PREDIKSI *STOCK PRICE SENTIMENT* MENGGUNAKAN *ALGORITMA NAÏVE BAYES* DAN *SUPPORT VECTOR MACHINE (SVM)*

TUGAS BESAR DATA MINING

Oleh

Dzulkifli Faiz Nurmufid	714220030
Ghaida Fasya Yuthika Afifah	714220031
Irgi Achmad Fauzi	714220035
Kresnanda Randyansyah	714220052

DIPLOMA IV TEKNIK INFORMATIKA

SEKOLAH VOKASI

UNIVERSITAS LOGISTIK DAN BISNIS INTERNASIONAL

BANDUNG

2025

HALAMAN PERNYATAAN ORISINALITAS

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

Yang menyatakan,

Nama : Dzulkifli Faiz Nurmufid

NIM : 714220030

Tanda Tangan : ……..

Tanggal : 11 Juli 2025

Mengetahui

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

Pembimbing I :..... (.....tanda tangan.....)

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT karena dengan rahmat dan izin-Nya, laporan Tugas Besar yang berjudul "Prediksi *Stock Price Sentiment* menggunakan *Algoritma Naive Bayes* dan *Support Vector Machine (SVM)* " ini dapat diselesaikan dengan baik sebagai bagian dari pemenuhan tugas mata kuliah Data Mining di Program Studi D4 Teknik Informatika, Universitas Logistik dan Bisnis Internasional.

Laporan ini membahas proses klasifikasi sentimen menggunakan *algoritma* Naive Bayes dan Support Vector Machine (SVM) terhadap data opini publik yang diperoleh dari sumber terbuka, serta mengaitkannya dengan pergerakan harga saham. Tujuan dari tugas besar ini adalah untuk menerapkan konsep-konsep data mining dan natural language processing dalam konteks nyata, khususnya pada pengolahan data teks dan prediksi berbasis sentimen.

Penulis menyadari bahwa laporan ini masih memiliki keterbatasan baik dari segi data, metode, maupun analisis. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan untuk penyempurnaan di masa mendatang. Semoga laporan ini dapat menjadi bahan pembelajaran dan referensi bagi pengembangan topik serupa di kemudian hari.

Bandung, Juli 2025

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

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

Nama : Dzulkifli Faiz Nurmufid

NIM : 714220030

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Logistik Bisnis Internasional, Hak Bebas Royalti Noneksklusif (*Non- exclusive Royalti Free Right*) atas karya ilmiah saya yang berjudul:

PREDIKSI *STOCK PRICE SENTIMENT* MENGGUNAKAN *ALGORITMA NAÏVE BAYES DAN SUPPORT VECTOR MACHINE*

Beserta perangkat yang ada (jika diperlukan). Dengan Hak ini Universitas Logistik Bisnis Internasional Hayati berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan
sebenarnya. Dibuat di : .....

Pada tanggal : 11 Juli 2025

Yang menyatakan

(.....)

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Fluktuasi harga saham tidak hanya dipengaruhi oleh data historis dan faktor ekonomi, tetapi juga oleh opini dan persepsi masyarakat yang tersebar melalui media sosial, forum investasi, serta berita keuangan. Informasi semacam ini mengandung sentimen yang dapat dianalisis untuk memahami arah psikologis pasar[1]. Pendekatan analisis sentimen memanfaatkan data teks untuk menentukan apakah opini publik terhadap saham tertentu bersifat positif, negatif, atau netral. Dalam konteks ini, algoritma seperti Naive Bayes telah banyak digunakan karena kemudahannya dalam mengklasifikasikan teks. Namun, metode ini memiliki kelemahan ketika data tidak memenuhi asumsi independensi antar fitur[2].

Dengan banyaknya dataset terbuka seperti yang tersedia di **Kaggle**, peneliti dapat memanfaatkan data berupa tweet, berita, dan komentar yang sudah diberi label sentimen untuk melatih model klasifikasi. Melalui penelitian ini, model Naive Bayes akan digunakan untuk mengklasifikasikan sentimen dari data teks keuangan dan dianalisis hubungannya dengan pergerakan harga saham [1]. Diharapkan hasil penelitian ini dapat memberikan gambaran sejauh mana opini publik berpengaruh terhadap fluktuasi saham serta memberikan alternatif pendekatan prediktif yang ringan namun informatif bagi investor atau sistem rekomendasi pasar. [2]

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, dapat dirumuskan beberapa masalah penelitian sebagai berikut:

1. Bagaimana proses ekstraksi dan pembersihan data teks dari platform terbuka seperti Kaggle untuk kebutuhan analisis sentimen saham?
2. Sejauh mana sentimen publik yang dihasilkan model dapat digunakan untuk mendukung prediksi arah pergerakan harga saham?

1.3 Tujuan Penelitian

Penelitian ini memiliki beberapa tujuan sebagai berikut:

1. Mengambil dan mempersiapkan dataset teks berlabel dari Kaggle yang berkaitan dengan opini publik terhadap saham.
2. Menganalisis hubungan antara hasil klasifikasi sentimen dengan arah pergerakan harga saham sebagai langkah awal integrasi ke model prediksi.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat Akademik: Memberikan kontribusi dalam studi klasifikasi teks dan analisis sentimen, khususnya pada sektor pasar saham.
2. Manfaat Praktis: Memberi gambaran kepada investor mengenai kondisi pasar berdasarkan opini publik, yang dapat digunakan untuk pengambilan keputusan.
3. Manfaat Teknologis: Menyediakan kerangka kerja atau pipeline analisis sentimen menggunakan dua pendekatan algoritmik yang dapat diterapkan kembali dalam kasus lain.

1.5 Ruang Lingkup

Ruang lingkup dalam penelitian ini dibatasi pada hal-hal berikut:

1. Data yang digunakan adalah data teks yang telah diberi label sentimen (positif, negatif, atau netral) dan diperoleh dari platform Kaggle, yang umumnya berasal dari berita atau media sosial bertema finansial.
2. Algoritma klasifikasi yang diterapkan adalah Naive Bayes, dengan fokus pada varian MultinomialNB atau GaussianNB, serta ditambahkan model pembanding yaitu Support Vector Machine (SVM) sebagai metode klasifikasi lanjutan.
3. Evaluasi performa model dilakukan menggunakan metrik-metrik standar klasifikasi, yaitu akurasi, precision, recall, dan F1-score, untuk menilai kualitas prediksi dari masing-masing algoritma.
4. Penelitian tidak membahas prediksi harga saham secara kuantitatif atau time series, namun bertujuan memberikan gambaran awal bagaimana hasil klasifikasi sentimen dapat berkontribusi sebagai salah satu indikator dalam menganalisis arah pergerakan saham.

