Jurnal Riset Aplikasi Ilmu Data dan Sistem 1 (1), 183–195 DOI: 0000
Juli 2026

Implementasi dan Evaluasi Model Klasifikasi Sentimen untuk Menganalisis Opini Publik Terhadap Aset Kripto pada Media Sosial X dan Youtube

Muhammad Habib Nur Aiman[1], M. Habiburrohman Al-Fathir[1], Bagus Arya Dwipangga[2], Ulfa Siti Nuraini[1,*], Dian Handayani[2]

[1]*Program Studi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Indonesia*
[2]*Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, Indonesia*

Abstrak

Media sosial seperti X dan YouTube telah menjadi platform utama bagi publik untuk mengekspresikan opini terkait aset kripto. Namun, analisis sentimen publik saat ini seringkali masih menggunakan ekstraksi fitur tradisional yang mengabaikan konteks semantik dari teks tersebut. Untuk mengisi celah tersebut, penelitian ini mengimplementasikan dan mengevaluasi model klasifikasi sentimen dengan memanfaatkan ekstraksi fitur kontekstual berbasis *Sentence-BERT* (*SBERT*). Ruang lingkup data dalam penelitian ini mencakup 7.461 teks opini publik yang dikumpulkan dari platform X dan YouTube selama periode Januari hingga November 2025. Evaluasi dilakukan terhadap dua algoritma *supervised learning*, yaitu *Logistic Regression* dan *Support Vector Machine* (*SVM*). Hasil pengujian berdasarkan metrik *F1-Score* menunjukkan bahwa algoritma *SVM* memberikan performa terbaik dengan skor 0,73, mengungguli *Logistic Regression* yang memperoleh skor 0,70. Lebih dari sekadar performa model, temuan utama dari distribusi sentimen menunjukkan dominasi opini netral (3.593 data), diikuti sentimen negatif (1.991 data) dan positif (1.877 data). Hal ini mengindikasikan bahwa diskursus publik terhadap aset kripto sebagian besar bersifat informatif, namun dengan kecenderungan publik yang lebih waspada dan berhati-hati terhadap risiko pasar dibandingkan bersikap optimis. Seluruh hasil analisis ini divisualisasikan dalam bentuk dasbor interaktif sebagai alat bantu pemantauan sentimen secara *real-time*.

Kata Kunci *analisis sentimen; aset kripto; machine learning; Sentence-BERT; support vector machine*

Abstract

Social media platforms such as X and YouTube have become primary channels for the public to express opinions regarding crypto assets. However, current sentiment analyses often rely on traditional feature extraction methods that ignore the semantic context of the text. To address this gap, this study implements and evaluates sentiment classification models utilizing contextual feature extraction based on Sentence-BERT (SBERT). The data scope encompasses 7,461 public opinion texts collected from X and YouTube platforms during the January to November 2025 period. Two supervised learning algorithms, Logistic Regression and Support Vector Machine

[*]Corresponding author. Email: ulfanurani@unesa.ac.id.

Received 18 Mei 2026; Revised 8 Juli 2026; Accepted 8 Juli 2026

(SVM), were evaluated. The evaluation results, based on the F1-Score metric, demonstrate that the SVM algorithm achieved the best performance with a score of 0.73, outperforming Logistic Regression (0.70). Beyond model performance, the key findings from the sentiment distribution reveal a dominance of neutral opinions (3,593 data), followed by negative (1,991 data) and positive (1,877 data) sentiments. This indicates that public discourse on crypto assets is predominantly informational, with a tendency for the public to be more cautious and wary of market risks rather than overly optimistic. All analysis results are visualized in an interactive dashboard serving as a real-time sentiment monitoring tool.

Keywords *cryptocurrency; machine learning; sentiment analysis; Sentence-BERT; support vector machine*

1 Pendahuluan

Perkembangan teknologi informasi dan komunikasi telah membawa perubahan besar pada cara masyarakat berinteraksi di dunia digital. Media sosial kini menjadi wadah utama bagi publik untuk mengekspresikan pendapat terhadap berbagai macam isu, termasuk fenomena ekonomi dan mata uang digital seperti mata uang kripto (*cryptocurrency*). Platform seperti X dan YouTube merupakan dua dari banyak media sosial yang paling sering digunakan oleh pengguna untuk berdiskusi, bertukar informasi, serta membagikan pandangan mereka terkait mata uang kripto seperti Bitcoin, Ethereum, Solana, dan lain sebagainya. Mata uang kripto semakin populer seiring berkembangnya teknologi *blockchain* serta potensi keuntungan yang ditawarkan. Kendati demikian, volatilitas harga, isu regulasi pemerintah, serta risiko investasi yang tinggi membuat opini publik terhadap hadirnya mata uang kripto sangat beragam dan terpolarisasi.

Memahami pola opini publik terhadap mata uang kripto menjadi penting untuk banyak pihak, seperti investor, peneliti, dan pembuat regulasi. Salah satu pendekatan yang terbukti andal untuk menganalisis opini publik secara kuantitatif adalah analisis sentimen. Analisis ini sangat bergantung pada penerapan disiplin ilmu *Natural Language Processing* (*NLP*), yaitu sebuah cabang ilmu kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada interaksi antara komputer dan bahasa manusia untuk memproses teks dalam jumlah besar. Dengan bantuan teknik *NLP* dan *Machine Learning*, data teks tidak terstruktur dari media sosial dapat dikonversi menjadi informasi yang terukur (Sarkar, 2019). Penggunaan media sosial sebagai sumber data sentimen telah menjadi praktik umum, seperti dalam konteks generasi Z (Nurbaiti, 2023), pengungkapan diri mahasiswa (Tamaraya and Ubaedullah, 2021), hingga sentimen pariwisata (Yahya and Wahyuni, 2024).

