

Investasi adalah sebuah komitmen untuk menempatkan sejumlah dana kepada suatu proyek dengan adanya harapan untuk mendapatkan keuntungan di masa yang akan datang[14]. Investasi melibatkan penempatan dana pada berbagai instrumen keuangan atau aset riil-mulai dari saham, obligasi, dan reksa dana hingga logam mulia seperti emas, properti tanah, bahkan aset digital seperti cryptocurrency. Masing-masing kelas aset ini memiliki karakteristik risiko, likuiditas, dan potensi imbal hasil yang berbeda. Saham bisa memberi pertumbuhan modal dan pembagian keuntungan secara rutin, tetapi juga bisa berubah-ubah tergantung pasar. Emas biasanya dianggap sebagai pilihan aman ketika ekonomi tidak stabil, meskipun tidak memberi hasil uang secara berkala. Sementara itu, *cryptocurrency* berpotensi memberi keuntungan besar dalam waktu singkat, tetapi sangat fluktuatif dan membutuhkan pemahaman tentang teknologi serta aturan yang dalam.

Tujuan utama berinvestasi adalah mencapai imbal hasil yang melebihi besaran modal awal, sehingga membangun aliran pendapatan pasif-seperti dividen dari saham atau sewa properti-yang dapat menambah atau bahkan menggantikan penghasilan aktif. Dengan pendapatan pasif yang terkelola baik, investor dapat memanfaatkan efek bunga majemuk, di mana keuntungan yang diperoleh di periode pertama diinvestasikan kembali untuk menghasilkan keuntungan tambahan pada periode berikutnya. Strategi ini sering dijadikan tolok ukur bagi pencapaian kebebasan finansial (*financial freedom*), di mana seseorang memiliki aset dan arus kas cukup untuk menutup biaya hidup tanpa ketergantungan pada gaji pekerjaan tetap.

Namun, mencapai tujuan keuangan jangka panjang membutuhkan perencanaan yang matang, seperti mencermati diversifikasi portofolio untuk mengurangi risiko, menentukan masa investasi sesuai dengan tingkat kemampuan mengambil risiko, serta memiliki disiplin untuk tidak tergoda melakukan spekulasi berdasarkan perkiraan pasar. Investor harus memahami bahwa setiap instrumen memiliki siklus dan reaksi yang berbeda terhadap kondisi ekonomi, geopolitik, atau inovasi teknologi. Dengan menyesuaikan alokasi dana berdasarkan tujuan, toleransi risiko,

dan periode investasi, seorang investor dapat menavigasi berbagai tantangan pasar dan memaksimalkan peluang untuk meraih kebebasan finansial yang berkelanjutan. [15]. Investasi jangka pendek umumnya memiliki horizon waktu kurang dari satu hingga dua tahun dan sering kali dipilih oleh investor yang menginginkan likuiditas tinggi serta peluang keuntungan cepat. Dalam praktiknya, investor akan memantau pergerakan pasar saham harian-misalnya menargetkan saham perusahaan dengan tren pertumbuhan positif-lalu menjual kembali begitu harga telah naik sesuai ekspektasi. Strategi ini memungkinkan realisasi keuntungan yang jauh lebih besar dibandingkan simpanan deposito, meski memerlukan ketelitian dalam membaca grafik dan kesiapan menghadapi fluktuasi harga yang tinggi.

Investasi jangka menengah mencakup periode tiga hingga lima tahun dan cocok untuk tujuan finansial yang telah direncanakan, seperti membiayai pendidikan anak. Pada jenis investasi ini, pilihan instrumen sering jatuh pada produk yang menawarkan keseimbangan antara risiko dan imbal hasil, seperti reksa dana campuran atau obligasi korporasi berkualitas. Dengan jangka waktu yang lebih panjang, investor dapat meredam volatilitas pasar dan memanfaatkan efek akumulasi bunga atau dividen, sehingga modal awal berpotensi tumbuh lebih stabil dan terukur.

Sementara itu, investasi jangka panjang biasanya berlangsung lebih dari sepuluh tahun dan menekankan pertumbuhan nilai aset secara signifikan. Contoh klasiknya adalah pembelian properti atau tanah, di mana investor memperoleh passive income dari penyewaan sekaligus meraup keuntungan atas kenaikan nilai pasar properti dalam dekade berikutnya. Dengan sabar menahan periode penyesuaian harga dan memanfaatkan tren kenaikan nilai aset, investor jangka panjang berpeluang mencapai akumulasi kekayaan yang substansial dan mendekatkan diri pada kebebasan finansial [16].

