Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

Analisis Sentimen Reddit *WallStreetBets* Menggunakan Multinomial Naive Bayes pada Dinamika Pasar Saham

Rangga Wisnu Sena*[1], Agung Perdananto[2]
[1,2] Universitas Pamulang; Jl. Surya Kencana No.1, Pamulang Barat, (021) 741-2566 atau 7470 9855
[1,2]Jurusan Teknik Informatika, Fakultas Teknik, Universitas Pamulang

e-mail: [1]ranggawisnusena@gmail.com, [2]dosen00287@unpam.ac.id

Abstrak

Penelitian ini menyajikan investigasi komputasional mendalam mengenai pengembangan model analisis sentimen untuk mengklasifikasikan opini investor ritel pada platform Reddit, khususnya subreddit WallStreetBets. Fenomena social trading telah merombak lanskap pasar modal modern, di mana sentimen kolektif memiliki kapasitas untuk memicu volatilitas harga saham secara signifikan. Penelitian ini mengadopsi algoritma Multinomial Naive Bayes yang diperkuat dengan tahapan pra-pemrosesan data ekstensif, mencakup penanganan slang finansial, konversi emoji, serta normalisasi teks. Data dikumpulkan melalui Python Reddit API Wrapper (PRAW) dan diekstraksi menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Evaluasi empiris menunjukkan bahwa model mencapai akurasi keseluruhan sebesar 72%. Analisis granular menunjukkan presisi tinggi pada kelas positif (79%) dan negatif (86%), namun mengungkapkan kelemahan kritis pada recall kelas negatif yang hanya mencapai 46%. Hal ini dikarenakan keterbatasan asumsi independensi fitur dalam mendeteksi nuansa sarkasme dan ekspresi negatif implisit yang menjadi ciri khas kultur komunikasi komunitas tersebut. Temuan ini memberikan kontribusi pada pemahaman efektivitas metode bag-of-words dalam analisis teks finansial informal.

Kata kunci: Analisis Sentimen, Investasi Saham, Multinomial Naive Bayes, Reddit, WallStreetBets.

I. PENDAHULUAN

Transformasi digital telah menciptakan paradigma baru dalam dinamika pergerakan harga aset finansial melalui demokratisasi informasi di media sosial. Platform media sosial kini mendominasi sebagai sumber informasi utama yang membentuk opini publik, termasuk dalam konteks pengambilan keputusan investasi (Judijanto et al., 2023). Salah satu entitas yang paling menonjol dalam ekosistem ini adalah subreddit r/*WallStreetBets* (WSB), sebuah komunitas daring yang telah mendemonstrasikan kapasitas signifikan dalam memengaruhi volatilitas pasar melalui koordinasi kolektif investor ritel (Desiderio et al., 2024). Fenomena ini mencapai puncaknya pada peristiwa "short squeeze" saham GameStop (GME) awal tahun 2021, membuktikan bahwa sentimen media sosial merupakan variabel pasar fundamental yang berhubungan erat dengan volume perdagangan dan pengembalian saham (Bongini et al., 2025).

Dalam perspektif *Behavioral Finance*, media sosial memperkuat bias kognitif seperti herding behavior (perilaku ikut-ikutan) yang dapat memutus hubungan antara harga saham dengan nilai fundamentalnya (Sudaryati, 2021). Diskusi di forum seperti WSB memiliki kekuatan untuk memengaruhi likuiditas dan arah tren pasar, terutama pada saham-saham berkapitalisasi kecil yang menjadi target spekulasi investor ritel yang terkoneksi secara digital (Lizetha & Prawadika, 2021). Namun, menganalisis data tekstual dari WSB menghadirkan tantangan komputasional yang masif. Volume data yang sangat besar tidak memungkinkan untuk diproses secara manual dalam kerangka waktu yang relevan bagi pengambilan keputusan investasi (Wang et al., 2024).

Tantangan linguistik juga menjadi hambatan substantif bagi pendekatan *Natural Language Processing* (NLP) konvensional. Data media sosial bersifat tidak terstruktur, penuh dengan kesalahan tata bahasa, singkatan, dan variasi ejaan yang ekstrem (Permatasari et al., 2021). Penggunaan istilah slang

JATIMIKA
Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

finansial dan simbol visual seperti emoji secara intensif menciptakan hambatan semantik bagi model klasifikasi tradisional yang umumnya dilatih menggunakan korpus formal (Supriani et al., 2021). Selain itu, gaya bahasa yang dipenuhi sarkasme, ironi, dan humor yang merendahkan diri sendiri sering kali menyebabkan kesalahan klasifikasi fatal pada model yang tidak mempertimbangkan konteks unik komunitas digital (Sakthivel et al., 2025).