Terkait dengan mata uang kripto, sejumlah studi internasional telah membuktikan efektivitas analisis sentimen dalam menangkap respons pasar (Lopez et al., 2023). Hassan et al. (2022) menganalisis 15.000 *tweet* menggunakan *sentiment lexicon* dan menemukan dominasi sentimen positif sebesar 33%. Aslam et al. (2022) menggunakan *deep learning* gabungan *LSTM-GRU* pada 40.000 teks terkait kripto dengan akurasi 0,99, sedangkan Bhowmik et al. (2025) membuktikan bahwa algoritma klasifikasi dapat memprediksi volatilitas Bitcoin menggunakan *Support Vector Machine* (*SVM*). Meskipun literatur tersebut menegaskan urgensi analisis sentimen pada domain kripto, sebagian besar studi terdahulu masih mengandalkan ekstraksi fitur tradisional (seperti *TF-IDF*) yang cenderung mengabaikan konteks semantik dari teks, serta seringkali hanya berfokus pada satu platform, khususnya Twitter.

Berangkat dari batasan penelitian terdahulu, penelitian ini menetapkan ruang lingkup studi kasus pada opini publik masyarakat Indonesia terhadap fenomena aset kripto selama tahun

2025 dengan menggabungkan komparasi data dari dua platform, yaitu X dan YouTube. Untuk mengatasi masalah hilangnya konteks semantik pada fitur teks konvensional, penelitian ini mengimplementasikan *Sentence-BERT* (*SBERT*) pada tahap rekayasa fitur (*feature engineering*). Penelitian terbaru menunjukkan bahwa model berbasis *transformer* sangat superior dalam menangani data finansial dan teks mikrobog (Chen et al., 2025; Naseem et al., 2024). Metode *SBERT* dipilih secara spesifik karena kemampuannya dalam merepresentasikan makna kalimat secara keseluruhan ke dalam bentuk vektor yang padat (*contextual embedding*), sehingga jauh lebih tangguh dalam memahami bahasa informal dan slang di media sosial dibandingkan metode tradisional (Reimers and Gurevych, 2019).

Sebagai metode pengklasifikasi, penelitian ini mengimplementasikan dan mengevaluasi dua algoritma pembelajaran mesin terarah (*supervised learning*), yaitu *Logistic Regression* dan *Support Vector Machine* (*SVM*). Kedua metode ini dipilih secara spesifik karena efisiensi komputasi dan ketangguhannya (*robustness*) dalam menangani data ruang vektor berdimensi tinggi (*dense vector*) yang dihasilkan oleh *SBERT*. Secara khusus, *SVM* sangat handal dalam menemukan *hyperplane* pemisah antar-kelas di ruang dimensi tinggi, sementara *Logistic Regression* sangat efektif dalam memodelkan probabilitas untuk klasifikasi *multiclass*. Melalui implementasi *SBERT* dan evaluasi kedua model klasifikasi tersebut, penelitian ini diharapkan mampu memberikan kerangka kerja yang komprehensif dalam memetakan opini publik terhadap aset kripto, yang hasilnya juga diintegrasikan ke dalam sebuah dasbor publik interaktif.

2 Metodologi

Penelitian ini mengikuti alur metodologi yang sistematis, dimulai dari inisialisasi, pengumpulan data, pra-pemrosesan teks, pelabelan sentimen, rekayasa fitur, pelatihan model, evaluasi, hingga penyajian hasil. Diagram alur keseluruhan proses penelitian ditunjukkan pada Gambar 1.

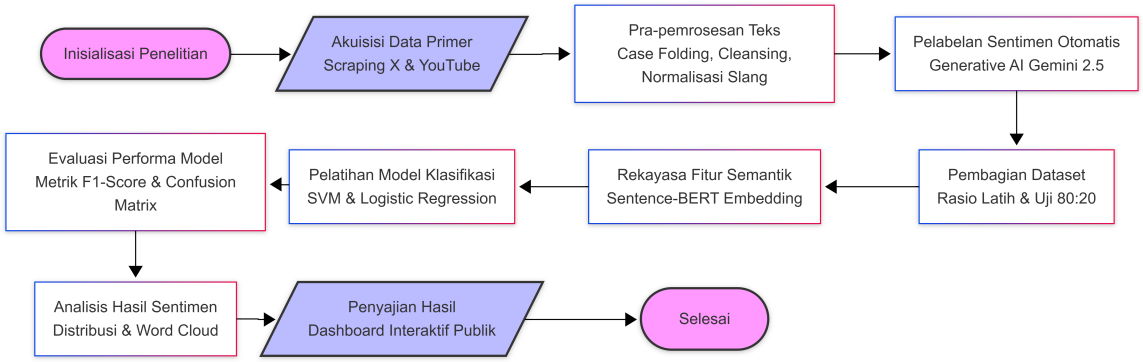


Figure 1: Diagram Alur Metodologi Penelitian.

