COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR BUMN BANK STOCK SENTIMENT CLASSIFICATION DURING DANANTARA FORMATION PERIOD

ANALISIS KOMPARATIF MODEL MACHINE LEARNING UNTUK KLASIFIKASI SENTIMEN SAHAM BANK BUMN PADA PERIODE PEMBENTUKAN DANANTARA

Hafizha Nurul Qolby[1], Rangga Gelar Guntara[2], Syti Sarah Maesaroh[3]
Universitas Pendidikan Indonesia, Jl. Dadaha No.18, Kahuripan, Tawang, Tasikmalaya
hafizhaqolby@upi.edu[1], ranggagelar@upi.edu[2], sytisarah@upi.edu[3]

Abstract *- Discussions about state-owned bank stocks (BBRI, BBNI, and BMRI) on platform X intensified during the formation of Danantara. However, the correlation between social media sentiment and stock movements remains weak due to high noise levels and potential buzzer activity. This study combines sentiment and text similarity analyses (cosine similarity) to identify repeated communication patterns in discussions related to state-owned bank stocks. A total of 1,086 tweets were manually labeled and verified by two independent validators Text features were represented using TF–IDF and evaluated through four classical machine learning algorithms: Naïve Bayes, Logistic Regression, Support Vector Machine, and XGBoost. The model was validated using a hold-out scheme (80:20) and assessed with a confusion matrix. The sentiment distribution of the dataset shows 53% negative and 47% positive tweets Logistic Regression achieved the highest accuracy of 66%. The cosine similarity analysis identified 1.8% of tweets with similarity ≥0.90, indicating limited recurring communication patterns. These findings suggest that integrating sentiment and text similarity analyses can serve as an initial approach to detect indications of coordinated activity and to understand public opinion dynamics toward state-owned bank stocks.*

Keywords - *sentiment analysis, cosine similarity, machine learning, stocks, SOE.*

Abstrak - Diskusi mengenai saham bank BUMN (BBRI, BBNI, dan BMRI) di *platform* X meningkat pesat saat pembentukan Danantara. Namun, korelasi antara sentimen media sosial dan pergerakan harga saham masih lemah akibat tingginya *noise* dan potensi aktivitas *buzzer*. Penelitian ini menggabungkan analisis sentimen dan analisis kemiripan teks (*cosine similarity*) untuk mengidentifikasi pola komunikasi yang berulang dalam percakapan saham bank BUMN. Sebanyak 1.086 *tweet* dilabeli secara manual dan diverifikasi oleh dua validator independen. Fitur teks dibangun dengan TF–IDF dan algoritma *machine learning* klasik, yaitu Naïve Bayes, Logistic Regression, Support Vector Machine, dan XGBoost. Model divalidasi dengan skema *hold-out* (80:20) dan dievaluasi menggunakan *confusion* matrix. Distribusi sentimen menunjukkan 53% *tweet* bernada negatif dan 47% positif. Hasil evaluasi menunjukkan Logistic Regression memiliki akurasi tertinggi sebesar 66%. Analisis *cosine similarity* mengidentifikasi 1,8% *tweet* dengan kemiripan ≥0,90, mengindikasikan pola komunikasi berulang meski tidak masif. Temuan ini menegaskan bahwa kombinasi analisis sentimen dan kemiripan teks dapat digunakan sebagai langkah awal untuk mendeteksi indikasi aktivitas terkoordinasi serta memahami dinamika opini publik terhadap saham bank BUMN.

Kata Kunci - analisis sentimen, *cosine similiarity*, *machine learning*, saham, bank BUMN.

I. PENDAHULUAN

Popularitas pasar modal Indonesia terus meningkat dalam beberapa tahun terakhir, ditandai dengan lonjakan signifikan pada jumlah investor ritel yang telah menembus angka 16 juta *Single Investor Identification* (SID) per Juni 2025. Pertumbuhan ini sejalan dengan ekspansi media sosial yang mendorong maraknya diskusi saham di media sosial, khususnya pada *platform* X (Twitter)[1]. Berbeda dengan investor institusional yang biasanya yang mengandalkan analisis terstruktur, investor ritel lebih dipengaruhi oleh faktor emosional dan informasi yang bersumber dari opini publik [2], [3]. Akibatnya, sentimen di media sosial berpotensi membentuk perilaku dan arah keputusan investasi.

Berbagai studi menunjukkan bahwa media sosial, khususnya X, mencerminkan persepsi pasar dan sentimen investor yang berkaitan dengan pergerakan harga saham[4], [5]. Penelitian Liu dkk[6] juga menemukan bahwa hubungan antara harga saham, *return*, dan sentimen dapat dikaitkan dengan kejadian pasar atau berita tertentu. Pembentukan Danantara menjadi salah satu peristiwa ekonomi yang memicu lonjakan percakapan publik mengenai saham bank BUMN. Kebijakan ini menuai beragam reaksi, mulai dari optimisme terhadap efisiensi pengelolaan investasi [7] hingga kekhawatiran terkait transparansi dan pengawasan kelembagaan [8]. Puncak diskusi publik terjadi pada 24 Februari 2025, ketika topik ini menempati posisi *trending* pertama di *platform* X dengan lebih dari 65 ribu cuitan. Ketidakpastian yang tercermin dalam sentimen publik dapat berimplikasi pada dinamika pasar modal, sejalan dengan temuan Cevik dkk [9] yang menyatakan bahwa sentimen negatif dapat meningkatkan volatilitas saham.