Beberapa penelitian terdahulu umumnya berfokus pada akurasi global tanpa memberikan perhatian mendalam pada elemen visual dan jargon spesifik yang mendominasi komunikasi di komunitas investor ritel (Permatasari et al., 2021). Sebagian besar studi menggunakan pembersihan data standar yang justru berisiko menghapus indikator sentimen penting seperti emoji atau istilah slang yang memiliki makna finansial tertentu. Kesenjangan inilah yang menjadi landasan bagi penelitian ini untuk menerapkan pendekatan yang lebih kontekstual dalam menangkap dinamika opini di platform media sosial yang bising.

Sisi pembeda utama dalam penelitian ini terletak pada pengolahan data yang dilakukan secara khusus untuk mengenali bahasa khas komunitas WallStreetBets. Prosedur yang diterapkan tidak hanya membersihkan teks secara umum, tetapi juga menerjemahkan arti emoji dan istilah unik seperti "tendies", "stonks", atau "diamond hands" agar pesan emosional investor tetap terjaga. Hal ini krusial karena dalam dunia social trading, satu simbol visual sering kali membawa makna yang lebih dalam dibandingkan teks formal (Sakthivel et al., 2025). Berbeda dengan pendekatan klasifikasi teks umum, penelitian ini memberikan kontribusi ilmiah melalui analisis terhadap efektivitas algoritma Multinomial Naive Bayes dalam menghadapi pola komunikasi investor ritel yang sangat dinamis.

Hasil analisis memberikan gambaran nyata bahwa algoritma Naive Bayes efektif dalam mendeteksi sentimen positif, namun memiliki tantangan besar dalam mengenali sindiran atau sarkasme. Hal ini terbukti dari temuan penelitian di mana sistem menghadapi kesulitan dalam menjaring seluruh komentar negatif secara tepat, yang ditunjukkan oleh nilai recall kelas negatif sebesar 46%. Poin penting yang dihasilkan adalah bahwa pemahaman terhadap budaya komunikasi komunitas sangat menentukan akurasi analisis di media sosial finansial, mengingat keterbatasan asumsi independensi fitur dalam menangkap konteks semantik yang kontradiktif (Permatasari et al., 2021). Temuan ini diharapkan menjadi masukan berharga bagi para investor dan pembuat kebijakan dalam memahami volatilitas pasar di era digital.

II. METODE PELAKSANAAN

Penelitian ini mengadopsi pendekatan kuantitatif berbasis text mining melalui kerangka kerja Knowledge Discovery in Databases (KDD) yang dipilih karena kemampuannya dalam menyediakan alur sistematis untuk mengekstraksi informasi berharga dari kumpulan data besar. Alur utama KDD dalam penelitian ini dimulai dengan tahap pemilihan data (data selection) dari komunitas r/*WallStreetBets*, yang kemudian dilanjutkan dengan fase pra-pemrosesan dan pembersihan (preprocessing & cleaning) untuk menangani tingkat noise yang tinggi pada data media sosial. Langkah ini menjadi krusial untuk memastikan kualitas data melalui konversi emoji dan normalisasi jargon finansial sebelum masuk ke tahap transformasi data menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Selanjutnya, tahap data mining dilakukan untuk mengekstraksi pola sentimen, yang diakhiri dengan evaluasi mendalam guna mengukur akurasi serta efektivitas model yang dikembangkan (Sakthivel et al., 2025).

Algoritma Multinomial Naive Bayes (MNB) dipilih sebagai metode utama klasifikasi karena efisiensinya yang sangat tinggi dalam menangani data teks berdimensi tinggi dengan distribusi frekuensi diskrit (Permatasari et al., 2021). Varian Multinomial dinilai sangat tepat digunakan bersama pembobotan TF-IDF karena mampu menghitung probabilitas berdasarkan jumlah kemunculan kata dalam dokumen, yang sangat krusial dalam menangkap bias atau sikap investor ritel. Selain itu, pemilihan MNB didasarkan pada ketangguhannya dalam memberikan kinerja yang optimal meskipun diterapkan pada dataset yang memiliki sebaran kelas tidak seimbang (imbalanced data) (Fauzi et al., 2023). Keunggulan ini menjadikannya solusi yang lebih efisien untuk memetakan sentimen media sosial yang bersifat dinamis dan bising dibandingkan dengan algoritma klasifikasi tradisional lainnya (Permatasari et al., 2021).

2.1. Analisis Kebutuhan Sistem

Untuk memastikan implementasi model berjalan dengan optimal, dilakukan identifikasi kebutuhan perangkat keras dan perangkat lunak yang spesifik. Penggunaan spesifikasi yang memadai sangat krusial mengingat besarnya volume data yang diambil dari platform Reddit.

Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

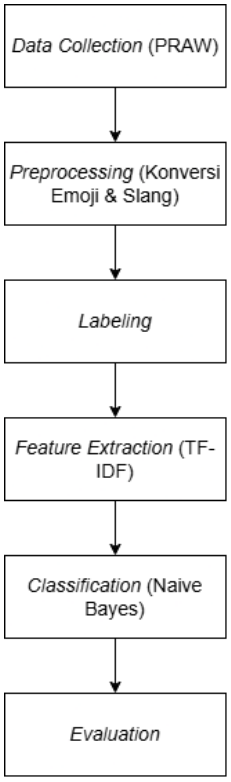
Tabel 1. Analisa Kebutuhan Perangkat Keras dan Perangkat Lunak

Perangkat Keras	**Spesifikasi**
Processor	Intel Core i5 atau setara
Penyimpanan	Minimum 256 GB SSD
RAM	Minimum 8GB
Koneksi Internet	Stabil (untuk pengambilan data Reddit)
Perangkat Lunak	**Spesifikasi**
Sistem Operasi	Windows 10
Aplikasi Simulator	Visual Studio Code
Bahasa Pemrograman	Python 3.x
Library Python	Pandas Re Nltk Matplotlib Sseaborn Scikit-learn Wordcloud

Secara teknis, sistem dibangun menggunakan bahasa pemrograman Python 3.8 dengan dukungan berbagai pustaka pendukung seperti Pandas untuk manipulasi data, Scikit-learn untuk implementasi algoritma machine learning, serta NLTK untuk pemrosesan bahasa alami.

2.2. Perancangan Penelitian

Rancangan penelitian disusun sebagai panduan operasional dalam melaksanakan setiap tahap eksperimen. Alur ini memastikan bahwa data mentah yang bising dari media sosial dapat ditransformasi menjadi informasi yang dapat diinterpretasikan oleh algoritma klasifikasi



Gambar 1. Alur Penelitian (*Flowchart*)

2.3. Pengumpulan Data dan Dataset

Data primer dalam penelitian ini diekstraksi langsung dari subreddit r/WallStreetBets menggunakan pustaka PRAW (Python Reddit API Wrapper). PRAW merupakan antarmuka pemrograman aplikasi resmi yang memungkinkan akses data publik Reddit melalui autentikasi berbasis OAuth2, sehingga menjamin validitas dan kepatuhan etika penelitian.

Dataset yang dikumpulkan mencakup periode aktivitas pasar yang signifikan antara Oktober hingga November 2024. Dari total data awal sebanyak 5.032 entri, setiap baris data merepresentasikan satu postingan yang mencakup atribut penting seperti ID unik, judul (title), isi badan teks (body), skor upvote, jumlah komentar, serta stempel waktu (timestamp)

Tabel 2. Rincian Dataset r/WallStreetBets

Judul Dataset	Reddit Posts from r/WallStreetBets (Custom Dataset)
Sumber	Reddit API - Subreddit: r/WallStreetBets
Deskripsi	Dataset ini berisi kumpulan postingan dari *subreddit* r/WallStreetBets yang berkaitan

JATIMIKA
Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

	dengan diskusi saham. Data ini digunakan untuk analisis sentimen guna mengetahui kecenderungan opini investor terhadap saham tertentu berdasarkan konten yang diposting.
Jumlah Sampel	5032
Jumlah Fitur	8 (id, title, score, url, comms_num, created, body, timestamp
Informasi Fitur	Id: ID unik postingan
	Title: judul atau isi postingan
	Score: jumlah upvote
	Url: Link postingan terkait
	Comms_num: Jumlah komentar postingan
	Created: Waktu pembuatan postingan dalam format asli
	Body: Isi badan postingan
	Timestamp: Waktu pembuatan postingan yang sudah diformat
Distribusi Kelas	Ditentukan melalui validasi manual (manual labeling). Setiap data diklasifikasikan secara mendalam oleh peneliti ke dalam kategori Positif, Negatif, dan Netral guna membentuk ground truth yang akurat sebagai acuan pembelajaran model (supervised learning).
Penggunaan	Dataset digunakan dalam penelitian ini sebagai dasar dalam membangun sistem analisis sentimen berbasis *machine learning* untuk mengklasifikasikan opini pengguna Reddit terhadap saham.
Sitasi Utama	Dataset ini tidak memiliki publikasi resmi karena dibuat sendiri melalui pengambilan data langsung dari Reddit menggunakan skrip Python.
Penelitian Lain	Belum ada karena merupakan dataset orisinal. Namun, pendekatan serupa telah dilakukan dalam penelitian-penelitian analisis sentimen berbasis media sosial.
Bukti Validitas	1. Data berasal dari platform real-time dengan komunitas aktif
	2. Diperoleh melalui API resmi
	3. Tidak mengandung data pribadi pengguna sehingga memenuhi prinsip etika penelitian