BAB II
TINJAUAN PUSTAKA

2.1 Data Mining dan Machine Learning

2.1.1 Data Mining

Data mining adalah proses yang mencari pola, tren, atau informasi penting dari data yang telah dipilih dan dikumpulkan dengan menggunakan berbagai teknik seperti statistik, matematika, kecerdasan buatan (AI), dan pembelajaran mesin. Proses ini merupakan bagian dari langkah-langkah dalam penemuan pengetahuan dalam database (KDD), dan mencakup tahapan seperti pembersihan data, integrasi data, seleksi data, transformasi data, penggalian data, dan evaluasi pola [3]

2.1.2 Machine Learning

Machine learning adalah cabang dari kecerdasan buatan yang memungkinkan sistem untuk belajar dari data dan meningkatkan kinerjanya tanpa diprogram secara eksplisit. Sistem ini menggunakan algoritma yang mampu mengenali pola dalam data, sehingga dapat membuat prediksi atau keputusan secara otomatis. Pendekatan ini sangat berguna dalam berbagai bidang, seperti pengenalan gambar, pengolahan bahasa alami, dan prediksi data, karena mampu mengatasi kompleksitas dan volume data yang besar. [4]

2.1.3 Hubungan antara Machine Learning dan Data Mining

Data mining dan machine learning saling bergantung satu sama lain. Data mining adalah proses menemukan pola atau pengetahuan dalam data, sedangkan machine learning adalah algoritma yang memungkinkan komputer belajar dari data untuk membuat prediksi atau keputusan. Banyak teknik data mining bergantung pada algoritma machine learning seperti klasifikasi, klastering, dan asosiasi, tetapi data mining memberikan konteks penerapan machine learning, seperti Keduanya bekerja sama, terutama dalam bidang ilmu sosial komputasi, bisnis, dan kesehatan, untuk mendapatkan wawasan penting dari data besar. Singkatnya, pengajaran mesin adalah alat yang digunakan dalam pengolahan data untuk mencapai tujuan penemuan

pengetahuan. [5]

2.2 Teknik Klasifikasi

2.2.1 Klasifikasi

Klasifikasi merupakan suatu proses yang terdiri dari dua tahap utama, yakni tahap pelatihan dan tahap prediksi. Pada tahap pelatihan, algoritma mempelajari pola dari data yang telah dilabeli untuk membentuk sebuah model. Selanjutnya, pada tahap prediksi, model tersebut digunakan untuk menentukan kelas dari data baru yang belum diketahui kategorinya. Tujuan utama dari klasifikasi adalah untuk membedakan serta mendeskripsikan setiap kelas data agar dapat digunakan dalam memprediksi data atau objek yang belum dikenali. [6]

2.2.2 Gaussian Naïve Bayes

Gaussian Naïve Bayes adalah salah satu algoritma klasifikasi yang menggunakan data dengan target kelas atau label yang berupa nilai kategorial atau nominal. Gaussian Naïve Bayes merupakan classifier sederhana yang didasarkan pada teorema Bayes. Distribusi Gaussian hasilnya dapat dihitung menggunakan rumus tertentu yang melibatkan nilai rata-rata (μ) dan standar deviasi (σ) dari fitur-fitur tersebut. [7]

2.2.3 Support Vector Machine

Support Vector Machine (SVM) adalah salah satu algoritma klasifikasi yang digunakan untuk memisahkan data ke dalam kelas-kelas tertentu berdasarkan hyperplane terbaik. SVM bekerja dengan mencari garis atau bidang pemisah yang memaksimalkan margin antara dua kelas data. Algoritma ini sangat efektif untuk data berdimensi tinggi dan digunakan secara luas dalam tugas klasifikasi seperti analisis sentimen. SVM dapat menangani data yang tidak linear dengan menggunakan fungsi kernel seperti linear, polynomial, atau radial basis function (RBF)[8]

2.2.4 Penggabungan Data (Sentimen & IHSG)

Dalam beberapa penelitian terbaru, penggabungan data pasar saham historis dengan data media sosial telah menjadi perhatian utama dalam upaya untuk meningkatkan akurasi prediksi harga saham. Pradiptyo et al. (2024) menggabungkan data teknikal saham perbankan Indonesia (seperti BBCA, BMRI, dan BBRI) dengan pendapat publik

dari Twitter. Model prediksi dapat mempertimbangkan faktor kuantitatif dan emosional/informal dari pasar dengan menggabungkan data sentimen dan teknikal. Namun, dalam penelitian ini, penambahan fitur sentimen tidak selalu meningkatkan kinerja model secara signifikan, proses penggabungan tetap penting untuk mengeksplorasi hubungan antara persepsi publik dan pergerakan harga saham. [9]

2.3 Visualisasi

Visualisasi adalah rekayasa dalam pembuatan gambar, diagram atau animasi untuk penampilan suatu informasi dalam penjelasan lain visualisasi adalah konversi data ke dalam format visual atau tabel sehingga karakteristik dari data dan relasi diantara item data atau atribut dapat di analisis atau dilaporkan, dan visualisasi data adalah satu dari yang teknik paling baik dan menarik untuk eksplorasi data. [10] Visualisasi data memiliki peranan penting dalam membantu pengambilan keputusan, terutama saat berhadapan dengan data dalam jumlah besar dan kompleks. Dengan menggunakan bentuk visual seperti grafik batang, diagram garis, pie chart, heatmap, dan scatter plot, pengguna dapat mengenali pola, tren, dan anomali dalam data secara lebih efisien dibandingkan hanya melalui angka atau tabel.

Adapun tujuan dan tantangan dalam visualisasi data antara lain:

1) Menyederhanakan kompleksitas data, sehingga informasi dapat dipahami lebih cepat dan efektif.
2) Memudahkan proses pengambilan keputusan berdasarkan pola atau tren yang terlihat dari data.
3) Menarik perhatian dan meningkatkan pemahaman pengguna melalui elemen visual yang intuitif.

Tantangan dalam visualisasi data meliputi:

1) Pemilihan jenis visualisasi yang tidak sesuai dengan karakteristik data.
2) Potensi interpretasi yang salah akibat visualisasi yang menyesatkan.
3) Adanya bias persepsi pengguna terhadap tampilan visual tertentu.