Analisis sentimen berbasis *machine learning* telah banyak diterapkan untuk memahami opini publik di berbagai domain media sosial, mulai dari isu politik dan sosial [10], [11], [12], [13]. Penelitian terdahulu menunjukkan bahwa algoritma klasik seperti Naïve Bayes, Logistic Regression, dan SVM mampu memberikan hasil yang kompetitif pada klasifikasi teks NLP [14], [15]. Berbeda dari penelitian sebelumnya yang umumnya hanya menggunakan algoritma linear, penelitian ini juga menguji XGBoost, yaitu algoritma *ensemble* yang menggabungkan sejumlah *decision tree* secara bertahap melalui proses *gradient boosting* yang memungkinkan model untuk memperbaiki kesalahan sehingga mampu menghasilkan prediksi yang lebih akurat [16]. Masing-masing algoritma memiliki karakteristik berbeda, Naïve Bayes unggul dalam kecepatan dan efisiensi, Logistic Regression efektif dalam klasifikasi biner dengan data linier [17], SVM memiliki *margin* pemisah yang kuat untuk data berdimensi tinggi, sedangkan XGBoost menawarkan performa tinggi melalui optimasi berbasis *gradient boosting* [18].

Seluruh model menggunakan representasi *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengubah teks menjadi fitur numerik yang mencerminkan bobot setiap kata [19]. Selain itu, penelitian ini juga menerapkan metrik *cosine similarity* untuk mengukur tingkat kemiripan antar-*tweet*. Pendekatan ini digunakan untuk mengidentifikasi kemunculan pola komunikasi yang serupa atau berulang, yang dapat memberikan wawasan awal mengenai potensi penyebaran informasi terkoordinasi (*buzzer*) di ruang diskusi saham. Berdasarkan latar belakang tersebut, penelitian ini berupaya menjawab dua pertanyaan utama:

(1) Bagaimana performa empat algoritma *machine learning* klasik dalam mengklasifikasikan sentimen *tweet* terkait saham bank BUMN di *platform* X, dan
(2) Sejauh mana tingkat kemiripan teks antar-*tweet* dapat mengindikasikan pola komunikasi berulang yang berpotensi mencerminkan aktivitas terkoordinasi (*buzzer*).

Selanjutnya, penelitian ini bertujuan untuk membandingkan performa empat algoritma *machine learning* klasik dalam klasifikasi sentimen *tweet* saham bank BUMN di platform X. Hasil penelitian diharapkan dapat memberikan kontribusi empiris dalam pengembangan model analisis sentimen berbahasa Indonesia yang lebih andal serta menjadi dasar bagi penelitian lanjutan mengenai pengaruh opini publik terhadap dinamika pasar modal di Indonesia.

II. SIGNIFIKANSI STUDI

A. Studi Literatur

Analisis sentimen merupakan cabang *Natural Language Processing* (NLP) yang berfokus pada identifikasi dan klasifikasi emosi atau opini dari teks, seperti positif, negatif, dan netral [20]. Dalam konteks pasar modal, sentimen di media sosial berperan penting dalam mencerminkan persepsi dan perilaku investor [3], [21], [22]. Sejumlah penelitian terdahulu menunjukkan bahwa opini publik di *platform* X memiliki korelasi kuat dengan dinamika harga saham karena sentimen cenderung cepat direspon oleh pasar [23], [24], [25]. Pembentukan Danantara menjadi peristiwa ekonomi yang memicu polarisasi opini terhadap saham bank BUMN, menjadikannya periode ideal untuk menganalisis dinamika sentimen publik.

Penelitian ini berfokus pada perbandingan empat model *machine learning* klasik untuk mengidentifikasi model paling sesuai bagi teks pendek berbahasa Indonesia. Model linear seperti Naïve Bayes, Logistic Regression, dan SVM dikenal memiliki efisiensi komputasi tinggi dan stabilitas parameter yang baik untuk teks berdimensi tinggi dengan representasi TF-IDF [14], [26], [27]. Beberapa studi menunjukkan bahwa model linear tersebut cenderung memberikan hasil optimal pada *dataset* kecil hingga menengah yang bersifat *sparse* karena mampu menangkap pola linier dengan presisi yang memadai [28], [29].

Sejumlah studi komparatif juga menyoroti konteks perbedaan performa antara model linear dan *ensemble*. Logistic Regression dan SVM cenderung unggul pada teks pendek dengan Tfitur TF-IDF yang linier dan terdistribusi jarang karena kestabilan parameter serta efisiensi pelatihan [29], [30]. Sebaliknya, XGBoost lebih unggul Ketika data berukuran besar dan memiliki pola kompleks atau interaksi non-linear antarfitur [16], [31], [32] Pemetaan ini menjadi penting untuk memahami konteks keunggulan relatif tiap algoritma sekaligus memperkuat dasar pemilihan model dalam penelitian ini.

Tingginya *social noise* di media sosial tidak hanya disebabkan oleh opini emosional, tetapi juga potensi aktivitas terkoordinasi oleh akun *buzzer* yang menyebarkan konten berulang untuk memengaruhi harga saham. Penelitian sebelumnya telah mengembangkan metode di luar klasifikasi sentimen, yaitu analisis kemiripan teks menggunakan *cosine similarity*. Metode ini diimplementasikan untuk mendeteksi *tweet* berdasarkan kemiripan konten pengguna dan resonansi pesan strategis [33]. Dalam penelitian ini, nilai kemiripan $\geq 0{,}90$ diinterpretasikan sebagai indikasi kuat adanya koordinasi pesan, mengacu pada penelitian Al Rasyid dkk. [34] yang menetapkan ambang tersebut sebagai batas antara kemiripan semantik dan duplikasi literal. Analisis ini membantu membedakan kesamaan topik alami dari pola komunikasi yang berulang.

B. Metode Penelitian

Penelitian ini mengadopsi *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai kerangka metodologi utama karena sifatnya yang sistematis, fleksibel, dan telah banyak digunakan dalam proyek *data mining* modern [35], [36]. Pendekatan ini membagi proses pengembangan menjadi enam fase, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Penelitian ini menggunakan pendekatan kuantitatif-komparatif dengan data sekunder berupa *tweet* di *platform* X yang membahas saham bank BUMN, khususnya BRI, BNI, dan Mandiri. Kerangka metodologi CRISP-DM yang digunakan dalam penelitian ini ditunjukkan pada Gambar 1.

