Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 1 (2026) pp: 7346-7354
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Analisis Sentimen untuk Deteksi Penipuan pada Twitter Menggunakan TF-IDF dan Naive Bayes

Najwa Meyda, Muhammad Tafarel Akbar, Muhamad Nurdin, Septian Handita Surya, Alghozi Irsyadul Ibad, Ahmad Nursodiq
1Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang
najwameyda155@gmail.com, akbarfarel8202@gmail.com, nurdin081198@gmail.com, septiansin@gmail.com, alghozi748@gmail.com, dosen02526@unpam.ac.id

Abstrak

Penipuan digital melalui media sosial, khususnya Twitter, semakin marak dan menimbulkan kerugian bagi masyarakat. Penelitian ini bertujuan untuk menganalisis sentimen dan mendeteksi konten penipuan pada tweet berbahasa Indonesia menggunakan pendekatan pembelajaran mesin. Dataset diperoleh dari Twitter pada periode 2022–2025 dengan proses pengambilan data berdasarkan kata kunci terkait penipuan digital. Sebanyak 2.593 tweet digunakan setelah melalui proses pembersihan dan penghapusan duplikasi, dengan distribusi 1.829 tweet non-penipuan dan 764 tweet penipuan. Data kemudian dilabeli secara manual menjadi dua kelas, yaitu penipuan dan non-penipuan. Tahapan praproses meliputi pembersihan teks, tokenisasi, stemming menggunakan Sastrawi Stemmer, filtering stopword, dan normalisasi huruf. Ekstraksi fitur dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) dengan konfigurasi unigram dan bigram. Model klasifikasi utama dibangun menggunakan algoritma Multinomial Naive Bayes dengan penanganan ketidakseimbangan kelas melalui pembobotan sampel (sample weight). Sebagai pembanding, dilakukan pula eksperimen menggunakan algoritma Decision Tree dengan pengaturan class weight balanced. Hasil pengujian menunjukkan bahwa model Naive Bayes mencapai akurasi sebesar 87,09% dengan recall kelas penipuan sebesar 94,12%, yang menunjukkan kemampuan model dalam mendeteksi tweet penipuan secara efektif. Model Decision Tree menghasilkan akurasi lebih tinggi sebesar 96,53% dengan recall penipuan 92,81%. Analisis fitur menunjukkan bahwa kata-kata seperti "tipu", "scam", "modus", dan "deepfake" memiliki pengaruh kuat dalam klasifikasi penipuan. Penelitian ini diharapkan dapat menjadi dasar pengembangan sistem deteksi penipuan berbasis teks pada media sosial.

Kata kunci: Penipuan Digital, Analisis Sentimen, Twitter, TF-IDF, Naive Bayes

1. Latar Belakang

Perkembangan media sosial telah mengubah cara masyarakat berkomunikasi dan berbagi informasi secara cepat dan luas. Twitter sebagai salah satu platform media sosial populer memungkinkan pengguna untuk menyampaikan opini, pengalaman, dan informasi secara real time. Namun, kemudahan tersebut juga dimanfaatkan oleh pihak tidak bertanggung jawab untuk melakukan berbagai bentuk penipuan digital, seperti penipuan transaksi daring, akun palsu, dan modus phishing yang menyasar pengguna secara masif.

Penipuan yang tersebar melalui Twitter sering kali sulit dikenali karena disamarkan dalam bentuk percakapan sehari-hari, promosi, atau laporan korban. Hal ini menyebabkan pengguna awam kesulitan membedakan antara informasi yang valid dan konten penipuan. Oleh karena itu, diperlukan suatu pendekatan otomatis yang mampu membantu mengidentifikasi dan mendeteksi tweet yang mengandung unsur penipuan secara efektif.

Beberapa penelitian sebelumnya telah menerapkan analisis sentimen dan klasifikasi teks untuk mendeteksi konten negatif di media sosial dengan menggunakan algoritma pembelajaran mesin. Namun, sebagian penelitian lebih berfokus pada perbandingan algoritma tanpa memberikan analisis mendalam terhadap karakteristik kata yang berpengaruh dalam mendeteksi penipuan. Selain itu, penelitian yang secara khusus membahas penipuan berbahasa Indonesia di Twitter masih relatif terbatas.

Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk menerapkan metode Term Frequency–Inverse Document Frequency (TF-IDF) sebagai ekstraksi fitur dan algoritma Multinomial Naive Bayes sebagai metode klasifikasi dalam mendeteksi tweet penipuan. Penelitian ini bertujuan untuk mengetahui performa model dalam mengklasifikasikan tweet penipuan dan non-penipuan serta menganalisis kata-kata yang paling berpengaruh dalam proses deteksi. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi penipuan berbasis teks pada media sosial.

2. Tinjauan Pustaka

2.1. Analisis Sentimen

Analisis sentimen merupakan proses komputasional untuk mengidentifikasi dan mengekstraksi opini, sikap, atau perasaan yang terkandung dalam suatu teks (Asrumi et al., 2023). Dalam konteks media sosial, analisis sentimen digunakan untuk memahami pandangan publik terhadap suatu topik berdasarkan konten yang diunggah pengguna. Penelitian oleh Aprilia dan Isnain (2025) menerapkan Naive Bayes untuk analisis sentimen terkait kampanye anti-korupsi di Twitter dan memperoleh akurasi yang kompetitif, membuktikan bahwa pendekatan berbasis teks efektif untuk mengklasifikasikan opini dalam bahasa Indonesia. Lebih lanjut, Nugroho et al. (2025) menggunakan Naive Bayes Classifier untuk menganalisis sentimen terkait dugaan kecurangan Pemilu 2024 di Twitter dan menunjukkan bahwa fitur berbasis kata kunci memiliki daya diskriminasi yang kuat dalam konteks penipuan. Hal ini mendukung pendekatan yang digunakan dalam penelitian ini, yaitu menggunakan TF-IDF sebagai representasi fitur teks sebelum proses klasifikasi.

2.2. Deteksi Penipuan Digital

Deteksi penipuan digital merupakan tantangan yang semakin kompleks seiring meningkatnya variasi modus dan bahasa yang digunakan oleh pelaku penipuan di media sosial. Machova et al. (2022) membandingkan beberapa pendekatan pembelajaran mesin dan analisis sentimen untuk mendeteksi ulasan mencurigakan secara daring dan menemukan bahwa kombinasi fitur linguistik dengan algoritma klasifikasi menghasilkan performa deteksi yang lebih baik dibandingkan pendekatan tunggal. Penelitian Sovia et al. (2025) yang menggunakan arsitektur Long Short-Term Memory (LSTM) untuk mendeteksi berita palsu di Twitter menunjukkan bahwa model deep learning mampu menangkap konteks sekuensial yang tidak dapat ditangkap oleh model klasik. Meskipun demikian, model berbasis statistik seperti Naive Bayes tetap relevan karena sifatnya yang efisien, interpretatif, dan tidak membutuhkan data latih dalam jumlah sangat besar. Penelitian Lutfiyani dan Retnowati (2021) yang menerapkan Naive Bayes dan Decision Tree J48 untuk deteksi spam email menunjukkan bahwa Naive Bayes memiliki keunggulan dalam hal kecepatan komputasi meskipun akurasinya sedikit lebih rendah dibandingkan Decision Tree pada beberapa skenario.

2.3. TF-IDF sebagai Ekstraksi Fitur

Term Frequency–Inverse Document Frequency (TF-IDF) adalah metode pembobotan kata yang banyak digunakan dalam Natural Language Processing (NLP) untuk merepresentasikan teks sebagai vektor numerik. TF-IDF memberikan bobot tinggi pada kata yang sering muncul dalam satu dokumen tetapi jarang muncul di keseluruhan korpus, sehingga kata-kata yang benar-benar khas dan bermakna untuk suatu dokumen akan mendapat perhatian lebih dari model. Ghozali et al. (2025) dan Lestari et al. (2025) sama-sama menggunakan TF-IDF sebagai representasi fitur dalam penelitian analisis sentimen berbasis Naive Bayes dan menunjukkan bahwa kombinasi ini menghasilkan representasi yang cukup informatif untuk tugas klasifikasi teks pendek seperti tweet. Dalam penelitian ini, TF-IDF dikonfigurasi dengan konfigurasi unigram dan bigram (ngram_range=(1,2)) untuk menangkap pola kata tunggal maupun pasangan kata yang relevan dengan konteks penipuan, seperti “akun palsu” dan “transfer uang”.