### **2.2.3 Deep Learning**

Deep learning dengan *Long Short-Term Memory* (LSTM) unggul dalam menangkap ketergantungan jangka panjang pada data deret waktu-misalnya pola



musiman atau momentum harga-karena arsitekturnya yang mampu mengingat informasi dari rentang waktu yang luas dan mengurangi masalah vanishing gradient pada RNN konvensional. Sementara itu, *Extreme Gradient Boosting* (XGBoost) memanfaatkan ensemble pohon keputusan untuk memperbaiki prediksi dengan fokus pada meminimalkan residual dari model sebelumnya, menjadikannya sangat efektif dalam mengolah fitur teknis dan sentimen yang bersifat tabular. Dengan menggabungkan LSTM sebagai model urutan (*sequence model*) dan XGBoost untuk menangani sisa kesalahan (*residual*), studi seperti yang dilakukan Pham et al. (2021) menunjukkan peningkatan akurasi prediksi harga saham dibandingkan metode tradisional seperti ARIMA maupun model Tunggal. Optimalisasi hyperparameter pada LSTM juga terbukti meningkatkan kinerja dalam tugas-tugas prediksi risiko keuangan, bahkan mengungguli Random Forest dan XGBoost murni pada beberapa metrik seperti AUC. Teknologi ini memanfaatkan jaringan saraf tiruan (*artificial neural networks*) yang memiliki banyak lapisan untuk memproses data yang sangat besar dan kompleks. Dalam konteks ini, deep learning sangat efektif dalam mengekstraksi fitur-fitur tersembunyi dari data, yang sebelumnya sulit dikenali dengan metode analisis konvensional. Dengan kemampuannya untuk belajar secara otomatis dari data, deep learning tidak memerlukan ekstraksi fitur manual, yang menjadikannya alat yang sangat kuat dalam analisis keuangan, terutama untuk mendeteksi pola dan tren yang mungkin tidak terlihat oleh analis manusia.[13]

Penggunaan deep learning dalam analisis data keuangan juga sangat membantu dalam meningkatkan transparansi dan akurasi dalam pengambilan keputusan. Dalam hal ini, model-model deep learning dapat mengidentifikasi anomali atau potensi kecurangan dalam laporan keuangan, yang dapat membantu perusahaan dan investor mengurangi risiko dan membuat keputusan yang lebih terinformasi. Selain itu, dengan kemampuan untuk menganalisis data dalam jumlah besar dan mengidentifikasi pola tersembunyi, teknologi ini juga memungkinkan prediksi yang lebih akurat tentang pergerakan harga pasar atau perilaku keuangan lainnya. Seiring dengan semakin banyaknya data yang tersedia, penerapan deep learning



dalam bidang keuangan diharapkan akan terus berkembang, meningkatkan efisiensi dan efektivitas analisis dalam berbagai aspek keuangan dan ekonomi.[14]

#### 2.2.4 Analisis Sentimen

Secara kegunaan, analisis sentimen merupakan pengetahuan atau teknik yang digunakan untuk mencari emosi berdasarkan ulasan, baik dalam bentuk teks maupun verbal [23]. Teknik ini sering digunakan untuk mengeksplorasi perasaan yang terkandung dalam ulasan pelanggan terhadap produk tertentu. Dalam menggunakan analisis sentimen, maka diperlukan sebuah dataset yang telah memiliki label. Pendekatan dengan metode pembelajaran supervised merupakan pendekatan menggunakan pembelajaran mesin dan kecerdasan buatan yang memanfaatkan data dengan sifat telah terlabeli. Model tersebut akan mampu menghasilkan output yang akurat dan sesuai dengan label yang diinginkan [4]. Untuk memenuhi pelabelan terkait analisis sentimen, maka diperlukan suatu ulasan sebagai sumber data. Ulasan ini seringkali mengandung opini positif dan negatif tentang perubahan yang terjadi [24]. Pelabelan perlu untuk dilakukan, sebab dalam penelitian analisis sentimen, maka penelitian ini mencoba untuk mengklasifikasikan terkait apakah teks ulasan yang ditulis pengguna bersifat positif atau negatif [25]. Ulasan tersebut memberikan wawasan tentang dampak yang dihasilkan. Ini membantu dalam konteks analisis sentimen terkait ulasan teks untuk memahami pendapat yang terkait dengan masalah atau objek tertentu, apakah cenderung positif atau negatif [5]. Serta, dalam analisis sentimen merupakan teknik yang baik untuk memeriksa bagaimana masyarakat merespon akan terjadinya perubahan [26].