Distribusi Kelas	Belum tersedia; akan ditentukan melalui proses klasifikasi sentimen menjadi Positif, Negatif, dan Netral
Penggunaan	Dataset digunakan dalam penelitian ini sebagai dasar dalam membangun sistem analisis sentimen berbasis *machine learning* untuk mengklasifikasikan opini pengguna Reddit terhadap saham.
Sitasi Utama	Dataset ini tidak memiliki publikasi resmi karena dibuat sendiri melalui pengambilan data langsung dari Reddit menggunakan skrip Python.
Penelitian Lain	Belum ada karena merupakan dataset orisinal. Namun, pendekatan serupa telah dilakukan dalam penelitian-penelitian analisis sentimen berbasis media sosial.
Bukti Validitas	1. Data berasal dari platform real-time dengan komunitas aktif
	2. Diperoleh melalui API resmi
	3. Tidak mengandung data pribadi pengguna sehingga memenuhi prinsip etika penelitian

2.4. Pra-pemrosesan Teks

Tahap pra-pemrosesan merupakan fase kritis untuk menangani tingkat *noise* yang tinggi pada data media sosial (Kavanagh et al., 2023). Mengingat karakteristik unik komunitas WallStreetBets yang sangat bergantung pada jargon finansial dan ekspresi visual, diimplementasikan beberapa langkah khusus:

a. Konversi Emoji

Emoji seperti 🚀 atau 💎🙌 tidak dihilangkan, melainkan diubah menjadi representasi tekstual (misalnya "extremely bullish" atau "strong hold"). Hal ini dilakukan karena inklusi emoji terbukti secara signifikan meningkatkan akurasi model dalam menangkap sentimen emosional pengguna dibandingkan dengan mengeliminasinya (Doan et al., 2024).

Tabel 3. Sample Konversi Emoji

Teks Asli	Teks Hasil Konversi	Interpretasi Sentimen
🚀	extremely bullish	Optimisme Tinggi

JATIMIKA
Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

💎🙌	strong hold	Komitmen Investasi
🦍	retail investor	Soliaritas Komunitas
📉	price decrease	Sentimen Negatif
📈	price increase	Keyakinan Tren
🌕	huge increase	Euforia Spekulatif
💩	very negative	Penolakan Total
🔥	very positive	Antusiasme Masif

b. Normalisasi Slang

Menggunakan kamus normalisasi khusus yang disusun berdasarkan terminologi unik komunitas WSB. Istilah seperti "tendies", "stonks", dan "YOLO" dipetakan ke dalam bahasa standar guna menjaga preservasi semantik selama proses ekstraksi fitur (Kulmanov et al., 2025).

Tabel 4.Sampel Konversi Slang

Teks Asli	Teks Hasil Konversi	Interpretasi Sentimen
stonks	stocks	Finansial
tendies	profit gain money	Keuntungan
ath	all time high	Euforia Puncak
atl	all time low	Kapitulasi
fomo	fear of missing out	Kecemasan Impulsif
cooked	struggling	Kekalahan Telak
dip	price decrease	Spekulasi Peluang
squeeze	price manipulation	Agresi Kolektif

c. Pembersihan Dasar

Mencakup konversi teks ke huruf kecil (*case folding*), penghapusan URL, *mention* pengguna, karakter khusus, serta penghapusan angka dan referensi tanggal yang tidak mengandung makna sentimen (Kavanagh et al., 2023).

d. Penghapusan *Stopwords*

Menggunakan daftar gabungan dari pustaka NLTK dan daftar kata tambahan yang spesifik untuk domain finansial guna mengurangi dimensionalitas fitur (Kavanagh et al., 2023)

e. Tokenisasi dan Lemmatisasi

Teks dipecah menjadi unit-unit kata (token) dan kemudian diubah ke bentuk dasarnya (lemma) menggunakan WordNetLemmatizer yang dipadukan dengan Part-of-Speech (POS) tagging untuk memastikan akurasi kontekstual. (Setiadi et al., 2025).

2.5. Ekstraksi Fitur TF-IDF

Teks yang telah bersih ditransformasi menjadi representasi numerik menggunakan skema pembobotan Term Frequency-Inverse Document Frequency (TF-IDF). Metode ini mengevaluasi kepentingan relatif setiap kata dalam dokumen terhadap keseluruhan korpus (Yang et al., 2023). Bobot kata t dalam dokumen d dihitung dengan rumus

$$W_{t,d} = TF_{t,f} \times \log\left(\frac{N}{DF_t}\right)$$

Dimana $TF_{t,f}$ adalah frekuensi kemunculan kata dalam dokumen, N adalah total jumlah dokumen, dan $TF_{t,f}$ adalah jumlah dokumen yang mengandung kata t. Representasi ini memungkinkan model untuk memberikan bobot lebih pada istilah-istilah diskriminatif yang menjadi ciri khas sentimen tertentu.