2.4. Naive Bayes untuk Klasifikasi Teks

Algoritma Naive Bayes merupakan salah satu metode klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi antar fitur. Meskipun asumsi ini jarang terpenuhi sepenuhnya dalam data teks nyata, Naive Bayes tetap menunjukkan performa yang kompetitif dalam berbagai tugas klasifikasi teks. Varian Multinomial Naive Bayes secara khusus dirancang untuk data yang direpresentasikan dalam bentuk frekuensi atau bobot kata, menjadikannya pilihan yang sesuai untuk dipadukan dengan TF-IDF. Supriadi dan Fatmasari (2021)

DOI: https://doi.org/10.31004/riggs.v5i1.4991

menerapkan Naive Bayes dalam sistem analisis opini pengguna Twitter berbasis web dan membuktikan bahwa algoritma ini dapat diimplementasikan secara efisien bahkan pada lingkungan dengan sumber daya terbatas. Asmara et al. (2020) juga menggunakan Naive Bayes Classifier untuk analisis sentimen masyarakat terhadap Pemilu 2019 di Twitter, yang menunjukkan konsistensi performa algoritma ini pada dataset berbahasa Indonesia. Hadianti dan Tember (2022) membandingkan Naive Bayes dan SVM untuk analisis sentimen terkait Covid-19 dan menemukan bahwa Naive Bayes lebih unggul dalam hal efisiensi waktu pelatihan, meskipun SVM sedikit lebih akurat pada beberapa konfigurasi. Temuan-temuan ini memperkuat alasan pemilihan Multinomial Naive Bayes sebagai algoritma utama dalam penelitian ini.

2.5. Praproses Teks Bahasa Indonesia

Praproses teks merupakan tahap krusial dalam pipeline NLP yang bertujuan untuk mengurangi noise dan meningkatkan konsistensi representasi data. Untuk teks berbahasa Indonesia, tantangan utama praproses meliputi penanganan kata tidak baku, singkatan, dan variasi morfologi yang tinggi. Penggunaan Sastrawi Stemmer telah terbukti efektif dalam mereduksi variasi morfologi kata bahasa Indonesia ke bentuk dasarnya, sehingga kata-kata dengan makna yang sama tetapi bentuk berbeda dapat dikelompokkan bersama oleh model. Rafdi et al. (2021) yang menerapkan Naive Bayes dengan feature selection PSO dan Genetic Algorithm menunjukkan bahwa kualitas fitur yang baik, termasuk hasil stemming yang konsisten, memberikan kontribusi signifikan terhadap performa akhir model. Ponmani dan Thangaraj (2022) juga menegaskan pentingnya praproses yang menyeluruh dalam penelitian analisis sentimen berbasis kluster untuk data Twitter, di mana noise yang tidak dibersihkan dapat menurunkan kualitas representasi fitur secara signifikan. Dalam penelitian ini, praproses dilakukan secara berurutan mulai dari pembersihan teks, tokenisasi, stemming dengan Sastrawi, hingga filtering stopword untuk menghasilkan representasi teks yang optimal.

3. Metode Penelitian

Penelitian ini menggunakan metode eksperimen dengan pendekatan klasifikasi teks untuk mendeteksi tweet penipuan pada media sosial Twitter. Tahapan penelitian yang dilakukan meliputi pengumpulan data, pelabelan data, praproses teks, ekstraksi fitur menggunakan TF-IDF, pembangunan model klasifikasi menggunakan algoritma Multinomial Naive Bayes, serta evaluasi performa model. Alur tahapan penelitian secara lengkap ditunjukkan pada Gambar 1.