#### 2.2.5 Text Mining

Pada tahap *Text Mining* pertama-tama kita membuat salinan data mentah (`merged_df`) ke variabel `df_cleaned` agar perubahan tidak merusak data asli. Selanjutnya, setiap entri teks cuitan (`'full_text'`) dibersihkan melalui fungsi `clean_text()`, yang mencakup:

1. Penghapusan URL (semua pola `http...`, `www...`, atau `https...`) agar tautan tidak memengaruhi analisis,
  2. Penghapusan mention (`@username`) dan hashtag (`#tag`) karena kedua elemen ini bersifat metadata media sosial, bukan konten opini,
  3. Penghapusan angka untuk memfokuskan pada kata-kata bermakna,
  4. Penghapusan tanda baca non-alfanumerik supaya tokenisasi berjalan lebih bersih,
  5. Konversi seluruh teks menjadi huruf kecil dan pemangkasan spasi di kedua ujung teks.
- Hasilnya, kolom `clean_text` berisi versi teks yang siap di-tokenize dan diekstraksi fiturnya.

Setelah pembersihan, langkah berikutnya adalah inisialisasi *sentiment analyzer* menggunakan pustaka VADER (SentimentIntensityAnalyzer dari NLTK). VADER dipilih karena keandalannya dalam menganalisis teks singkat dan informal seperti Twitter, termasuk penanganan emotikon, intensifier, dan variasi huruf kapital. Dengan objek `sia` ini, setiap `clean_text` dihitung skor *compound*-nya, yaitu nilai kontinu antara  $-1$  (sangat negatif) hingga  $+1$  (sangat positif).

Untuk mengubah skor kontinu menjadi kategori sentiment yang mudah diinterpretasi, dibuat fungsi `get_sentiment(text)`:

1. Jika skor  $\geq 0.05$  maka dikategorikan “positif”,
2. Jika skor  $\leq -0.05$  maka “negatif”,
3. Sebaliknya masuk kategori “netral”.

Pemilihan ambang batas ini mengikuti rekomendasi default VADER untuk memisahkan sentimen kuat dan lemah.

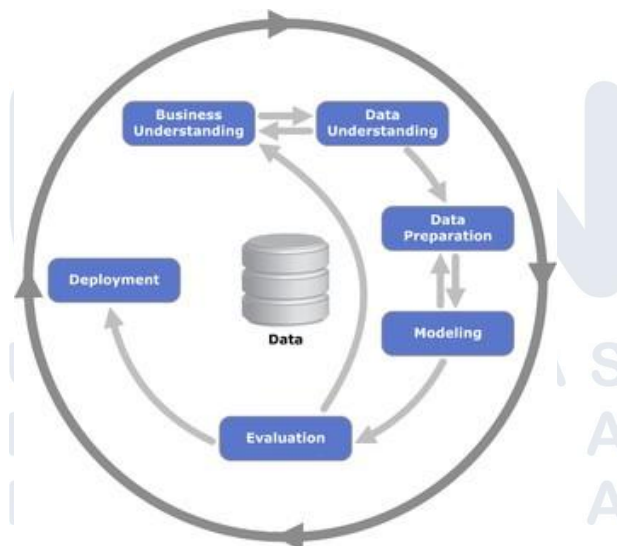
Terakhir, fungsi `get_sentiment` diterapkan ke seluruh kolom `clean_text`, menghasilkan kolom baru sentiment di `df_cleaned`. Untuk meninjau hasil awal, kita menampilkan 100 baris pertama dari kolom `clean_text` dan sentiment. Distribusi awal ini akan menjadi dasar validasi kualitas ekstraksi sentimen sebelum digabungkan sebagai fitur dalam model prediksi harga saham. Dengan demikian,

seluruh rangkaian kode di atas membentuk fondasi text mining-dari pembersihan teks hingga pengkodean opini-yang siap diintegrasikan ke dalam analisis kuantitatif selanjutnya.