2.1 *Data Acquisition*

Data yang digunakan dalam penelitian ini merupakan data primer berupa opini publik dalam format teks. Data diambil dari dua platform media sosial utama, yaitu X (sebelumnya Twitter) dan YouTube, yang dikenal sebagai platform aktif untuk diskusi publik dan pembentukan opini terkait pasar aset kripto. Proses akuisisi data dilakukan dengan teknik *web scraping* menggunakan *library* Python (seperti *twikit* untuk X dan YouTube Data API v3 untuk YouTube) secara *batch*

untuk mengumpulkan data historis. Rentang waktu pengambilan data difokuskan pada periode 11 bulan terakhir (Januari–November 2025) untuk menangkap sentimen publik terhadap fluktuasi pasar terkini. Kata kunci pencarian yang digunakan antara lain "Bitcoin", "Ethereum", "aset kripto", serta *slang* finansial dan kripto yang relevan seperti "cuan", "*fud*", "boncos", "RDT", dan "sangkut".

Pada platform YouTube, pengumpulan data dilakukan menggunakan YouTube Data API v3 dengan menargetkan 22 video yang setiap 2 video mewakili satu bulan dari Januari hingga November 2025. Komentar dari setiap video diambil berdasarkan relevansi, dengan batasan maksimal 350 komentar per video, sehingga terkumpul total 5.544 komentar. Untuk platform X, proses pencarian data menerapkan strategi kata kunci yang komprehensif mencakup istilah aset utama serta terminologi *slang* yang populer di komunitas kripto. Pengumpulan data dilakukan secara periodik dengan batasan kuota maksimal 350 *tweet* untuk setiap bulannya.

2.2 *Text Pre-processing*

Data teks yang diperoleh dari media sosial bersifat tidak terstruktur (*unstructured*) dan mengandung banyak *noise*. Tahapan *pre-processing* dilakukan untuk membersihkan dan menstandarisasi data sebelum masuk ke tahap *feature engineering*. Berbeda dengan pendekatan *NLP* klasik, *pre-processing* untuk model *transformer* (seperti *SBERT*) difokuskan pada pembersihan *noise* sambil mempertahankan konteks kalimat.

Tahapan *pre-processing* yang dilakukan meliputi *case folding* untuk menyeragamkan seluruh teks ke format huruf kecil (*lowercase*), *data cleansing* untuk mengeliminasi elemen *noise* seperti *username*, *hashtag*, *URL*, *emoji*, serta karakter non-alfanumerik, dan normalisasi teks. Tahap normalisasi ini krusial untuk menangani data informal, di mana sebuah kamus normalisasi (*slang dictionary*) dibangun secara manual untuk mengubah kata-kata tidak baku, singkatan, dan *slang* ke dalam bentuk baku Bahasa Indonesia. Berbagai teknik normalisasi teks telah dikaji secara sistematis dalam literatur sebelumnya (Aliero et al., 2023). Namun, teknik *stopword removal* dan *stemming* tidak diterapkan dalam penelitian ini karena model *SBERT* yang digunakan bekerja secara kontekstual dan membutuhkan struktur kalimat yang relatif utuh, termasuk *stopwords*, untuk memahami makna dan menghasilkan representasi vektor yang akurat (Pota et al., 2021).

2.3 *Feature Engineering*

Setelah teks dibersihkan, tahapan *feature engineering* dilakukan untuk mengubah data teks kualitatif menjadi representasi vektor numerik yang padat (*dense vector*). Penelitian ini menggunakan model *embedding* kontekstual berbasis *transformer*, yaitu *Sentence-BERT* (*SBERT*), dengan memanfaatkan model *pre-trained* yang telah dioptimalkan untuk Bahasa Indonesia, yaitu indobenchmark/indobert-base-p1. *SBERT* merupakan adaptasi dari arsitektur *BERT* yang dioptimasi untuk menghasilkan representasi kalimat yang bermakna secara semantik melalui jaringan *siamese* (Reimers and Gurevych, 2019).

Berbeda dengan *embedding* statis (*Word2Vec*) atau *sparse vector* (*TF-IDF*), *SBERT* menghasilkan satu vektor untuk keseluruhan kalimat dengan memperhitungkan konteks setiap kata di dalamnya. Dengan demikian, setiap dokumen teks dalam korpus ditransformasikan menjadi sebuah vektor berdimensi 768 yang kemudian digunakan sebagai fitur masukan untuk model klasifikasi. Implementasi dilakukan dengan memuat model *pre-trained* tersebut menggunakan *library sentence-transformers*. Proses *embedding* kemudian dijalankan secara terpisah pada data latih (X_train) dan data uji (X_test).

2.4 *Sentiment Labeling*

Proses pelabelan sentimen dilakukan secara menyeluruh terhadap total 7.461 baris data yang telah dikumpulkan dari platform X dan YouTube pada periode Januari hingga November 2025. Penelitian ini menerapkan metode anotasi otomatis (*automated labeling*) dengan memanfaatkan kecerdasan buatan, yaitu model Gemini 2.5. Penerapan Gemini 2.5 untuk melabeli keseluruhan dataset bertujuan untuk menjaga konsistensi standar penilaian pada setiap baris data dan meminimalisir bias subjektivitas antar-anotator (*inter-annotator bias*).

2.5 *Classification Models*

Penelitian ini menggunakan pendekatan *supervised learning* untuk klasifikasi sentimen. Dua model *machine learning* klasik dipilih untuk dilatih dan dievaluasi kinerjanya berdasarkan fitur *embedding* yang dihasilkan oleh *SBERT*, yaitu *Logistic Regression* dan *Support Vector Machine* (*SVM*).