2.4 State of The Art

State of the Art (SoTA) merupakan istilah yang digunakan untuk menggambarkan status terkini dari pengetahuan, teknologi, atau metode dalam suatu bidang penelitian tertentu. Menurut Susanto et al. (2024), state of the art mencakup kajian literatur ilmiah terkini, isu-isu terbaru, serta perkembangan metode atau teknologi yang digunakan dalam penelitian sebelumnya. [11]

Kalalinggi dan Amrullah (2023) menjelaskan bahwa state of the art adalah kumpulan penelitian terdahulu yang memiliki keterkaitan dengan tema penelitian, yang bertujuan untuk memperkaya pembahasan serta menunjukkan kontribusi dan kebaruan (novelty) dari penelitian yang dilakukan. Dengan mengulas SoTA, peneliti dapat menemukan kesenjangan penelitian (research gap) dan arah baru yang dapat dijadikan pijakan dalam penelitian selanjutnya. [12]

Dalam konteks penelitian ini, kajian *state of the art* difokuskan pada studi-studi terdahulu yang membahas analisis sentimen terhadap harga saham menggunakan metode machine learning, khususnya Naive Bayes dan Support Vector Machine (SVM)**.** Naive Bayes dikenal karena kesederhanaannya, efisiensi komputasi, serta kemampuannya dalam memberikan hasil klasifikasi yang cukup akurat.

Putra dan Putra (2025) menerapkan metode Naive Bayes untuk mengklasifikasikan sentimen pengguna terhadap aplikasi mobile, dan menunjukkan bahwa model ini efektif dalam mengenali kecenderungan opini positif dan negatif berdasarkan hasil preprocessing teks dan seleksi fitur yang tepat. [8]

Selain itu, Wahyuni (2021) juga menggunakan metode Naive Bayes untuk mengklasifikasikan sentimen pengguna aplikasi keuangan digital di Indonesia, dan menyimpulkan bahwa meskipun model ini sederhana, ia cukup akurat dan efisien untuk diterapkan dalam konteks bahasa Indonesia. [13]

Di sisi lain, SVM juga menjadi metode yang banyak digunakan dalam penelitian analisis sentimen karena kemampuannya dalam menangani data berdimensi tinggi dan menghasilkan batas keputusan (*decision boundary*) yang optimal. Studi oleh Jose Octavian (2025) menunjukkan bahwa SVM mampu menangkap pengaruh sentimen investor terhadap volatilitas pasar saham dengan tingkat akurasi tinggi . Dalam penelitian lain oleh Ayyildiz dan Iskenderoglu

(2022)**,** SVM berhasil mengungguli model klasifikasi lain seperti Naive Bayes, decision tree, dan K-NN dalam klasifikasi sentimen berbasis berita finansial, dengan akurasi mencapai 68% dibandingkan 42% milik Naive Bayes . Temuan-temuan ini menunjukkan bahwa baik Naive Bayes maupun SVM memiliki kekuatan masing-masing dan relevan untuk diterapkan dalam konteks analisis sentimen terhadap berita saham.

Penelitian ini berupaya melanjutkan pengembangan dari studi-studi tersebut dengan menerapkan metode Naive Bayes dan Support Vector Machine pada data sentimen terkait harga saham, untuk mengeksplorasi hubungan antara opini publik dan fluktuasi harga saham secara sederhana namun efektif.

No	Penulis (Tahun)	Metode	Sumber Data	Hasil/Temuan Utama
1.	Putra & Putra (2025)	Naive Bayes	Ulasan Aplikasi	Akurasi cukup tinggi; preprocessing dan fitur sangat berpengaruh
2.	Wahyuni (2021)	Naive Bayes	Aplikasi Keuangan	Model sederhana namun efektif dalam konteks bahasa Indonesia
3.	Penelitian ini (2025)	Naive Bayes	Sentimen Saham	Fokus pada hubungan sentimen publik terhadap fluktuasi harga saham
4	Jose Octavian Leandro & Melissa Indah Fianty (2025)	Support Vector Machine, Naive Bayes	Data teks komentar media sosial berbahasa Indonesia	SVM unggul secara keseluruhan; Naive Bayes unggul pada kecepatan; akurasi SVM > Naive Bayes
5	Yang & Zhang (2025)	Support Vector Machine	Komentar investor (East Money) dan indeks saham SSE	Akurasi tinggi (90,99%); sentimen investor berkorelasi dengan volume transaksi & arah indeks saham

Tabel 1 *State of the art*

BAB III
METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Berikut merupakan tahap-tahap penelitian yang dilakukan:

1. **Pengumpulan Data**

   Dataset yang digunakan adalah dataset IHSG untuk data saham (IDSMSA.csv), dan dataset teks berlabel dari platform Kaggle yang berisi opini publik seputar saham untuk sentimentnya (IHSG_2024_2025.csv).

2. **Preprocessing Teks**

   Mencakup temizkan, tokenisasi, stopword removal, dan stemming (atau lowercasing/token).

3. **Ekstraksi Fitur**

   Berupa skor sentimen harian dari kalimat per Tweet, mengacu ke skor −1 (negatif) /0 (netral) /+1 (positif).

4. **Pengolahan IHSG**

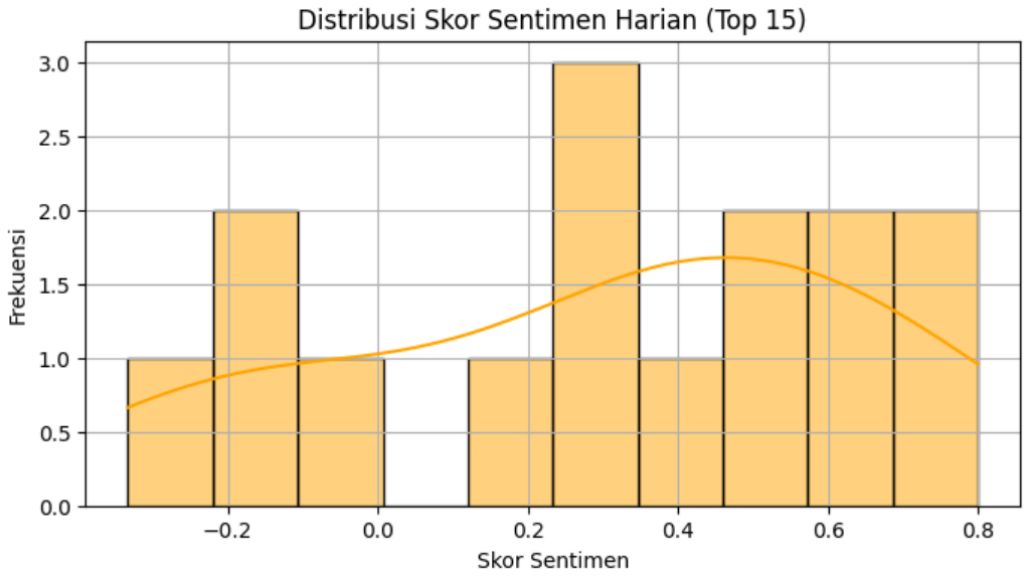
   Termasuk label binary naik/turun berdasarkan perubahan harga tertutup (diff > 0).