Gambar 1. Kerangka Metodologi CRISP-DM

Business Understanding

Tahap ini berfokus pada pemahaman konteks pasar modal Indonesia dan perumusan masalah penelitian [37]. Media sosial kini berperan penting dalam membentuk persepsi investor ritel, meskipun korelasi antara sentimen dan harga saham masih rendah akibat *social noise,* yaitu gangguan informasi yang muncul dari misinformasi atau aktivitas akun yang menyebarkan konten berulang seperti *buzzer* [38]. Kondisi ini dapat mengaburkan sinyal sentimen yang sebenarnya tercermin dari opini publik. Oleh karena itu, penelitian ini mengembangkan model analisis sentimen berbasis *machine learning* untuk mengklasifikasikan opini publik terkait saham bank BUMN, serta mengevaluasi potensi pola komunikasi terkoordinasi.

Data Understanding

Tahap ini bertujuan memahami karakteristik data yang digunakan. Data dikumpulkan melalui *web scraping* dari *platform* X menggunakan kata kunci "BBRI OR BBNI OR BMRI" selama satu bulan, menghasilkan 1.498 *tweet* berbahasa Indonesia. Analisis difokuskan pada kolom "full_text" sebagai representasi opini pengguna. Sebagian besar *tweet* menggunakan bahasa informal dan singkatan dengan panjang kurang dari 25 kata. Istilah yang paling sering muncul meliputi "bbri", "saham", "turun", "beli", dan "ihsg", yang menunjukkan fokus percakapan publik pada saham bank BUMN dan kondisi pasar modal nasional. Proses pengumpulan data dilakukan sesuai *Terms of Service* (ToS) *platform* X dengan menjaga prinsip etika penelitian digital. Seluruh data bersifat publik dan informasi identitas pengguna seperti nama akun serta tautan dihapus sebelum dianalisis untuk memastikan privasi dan kepatuhan etis.

Data Preparation

Tahap *data preparation* bertujuan untuk menyiapkan data mentah agar dapat diolah secara optimal oleh model *machine learning*. Dalam penelitian ini, proses *data preparation* terdiri atas tiga tahap utama, yaitu *data cleansing*, *sentiment labeling*, dan *feature extraction*. Ketiga tahap ini berfungsi untuk memastikan data yang digunakan telah bersih, terstruktur, dan dapat direpresentasikan secara numerik sehingga siap untuk proses pemodelan. Tahap *data cleansing* dilakukan untuk menghilangkan elemen yang tidak relevan dari teks *tweet* agar siap digunakan dalam analisis. *Tweet* di media sosial umumnya mengandung *noise* yang tidak berkontribusi terhadap makna sentimen. Proses pembersihan dilakukan melalui beberapa langkah: menghapus karakter khusus, tanda baca berlebih, URL dan tagar; mengubah seluruh huruf menjadi huruf kecil (*case folding*); menghapus kata umum (*stopwords*) seperti "yang", "dan", "di", serta "ke"; serta melakukan *tokenizing* dan

stemming menggunakan pustaka Sastrawi untuk mengembalikan setiap kata ke bentuk dasarnya [39]. Tahapan ini penting untuk meningkatkan kualitas data dan memastikan hasil analisis sentimen lebih akurat [40]

Tahap *sentiment labeling* bertujuan untuk mengklasifikasikan setiap *tweet* berdasarkan arah emosional terhadap saham yang dibahas. Dalam konteks pasar modal, sentimen dibagi menjadi dua kategori utama: *bullish* (positif) yang mencerminkan optimisme terhadap kenaikan harga saham dan *bearish* (negatif) yang menunjukkan pesimisme terhadap penurunan harga [41]. Penelitian ini menggunakan pendekatan klasifikasi biner dengan menghapus kategori netral agar tidak menimbulkan *class imbalance*. Proses pelabelan dilakukan secara manual dan divalidasi oleh dua praktisi pasar berpengalaman, yaitu seorang *Chief Investment Officer* dan seorang dosen sekaligus pelaku saham aktif. Validasi ini dilakukan untuk memastikan kesesuaian interpretasi dengan konteks pasar modal Indonesia.

Tahap terakhir dalam *data preparation* adalah *feature extraction*, yaitu proses mengubah data teks yang telah dibersihkan menjadi representasi numerik yang dapat diproses oleh algoritma *machine learning*. Penelitian ini menggunakan metode TF–IDF, yang memberikan bobot numerik pada setiap kata berdasarkan tingkat kemunculannya dalam sebuah dokumen (*term frequency*) dan kelangkaannya di seluruh korpus (*inverse document frequency*). Nilai TF–IDF yang lebih tinggi menunjukkan bahwa kata tersebut memiliki makna penting dan berkontribusi dalam membedakan sentimen positif dan negatif dalam teks.

Dataset kemudian dibagi menjadi 80% data latih dan 20% data uji menggunakan metode *train-test split* untuk memastikan validitas hasil pemodelan. Rasio 80:20 dipilih karena secara empiris dianggap seimbang untuk dataset berukuran menengah, dimana 80% data memberikan stabilitas parameter model, sedangkan 20% data uji cukup representatif untuk mengevaluasi kinerja model secara objektif tanpa menyebabkan *data leakage* atau bias evaluasi.

Modeling

Pada tahap ini, empat model *machine learning* klasik digunakan untuk mengklasifikasikan sentimen *tweet* menjadi dua kategori: positif dan negatif. Pemilihan model didasarkan pada pertimbangan efisiensi komputasi, interpretabilitas, serta relevansi terhadap dataset berukuran menengah dan berdimensi tinggi seperti teks hasil TF-IDF Tabel berikut merangkum kelebihan dan keterbatasan utama masing-masing algoritma berdasarkan karakteristik data teks dan hasil studi terdahulu.