Logistic Regression dipilih karena efisiensinya secara komputasi dan performanya yang baik dalam menangani data *dense vector* berdimensi tinggi (Agresti, 2019). Secara matematis, probabilitas prediksi kelas k untuk vektor masukan $\mathbf{x}$ pada regresi logistik multinomial dirumuskan menggunakan fungsi *Softmax* sebagai berikut:

$$P(y = k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x} + b_k)}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x} + b_j)} \tag{1}$$

di mana $\mathbf{w}_k$ adalah vektor bobot, b_k adalah bias untuk kelas k, dan K adalah total jumlah kelas sentimen. Parameter *class_ weight='balanced'* diterapkan untuk mengatasi ketidakseimbangan jumlah data antar kelas, dengan jumlah iterasi maksimum (*max_ iter*) diatur pada 1000. Implementasi regresi logistik untuk klasifikasi berita di Indonesia telah menunjukkan hasil yang menjanjikan dalam literatur terdahulu (Fahmuddin et al., 2023).

SVM dipilih karena dikenal kuat dalam menemukan *hyperplane* optimal yang memisahkan kelas-kelas data, menjadikannya sangat efektif untuk tugas klasifikasi di ruang fitur yang kompleks seperti *embedding SBERT* (Cortes and Vapnik, 1995). Fungsi keputusan klasifikasi *SVM* dirumuskan sebagai:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{2}$$

Model ini menggunakan kernel *Radial Basis Function* (*kernel='rbf'*) yang diformulasikan sebagai $K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}||^2)$ untuk menangkap hubungan non-linear pada data vektor. Parameter *class_ weight='balanced'* digunakan untuk menangani *imbalance* data, dan *probability=True* diaktifkan agar model dapat menghasilkan estimasi probabilitas.

Selain kedua model tersebut, diterapkan juga teknik *Ensemble Learning* menggunakan metode *Voting Classifier* dengan mekanisme *Soft Voting* (*voting='soft'*), di mana keputusan akhir klasifikasi didasarkan pada rata-rata probabilitas yang diprediksi oleh kedua model. Pendekatan ensemble untuk klasifikasi *multiclass* telah terbukti efektif dalam berbagai domain aplikasi (Fu et al., 2021; Zainottah et al., 2025).

2.6 *Evaluation Metrics*

Untuk mengevaluasi performa dan menentukan model klasifikasi terbaik, digunakan beberapa metrik evaluasi standar yang berbasis pada *Confusion Matrix* (Grandini et al., 2020), yang

meliputi *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Keempat metrik tersebut dihitung menggunakan persamaan matematika berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

F1-Score dijadikan acuan utama dalam penentuan model terbaik, karena kemampuannya memberikan gambaran yang seimbang antara *Precision* dan *Recall*, yang sangat penting terutama jika distribusi data sentimen tidak seimbang (*imbalanced*). Dataset yang telah melalui tahap pra-pemrosesan dan pelabelan selanjutnya dibagi menjadi data latih (*training set*) dan data uji (*testing set*) dengan rasio 80:20.

3 *Results and Discussion*

3.1 *Data Acquisition Results*

Akuisisi data dilakukan dari dua platform utama, yaitu YouTube dan X (Twitter), dengan tujuan memperoleh pandangan publik terkait topik *bitcoin* dan *crypto* sepanjang tahun 2025. Pada platform YouTube, terkumpul total 5.544 komentar dari 22 video. Untuk platform X, pengumpulan data dilakukan secara periodik dalam rentang waktu Januari hingga November 2025 dengan batasan kuota yang konsisten maksimal 350 *tweet* untuk setiap bulannya. Total keseluruhan data yang terkumpul dari kedua platform adalah 7.461 baris data.

3.2 *Pre-processing Results*

Dataset awal yang dihasilkan dari proses *scraping* masih mengandung berbagai elemen yang berpotensi mengganggu, seperti *emoji*, karakter khusus, penggunaan *slang words*, serta *typo*. Setelah dilakukan tahapan *case folding*, penghapusan *URL*, *mention*, *hashtag*, *timestamp*, penggantian *emoji* menjadi bentuk teks, pembersihan *punctuation*, normalisasi *whitespace*, filter panjang kata (5–25 kata), dan normalisasi *slang words* menggunakan kamus khusus untuk konteks *crypto* dan bahasa percakapan Indonesia, data teks menjadi lebih bersih dan siap untuk dilakukan analisis lebih lanjut. Untuk memberikan gambaran yang jelas mengenai transformasi data, Tabel 1 menyajikan contoh ilustrasi perubahan teks opini sebelum dan sesudah melalui proses *text pre-processing* NLP.

Selain proses pembersihan teks, tahapan pelabelan otomatis menggunakan Gemini 2.5 juga menghasilkan anotasi sentimen yang konsisten secara kontekstual. Contoh representatif dari hasil pelabelan data untuk masing-masing kelas sentimen (Positif, Negatif, dan Netral) ditampilkan pada Tabel 2.

Table 1: Contoh Ilustrasi Hasil Pre-processing Teks.

No.	**Teks Mentah (*Raw Text*)**	**Hasil *Pre-processing***
1	rungkad mulu di kripto karena takut ketinggalan kenapa jiwa takut ketinggalan ku selalu bergejolak setiap lihat coin tren	rungkad mulu kripto takut ketinggalan jiwa takut ketinggalan ku selalu bergejolak setiap lihat coin tren
2	yang saya lihat dari kemarin chart beberapa koin sudah oversold kalau fomc turun dipause malem ini apa bakal jadi pemicunya altcoin season awal february	lihat kemarin beberapa koin oversold fomc turun dipause malem bakal jadi pemicunya altcoin season awal february
3	saham us harga turun buruk yakali market kripto bakalan di bikin harga turun juga	saham us harga turun buruk yakali market kripto bakalan bikin harga turun

Table 2: Contoh Hasil Pelabelan Sentimen Dataset.