## 2.3 Framework/Algoritma/SDLC Penelitian

### 2.3.1 CRISP-DM

*Cross Industry Standard Process For Data Mining* atau disingkat menjadi CRISP-DM dikembangkan pada tahun 1996 yang bertujuan untuk dijadikan *methodology standard non-proprietary* bagi data mining, ini ditujukan untuk melakukan proses analisis strategi pemecahan masalah dari suatu bisnis atau penelitian di suatu industri[15]. Berikut merupakan enam tahap dalam metodologi CRISP-DM, seperti business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Berikut merupakan langkah-langkah dalam mengimplementasikan CRISP-DM, yaitu:



Gambar 2. 1 Framework CRISP-DM

#### 1. Business Understanding

Tahap Business Understanding adalah langkah pertama yang sangat krusial dalam pengembangan model data mining. Pada tahap ini, dilakukan pemahaman mendalam terhadap masalah bisnis yang ingin diatasi, termasuk mengidentifikasi tujuan dan kebutuhan dari model yang akan dikembangkan. Proses ini bertujuan untuk memastikan bahwa model yang dibangun benar-benar relevan dan dapat memberikan solusi terhadap masalah yang dihadapi bisnis. Selain itu, tahap ini juga melibatkan penentuan ruang lingkup proyek, serta pemahaman tentang ketersediaan data yang ada dan sumber data yang bisa dimanfaatkan. Poin penting lainnya adalah pemetaan peran dan dampak model terhadap tujuan bisnis yang lebih besar, sehingga hasil akhirnya dapat memberikan kontribusi nyata terhadap keputusan strategis perusahaan.

## **2. Data Understanding**

Pada tahap Data Understanding, fokus utama adalah untuk memperoleh pemahaman yang lebih mendalam tentang data yang tersedia. Proses ini mencakup analisis awal terhadap dataset yang dimiliki, dengan menggunakan statistik deskriptif untuk mendapatkan gambaran umum tentang distribusi data, hubungan antar variabel, dan potensi pola yang bisa muncul. Visualisasi data digunakan untuk menggali insight lebih lanjut dan membantu dalam mendeteksi outlier yang mungkin mempengaruhi kualitas model. Selain itu, tahap ini juga melibatkan identifikasi masalah dalam data seperti inkonsistensi, duplikasi, atau data yang hilang, yang dapat mempengaruhi efektivitas analisis. Data yang telah dianalisis akan memberikan gambaran lebih jelas tentang kualitas dan relevansi data yang digunakan untuk penelitian selanjutnya.

## **3. Data Preparation**

Data Preparation adalah tahap penting yang memastikan bahwa data yang digunakan dalam pemodelan berada dalam kondisi terbaik. Pada tahap ini, dilakukan pembersihan data untuk menghilangkan kesalahan, inkonsistensi, atau duplikasi yang mungkin ada dalam dataset. Proses ini juga mencakup

penggabungan data dari berbagai sumber yang relevan, serta pengisian data yang hilang atau penghapusan data yang tidak relevan agar analisis berikutnya lebih akurat. Selain itu, transformasi data seperti normalisasi, encoding, atau pembuatan fitur baru juga dilakukan untuk memudahkan model dalam mengidentifikasi pola dan hubungan yang signifikan. Data yang sudah dipersiapkan dengan baik akan mempengaruhi kualitas dan akurasi dari model yang akan dibangun, sehingga tahap ini sangat penting untuk memastikan hasil yang optimal.

#### **4. Modelling**

Tahap Modeling adalah inti dari proses pengembangan model data mining. Pada tahap ini, dilakukan pemilihan algoritma yang paling sesuai dengan permasalahan yang ada, serta penyesuaian parameter yang diperlukan agar algoritma dapat bekerja secara maksimal. Proses ini melibatkan percobaan dengan berbagai model yang mungkin, seperti model regresi, klasifikasi, atau clustering, tergantung pada jenis masalah yang ingin diselesaikan. Selain itu, dilakukan pengujian terhadap performa model pada dataset yang tersedia untuk menentukan apakah model tersebut sudah memberikan hasil yang diinginkan. Penyesuaian dan optimasi terus dilakukan pada setiap iterasi untuk mendapatkan model terbaik yang dapat menggambarkan pola dalam data secara akurat. Keberhasilan tahap ini sangat bergantung pada pemilihan algoritma yang tepat dan kemampuan untuk mengoptimalkan parameter model.

#### **5. Evaluation**

Pada tahap Evaluation, model yang telah dibangun diuji untuk memastikan bahwa kinerjanya sesuai dengan ekspektasi dan dapat diaplikasikan pada data yang belum pernah dilihat sebelumnya (data uji). Proses evaluasi ini melibatkan pengukuran berbagai metrik kinerja, seperti akurasi, precision, recall, atau F1-score, tergantung

pada jenis masalah yang dihadapi. Selain itu, dilakukan juga analisis terhadap kekuatan dan kelemahan model dengan menggunakan data yang beragam untuk melihat apakah model mampu generalisasi dengan baik. Evaluasi ini bertujuan untuk memastikan bahwa model yang dibangun tidak hanya bekerja dengan baik pada data latihan, tetapi juga dapat diandalkan pada data nyata yang akan dihadapi di dunia bisnis. Hasil evaluasi ini akan memberikan gambaran tentang seberapa efektif model dalam memecahkan masalah yang ada.