2.6. Klasifikasi Multinomial Naïve Bayes

Algoritma Multinomial Naïve Bayes dipilih karena efisiensinya dalam menangani data teks berdimensi tinggi dengan distribusi frekuensi diskrit (Fauzi et al., 2023). Model bekerja berdasarkan Teorema Bayes yang menghitung probabilitas posterior kelas c (Positif, Negatif, Netral) diberikan dokumen d sebagai berikut

$$P(c|d) = \frac{P(c)\prod_{i=1}^{n} P(t_i|c)}{P(d)}$$

Karena penyebut $P(d)$ konstan untuk semua kelas, keputusan klasifikasi didasarkan pada pembilang tertinggi. Untuk menangani masalah zero probability (ketika kata baru muncul di data uji yang tidak ada di

JATIMIKA
Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

data latih), diterapkan teknik Laplace Smoothing ($\alpha = 1$) dalam estimasi probabilitas likelihood

$$\hat{P}(t|c) = \frac{N_{tc} + 1}{N_c + |V|}$$

Dimana N_{tc} adalah frekuensi kata t di kelas c, N_c adalah total kata di kelas c, dan $|V|$ adalah ukuran kosakata unik dalam korpus latih. Keputusan klasifikasi diambil berdasarkan probabilitas posterior tertinggi yang dihitung dalam ruang log-probabilitas untuk menghindari kesalahan numerik (*underflow*).

2.7. Evaluasi dan Skenario Pengujian

Untuk mengukur kinerja model secara objektif, dataset dibagi menjadi data latih (training set) sebanyak 80% dan data uji (testing set) sebanyak 20%. Evaluasi dilakukan menggunakan metrik standar industri sebagai berikut.

a. Akurasi

Mengukur rasio prediksi benar terhadap keseluruhan data.

b. Presisi

Mengukur ketepatan prediksi untuk setiap kelas sentimen.

c. Recall

Mengukur kemampuan model dalam menjaring kembali semua instansi dari kelas tertentu.

d. F1-Score

Memberikan gambaran keseimbangan antara presisi dan recall, terutama pada kasus sebaran kelas yang tidak seimbang.

III. HASIL DAN PEMBAHASAN

Implementasi model analisis sentimen pada komunitas WallStreetBets memberikan gambaran mendalam mengenai bagaimana arus informasi digital bertransformasi menjadi kekuatan pasar yang nyata. Hasil penelitian ini menunjukkan bahwa aktivitas diskusi daring bukan sekadar pertukaran opini, melainkan cerminan dari psikologi massa yang memiliki implikasi langsung terhadap dinamika pasar saham (Judijanto et al., 2023). Fenomena koordinasi kolektif investor ritel di media sosial terbukti dapat memutus hubungan antara harga aset dengan nilai fundamentalnya, sehingga menciptakan pola volatilitas baru yang didorong oleh sentimen emosional (Sudaryati, 2021). Melalui pemetaan sentimen positif, negatif, dan netral, penelitian ini mengungkap bagaimana persepsi publik dapat memengaruhi likuiditas dan arah tren harga saham secara signifikan, terutama pada saham-saham yang menjadi target spekulasi ritel (Desiderio et al., 2024).

3.1. Persiapan Data dan Karakteristik Dataset

Dataset yang berhasil diekstraksi melalui PRAW berjumlah 5.032 postingan mentah dari subreddit r/WallStreetBets. Setelah melalui tahap pembersihan data yang tidak memiliki teks (null) dan postingan yang dihapus oleh moderator, diperoleh 2.689 entri valid yang digunakan dalam eksperimen. Analisis statistik deskriptif menunjukkan variasi yang signifikan pada panjang teks, di mana beberapa postingan merupakan analisis fundamental yang sangat panjang sementara yang lain hanya berupa judul singkat berisi spekulasi.

Tabel 5. Statistik Deskriptif Karakteristik Data Teks

Atribut	Rata-rata	Minimum	Maksimum	Median	Standar Deviasi
Panjang Title (karakter)	49,14	3	00	41	34,65
Panjang Body (karakter)	1.095,73	1	29.703	509	1.757,99
Jumlah Kata (title)	8,93	1	60	7	6,24
Jumlah Kata (body)	177,22	0	4.591	83	278,55

3.2. Sampel Dokumen untuk Analisis

Untuk memberikan gambaran transparan mengenai mekanisme pembobotan TF-IDF dan cara kerja algoritma Multinomial Naive Bayes, dilakukan simulasi perhitungan menggunakan 4 dokumen sampel (D1-D4). Penting untuk dicatat bahwa pemilihan empat dokumen ini hanya berfungsi sebagai unit peraga atau ilustrasi matematis guna memudahkan pemahaman terhadap rumus yang diterapkan pada sistem, dan bukan

JATIMIKA
Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

merepresentasikan keseluruhan dataset yang berjumlah 2.689 entri tersebut.