No.	**Teks Hasil *Pre-processing***	**Label Sentimen**
1	wis menang lur isih do denial bitcoiner saiki lagi do asyik menikmati tontonan wong wong intitusi negara bankir do lur one way ticket lur	Positif
2	saham us harga turun buruk yakali market kripto bakalan bikin harga turun	Negatif
3	lihat kemarin beberapa koin oversold fomc turun dipause malem bakal jadi pemicunya altcoin season awal february	Netral

3.3 *Model Evaluation Results*

Setelah proses pelatihan selesai, kinerja ketiga model (*Logistic Regression*, *SVM*, dan *Voting Classifier*) dievaluasi menggunakan data uji (*X_ test_ emb*) yang tidak dilihat oleh model selama proses pelatihan. Tabel 3 menunjukkan perbandingan hasil evaluasi dari ketiga model yang diuji.

Table 3: Perbandingan Performa Model Klasifikasi.

Model	*Logistic Regression*	*SVM*	*Ensemble*
Accuracy	0.71	**0.73**	**0.73**
Precision	0.70	**0.73**	0.72
Recall	0.71	**0.73**	0.72
F1-Score	0.70	**0.73**	0.72

Berdasarkan hasil evaluasi pada Tabel 3, model *Support Vector Machine* (*SVM*) menunjukkan kinerja paling superior dibandingkan model lainnya dengan mencatatkan akurasi sebesar 0,73 dan *F1-Score* 0,73. Hal ini mengindikasikan bahwa algoritma *SVM* mampu menangani kompleksitas data vektor hasil *embedding SBERT* secara efektif dalam memisahkan kelas sentimen daripada *Logistic Regression*. Sebaliknya, model *Logistic Regression* mencatatkan performa

terendah dengan *F1-Score* 0,70, yang menandakan keterbatasannya dalam menangkap pola data yang kompleks.

Menariknya, metode *Ensemble* (*Voting Classifier*) justru menghasilkan performa menengah (*F1-Score* 0,72), di mana penggabungan dengan *Logistic Regression* justru menurunkan potensi akurasi maksimal yang bisa dicapai oleh *SVM* secara tunggal. Hal ini disebabkan oleh model *Logistic Regression* memberikan probabilitas yang *overconfident*, sehingga menarik turun rata-rata probabilitas model *SVM* yang benar. Oleh karena itu, *SVM* ditetapkan sebagai metode terbaik untuk digunakan pada tahap analisis selanjutnya.

Keunggulan performa *SVM* secara tunggal dalam penelitian ini menunjukkan konsistensi dengan temuan Bhowmik et al. (2025), yang menegaskan bahwa karakteristik algoritma *SVM* sangat tangguh dan adaptif dalam memetakan batas keputusan pemisah (*decision boundary*) pada ruang fitur bermutu tinggi untuk domain opini pasar kripto. Konfigurasi kernel *Radial Basis Function* (RBF) yang diterapkan terbukti mampu mentransformasikan fitur *SBERT* berdimensi 768 ke ruang dimensi yang lebih tinggi secara optimal, sehingga pemisahan kelas sentimen yang rumit dapat dicapai dengan galat minimal.

Namun, temuan mengenai penurunan performa pada metode *Ensemble* menyajikan anomali dan perbedaan yang kontras jika dikomparasikan dengan studi oleh Zainottah et al. (2025). Dalam studi tersebut, integrasi ensemble *SVM-Logistic* berhasil mendongkrak akurasi karena basis ekstraksi fiturnya menggunakan *TF-IDF* yang bersifat jarang (*sparse vector*). Kontribusi ilmiah unik (*novelty*) dari penelitian ini membuktikan bahwa ketika berhadapan dengan vektor padat kontekstual (*dense contextual embedding*) besutan *SBERT*, fungsi *Softmax* regresi logistik menghasilkan estimasi probabilitas multiclass yang terlalu berani (*overconfident*). Akibatnya, mekanisme *soft voting* justru mengonstruksi bias baru yang mereduksi ketajaman prediksi *hyperplane* milik *SVM*. Narasi komparatif ini menjadi landasan penting untuk menjembatani transisi dari pembuktian matematis model menuju pembahasan substantif mengenai bagaimana model *SVM* pilihan ini mengekstraksi dan memetakan dinamika psikologis kelompok investor secara riil di lapangan.

3.4 *Sentiment Distribution Analysis*

Berdasarkan hasil pelabelan sentimen terhadap keseluruhan dataset yang digabungkan dari platform X dan YouTube, diperoleh distribusi sentimen seperti ditunjukkan pada Gambar 2. Hasil analisis menunjukkan bahwa sentimen netral merupakan kelas yang paling dominan dengan jumlah sebanyak 3.593 data, diikuti oleh sentimen negatif sebanyak 1.991 data dan sentimen positif sebanyak 1.877 data.

Dominasi sentimen netral yang sangat signifikan dalam penelitian ini mengungkap karakteristik penting dari diskursus aset kripto di Indonesia. Faktor utama penyebab tingginya sentimen netral adalah sifat dari interaksi pengguna di platform X dan YouTube yang didominasi oleh penyebaran informasi faktual, seperti pembaruan harga harian (*price action*), analisis teknikal grafik tanpa opini emosional, serta penyebaran tautan terkait proyek *airdrop* maupun jadwal *token unlock*. Karakteristik data ini menunjukkan bahwa media sosial tidak hanya berfungsi sebagai wadah peluapan emosi investor retail, melainkan telah bertransformasi menjadi infrastruktur distribusi informasi dan edukasi ekonomi digital yang masif.