## 6. Deployment

Tahap terakhir adalah Deployment, yaitu implementasi model yang telah dikembangkan ke dalam lingkungan produksi. Pada tahap ini, model yang sudah diuji dan dievaluasi diintegrasikan ke dalam sistem yang ada dan digunakan untuk membuat prediksi pada data baru secara real-time. Selain itu, dilakukan pemantauan terhadap performa model untuk memastikan bahwa model tetap berjalan dengan baik dan memberikan hasil yang akurat seiring berjalannya waktu. Pada tahap deployment ini, sering kali diperlukan pengaturan pemeliharaan dan pembaruan model secara periodik agar model tetap relevan dengan perubahan kondisi bisnis atau data yang baru. Keberhasilan tahap deployment akan mengarah pada penggunaan model dalam pengambilan keputusan bisnis yang nyata, memberikan manfaat langsung bagi organisasi atau perusahaan.

### 2.3.2 *XGBoost*

*XGBoost* atau *Extreme Gradient Boosting* adalah salah satu algoritma Deep Learning berbasis pohon keputusan (decision tree) yang sangat populer karena kecepatan dan akurasi prediksinya. Algoritma ini dikembangkan berdasarkan teknik boosting, di mana beberapa model lemah (weak learners)-biasanya berupa decision tree-digabungkan secara berurutan untuk membentuk model yang kuat (strong learner). Tujuannya adalah untuk memperbaiki kesalahan prediksi dari model sebelumnya, hingga tercipta model akhir yang lebih akurat dan andal.[16]

Cara kerja XGBoost dapat diibaratkan sebagai proses kolaboratif di mana model-model sederhana saling melengkapi untuk meningkatkan akurasi. Model pertama akan memprediksi hasil dasar, kemudian model berikutnya fokus memperbaiki kesalahan yang dibuat oleh model sebelumnya. Proses ini berlangsung secara berulang hingga keseluruhan kesalahan dapat diminimalkan. Dalam konteks ini, setiap pohon keputusan bertugas untuk mempelajari kesalahan yang tersisa dan terus menyempurnakan prediksi akhir secara bertahap. [17] [18]

XGBoost dikenal unggul dalam hal kecepatan dan kinerja, karena mampu menangani data berukuran besar dengan efisien dan tetap menghasilkan prediksi yang akurat. Algoritma ini sering digunakan dalam berbagai kompetisi data science karena keandalannya. Selain itu, XGBoost dilengkapi dengan mekanisme regularisasi yang efektif, sehingga model yang dihasilkan lebih stabil dan tidak mudah mengalami overfitting, yaitu kondisi saat model terlalu menyesuaikan diri dengan data pelatihan sehingga kurang mampu menggeneralisasi pada data yang belum pernah dilihat.

Dalam penelitian ini, XGBoost digunakan untuk membantu memprediksi harga saham PGAS dengan memperhitungkan berbagai faktor teknikal dan eksternal, termasuk sentimen berita atau sosial media. Tidak seperti LSTM yang fokus pada pola berurutan dalam waktu, XGBoost lebih fleksibel dalam mengolah berbagai jenis fitur sekaligus, bahkan yang tidak berurutan. Oleh karena itu, XGBoost menjadi pelengkap yang ideal bagi LSTM, karena dapat memperkaya prediksi dengan memasukkan variabel-variabel penting yang tidak ditangkap oleh LSTM.

### 2.3.3 LSTM (*Long Short Term Memory*)

*LSTM* atau *Long Short-Term Memory* merupakan salah satu bentuk dari *Recurrent Neural Network* (RNN) yang dirancang khusus untuk mengingat informasi dalam jangka waktu yang lebih panjang. Berbeda dengan RNN biasa yang mudah lupa terhadap informasi lama karena masalah *vanishing gradient*, LSTM mengatasi kelemahan ini dengan cara mengatur informasi mana yang perlu diingat dan mana yang perlu dilupakan melalui sistem gerbang (*gates*). [5], [19], [20]



Secara sederhana, LSTM bekerja seperti memori manusia: kita tidak mungkin mengingat semua hal, tapi kita bisa menyimpan informasi penting untuk digunakan di masa depan. Nah, dalam LSTM, hal ini dilakukan melalui tiga komponen utama, yaitu *forget gate*, *input gate*, dan *output gate*. Ketiganya bekerja sama untuk menyaring informasi-misalnya, apakah informasi dari hari kemarin masih penting untuk hari ini? Atau apakah pola dua minggu lalu punya pengaruh terhadap prediksi harga esok hari?