Tabel 6. Sampel Dokumen untuk Simulasi Klasifikasi

Dokumen	Term (Kata)	Label
1	move tomorrow	Netral
2	weekly earnings thread	Netral
3	boeing raise billion share sale one large public company	Positif
4	israel strike military target iran	Negatif

3.3. Analisis Efektivitas Pra-pemrosesan Teks (Preprocessing)

Tahap pra-pemrosesan bertujuan mentransformasi teks informal menjadi data bersih. Dengan merujuk pada D3 (sampel Positif) dari Tabel 5, berikut adalah ilustrasi tahapan pembersihannya:

Tabel 7. Contoh Transformasi Preprocessing pada Dokumen D3

Tahapan	Output Teks / Hasil Transformasi	Keterangan Perubahan
Data Asli (Raw)	Boeing Raises $21 Billion in Share Sale, One of the Largest Ever by a Public Company	Data mentah hasil ekstraksi PRAW.
1. Konversi Emoji	Boeing Raises $21 Billion in Share Sale, One of the Largest Ever by a Public Company	Tidak ada perubahan karena tidak terdapat emoji pada sampel ini.
2. Case Folding	boeing raises $21 billion in share sale, one of the largest ever by a public company	Seluruh karakter diubah menjadi huruf kecil (*lowercase*).
3. Menghilangkan Tanggal	boeing raises $21 billion in share sale, one of the largest ever by a public company	Menghapus referensi waktu untuk mengurangi *noise*.
4. Normalisasi Slang	boeing raises $21 billion in share sale, one of the largest ever by a public company	Mengonversi jargon komunitas ke bahasa standar.
5. Hapus URL & Mention	boeing raises $21 billion in share sale, one of the largest ever by a public company	Menghapus tautan eksternal dan *username* Reddit.
6. Hapus Special Character	boeing raises 21 billion in share sale one of the largest ever by a public company	Menghapus tanda baca (koma) dan simbol mata uang ($).
7.Hapus Stopwords	boeing raises billion share sale one largest public company	Menghapus kata umum seperti "*in*", "*of*", "*the*", "*by*", "*a*".
8. Tokenisasi	['boeing', 'raises', 'billion', 'share', 'sale', 'one', 'largest', 'public', 'company']	Pemecahan kalimat menjadi unit kata (token).
9. Lematisasi	boeing raise billion share sale one large public company	Mengembalikan kata ke bentuk dasar (*raises* → *raise*, *largest* → *large*).

Penggunaan lematisasi pada D3 terbukti krusial karena mengubah "raises" menjadi "raise" dan "largest"

JATIMIKA
Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

menjadi "large", sehingga model dapat mengenali bentuk dasar kata tersebut di dokumen lain (Sakthivel et al., 2025).

3.4. Simulasi Perhitungan Manual Multinomial Naive Bayes

Setelah dokumen D1-D4 dibersihkan, dilakukan simulasi perhitungan untuk memvalidasi algoritma.

a. Pembobotan TF-IDF

Matriks ini menunjukkan bobot kepentingan kata terhadap masing-masing dokumen sampel (D1-D4).

Tabel 8. Matriks Final TF-IDF pada Dokumen Sampel

Kata (Term)	D1 (TF-IDF)	D2 (TF-IDF)	D3 (TF-IDF)	D4 (TF-IDF)
billion	0	0	0.6020	0
boeing	0	0	0.6020	0
company	0	0	0.6020	0
earnings	0	0.6020	0	0
iran	0	0	0	0.6020
israel	0	0	0	0.6020
large	0	0	0.6020	0
military	0	0	0	0.6020
move	0.6020	0	0	0
one	0	0	0.6020	0
public	0	0	0.6020	0
raise	0	0	0.6020	0
sale	0	0	0.6020	0
share	0	0	0.6020	0
strike	0	0	0	0.6020
target	0	0	0	0.6020
thread	0	0.6020	0	0
tomorrow	0.6020	0	0	0
weekly	0	0.6020	0	0

b. Probabilitas Likelihood

Tabel 9. Hasil Perhitungan Probabilitas Likelihood per Kelas

Kata (Term)	P(kata|Netral)	P(kata|Positif)	P(kata|Negatif)
billion	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
boeing	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
company	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
earnings	$\frac{2}{24} = (0.0833)$	$\frac{1}{28} = (0.0357)$	$\frac{1}{24} = (0.0417)$
iran	$\frac{1}{24} = (0.0417)$	$\frac{1}{28} = (0.0357)$	$\frac{2}{24} = (0.0833)$
israel	$\frac{1}{24} = (0.0417)$	$\frac{1}{28} = (0.0357)$	$\frac{2}{24} = (0.0833)$
large	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
military	$\frac{1}{24} = (0.0417)$	$\frac{1}{28} = (0.0357)$	$\frac{2}{24} = (0.0833)$
move	$\frac{2}{24} = (0.0833)$	$\frac{1}{28} = (0.0357)$	$\frac{1}{24} = (0.0417)$
one	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
public	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
raise	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
sale	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$
share	$\frac{1}{24} = (0.0417)$	$\frac{2}{28} = (0.0714)$	$\frac{1}{24} = (0.0417)$

Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:

strike	$\frac{1}{24} = (0.0417)$	$\frac{1}{28} = (0.0357)$	$\frac{2}{24} = (0.0833)$
target	$\frac{1}{24} = (0.0417)$	$\frac{1}{28} = (0.0357)$	$\frac{2}{24} = (0.0833)$
thread	$\frac{2}{24} = (0.0833)$	$\frac{1}{28} = (0.0357)$	$\frac{1}{24} = (0.0417)$
tomorrow	$\frac{2}{24} = (0.0833)$	$\frac{1}{28} = (0.0357)$	$\frac{1}{24} = (0.0417)$
weekly	$\frac{2}{24} = (0.0833)$	$\frac{1}{28} = (0.0357)$	$\frac{1}{24} = (0.0417)$

3.5. Hasil Evaluasi Model

Proses pengujian dilakukan dengan membagi dataset valid ke dalam data latih dan data uji. Berdasarkan hasil klasifikasi yang telah dilakukan terhadap data yang telah dilabeli secara manual, model menunjukkan performa akurasi sebesar 72%. Analisis lebih lanjut pada bagian ini akan membedah bagaimana model merespons data uji serta faktor-faktor yang memengaruhi nilai recall pada setiap kelas sentimen.

Tabel 10. Confusion Matrix

Kelas	Precision	Recall	F1-Score
Negatif	86%	46%	60%
Netral	67%	92%	78%
Positif	79%	64%	71%

Hasil evaluasi menunjukkan pola yang menarik di mana kelas negatif memiliki presisi tertinggi (86%). Hal ini berarti ketika model memprediksi suatu postingan sebagai negatif, prediksi tersebut memiliki probabilitas kebenaran yang sangat tinggi. Namun, terjadi bias sistematis terhadap kelas netral yang memiliki recall tertinggi (92%) akibat dominasi proporsi kelas tersebut dalam dataset pelatihan (Fauzi et al., 2023).

3.6. Analisis Kelemahan Recall Negatif (46%)

Pola misklasifikasi pada kelas negatif menunjukkan bahwa hampir separuh dari sentimen negatif gagal dideteksi dan justru dikategorikan sebagai netral. Analisis mendalam mengungkapkan bahwa hal ini disebabkan oleh karakteristik linguistik unik di WallStreetBets:

a. Sarkasme dan Ironi

Ekspresi negatif yang dibungkus dengan bahasa literal positif sering kali membingungkan model *machine learning* tradisional (Sakthivel et al., 2025). Contohnya, frasa "Great job losing my money" mengandung kata positif ("Great") yang mendominasi probabilitas kelas.

b. Asumsi Independensi Fitur

Keterbatasan Naive Bayes dalam menangkap konteks sekuensial menyebabkan model gagal mendeteksi kontradiksi semantik dalam kalimat kompleks (Alaramma et al., 2023).

c. Representasi TF-IDF

Metode TF-IDF tradisional tidak mempertimbangkan hubungan antar kata atau konteks semantik yang lebih dalam (Alammary, 2021).

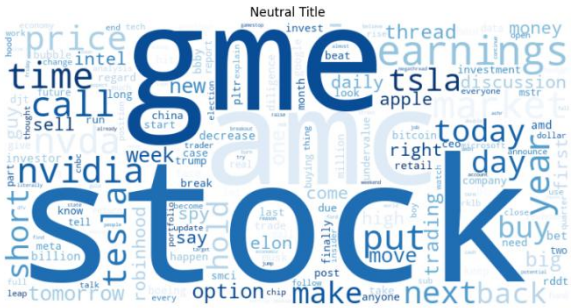
3.7. Visualisasi WordCloud

Visualisasi WordCloud memperjelas dominasi kata kunci pada masing-masing kategori sentimen setelah melalui tahap pra-pemrosesan.

Gambar 2. Visualisasi WordCloud Kelas Positif

Gambar 3. Visualisasi WordCloud Kelas Negatif

Jurnal Kreativitas Mahasiswa Informatika
Volume 7 Nomor 1 bulan April Tahun 2026 Page 86 - 96
p-ISSN: 2797-6327 e-ISSN:



Gambar 4. Visualisasi WordCloud Kelas Netral

Sentimen positif didominasi oleh kata *'gain', 'moon',* dan *'buy'*, yang menunjukkan ekspektasi keuntungan. Sebaliknya, sentimen negatif menonjolkan kata *'loss', 'down',* dan *'risk'*. Ukuran kata dalam visualisasi menggambarkan frekuensi kemunculannya, yang mengonfirmasi bahwa model paling efektif dalam mengenali sentimen yang diekspresikan secara eksplisit dan konsisten dalam dataset.