TABEL I
KELEBIHAN DAN KETERBATASAN MODEL ALGORITMA

Algoritma	Kelebihan	Keterbatasan
Naïve Bayes	- Sederhana dan efisien untuk dataset besar. - Cocok untuk teks berdimensi tinggi dan *sparse*. - Mampu menghasilkan prediksi probabilistik.	- Asumsi independensi antar fitur sering tidak terpenuhi. - Sensitif terhadap korelasi antar kata. - Akurasi menurun jika distribusi data tidak normal.
Logistic Regression (LR)	- Mudah diimplementasikan dan diinterpretasikan. - Stabil untuk dataset kecil hingga menengah. - Memiliki dasar probabilistik yang kuat untuk klasifikasi biner. - Tidak memerlukan asumsi distribusi tertentu.	- Kurang mampu menangkap hubungan non-linear. - Rentan terhadap multikolinearitas antar fitur. - Akurasi menurun pada data yang kompleks atau tidak seimbang.

Algoritma	Kelebihan	Keterbatasan
Support Vector Machine (SVM)	- Performa kuat pada data berdimensi tinggi seperti TF-IDF. - Risiko *overfitting* rendah dengan pemilihan kernel yang tepat - Efektif untuk data teks pendek atau semi-terstruktur.	- Komputasi relatif mahal untuk dataset besar. - Sulit diinterpretasikan (*black-box model*). - Sensitif terhadap data tidak terstruktur (*noisy data*)
Extreme Gradient Boosting (XGBoost)	- Akurasi tinggi berkat pendekatan *ensemble*. - Mampu menangani relasi non-linear dan interaksi antar fitur.	- Memerlukan *tuning hyperparameter* yang kompleks. - Waktu pelatihan lebih lama dibanding model linear. - Interpretasi hasil relatif lebih sulit.

Setiap model dikonfigurasi dengan *hyperparameter* yang disesuaikan dengan karakteristik teks berbahasa Indonesia yang pendek dan informal. Fitur diekstraksi menggunakan kombinasi TF-IDF *word-level* dan *character-level* melalui FeatureUnion untuk menangkap konteks semantik dan pola penulisan khas di *tweet*. Selain itu, dilakukan analisis *cosine similarity* antar-*tweet* guna mengidentifikasi pola komunikasi berulang yang berpotensi menunjukkan aktivitas *buzzer*, dengan nilai kemiripan ≥0,90 dianggap sebagai indikasi kuat adanya pola teks identik. Adapun persamaan cosine similiarity disajikan pada persamaan 1 sebagai berikut:

$$Cosine\ similiarity = \cos(\theta) = \frac{\mathrm{A} \bullet \mathrm{B}}{||\mathrm{A}|| \times ||\mathrm{B}||} \quad (1)$$

Evaluation

Evaluasi model dilakukan menggunakan empat metrik utama, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*, serta visualisasi *confusion matrix* untuk menilai sebaran hasil prediksi. Model dengan performa terbaik berdasarkan nilai *F1-score* dan stabilitas hasil *cross-validation* dipilih sebagai model utama untuk analisis sentimen lebih lanjut. Tabel *confusion matrix* ditunjukkan pada Tabel 2.

TABEL II
CONFUSION MATRIX

Kelas Sebenarnya	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Keterangan:
True Positive (TP) : *Tweet* positif yang berhasil diprediksi benar sebagai positif.
True Negative (TN) : *Tweet* negatif yang berhasil diprediksi benar sebagai negatif.
False Positive (FP) : *Tweet* negatif yang salah diprediksi sebagai positif.
False Negative (FN) : *Tweet* positif yang salah diprediksi sebagai negatif.

Deployment

Tahap ini bertujuan untuk menginterpretasikan hasil klasifikasi sentimen dan memahami pola persebaran opini publik terkait saham bank BUMN selama periode penelitian. Model dengan performa terbaik digunakan untuk memetakan distribusi sentimen positif dan negatif guna melihat kecenderungan persepsi pasar. Analisis tambahan menggunakan *cosine similarity* diterapkan untuk mengamati tingkat kemiripan antar-*tweet*, yang dapat memberikan gambaran mengenai pola komunikasi berulang atau penyebaran informasi dengan struktur serupa di media sosial.

III. HASIL DAN PEMBAHASAN

Penelitian ini menjalankan dua rangkaian eksperimen utama. Eksperimen pertama berfokus pada mengembangkan model analisis sentimen menggunakan empat algoritma *machine learning*. Eksperimen kedua menambahkan analisis kemiripan teks untuk mendeteksi kemungkinan adanya pola komunikasi terkoordinasi yang menunjukkan aktivitas *buzzer*.

A. Dataset

Dataset dikumpulkan dari *platform* X (Twitter) menggunakan pustaka *tweet-harvest*, dengan kata kunci yang relevan terhadap saham bank BUMN. Periode pengambilan data dipilih selama satu bulan setelah pembentukan Danantara, karena peristiwa tersebut memicu peningkatan aktivitas dan perubahan sentimen investor terhadap saham bank BUMN di media sosial.

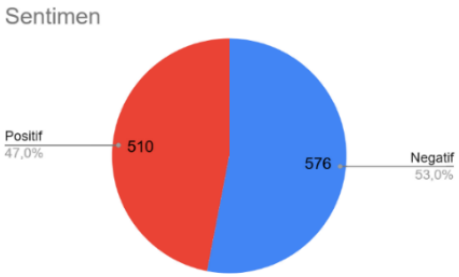
TABEL III
KRITERIA DATASET *TWEET*

Kriteria	Limitasi
Keyword/hastags	BBRI, BBNI, BMRI
Periode pengambilan data	24 Februari – 24 Maret 2025
Bahasa	Indonesia
Konten	Tweet yang berisi teks kosong, link berita tanpa opini, iklan, atau promosi non-saham dihapus.

Proses *scraping* menghasilkan total 1.498 tweet publik berbahasa Indonesia yang relevan dengan topik saham BUMN. Setiap entri memiliki sejumlah atribut utama seperti conversation_id_str (identitas percakapan), created_at (waktu unggahan), id_str dan user_id_str (pengenal unik), serta full_text yang berisi isi opini pengguna. Atribut tambahan seperti image_url, lang, location, dan tweet_url digunakan sebagai pelengkap metadata.