LSTM sangat cocok digunakan untuk kasus prediksi harga saham karena data saham bersifat berurutan (*time-series*), di mana nilai saat ini bergantung pada nilai sebelumnya. Maka dari itu, LSTM dapat mempelajari pola-pola dari waktu ke waktu dan melakukan prediksi dengan lebih akurat dibanding metode biasa. Dalam penelitian ini, LSTM digunakan untuk mempelajari pergerakan harga saham PGAS berdasarkan data historisnya. Dengan mempertimbangkan konteks waktu dan tren sebelumnya, LSTM dapat membantu menghasilkan prediksi harga yang lebih realistis.[21]

#### 2.3.4 VADER

VADER (*Valence Aware Dictionary and Sentiment Reasoning*) adalah model analisis sentimen yang dirancang khusus untuk teks media sosial. Model ini menggabungkan fitur leksikal yang dikurasi oleh manusia dengan aturan linguistik untuk menganalisis intensitas dan polaritas sentimen. Berbeda dengan alat analisis sentimen tradisional yang sering mengandalkan klasifikasi biner (positif atau negatif), VADER dapat mengukur intensitas sentimen dalam skala, memberikan analisis yang lebih rinci.[22]

Kunci efektivitas VADER terletak pada leksikonnya, yang mencakup lebih dari 7.500 fitur leksikal, seperti kata-kata, emotikon, akronim, dan istilah slang, yang semuanya diberi nilai intensitas sentimen. Fitur-fitur ini digabungkan dengan lima aturan umum yang memperhitungkan konvensi gramatikal dan sintaksis yang sering digunakan untuk menekankan sentimen, seperti tanda baca (misalnya, tanda seru), kapitalisasi (misalnya, huruf besar semua), dan pengubah derajat (misalnya, "sangat", "hampir").

Kinerja model ini diuji terhadap beberapa metode analisis sentimen yang telah mapan, termasuk LIWC, *SentiWordNet*, dan model pembelajaran mesin seperti *Naive Bayes* dan *Support Vector Machines*. Hasilnya menunjukkan bahwa VADER mengungguli model-model ini dalam beberapa konteks, khususnya dalam menganalisis konten media sosial seperti tweet, yang sering menggunakan bahasa informal dan emotikon. Menariknya, VADER bahkan mengalahkan pengklasifikasi manusia individu dalam mengklasifikasikan sentimen tweet.

Keunggulan VADER terletak pada kesederhanaannya, efisiensi komputasi, dan kemudahan penggunaannya. Model ini tidak memerlukan data pelatihan yang ekstensif atau model yang kompleks, menjadikannya sangat cocok untuk analisis sentimen secara real-time di platform media sosial. Selain itu, model ini mudah diperluas dan dipahami, yang sangat berguna bagi pengguna di luar bidang ilmu komputer, seperti sosiolog, psikolog, atau pemasar.

Secara keseluruhan, pendekatan berbasis aturan VADER menyediakan alat yang kuat untuk analisis sentimen, terutama untuk teks pendek dan informal yang khas di media sosial, dan mengungguli model-model lain dengan efisiensi dan kesederhanaan yang lebih besar.

## **2.4 Tools Penelitian**

Pada penelitian ini, beberapa perangkat lunak dipilih untuk mendukung pengolahan data, analisis sentimen, serta visualisasi alur kerja.

### **2.4.1 Python**

Python adalah bahasa pemrograman tingkat tinggi yang dirancang untuk memudahkan pengembangan perangkat lunak dengan sintaks yang jelas dan mudah dipahami. Dikenal karena kemampuannya yang serbaguna, Python digunakan dalam berbagai aplikasi mulai dari pengembangan web hingga analisis data, kecerdasan buatan (AI), dan machine learning. Kelebihan utama Python adalah sintaks yang bersih dan konsisten, yang memungkinkan pengembang untuk menulis kode yang lebih sedikit namun tetap kuat. Hal ini membuat Python sangat populer di kalangan pemula dan profesional di berbagai bidang. Salah satu fitur utama dari

Python adalah keberagaman pustaka dan framework yang mendukung pengembangan aplikasi, seperti *NumPy*, *Pandas*, *TensorFlow*, dan *Flask*, yang membuatnya menjadi pilihan utama di bidang analisis data dan pemrograman ilmiah.