Implikasi dari fenomena ini terhadap pasar finansial kripto mengindikasikan bahwa mayoritas komunitas investor di Indonesia sepanjang tahun 2025 berada dalam fase pengamatan yang rasional (*wait-and-see*). Investor cenderung bersikap objektif dalam menyerap informasi pasar sebelum mengambil keputusan transaksi. Namun, fakta bahwa proporsi sentimen negatif (1.991

data) sedikit lebih tinggi dibandingkan sentimen positif (1.877 data) merefleksikan adanya sensitivitas dan kecemasan laten di kalangan publik. Hal ini dipicu oleh tingginya volatilitas harga aset utama seperti Bitcoin, risiko kerugian finansial yang nyata, serta ketidakpastian regulasi makroekonomi global (seperti kebijakan suku bunga FOMC).

Jika dibandingkan dengan penelitian terdahulu, temuan dalam studi ini menyajikan kontribusi komparatif yang menarik. Hasil ini berbeda secara kontras dengan studi oleh Hassan et al. (2022) yang melaporkan adanya dominasi sentimen positif sebesar 33% pada data *tweet* kripto. Perbedaan mendasar ini disebabkan oleh faktor pemosisian waktu pengambilan data; penelitian Hassan dilakukan pada periode *bull-market* di mana euforia pasar sedang tinggi, sedangkan penelitian ini mengambil rentang waktu yang lebih luas dan dinamis sepanjang tahun 2025, sehingga berhasil menangkap respons publik yang lebih realistis dan berhati-hati saat pasar mengalami koreksi.

Di sisi lain, tingginya volume data netral dan negatif ini mendukung argumen dari Bhowmik et al. (2025) yang menyatakan bahwa dinamika volatilitas kripto sangat sensitif terhadap perubahan narasi mikro di media sosial. Pola pergeseran dari sentimen netral menuju negatif seringkali menjadi indikator awal terjadinya tekanan jual di pasar nyata. Dengan demikian, visualisasi distribusi ini mempertegas bahwa penggabungan data dari dua platform (X dan YouTube) mampu memberikan gambaran makro yang jauh lebih komprehensif mengenai kondisi psikologis pasar yang sebenarnya dibandingkan hanya berfokus pada analisis satu platform tunggal.

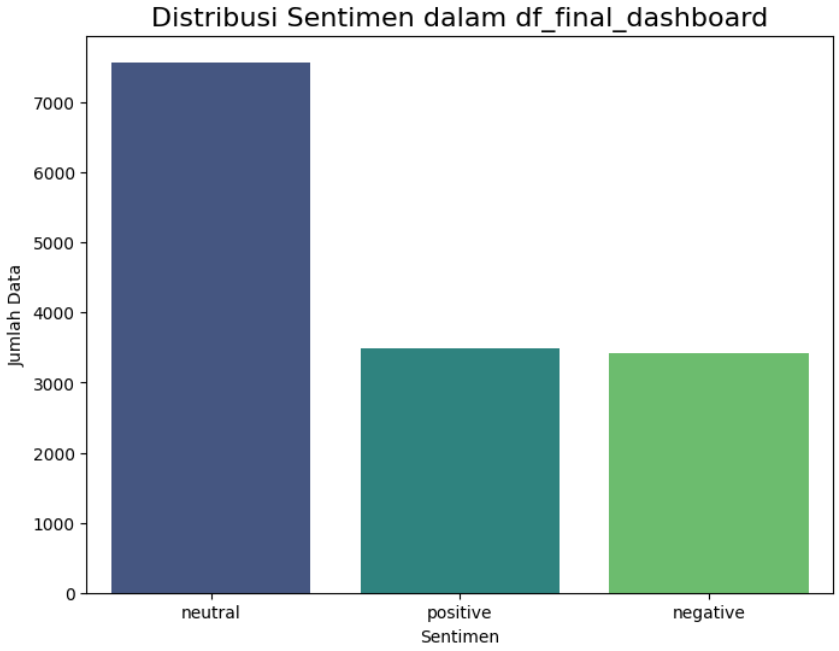
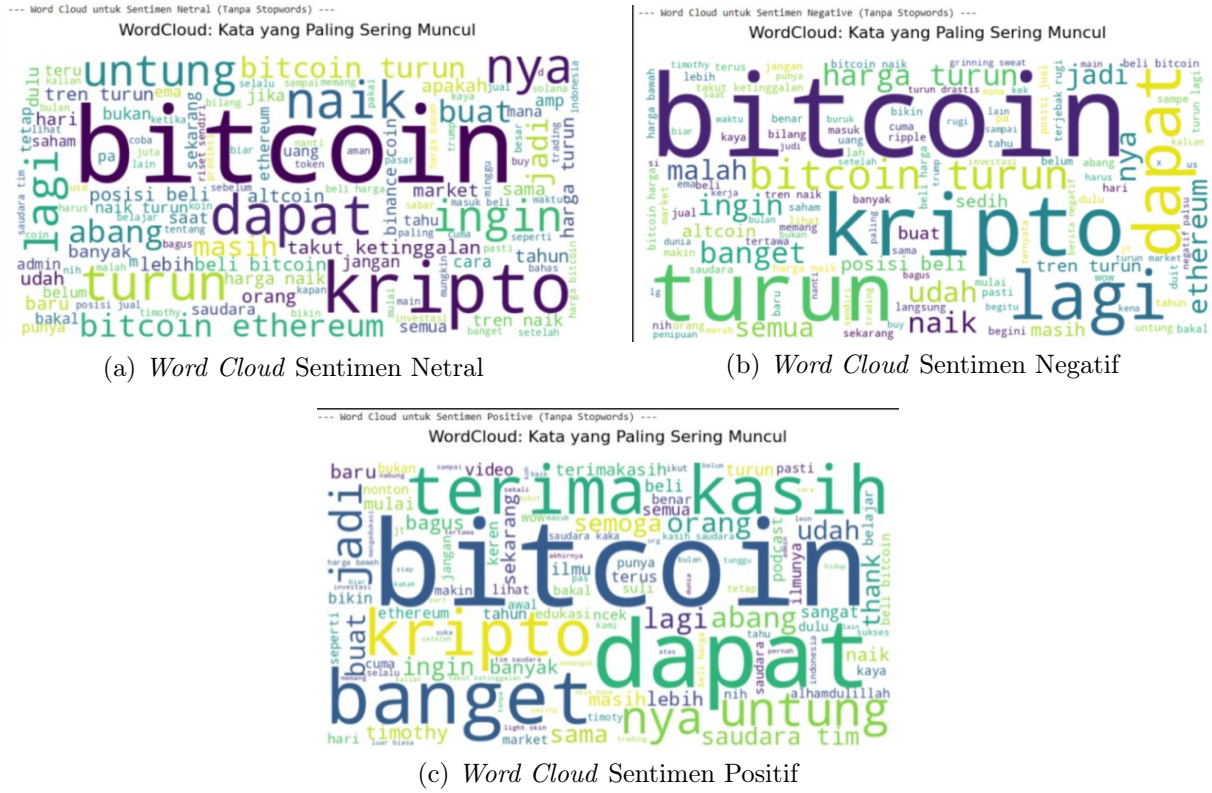