Data tersebut merepresentasikan percakapan publik di *platform* X mengenai saham bank BUMN selama periode pembentukan Danantara. Setelah dilakukan tahap *text preprocessing*, dilanjutkan proses pelabelan sentimen secara manual. Pada tahap ini, *tweet* dengan sentimen netral dihapus karena tidak menunjukkan arah emosi yang jelas terhadap saham, sehingga dapat menimbulkan *class imbalance* dalam pelatihan model.

Setelah penyaringan ini, jumlah data akhir yang digunakan adalah 1.086 *tweet*, terdiri atas 576 *tweet* (53,0%) diklasifikasikan sebagai sentimen negatif dan 510 *tweet* (47,0%) menunjukkan sentimen positif. *Tweet* bernada positif umumnya mencerminkan optimisme terhadap kinerja saham bank BUMN dan prospek ekonomi, sementara *tweet* bernada negatif cenderung berisi kekhawatiran terhadap kondisi pasar dan kebijakan pembentukan Danantara.

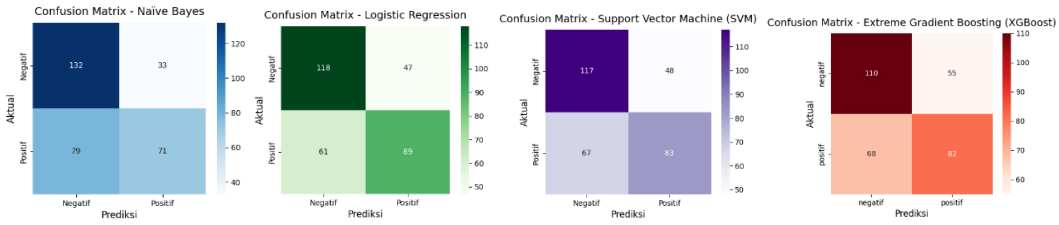


Gambar 2. Distribusi Sentimen Berdasarkan Pelabelan Manual

B. *Komparasi Model Algoritma*

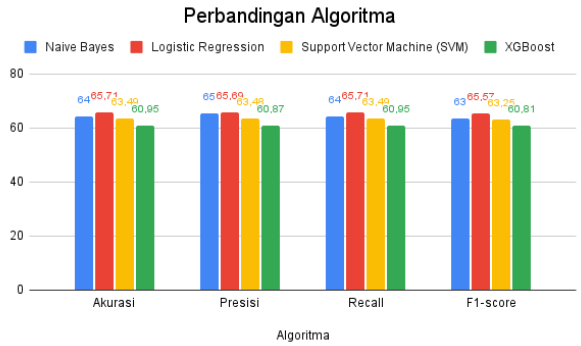
Evaluasi dilakukan terhadap empat algoritma *machine learning* dengan metrik evaluasi berupa *accuracy, precision, recall*, dan *F1-score*. Hasil evaluasi menunjukkan bahwa jumlah kesalahan klasifikasi lebih kecil dibanding prediksi aktual, yang menandakan bahwa model jarang salah mengidentifikasi sentimen positif sebagai negatif atau sebaliknya.

Berdasarkan *confusion matrix* pada Gambar 3, model Logistic Regression menunjukkan performa terbaik dengan 118 prediksi benar untuk kelas negatif dan 89 untuk kelas positif, serta total kesalahan klasifikasi (FP+FN) sebanyak 108 dari 315 data uji. Model SVM mencatat hasil serupa dengan 117 prediksi benar untuk kelas negatif dan 83 untuk kelas positif, sedangkan Naïve Bayes dan XGBoost menunjukkan tingkat kesalahan lebih tinggi. Distribusi kesalahan antar kelas relatif seimbang, menandakan tidak adanya bias signifikan terhadap salah satu kelas sentimen. Hasil ini memperkuat bahwa model linear seperti Logistic Regression dan SVM cenderung lebih stabil pada teks pendek berbasis TF-IDF dibandingkan model XGBoost.



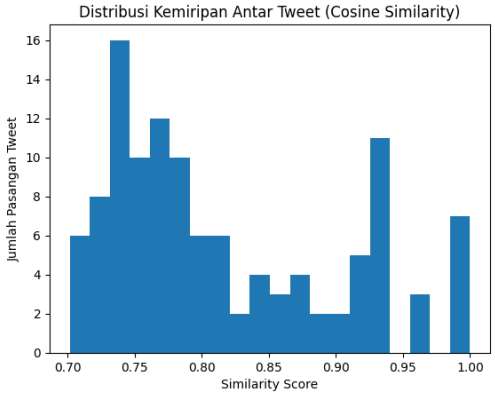
Gambar 3. *Heatmap Confusion Matrix*

Berdasarkan Gambar 4, Logistic Regression dan SVM menunjukkan performa paling stabil dengan nilai *F1-score* tertinggi, masing-masing sebesar 65,71% dan 63,49%, yang menandakan keseimbangan antara presisi dan kemampuan model dalam mendeteksi kelas dengan benar. Naïve Bayes menunjukkan performa yang sedikit lebih rendah, yaitu 63%, namun tetap unggul dari sisi efisiensi komputasi dan kesederhanaan implementasi, sehingga tetap sesuai untuk analisis teks pendek seperti *tweet* [30]. Sementara itu, XGBoost mencatat nilai akurasi terendah sebesar 60,95%, menunjukkan bahwa kompleksitas model kurang optimal untuk dataset dengan ukuran terbatas. Hasil ini konsisten dengan temuan Jahan dkk [29] yang menyimpulkan bahwa algoritma linear lebih unggul dalam klasifikasi sentimen berbasis TF–IDF pada *tweet*.