Gambar 2. 2 Logo Python

Python juga memiliki komunitas yang besar dan aktif, yang secara teratur memperbarui dan mendukung pustaka serta alat pengembangan. Penggunaan Python terus berkembang, dan berkat berbagai aplikasi, kemudahan penggunaan, serta berbagai pustaka dan modul yang tersedia, Python telah menjadi bahasa yang sangat penting dalam dunia teknologi dan pengembangan perangkat lunak modern.[23]

#### 2.4.2 Jupyter Notebook

Jupyter Notebook merupakan implementasi praktis dari konsep literate programming, di mana kode, dokumentasi, dan hasil eksekusi (seperti grafik atau tabel) digabung dalam satu dokumen interaktif. Dokumen ini disimpan dalam format JSON dengan ekstensi *.ipynb*, memuat urutan *cells* baik kode, Markdown, maupun output lengkap dengan metadata kernel (misalnya versi bahasa dan pustaka yang digunakan). Keunggulan utamanya terletak pada kemampuannya mendukung eksplorasi data secara iteratif: peneliti atau developer dapat menjalankan potongan kode (per *cell*) secara terpisah, memodifikasi, dan langsung melihat dampaknya, sehingga mempercepat proses *debugging* dan pengembangan prototipe.



Gambar 2. 3 Gambar Jupyter

Di balik antarmuka web-nya, Jupyter Notebook memisahkan dua komponen kunci: frontend dan kernel. Frontend, yang berjalan di browser, menampilkan editor kode, toolbar, dan panel hasil eksekusi; sementara kernel dijalankan sebagai proses terpisah di mesin lokal atau server bertugas mengeksekusi perintah pengguna. Komunikasi antara keduanya difasilitasi oleh *ZeroMQ* (ZMQ), pustaka *asynchronous messaging* yang menggunakan pola *publish-subscribe* untuk *broadcast* event seperti output baru, dan pola *request-reply* untuk permintaan eksekusi kode. Semua pesan ini dikemas dalam format JSON sehingga mudah diurai dan diperluas, memungkinkan penambahan protokol baru atau integrasi dengan layanan eksternal.

Ekosistem Jupyter Notebook didukung oleh ekstensifitas *kernel* dan *extension*. Kernel default mendukung Python, tetapi ada juga *kernel* untuk R, Julia, Scala, dan puluhan bahasa pemrograman lainnya. Selain itu, komunitas pengembang telah menciptakan ratusan *extension* mulai untuk visualisasi data (misalnya *interactive widgets*), integrasi Git, hingga *code linting* dan *autocomplete* yang canggih. Keterbukaan format dokumen dan arsitektur plugin-based ini menjadikan Jupyter Notebook tidak hanya alat analisis data, tetapi juga platform kolaborasi tertulis dari

laporan ilmiah, tutorial interaktif, hingga presentasi langsung yang dapat dijalankan ulang (*reproducible*) di berbagai lingkungan komputasi.

#### 2.4.3 Google Colab

Google Colab memperluas fungsionalitas dari Jupyter Notebook dengan menerapkan virtualisasi melalui Docker dalam infrastruktur *Google Cloud Platform* (GCP), memungkinkan pengguna untuk menjalankan notebook dalam lingkungan runtime yang terisolasi. Setiap instance notebook di Colab dikelola menggunakan Kubernetes, yang mengatur kontainer secara otomatis untuk memastikan skalabilitas dan ketersediaan. Lapisan keamanan diperkuat dengan sistem antivirus yang memindai kontainer sebelum eksekusi kode, sehingga pengguna dapat bekerja tanpa khawatir terhadap malware atau kode berbahaya yang tersembunyi.



Gambar 2. 4 Gambar Logo Google Colab

Komunikasi antara antarmuka pengguna yang diakses melalui browser dan mesin eksekusi di backend memanfaatkan protokol gRPC dan HTTPS. Ketika pengguna mengirim perintah eksekusi, frontend mengemas permintaan dalam format JSON dan mengirimkannya melalui gRPC ke service orchestrator. Dari sana, gRPC juga menghubungkan orchestrator dengan kernel Python yang berjalan di dalam kontainer, memungkinkan eksekusi kode secara real-time dengan latensi minimal.