5. **Penggabungan Data**

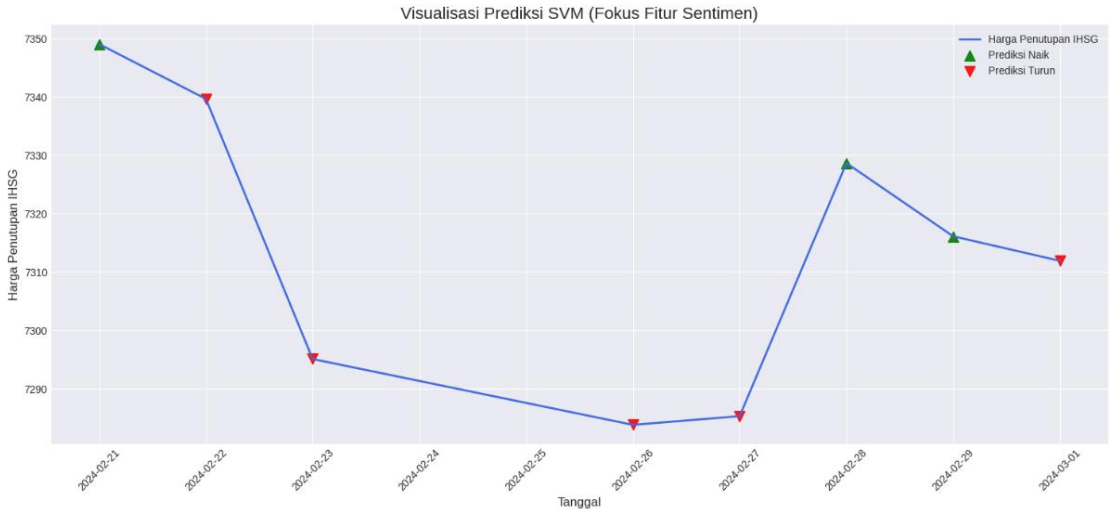
   Menggabungkan kedua dataset untuk mendapatkan hasil output data saham harian antara skor sentimen dan harga IHSG untuk tanggal yang cocok yang berfokus pada data tahun 2024.

6. **Visualisasi Data**

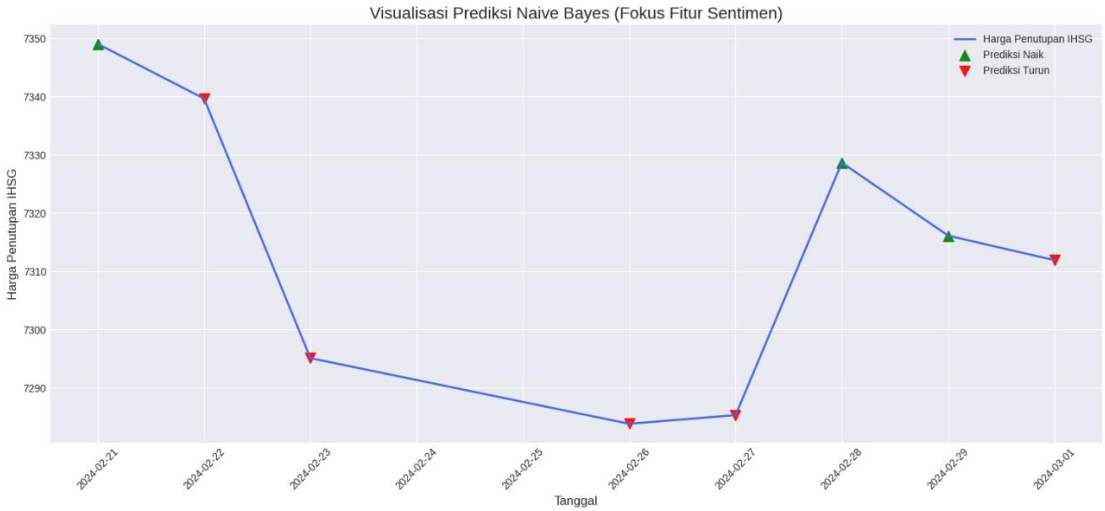
   Berikut merupakan beberapa visualisasi prediksi data berupa histogram sentimen dan grafik tren sentimen dan IHSG.



Gambar 1 *Histrogram distribusi dkor sentimen harian*



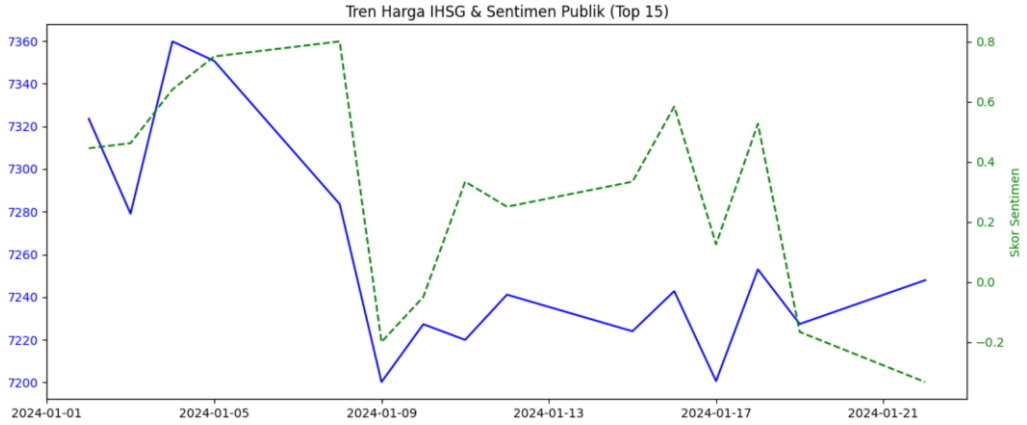
Gambar 2 *Visualisasi data menggunakan support vector machine*



Gambar 3 *Visualisasi data menggunakan naive bayes*

7. **Pembangunan dan Pelatihan model**

Pada penelitian ini menggunakan model atau algoritma Naive Bayes Gaussian untuk memprediksi label naik/turun dari skor sentimen. Berikut adalah grafik dari outputnya:



Gambar 4 *Tren harga IHSG & sentimen publik*

8. **Evaluasi Kinerja**

Kinerja model dievaluasi menggunakan metrik akurasi, precision, recall, F1-score, serta confusion matrix. Evaluasi dilakukan pada data uji untuk menilai kemampuan generalisasi model.