Gambar 4. Perbandingan Metrik Evaluasi Algoritma

C. Cosine Similiarity Untuk Mendeteksi Potensi Buzzer



Gambar 5. Cosine Similarity

Analisis kemiripan teks dilakukan menggunakan metrik *cosine similarity* untuk mengidentifikasi potensi adanya pola komunikasi terkoordinasi. Nilai kemiripan dihitung berdasarkan representasi vektor teks menggunakan metode TF-IDF, dengan rentang nilai antara 0 (tidak mirip) hingga 1 (identik). Untuk menghindari artefak data, seluruh *retweet* dan *quote tweet* telah dihapus pada tahap praproses sehingga hanya mencakup teks orisinal. Hasil observasi temporal menunjukkan bahwa *tweet* dengan nilai kemiripan ≥0,90 tidak tekonsentrasi pada satu tanggal tertentu, melainkan tersebar sepanjang periode pengamatan, sehingga kecil kemungkinan adanya aktivitas terkoordinasi dalam waktu bersamaan (*temporal burst*).

Hasil analisis menunjukkan bahwa *tweet* memiliki nilai kemiripan pada rentang 0,70–0,80, menandakan adanya kesamaan topik, namun tidak menunjukkan duplikasi langsung. Sebaliknya, *tweet* yang memiliki nilai ≥0,90, menunjukkan pola hampir identik dan berpotensi *buzzer*. Klasifikasi tingkat kemiripan ini sejalan dengan pendekatan yang digunakan oleh Al Rasyid dkk. [34], yang menetapkan nilai di atas 0,90 sebagai kemiripan tinggi, sementara rentang 0,60-0,80 menggambarkan kemiripan tematik tanpa indikasi duplikasi langsung.nSecara keseluruhan, tingkat kemiripan teks tergolong moderat, sehingga percakapan publik mengenai saham bank BUMN pada periode pembentukan Danantara cenderung bersifat organik, dengan indikasi aktivitas *buzzer* yang relatif rendah. Hasil ini sejalan dengan studi De Clerck dkk, yang menyebutkan bahwa aktivitas *buzzer* umumnya ditandai dengan pola teks berulang, penggunaan istilah identik, serta waktu unggahan yang berdekatan.

D. Diskusi

Hasil menunjukkan bahwa SVM dan Logistic Regression merupakan algoritma paling optimal dalam mengklasifikasikan sentimen *tweet* saham bank BUMN, dengan akurasi dan *F1-score* tertinggi dibandingkan model lainnya. Keduanya terbukti stabil pada skema validasi silang dan efektif dalam memproses teks pendek berbasis TF-IDF, sejalan dengan temuan Rumapea & Suria [18] yang menyatakan bahwa model linear memiliki kemampuan baik dalam memisahkan kelas sentimen tanpa memerlukan kompleksitas model tinggi.

Sebaliknya, Naïve Bayes menghasilkan performa moderat karena asumsi independensi fitur yang kuat, yaitu menganggap semua kata berkontribusi secara independen terhadap kelas sentimen. Asumsi ini sulit dipenuhi secara sempurna dalam NLP, di mana sentimen seringkali bergantung pada konteks dan kombinasi kata (*word dependency*). Lebih lanjut, XGBoost menunjukkan kinerja terendah, yang kemungkinan disebabkan oleh sensitivitas model *boosting* terhadap *social noise* dan kompleksitas dalam dataset yang relatif kecil. Hasil ini menguatkan bahwa, untuk analisis sentimen berbahasa Indonesia yang informal dan padat konteks, model linier seperti Logistic Regression lebih representatif karena menghindari kompleksitas berlebih yang dapat menyebabkan *overfitting*.

Selanjutnya, analisis *cosine similarity* digunakan untuk mengidentifikasi potensi pola komunikasi berulang dalam percakapan. Sebagian besar pasangan *tweet* memiliki tingkat kemiripan di bawah 0.90, yang menunjukkan bahwa diskusi publik bersifat organik. Namun, adanya sekitar 1,8% *tweet* dengan nilai kemiripan ≥ 0.90 mengindikasikan adanya kemiripan struktur teks yang berpotensi mencerminkan penyebaran pesan secara terkoordinasi. Temuan ini selaras dengan studi Tommasel & Rodriguez yang menyebutkan bahwa akun dengan pola linguistik berulang sering digunakan untuk memperkuat persepsi tertentu di media sosial.

Dengan demikian, integrasi antara analisis sentimen yang efisien dan analisis kemiripan teks terbukti efektif dalam memberikan gambaran menyeluruh terhadap dinamika opini publik, serta dapat menjadi pendekatan awal dalam mendeteksi indikasi manipulasi informasi atau diseminasi pesan terkoordinasi di ekosistem pasar modal di Indonesia. Penelitian selanjutnya disarankan untuk mengeksplorasi penggunaan model deep learning, seperti BERT atau LSTM, menggunakan korpus *tweet* berukuran lebih besar dan periode lebih panjang. Riset lanjutan juga dapat menerapkan validasi silang untuk meningkatkan reliabilitas hasil, serta memperdalam analisis terhadap profil pengguna yang berpotensi sebagai *buzzer* dengan meninjau karakteristik akun, seperti usia akun dan tingkat keaktifan di *platform* X.

IV. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa pendekatan *machine learning* seperti Logistic Regression dan Support Vector Machine (SVM) masih relevan dan efektif digunakan untuk analisis sentimen teks pendek, meskipun tingkat akurasi yang diperoleh berada pada kisaran 61-66%. Nilai tersebut tergolong wajar mengingat karakteristik data media sosial yang informal, mengandung singkatan, serta memiliki konteks semantik yang beragam. Analisis *cosine similarity* juga memberikan wawasan tambahan terkait adanya sebagian kecil *tweet* dengan kemiripan tinggi yang menunjukkan potensi penyebaran pesan *buzzer*, meskipun secara umum percakapan publik mengenai saham bank BUMN selama periode pembentukan Danantara bersifat organik. Penelitian ini berkontribusi dalam memperkuat pemahaman tentang dinamika sentimen saham di Indonesia serta memberikan dasar bagi pengembangan model deteksi konten berulang dan analisis sentimen yang lebih komprehensif pada studi mendatang. Meskipun demikian, penelitian ini memiliki beberapa keterbatasan, antara lain periode pengambilan data yang terbatas, penghapusan kelas sentimen netral yang dapat memengaruhi keseimbangan data, serta belum diterapkannya penyetelan *hyperparameter* dan validasi silang yang lebih mendalam. Selain itu, analisis kemiripan teks belum mencakup dimensi berbasis waktu maupun analisis akun. Oleh karena itu, riset lanjutan disarankan untuk memperluas cakupan data, menguji model berbasis *deep learning*, serta memperhatikan karakteristik pengguna agar pemetaan pola komunikasi dapat digambarkan secara lebih komprehensif.