Figure 2: *Bar Chart* Distribusi Sentimen.

3.5 *Word Cloud Visualization*



(a) *Word Cloud* Sentimen Netral

(b) *Word Cloud* Sentimen Negatif

(c) *Word Cloud* Sentimen Positif

Figure 3: *Word Cloud* untuk masing-masing kelas sentimen.

Untuk memahami karakteristik bahasa dan kata kunci yang dominan pada setiap kelas sentimen, dilakukan visualisasi *word cloud* secara terpisah untuk sentimen netral, negatif, dan positif.

Berdasarkan *word cloud* sentimen netral, kata-kata yang paling dominan adalah "*bitcoin*", "kripto", "naik", "turun", "harga", "*market*", serta nama aset seperti "*ethereum*". Dominasi istilah-istilah tersebut menunjukkan bahwa percakapan dalam kelas sentimen netral sebagian besar berfokus pada penyampaian informasi faktual mengenai pergerakan harga dan kondisi pasar kripto tanpa disertai ekspresi emosional yang kuat.

Pada *word cloud* sentimen negatif, kata-kata seperti "turun", "harga", "*bitcoin*", "rugi", dan "sedih" tampil dengan ukuran yang lebih dominan. Hal ini menunjukkan bahwa sentimen negatif dalam dataset sangat erat kaitannya dengan penurunan harga aset kripto dan kerugian finansial yang dialami pengguna. Kehadiran kata seperti "jual", "beli", dan "posisi" mengindikasikan bahwa emosi negatif sering muncul dalam konteks keputusan transaksi yang dianggap tidak menguntungkan.

Word cloud sentimen positif didominasi oleh kata-kata seperti "untung", "dapat", "banget", "naik", serta ekspresi apresiatif seperti "terima kasih" dan "alhamdulillah". Dominasi kata-kata tersebut menunjukkan bahwa sentimen positif terutama dipicu oleh pengalaman memperoleh keuntungan atau hasil yang dianggap memuaskan dari aktivitas investasi kripto. Selain itu, munculnya kata "semoga", "bagus", serta istilah terkait pembelajaran seperti "edukasi" dan "ilmu" mengindikasikan adanya optimisme dan apresiasi terhadap informasi yang bermanfaat.

3.6 *Dashboard Implementation*

Hasil dari *labeling* sentimen selanjutnya divisualisasikan dalam bentuk dasbor yang representatif. Digabungkan dengan informasi-informasi pendukung lainnya, dasbor ini menampilkan informasi yang dapat digunakan oleh investor, masyarakat, atau entitas lainnya untuk memantau perkembangan baik historikal maupun terbaru dari aset kripto.

Dasbor dibangun menggunakan modul *Uvicorn* dan layanan Cloudflare Zero Trust Tunnel untuk menghubungkan domain ke server. Server yang digunakan adalah dari penyedia layanan Tencent Cloud dengan spesifikasi 2 *vCPU* dan 2GB *RAM*. Informasi yang ditampilkan di dasbor di antaranya adalah sentimen dari X dan YouTube secara global, *word clouds* tiap sentimen, juga *bar chart* dari masing-masing sentimen yang dapat difilter berdasarkan tanggal mulai, tanggal selesai, juga platform yang digunakan.

Selain itu ada juga informasi lainnya seperti *tweet* terbaru dari *Key Opinion Leader* aset kripto yang diperbarui secara *realtime*, berita terbaru aset kripto yang diambil dari *API* CoinDesk, artikel terbaru aset kripto dari DropsTab, event *Airdrop* yang akan datang dari DropsTab, event *Unlock Token* yang akan datang dari DropsTab, juga harga aset kripto dengan kapitalisasi pasar tertinggi dari *API* CoinDesk.