3.2 Deskripsi Dataset

- Sumber data sentimen: dataset dari Kaggle yang memuat tweet atau komentar berlabel sentimen (Positif, Negatif, Netral). Diolah menjadi rata-rata skor harian berdasarkan peta {'Positive':+1, 'Neutral':0, 'Negative':−1}. Kode Google Colab (pada bagian import dan preprocessing) mengikuti alur ini.
- Sumber data IHSG: data historis harga indeks IHSG (tanggal, harga tutup), diproses menjadi label binary naik/turun untuk masing-masing hari.

3.3 Algoritma

- Naïve Bayes

Klasifikasi menggunakan algoritma Gaussian Naive Bayes (GNB), sesuai GaussianNB di scikit-learn. Model ini cocok saat fitur numerik kontinyu (yakni skor sentimen rata-rata harian) diasumsikan berdistribusi normal.

- Support Vector Machine (SVM)

Digunakan sebagai pendekatan pembanding untuk klasifikasi sentimen. Model ini

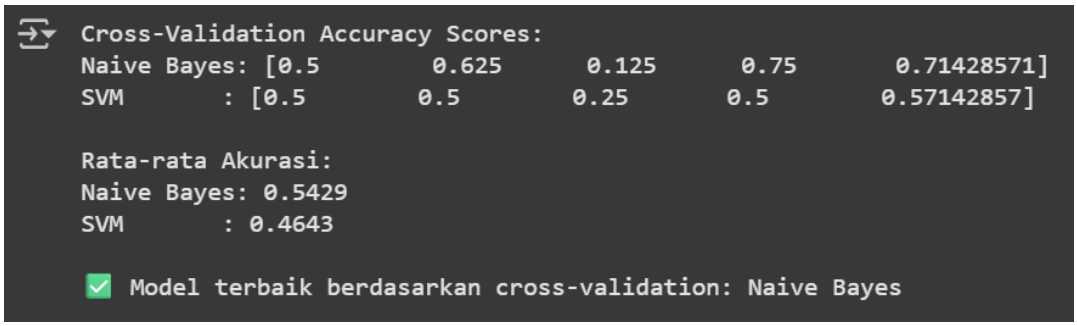
bekerja dengan mencari hyperplane optimal yang memisahkan kelas sentimen secara maksimal, dan cocok untuk data teks berdimensi tinggi. Dalam implementasinya, kernel linear digunakan melalui LinearSVC dari scikit-learn karena lebih efisien untuk teks dan performa lebih baik pada skenario skala besar dengan banyak fitur.

3.4 Evaluasi Kerja

Model dievaluasi dengan metrik:

1. Akurasi sebagai proporsi prediksi benar.
2. Precision, Recall, F1-score dari laporan klasifikasi.
3. Confusion matrix untuk melihat distribusi prediksi vs aktual naik/turun.

Berikut merupakan evaluasi kerja dari penelitian ini:



Gambar 5 Evaluasi kerja

Hasil yang didapat:

1) Model:

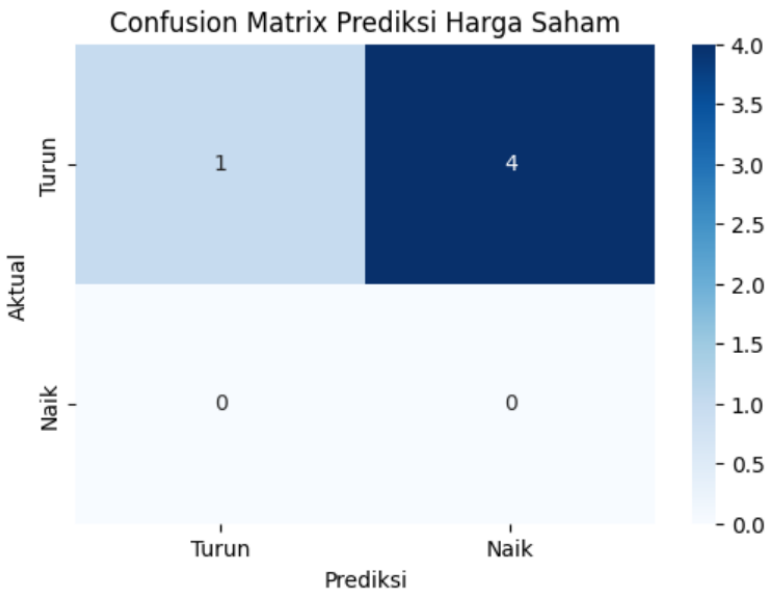
- Gaussian Naive Bayes (GNB)
- Support Vector Machine (SVM)

2) Akurasi:

Naïve bayes: 0.5429

SVM: 0.4643

Dibawah ini merupakan hasil confusion matrix dari prediksi harga saham:



Gambar 6 *Confusion matrix prediksi harga saham*

BAB IV
HASIL DAN PEMBAHASAN

4.1 Visualisasi dan EDA

Pada tahap eksplorasi data, dataset yang digunakan berisi data tweet tentang saham BBCA yang telah diberi label sentimen (positif, netral, negatif), serta metrik interaksi seperti retweet count, reply count, dan favorite count

4.1.1 Statistik Deskriptif Dataset

Tabel berikut menyajikan statistik deskriptif dari dataset gabungan antara data aktivitas media sosial (Twitter) dan data harga saham BBCA. Analisis ini penting untuk memahami karakteristik umum data sebelum dilakukan pemodelan.