REFERENSI

[1] G. Vicentini, A. Nucci, G. Caldarelli, dan E. Omodei, “Social media discussions anticipates financial market volumes,” *Physica A: Statistical Mechanics and its Applications*, vol. 661, Mar 2025, doi: 10.1016/j.physa.2025.130388.

[2] J. Cui, Q. Wei, dan X. Gao, “How Retail vs. Institutional Investor Sentiment Differ in Affecting Chinese Stock Returns?,” *Journal of Risk and Financial Management*, vol. 18, no. 2, Feb 2025, doi: 10.3390/jrfm18020095.

[3] I. G. N. A. Dananjaya, M. A. Prayudi, dan G. N. H. Wiguna, “The Influence of Retail Investor Activity and Sentiment on Social Media on Stock Market Dynamics in Bali,” *E-Jurnal Akuntansi*, vol. 35, no. 7, Agu 2025, doi: 10.24843/EJA.2025.v35.i07.p19.

[4] T. Adams, A. Ajello, D. Silva, dan F. Vazquez-Grande, "More than Words: Twitter Chatter and Financial Market Sentiment," *Finance and Economics Discussion Series*, no. 2023–034, hlm. 1–36, Mei 2023, doi: 10.17016/feds.2023.034.
[5] D. M. S. de Souza dan O. S. Martins, "Brazilian stock market performance and investor sentiment on Twitter," *Revista de Gestao*, vol. 31, no. 1, Jan 2024, doi: 10.1108/REGE-07-2021-0145.
[6] Q. Liu, W. S. Lee, M. Huang, dan Q. Wu, "Synergy between stock prices and investor sentiment in social media," *Borsa Istanbul Review*, vol. 23, no. 1, Jan 2023, doi: 10.1016/j.bir.2022.09.006.
[7] J. Agustine, "Pakar UGM Ungkap Dampak Positif dan Negatif Kemunculan Danantara ," Universitas Gadjah Mada. [Daring]. Tersedia pada: https://ugm.ac.id/id/berita/pakar-ugm-ungkap-dampak-positif-dan-negatif-kemunculan-danantara/
[8] P. N. Maula, E. V. Daniel, M. H. A. Irawan, dan S. R. L. Gaol, "Pengawasan dan Pertanggungjawaban Badan Pengelola Investasi Danantara dalam Pengelolaan Risiko Kerugian Investasi Keuangan Negara," *Jurnal Hukum Statuta*, no. 2, Apr 2025, [Daring].
[9] E. Cevik, B. Kirci Altinkeski, E. I. Cevik, dan S. Dibooglu, "Investor sentiments and stock markets during the COVID-19 pandemic," *Financial Innovation*, vol. 8, no. 1, Des 2022, doi: 10.1186/s40854-022-00375-0.
[10] N. Hadi dan D. Sugiarto, "Analisis Sentimen Pembangunan IKN pada Media Sosial X Menggunakan Algoritma SVM, Logistic Regression dan Naïve Bayes," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 1, hlm. 37–49, Jan 2025, doi: 10.30591/jpit.v10i1.7106.
[11] H. Aulia, M. Zulfadhilah, S. E. Prastya, dan M. S. Pebriadi, "Analisis Sentimen Masyarakat Terhadap Kesehatan Mental pada Media Sosial Twitter dengan Menggunakan Machine Learning," *Positif: Jurnal Sistem dan Teknologi Informasi*, vol. 10, no. 2, hlm. 75–81, 2024.
[12] W. B. Zulfikar, A. R. Atmadja, dan S. F. Pratama, "Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes," *Scientific Journal of Informatics*, vol. 10, no. 1, hlm. 25–34, Jan 2023, doi: 10.15294/sji.v10i1.39952.
[13] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, dan T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: using Latent Dirichlet Allocation for topic modeling on Twitter," *PLoS One*, vol. 15, no. 9, 2020, Diakses: 19 Oktober 2025. [Daring]. Tersedia pada: https://doi.org/10.1371/journal.pone.0239441
[14] R. B. Dahlian dan D. Sitanggang, "Sentiment Analysis of Digital Television Migration on Twitter Using Naïve Bayes Multinomial Comparison, Support Vector Machines, and Logistic Regression Algorithms," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 2, hlm. 280–288, Jul 2023, doi: 10.32736/sisfokom.v12i2.1668.
[15] P. Assiroj, A. Kurnia, dan S. Alam, "The performance of Naïve Bayes, support vector machine, and logistic regression on Indonesia immigration sentiment analysis," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, hlm. 3843–3852, Des 2023, doi: 10.11591/eei.v12i6.5688.
[16] Y. Yulistiani dan S. Styawati, "Analisis Sentimen Terhadap Calon Presiden Indonesia 2024 dengan Metode Extreme Gradient Boosting (XGBOOST)," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 3, hlm. 322–328, Des 2024, doi: 10.30591/jpit.v9i3.6127.
[17] Sutarman, R. Siringoringo, D. Arisandi, E. Kurniawan, dan E. B. Nababan, "Model Klasifikasi Dengan Logistic Regression Dan Recursive Feature Elimination Pada Data Tidak Seimbang," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, hlm. 735–742, Agu 2024, doi: 10.25126/jtiik.1148198.
[18] D. I. Rumapea dan O. Suria, "Instagram-Based Sentiment Analysis on the Oil Refinery Project in Batam Using SVM and XGBoost," *Jurnal Inovtek Polbeng - Seri Informatika*, vol. 10, 2025.
[19] A. D. M. Putri, N. Sulistianingsih, dan R. Rismayati, "Pengaruh Teknik Representasi Teks Bag of Words dan TF-IDF terhadap Akurasi Klasifikasi Sentimen Teks Multi-Domain," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 7, no. 4, hlm. 675–688, 2025.
[20] Y. Mao, Q. Liu, dan Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, Apr 2024, doi: 10.1016/j.jksuci.2024.102048.
[21] K. N. F. Syahnur, D. F. Surianto, dan Muhammad Try Dharsana, "Analysis Of Retail Sector Market Reaction In Indonesia On Social Media And Investor Sentiment," *Jurnal Manajemen Bisnis*, vol. 10, no. 2, hlm. 371–383, Sep 2023, doi: 10.33096/jmb.v10i2.72.
[22] N. Yang, A. Fernandez-Perez, dan I. Indriawan, "Social Media Sentiment, Investor Herding and Informational Efficiency," *SSRN Electronic Journal*, Jul 2023, doi: 10.2139/ssrn.4551493.