Setiap sesi notebook di Colab dilengkapi dengan manajemen otomatis terhadap penyimpanan dan versi dokumen. Google Drive diintegrasikan melalui API Drive v3, sehingga setiap perubahan-termasuk output grafik, tabel, dan checkpoint-disinkronkan secara otomatis ke akun Drive pengguna. Mekanisme checkpointing ini menjaga kontinuitas data: apabila koneksi terputus atau runtime kadaluarsa, pengguna dapat melanjutkan pekerjaan persis dari titik terakhir.

Salah satu fitur unggulan Colab adalah kemampuan untuk menggunakan akselerator perangkat keras-GPU dan TPU-yang dialokasikan secara dinamis melalui resource scheduler GCP. Pengguna tinggal memilih tipe akselerator yang dibutuhkan, lalu sistem akan menjadwalkan kontainer pada node yang dilengkapi perangkat keras tersebut. Kombinasi virtualisasi container dan penjadwalan sumber daya cloud memastikan bahwa komputasi intensif dapat dilakukan dengan efisien tanpa memerlukan infrastruktur lokal.

#### 2.4.4 Draw.io

Draw.io, yang kini dikenal sebagai diagrams.net, mengimplementasikan representasi internal berbasis XML untuk menyimpan diagram. Setiap elemen dalam diagram-termasuk shape, konektor, dan label-dikonversi menjadi node XML yang dilengkapi dengan atribut posisi, ukuran, dan gaya. Pendekatan ini memungkinkan pengelolaan data diagram yang terstruktur dan efisien. Di sisi *frontend*, *diagrams.net* menggunakan *JavaScript* dan *SVG (Scalable Vector Graphics)* untuk rendering vektor di browser, memanfaatkan model *Document Object Model (DOM)* untuk manipulasi elemen dan *Canvas API* untuk rendering grafis lebih kompleks. Dengan pendekatan berbasis web ini, pengguna dapat dengan mudah mengakses dan berinteraksi dengan diagram di dalam browser tanpa memerlukan perangkat lunak tambahan.

Untuk mendukung kolaborasi, diagrams.net memungkinkan penyimpanan backend melalui platform seperti Google Drive atau GitHub, menggunakan API RESTful atau protokol Git untuk commit dan pull data diagram. Fitur ini memungkinkan integrasi yang mulus dengan sistem kontrol versi dan kolaborasi real-time antara

pengguna. Untuk menghindari konflik saat pengeditan bersama, diagrams.net menerapkan konsep *optimistic locking*, yang memastikan bahwa hanya satu pengguna yang dapat mengedit bagian tertentu dari diagram pada waktu yang bersamaan, sehingga meningkatkan efisiensi dan mencegah masalah dalam proses kolaborasi.

#### 2.4.5 X (Twitter)

Twitter kini dikenal sebagai X merupakan platform media sosial yang memungkinkan pengguna berbagi informasi secara singkat dalam bentuk pesan singkat (*tweet*). Platform ini dirancang untuk menyampaikan berita, opini, dan interaksi secara cepat dalam skala global. Pengguna dapat mem-follow akun lain, membalas tweet, membagikan ulang (*retweet*), serta menyukai dan menyematkan konten. Twitter juga dikenal karena sistem hashtag dan trending topic-nya yang memungkinkan topik tertentu menjadi viral dan diperbincangkan secara luas. Hal ini menjadikan Twitter sebagai alat penting dalam membentuk opini publik dan menyebarkan informasi real-time.



Gambar 2. 5 Logo X

Rebranding Twitter menjadi X yang dilakukan oleh perusahaan induknya (di bawah kepemilikan Elon Musk) merupakan langkah besar dalam merombak identitas platform tersebut. Meskipun secara teknis fitur dasarnya tetap serupa, perubahan nama menjadi X menandai ambisi untuk menjadikan platform ini lebih dari sekadar



media sosial—yakni sebagai “super app” yang mencakup pembayaran, streaming, dan komunikasi multifungsi. Dalam konteks penelitian, X tetap berfungsi sebagai sumber data utama untuk analisis sentimen publik karena karakter tweet-nya yang padat, spontan, dan merefleksikan opini pasar secara langsung. Tweet-tweet ini sangat berharga dalam menganalisis persepsi publik terhadap isu tertentu, termasuk pergerakan harga saham seperti PGAS.

Sebagai salah satu sumber data real-time terbesar, X memainkan peran penting dalam penelitian sentimen media sosial, di mana tweet dianggap sebagai refleksi opini publik yang dapat dikoleksi, dianalisis, dan dikaitkan dengan pergerakan pasar atau fenomena sosial lainnya.