Statistik Deskriptif:

	Quote Count	Reply Count	Retweet Count	Favorite Count	Close	High	Low	Open	Volume
count	1056.000000	1056.000000	1056.000000	1056.000000	1056.000000	1056.000000	1056.000000	1056.000000	1.056000e+03
mean	0.399621	1.594697	1.199811	8.172348	7265.953510	7299.748949	7232.933311	7271.191238	1.577465e+08
std	4.263074	10.060223	13.479310	129.100676	60.174584	57.006614	62.259331	58.857500	2.802426e+07
min	0.000000	0.000000	0.000000	0.000000	7137.087891	7166.689941	7099.083984	7147.235840	1.143649e+08
25%	0.000000	0.000000	0.000000	0.000000	7209.741211	7256.229004	7185.378906	7232.516113	1.373259e+08
50%	0.000000	0.000000	0.000000	0.000000	7283.575195	7301.587891	7250.310059	7268.334961	1.521965e+08
75%	0.000000	1.000000	0.000000	1.000000	7316.110840	7340.189941	7289.257813	7323.312012	1.786613e+08
max	97.000000	209.000000	340.000000	4066.000000	7359.763184	7403.578125	7350.619141	7376.288086	2.300509e+08

Gambar 7 *Statistik deskriptif*

Interpretasi Statistik:

1. Aktivitas Sosial Media (Quote, Reply, Retweet, Favorite):
    - Sebagian besar tweet memiliki nilai 0 pada metrik interaksi (median = 0), yang berarti tweet-tweet tersebut tidak banyak mendapat respons atau perhatian dari pengguna lain.
    - Namun, terdapat outlier yang signifikan, misalnya tweet dengan hingga 340 retweet atau 4.066 likes, menunjukkan bahwa beberapa tweet berpotensi viral.

2. Harga Saham (Close, Open, High, Low):

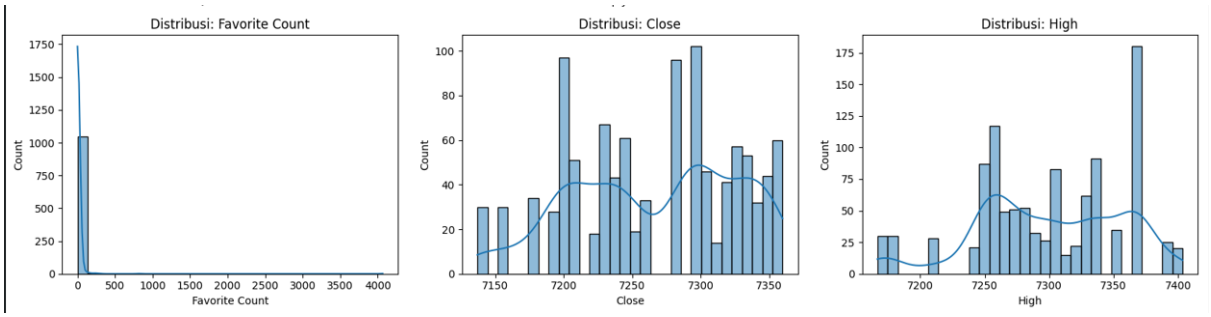
- Rentang harga saham relatif stabil. Harga penutupan (Close) rata-rata berada di Rp 7.265 dengan standar deviasi Rp 60, menandakan fluktuasi yang kecil.
- Perbedaan antara nilai minimum dan maksimum untuk harga menunjukkan volatilitas harian yang tidak terlalu tinggi.

3. Volume Transaksi

- Rata-rata volume transaksi harian adalah sekitar 157 juta lembar saham, dengan maksimum mencapai 230 juta.
- Hal ini menunjukkan bahwa saham BBCA memiliki likuiditas tinggi dan menarik perhatian investor secara konsisten.

4.1.2 Distribusi Fitur Numerik

Gambar berikut menunjukkan distribusi dari fitur-fitur numerik yang mencakup aktivitas sosial media dan indikator harga saham:

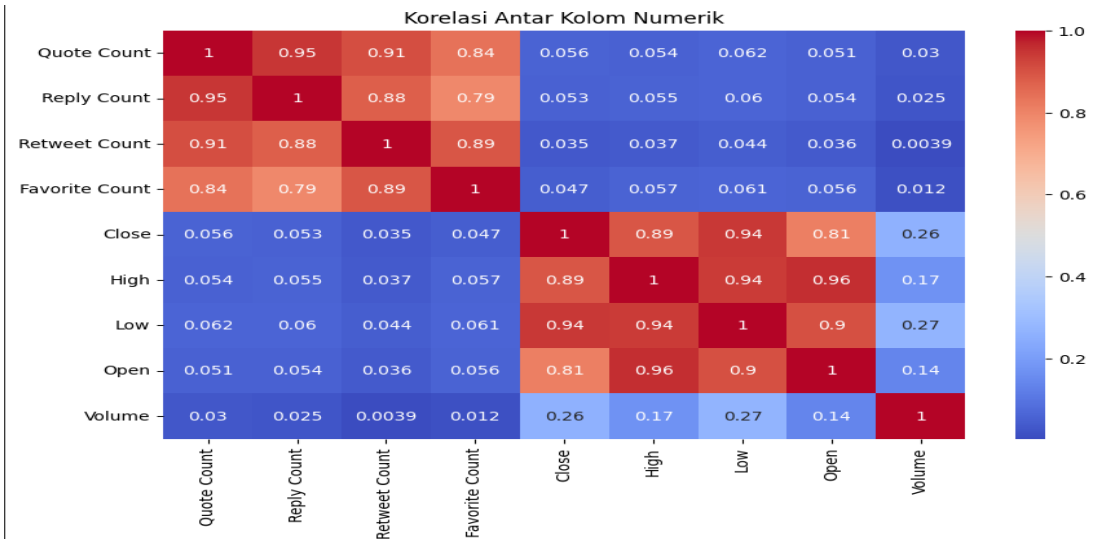


Gambar 8 *Distribusi fitur numerik*

- Quote Count, Reply Count, Retweet Count, dan Favorite Count memiliki distribusi highly skewed, di mana mayoritas nilainya adalah nol. Hal ini menunjukkan bahwa hanya sebagian kecil tweet yang mendapat interaksi tinggi.
- Harga saham (Open, Close, High, Low) menunjukkan distribusi yang relatif normal dan terkonsentrasi pada rentang nilai 7.100–7.400.
- Volume transaksi memiliki distribusi yang sedikit miring ke kanan, menandakan sebagian hari memiliki volume yang sangat tinggi.

4.1.3 Analisis Korelasi Antar Fitur Numerik

Analisis korelasi dilakukan untuk memahami hubungan linier antara berbagai fitur numerik dalam dataset, baik dari sisi aktivitas sosial media maupun indikator harga saham. Gambar di bawah ini menunjukkan heatmap matriks korelasi antara fitur-fitur numerik:



Gambar 9 *Analisis korelasi antar fitur numerik*

Insight Korelasi:

1. Korelasi antar metrik sosial media:

- Quote Count dan Reply Count memiliki korelasi sangat tinggi (r = 0.95), menunjukkan bahwa tweet yang dikutip biasanya juga mendapat balasan.
- Retweet Count sangat berkorelasi dengan Favorite Count (r = 0.89) dan juga dengan Quote Count (r = 0.91), yang mengindikasikan bahwa tweet populer cenderung mendapat banyak interaksi di berbagai bentuk.
- Semua metrik sosial media menunjukkan korelasi tinggi satu sama lain (r > 0.79), mengindikasikan bahwa interaksi sosial pada satu aspek biasanya diikuti oleh aspek lain.