IV. SIMPULAN

Penelitian menyimpulkan bahwa algoritma Multinomial Naive Bayes dengan ekstraksi fitur TF-IDF efektif untuk konfirmasi sentimen pada data media sosial finansial dengan tingkat presisi 86% pada kelas negatif. Tahapan pra-pemrosesan khusus, khususnya konversi emoji dan normalisasi slang, terbukti vital dalam mempertahankan informasi sentimen yang berharga. Namun, model memiliki kelemahan signifikan dalam menjaring opini negatif implisit akibat bias probabilitas prior dan ketidakmampuan menangkap nuansa sarkasme. Untuk penelitian selanjutnya, disarankan penggunaan teknik resampling seperti SMOTE untuk menyeimbangkan distribusi kelas (Fauzi et al., 2023) serta penggunaan model Deep Learning berbasis Transformer guna menangkap dependensi linguistik yang lebih kompleks (Dhanta et al., 2023).

V. DAFTAR PUSTAKA

Alaramma, S., Habu, A., Ya'u, B., & Madaki, A. 2023. Sentiment analysis of sarcasm detection in social media. *Gadau Journal of Pure and Allied Sciences*, 2(1).

Alammary, A. 2021. Arabic Questions Classification Using Modified TF-IDF. *IEEE Access*, 9.

Bongini, P., Rossolini, M., Maurino, A., & Osborne, F. 2025. The Information Power of Social Media for Investment Decisions: An AI-Driven Analysis of Reddit Posts. Journal of Financial Management, Markets and Institutions. World Scientific Publishing.

Desiderio, A., Aiello, L. M., Cimini, G., & Alessandretti, L. 2024. The dynamics of the Reddit collective action leading to the GameStop short squeeze. *arXiv preprint arXiv:2401.14999*.

Dhanta, R., Sharma, H., Kumar, V., & Singh, H. 2023. Twitter sentimental analysis using machine learning. *International Journal of Communication and Information Technology*, 4(1).

Doan, M., Dam, M., Doan, T., Nguyen, H., & Nguyen, B. 2024. Sentiment Classification in Mobile Gaming Reviews: Customized Transformer Models with Emojis Retained. *Customized Transformer Models*.

Fauzi, F., Ismatullah, I., & Nur, I. 2023. Unbalanced multiclass classification with adaptive synthetic multinomial naive bayes approach. *Informatyka Automatyka Pomiary W Gospodarce I Ochronie Środowiska*, 13(3).

Judijanto, L., Maulinda, R., Zulaika, S., & Tjahyadi, I. 2023. Pengaruh Sumber Informasi dan Interaksi Sosial di Media Sosial terhadap Pembentukan Opini Politik Masyarakat di Indonesia. *Sanskara Ilmu Sosial dan Humaniora*, 1(01).

Kavanagh, J., Greenhow, K., & Jordanous, A. 2023. Assessing the Effects of Lemmatisation and Spell Checking on Sentiment Analysis of Online Reviews. *IEEE International Conference on Semantic Computing*.

Kulmanov, S., Fazylzhanova, A., Bissengali, A., & Omarova, A. 2025. Slang Terms in the Field of Information Technology and Their Standardization. *Forum for Linguistic Studies*, 7(6).

Lizetha, V., & Prawadika, A. 2021. Konvergensi Setengah Hati: Invasi Konten Media Sosial dalam Program Berita Televisi di TVOne. *Jurnal Media Dan Komunikasi*, 2(1).

Permatasari, P. A., Linawati, L., & Jasa, L. 2021. Survei Tentang Analisis Sentimen Pada Media Sosial. *Majalah Ilmiah Teknologi Elektro*, 20(2).

Peruka, B. 2025. Sentemonet: A Comprehensive Framework for Multimodal Sentiment Analysis from Text and Emotions. *Journal of Information Systems Engineering & Management*.

Sakthivel, S., et al. 2025. NLP-Based Extended Lexicon Model For Sarcasm Detection With Tweets And Emojis. GJEIIR, 5(2).

Setiadi, D., et al. 2025. Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow. *J. Comput. Theor. Appl.*, 2(4).

Sudaryati, D. 2021. Relevansi Nilai Informasi Akuntansi Terhadap Harga Saham. *AKUNTANSI DEWANTARA*, 4(2).

Supriani, I., Fianto, B., Fauziah, N., & Maulayati, R. 2021. Revisiting the Contribution of Islamic Banks' Financing to Economic Growth: The Indonesian Experience. *Shirkah Journal of Economics and Business*, 6(1).

Veruswati, M., et al. 2022. Does It Still Show a Deficit? Arguing Post-COVID-19 Health Financing System in Bogor, Indonesia. *Kesmas National Public Health Journal*, 17(sp1).

Wang, Z., Hao, S., Zwetsloot, I. M., & Trimborn, S. 2024. Social Network Datasets on Reddit Financial Discussion. *arXiv preprint arXiv:2410.05002*.

Yang, Z., Xia, Y., & Xu, H. 2023. TGCN-Bert Emoji Prediction in Information Systems Using TCN and GCN Fusing Features Based on BERT. *International Journal on Semantic Web and Information Systems*, 19(1).