4 Kesimpulan

Berdasarkan hasil analisis dan evaluasi yang telah dilakukan, penelitian ini berhasil mengimplementasikan sistem analisis sentimen terhadap opini publik mengenai aset kripto di media sosial X dan YouTube secara menyeluruh. Proses pengolahan data mencakup akuisisi data, prapemrosesan teks, pelabelan otomatis, ekstraksi fitur, hingga pelatihan dan evaluasi model klasifikasi sentimen.

Hasil evaluasi menunjukkan bahwa algoritma *Support Vector Machine* (*SVM*) memberikan performa terbaik dengan nilai *F1-Score* sebesar 0,73, mengungguli *Logistic Regression* yang memperoleh *F1-Score* sebesar 0,70. Hal ini menunjukkan bahwa *SVM* lebih efektif dalam menangani representasi fitur berbasis *embedding SBERT* yang memiliki kompleksitas dan dimensi tinggi.

Analisis distribusi sentimen mengungkapkan bahwa opini publik terhadap aset kripto didominasi oleh sentimen netral sebanyak 3.593 data, yang umumnya bersifat informatif dan deskriptif. Selain itu, sentimen negatif (1.991 data) ditemukan lebih tinggi dibandingkan sentimen positif (1.877 data), yang mengindikasikan adanya respons publik yang cenderung reaktif terhadap penurunan harga, risiko kerugian, serta ketidakpastian pasar kripto.

Penelitian ini juga berhasil mengintegrasikan metode *Sentence-BERT* (*SBERT*) untuk ekstraksi fitur dan Gemini 2.5 untuk proses pelabelan sentimen ke dalam sebuah *dashboard* analisis sentimen *real-time* yang dapat diakses oleh publik. Integrasi ini menunjukkan bahwa pendekatan yang digunakan tidak hanya efektif dari sisi pemodelan, tetapi juga aplikatif dalam menyajikan hasil analisis sentimen secara interaktif.

Untuk pengembangan penelitian selanjutnya, disarankan agar cakupan sumber data diperluas dengan tidak hanya terbatas pada platform YouTube dan X, tetapi juga mencakup platform media sosial lain guna memperoleh gambaran opini publik yang lebih komprehensif. Selain itu, penelitian lanjutan dapat mengeksplorasi penggunaan model berbasis *deep learning*, seperti arsitektur *transformer* atau model *neural network* lainnya, guna meningkatkan performa klasifikasi sentimen serta menangkap pola linguistik yang lebih kompleks dalam data teks.

Ucapan Terima Kasih

Penelitian ini didukung oleh Program Studi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, dan Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta. Penulis juga menyampaikan terima kasih kepada Laboratorium Analisis Big Data, Sains Data Universitas Negeri Surabaya yang telah memberikan dukungan, bantuan, dan kontribusi dalam pelaksanaan penelitian ini.

References

Agresti, A. (2019). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, USA, 3rd edition.

Aliero, A. A. et al. (2023). Systematic review on text normalization techniques and approaches to non-standard words. *International Journal of Computer Applications*, 185:44–55.

Aslam, N. et al. (2022). Sentiment analysis and emotion detection on cryptocurrency-related tweets using ensemble lstm-gru model. *IEEE Access*, 10:39313–39324.

Bhowmik, P. K. et al. (2025). Ai-driven sentiment analysis for bitcoin market trends: A predictive approach to crypto volatility. *Journal of Ecohumanism*, 4(4):266–288.

Chen, Y., Wang, L., and Zhang, H. (2025). Enhancing sentiment classification of financial microblogs with sentence-level contextual embeddings. *Information Processing & Management*, 62(1):103820.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Fahmuddin, M., Aidid, M. K., and Taslim, M. J. (2023). Implementasi analisis regresi logistik dengan metode machine learning untuk mengklasifikasi berita di indonesia. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, 6(2):155–162.

Fu, L. et al. (2021). A machine learning based ensemble method for automatic multiclass classification of decisions. In *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Hassan, M. K. et al. (2022). Mining netizens' opinion on cryptocurrency: Sentiment analysis of twitter data. *Studies in Economics and Finance*, 39(3):365–385.

Lopez, J., Garcia, M., and Fernandez, R. (2023). Cryptocurrency market reaction to social media sentiment: A machine learning approach. *Finance Research Letters*, 52:103511.

Naseem, U. et al. (2024). Transformer-based deep learning models for sentiment analysis of cryptocurrency social media data. *Expert Systems with Applications*, 238:121689.

Nurbaiti, N. (2023). Characteristics of internet, smartphone, and social media usage among generation z in south jakarta after the covid-19 pandemic. *Journal of Health Sciences and Epidemiology*, 1(3):101–108.

Pota, M., Ventura, M., Fujita, H., and Esposito, M. (2021). Multilingual evaluation of preprocessing for bert-based sentiment analysis of tweets. *Expert Systems with Applications*, 181:115119.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sarkar, D. (2019). *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Apress, Berkeley, CA, 2nd edition.

Tamaraya, A. and Ubaedullah, D. (2021). Dampak penggunaan twitter terhadap pengungkapan diri mahasiswa. *Interaksi Peradaban: Jurnal Komunikasi dan Penyiaran Islam*, 1(1).

Yahya, S. and Wahyuni, S. (2024). Analisis sentimen publik terhadap pariwisata aceh di media sosial x menggunakan algoritma naive bayes classifier. *Bulletin of Information Technology (BIT)*, 5(4):269–278.

Zainottah, M. Z., Rengga, R. S., Yustian, Y. S., and Isa, I. R. (2025). Critical sentiment analysis of tokopedia electronic products using svm-logistic and tf-idf ensemble methods. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 4(3):2476–2482.