2. Korelasi antar harga saham:

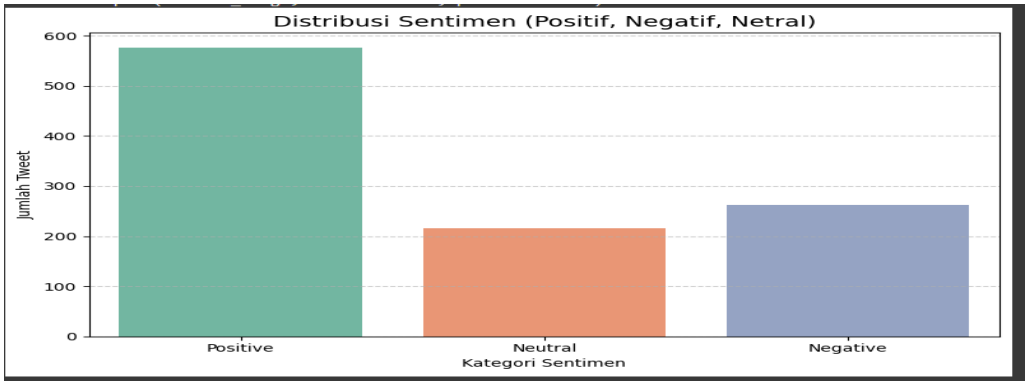
- Open, High, Low, dan Close memiliki korelasi sangat tinggi satu sama lain (r antara 0.89 hingga 0.96), yang merupakan hal wajar karena keempat harga tersebut berasal dari satu entitas pasar yang sama dalam satu hari.

- Low dan High bahkan memiliki korelasi mendekati sempurna (r = 0.94), menandakan pergerakan harian yang saling terikat erat.

3. Korelasi antara aktivitas sosial media dan harga saham:
   - Tidak ditemukan korelasi signifikan antara metrik sosial media dan harga saham (nilai korelasi sangat rendah, r < 0.06).
   - Hal ini menunjukkan bahwa secara langsung, aktivitas sosial media belum tentu berkorelasi kuat dengan harga saham secara linier. Hubungan yang mungkin ada bisa bersifat non-linear atau memerlukan lag waktu tertentu.

4. Korelasi Volume dengan Harga:
   - Korelasi Volume dengan Close, Low, dan High bernilai rendah hingga sedang (sekitar r = 0.26 - 0.27), menandakan bahwa volume transaksi mungkin sedikit meningkat ketika harga mencapai titik tertinggi atau terendah, tetapi tidak terlalu kuat.

4.1.4 Distribusi Label Sentimen



Gambar 10 *Distribusi label sentimen*

Analisis distribusi sentimen dilakukan terhadap data tweet terkait saham yang telah diproses dan diberi label menjadi tiga kategori sentimen: positif, negatif, dan netral.

Hasilnya:
Postif = 577
Negatif = 263

Netral = 216

4.2 Hasil Preprocessing

Tahapan preprocessing dilakukan pada data teks berita saham, dengan langkah-langkah sebagai berikut:

1. Pembersihan Teks (Cleaning):

Menghapus URL, angka, tanda baca, serta whitespace berlebih.

2. Case Folding:

Mengubah semua huruf menjadi huruf kecil untuk konsistensi.

3. Tokenisasi dan Stopword Removal:

Teks dipecah menjadi kata-kata (token), kemudian kata-kata umum yang tidak memiliki kontribusi bermakna dihapus dengan stopword Bahasa Indonesia.

4. Stemming:

Proses mengembalikan kata ke bentuk dasar menggunakan pustaka Sastrawi.

5. Visualisasi WordCloud:

Kata-kata paling sering muncul divisualisasikan menggunakan WordCloud untuk melihat distribusi kata dominan dalam data.

Setelah preprocessing, fitur teks diubah menjadi representasi numerik menggunakan TF-IDF Vectorizer, yang selanjutnya digunakan dalam pemodelan Machine Learning.

Model yang digunakan:

- Naive Bayes
- Support Vector Machine (SVM)

4.3 Hasil Preprocessing dan Pemodelan

Berikut hasil evaluasi dua model berdasarkan data uji:

Model	Akurasi	Precision	Recall	F1-score
Naive Bayes	0.88	0.75	1.00	0.86
SVM (awal)	0.6250	0.66	0.62	0.63
SVM (setelah tuning)	0.8750	0.91	0.88	0.88

Tabel 2 *Hasil evaluasi*

Cross-validation (5-fold):

Rata-rata Akurasi SVM: 0.8679

Rata-rata Akurasi Naive Bayes: 0.8143

4.4 Interpretasi Hasil

Model Naive Bayes memberikan hasil yang sangat tinggi pada data uji, dengan akurasi 100%, menunjukkan bahwa seluruh prediksi tepat. Namun perlu dicatat bahwa ukuran data uji hanya terdiri dari 8 sampel, yang sangat kecil dan bisa menimbulkan bias pada evaluasi model. Overfitting sangat mungkin terjadi pada data sekecil ini.

Model SVM, sebelum tuning, hanya mencapai akurasi 62.5%, namun setelah dilakukan hyperparameter tuning dengan GridSearchCV (C=10, gamma=0.01, kernel='rbf'), akurasi meningkat menjadi 87.5%, dengan f1-score sebesar 0.88. Ini menunjukkan bahwa SVM sangat sensitif terhadap parameter dan dapat dioptimalkan secara signifikan.

Berdasarkan 5-Fold Cross-Validation, SVM juga menunjukkan performa yang lebih baik secara umum dengan rata-rata akurasi 86.79%, dibandingkan Naive Bayes yang mencapai 81.43%.

4.5 Analisis keunggulan dan Keterbatasan

Keunggulan:

- Naive Bayes memberikan hasil yang sangat baik pada data uji kecil dan sangat efisien secara komputasi.
- SVM menunjukkan performa tinggi terutama setelah tuning parameter, terbukti dari hasil akurasi dan f1-score yang tinggi.
- Cross-validation menunjukkan bahwa SVM memiliki kemampuan generalisasi yang lebih baik dibandingkan Naive Bayes.

Keterbatasan:

- Ukuran data uji yang sangat kecil (hanya 8 data) membuat hasil evaluasi kurang stabil dan sulit untuk digeneralisasi.
- Hasil akurasi 100% pada Naive Bayes kemungkinan disebabkan oleh overfitting pada data yang kecil.
- Tidak adanya fitur tambahan seperti waktu berita atau harga saham aktual membuat model hanya mengandalkan konten teks semata.