[23] J. Cao, G. He, dan Y. Jiao, "Too Sensitive to Fail: The Impact of Sentiment Connectedness on Stock Price Crash Risk," *Entropy*, vol. 27, no. 4, Apr 2025, doi: 10.3390/e27040345.
[24] S. Li dan J. Kong, "News Sentiment and the Risk of a Stock Price Crash Risk: Based on Financial Dictionary Combined BERT-DCA," *Discrete Dyn Nat Soc*, 2022, doi: 10.1155/2022/8305947.
[25] Y. Burra, "The Influence of Social Media on Financial Markets: A Comprehensive Behavioral and Quantitative Analysis," *International Journal of Current Business and Social Sciences | IJCBSS*, vol. 10, no. 5, hlm. 2024, 2024, [Daring]. Tersedia pada: https://ijcbss.org
[26] M. Zhikri dan W. Istiono, "Handling Class Imbalance for Indonesian Twitter Sentiment Analysis A Comparative Study of Algorithms," *Journal of System and Management Sciences*, vol. 14, no. 10, Jun 2024, doi: 10.33168/jsms.2024.1010.
[27] D. S. Ramdan, R. D. Apnena, dan C. A. Sugianto, "Film Review Sentiment Analysis: Comparison of Logistic Regression and Support Vector Classification Performance Based on TF-IDF," *Journal of Applied Intelligent System*, vol. 8, no. 3, hlm. 341–352, 2023, [Daring]. Tersedia pada: http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/.
[28] R. Kholwal, "Text-Classify: A Comprehensive Comparative Study of Logistic Regression, Random Forest, and KNN Models for Enhanced Text Classification Performance," *International Journal of Advances in Engineering & Technology*, vol. 16, no. 5, hlm. 415–433, 2023.
[29] I. Jahan, M. N. Islam, M. M. Hasan, dan M. R. Siddiky, "Comparative analysis of machine learning algorithms for sentiment classification in social media text," *World Journal of Advanced Research and Reviews*, vol. 23, no. 3, hlm. 2842–2852, Sep 2024.
[30] E. Sihombing, M. H. Dar, dan F. A. Nasution, "Comparison Of Machine Learning Algorithms In Public Sentiment Analysis Of TAPERA Policy," *International Journal of Science*, hlm. 1089–1098, 2024, [Daring]. Tersedia pada: http://ijstm.inarah.co.id
[31] A. Chikhi, S. S. Mohammadi Ziabari, dan J. W. van Essen, "A Comparative Study of Traditional, Ensemble and Neural Network-Based Natural Language Processing Algorithms," *Journal of Risk and Financial Management*, vol. 16, no. 7, Jul 2023, doi: 10.3390/jrfm16070327.
[32] O. M. Ahmed, S. R. M. Zeebaree, dan S. Askar, "Comparative Analysis of XGBoost Performance for Text Classification with CPU Parallel and Non-Parallel Processing," *Indonesian Journal of Computer Science*, vol. 13, no. 2, hlm. 2024–1781, 2024.
[33] T. J. B. Cann, B. Dennes, T. Coan, S. O'Neill, dan H. T. P. Williams, "Using semantic similarity to measure the echo of strategic communications," *EPJ Data Sci*, vol. 14, no. 1, hlm. 20, Mar 2025, doi: 10.1140/epjds/s13688-025-00538-w.
[34] R. Al Rasyid, D. Handayani, dan U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, hlm. 2024, 2024, doi: 10.35870/jti.
[35] A. Rianti, N. W. A. Majid, dan A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," dalam *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) 2023*, 2023.
[36] C. Schröer, F. Kruse, dan J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," dalam *Procedia Computer Science*, Elsevier B.V., 2021, hlm. 526–534.
[37] J. Brzozowska, J. Pizoń, G. Baytikenova, A. Gola, A. Zakimova, dan K. Piotrowska, "Data Engineering in CRISP-DM Process Production Data - Case Study," *Applied Computer Science*, vol. 19, no. 3, hlm. 83–95, 2023, doi: 10.35784/acs-2023-26.
[38] N. P. Madali, M. Alsaid, dan S. Hawamdeh, "The impact of social noise on social media and the original intended message: BLM as a case study," *J Inf Sci*, vol. 50, no. 1, hlm. 89–103, Feb 2024, doi: 10.1177/01655515221077347.
[39] M. A. Hasanah, S. Soim, dan A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," 2021. [Daring]. Tersedia pada: http://jurnal.polibatam.ac.id/index.php/JAIC
[40] U. Naseem, I. Razzak, dan P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter," *Multimed Tools Appl*, vol. 80, no. 28–29, hlm. 35239–35266, Nov 2021, doi: 10.1007/s11042-020-10082-6.
[41] R. Abbas, M. Bilal Ijaz, dan A. Javeed, "Impact of Bullish and Bearish Trends on Investor Sentiment in Stock Market: A Study from Pakistan," *Contemporary Journal of Social Science Review*, vol. 02, no. 04, 2024.