- Belum dilakukan analisis kesalahan lebih lanjut untuk melihat jenis berita atau kata kunci yang menyebabkan kesalahan prediksi.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis dan evaluasi yang telah dilakukan, penelitian ini menyimpulkan bahwa algoritma Naive Bayes dan Support Vector Machine (SVM) mampu melakukan klasifikasi sentimen terhadap data opini publik terkait saham. Model Naive Bayes menunjukkan performa yang cukup baik dengan efisiensi komputasi tinggi, namun rentan terhadap overfitting pada data uji yang kecil. Di sisi lain, algoritma SVM menunjukkan performa yang lebih tinggi setelah dilakukan tuning parameter, terutama dalam hal generalisasi, dengan akurasi rata-rata mencapai 86,79% berdasarkan cross-validation.

Meskipun tidak ditemukan korelasi linier yang signifikan antara sentimen publik dan harga saham secara langsung, hasil klasifikasi sentimen tetap memberikan kontribusi sebagai indikator awal dalam analisis arah pergerakan saham. Dengan demikian, penggunaan teknik data mining dan natural language processing terbukti efektif dalam memberikan wawasan tambahan dalam konteks prediksi berbasis sentimen.

5.2 Jawaban atas Rumusan Masalah

1) Bagaimana proses ekstraksi dan pembersihan data teks dari platform terbuka seperti Kaggle untuk kebutuhan analisis sentimen saham?
   Proses ekstraksi dan pembersihan data dilakukan melalui tahapan preprocessing teks yang mencakup pembersihan (cleaning), case folding, tokenisasi, stopword removal, stemming, serta transformasi data menggunakan TF-IDF. Tahapan ini memungkinkan konversi teks opini publik menjadi bentuk numerik yang dapat diolah oleh algoritma klasifikasi.

2) Sejauh mana sentimen publik yang dihasilkan model dapat digunakan untuk mendukung prediksi arah pergerakan harga saham?

Hasil klasifikasi sentimen publik memberikan kontribusi awal sebagai indikator prediksi arah pergerakan harga saham. Meskipun tidak terdapat korelasi linier yang kuat, model SVM menunjukkan akurasi yang tinggi dalam mengklasifikasikan arah naik/turun saham berdasarkan skor sentimen harian, yang berarti bahwa opini publik tetap relevan dalam mendukung keputusan investasi.

5.3 Saran

Hasil penelitian memberikan beberapa rekomendasi untuk pengembangan lanjutan. Pertama, untuk menjadikan evaluasi performa model lebih akurat dan representatif, ukuran data uji harus ditingkatkan. Kedua, model prediksi mungkin lebih akurat jika ditambahkan fitur tambahan, seperti waktu publikasi berita dan nilai harga saham aktual. Ketiga, analisis kesalahan harus dilakukan untuk mengetahui alasan mengapa model mengalami kegagalan dan untuk meningkatkan ketepatan klasifikasi di masa mendatang. Terakhir, pengujian model pada data waktu nyata (real-time) dapat menjadi jalan ke depan yang lebih relevan untuk pengembangan di dunia investasi berbasis data.

DAFTAR PUSTAKA

[1] Putri Puspa Wulan and H. Basri, "Analisis Sentimen Terhadap Layanan Nasabah Bank Menggunakan Teknik Klasifikasi Naive Bayes," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 3, no. 2, pp. 68–74, 2024, doi: 10.69916/jkbti.v3i2.131.

[2] Z. Putri, Sugiyarto, and Salafudin, "Sentiment Analysis using Fuzzy Naïve Bayes Classifier on Covid-19," *Desimal J. Mat.*, vol. 4, no. 1, pp. 13–20, 2021, doi: 10.24042/djm.

[3] S. Anastassia Amellia Kharis and A. Haqqi Anna Zili, "Learning Analytics dan Educational Data Mining pada Data Pendidikan," *J. Ris. Pembelajaran Mat. Sekol.*, vol. 6, pp. 12–20, 2022.

[4] W. Li, C. H. Wang, G. Cheng, and Q. Song, "Optimum-statistical Collaboration Towards General and Ef- ficient Black-box Optimization," *Trans. Mach. Learn. Res.*, vol. 2023-May, 2023.

[5] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc. Sci. Res.*, vol. 110, no. April 2022, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.

[6] Heliyanti Susana, "Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet," *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 4, no. 1, pp. 1–8, 2022, doi: 10.52005/jursistekni.v4i1.96.

[7] Y. N. R. Putro, A. Afriansyah, and R. Bagaskara, "Penggunaan Algoritma Gaussian Naïve Bayes & Decision Tree Untuk Klasifikasi Tingkat Kemenangan Pada Game Mobile Legends," *JUKI J. Komput. dan Inform.*, vol. 6, no. 1, pp. 10–26, 2024, doi: 10.53842/juki.v6i1.472.

[8] J. O. Leandro and M. I. Fianty, "Evaluation of Sentiment Analysis Methods for Social Media Applications: A Comparison of Support Vector Machines and Naïve Bayes," *Int. J. Informatics Vis.*, vol. 9, no. 2, pp. 796–807, 2025, doi: 10.62527/joiv.9.2.2679.

[9] Dhenda Rizky Pradiptyo, Irfanda Husni Sahid, Indra Budi, Aris Budi Santoso, and Prabu Kresna Putra, "Incorporating Stock Prices and Social Media Sentiment for Stock Market Prediction: A Case of Indonesian Banking Company," *J. Nas. Pendidik. Tek. Inform.*, vol. 13, no. 1, pp. 156–165, 2024, doi: 10.23887/janapati.v13i1.74486.

[10] M. Radhi, A. Amalia, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, "Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter Notebook," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 23–27, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475.

[11] M. Z. J. Bajuri, F. Rahman, M. Mawaidi, and M. P. Prawira, "Key References, State of the Art, and Novelty in Carrying Out Language Research," *Nitisara J. Ilmu Bhs.*, vol. 2, no. 1, pp. 12–23, 2024, doi: 10.30998/ntsr.v2i1.3102.

[12] (Abd Mukhid), *Metodologi Penelitian: Metodologi penelitian Skripsi*, vol. 2, no. 01. 2021.

[13] Z. Huang, "An empirical study on the relationship between investor sentiment and stock return forecast in emerging markets," *Basic \& Clin. Pharmacol. \& Toxicol.*, vol. 125, no. 9, SI, p. 165, 2019.