Gambar 1. Diagram alur proses deteksi tweet penipuan

DOI: https://doi.org/10.31004/riggs.v5i1.4991

3.1. Pengumpuan dan Pelabelan Data

Data penelitian diperoleh dari media sosial Twitter pada rentang waktu tahun 2022 hingga 2025. Pengambilan data dilakukan berdasarkan kata kunci yang berkaitan dengan penipuan digital, seperti penipuan transaksi daring, akun palsu, dan laporan korban. Data yang diperoleh berupa teks tweet berbahasa Indonesia beserta atribut pendukung lainnya.

Proses pelabelan dilakukan secara manual dengan dua kelas, yaitu penipuan dan non-penipuan. Label penipuan diberikan pada tweet yang mengandung laporan korban, indikasi modus penipuan, atau peringatan terhadap akun mencurigakan. Sebaliknya, label non-penipuan diberikan pada tweet yang bersifat informatif, percakapan umum, atau tidak mengandung unsur penipuan. Setelah proses pembersihan data dan penghapusan duplikasi, diperoleh sebanyak 2628 tweet yang digunakan sebagai dataset penelitian.

3.2. Praproses Data Teks

Tahap praproses bertujuan untuk meningkatkan kualitas data teks sebelum dilakukan pemodelan. Proses ini meliputi pembersihan URL, mention pengguna, tanda pagar (hashtag), angka, simbol, serta tanda baca yang tidak relevan. Selanjutnya dilakukan normalisasi huruf dengan mengubah seluruh teks menjadi huruf kecil.

Teks yang telah dibersihkan kemudian melalui proses tokenisasi untuk memecah kalimat menjadi kata-kata penyusun. Selanjutnya dilakukan proses stemming menggunakan Sastrawi Stemmer untuk mengubah setiap kata ke bentuk dasarnya, sehingga variasi morfologi kata seperti “menipu”, “ditipu”, dan “penipuan” dapat direpresentasikan secara seragam menjadi kata dasar “tipu”. Filtering dilakukan dengan menghapus kata tidak bermakna menggunakan daftar stopword bahasa Indonesia serta kata-kata noise tambahan yang tidak relevan. Proses praproses ini menghasilkan representasi teks yang lebih bersih, konsisten, dan relevan untuk tahap ekstraksi fitur.

3.3. Ekstraksi Fitur

Ekstraksi fitur dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Metode ini digunakan untuk mengubah data teks menjadi representasi numerik dengan memberikan bobot pada setiap kata berdasarkan tingkat kemunculan kata tersebut dalam dokumen dan keseluruhan korpus. Parameter TF-IDF diatur untuk membatasi jumlah fitur dan mengurangi pengaruh kata yang terlalu jarang maupun terlalu sering muncul. Hasil dari tahap ini berupa matriks fitur yang merepresentasikan setiap tweet sebagai vektor numerik.

$$TF(t,d) = f_{t,d} / \sum_{k} f_{k,d} \quad (1)$$

$$IDF(t) = log(N/df_t) \quad (2)$$

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (3)$$

$TF(t,d)$ adalah nilai term frequency kata *t* pada dokumen *d*;
$f_{t,d}$ adalah jumlah kemunculan kata *t* pada dokumen *d*;
$\sum_k f_{k,d}$ adalah jumlah seluruh kata dalam dokumen *d*;
$IDF(t)$ adalah inverse document frequency dari kata *t*;
N adalah jumlah total dokumen;
df_t adalah jumlah dokumen yang mengandung kata *t*.

3.4. Klasifikasi Menggunakan Naive Bayes

Model klasifikasi utama dibangun menggunakan algoritma Multinomial Naive Bayes. Algoritma ini dipilih karena sesuai untuk data teks yang direpresentasikan dalam bentuk frekuensi atau bobot kata, serta bersifat interpretatif dan efisien secara komputasi. Untuk menangani ketidakseimbangan distribusi kelas, diterapkan pembobotan sampel (sample weight) dengan skema balanced pada saat pelatihan model. Sebagai eksperimen pembanding, digunakan pula algoritma Decision Tree dengan parameter class_weight=’balanced’ untuk mengevaluasi

DOI: https://doi.org/10.31004/riggs.v5i1.4991

perbandingan performa antar algoritma. Dataset dibagi menjadi data latih dan data uji menggunakan metode train-test split dengan proporsi 80% data latih dan 20% data uji menggunakan stratified sampling untuk memastikan distribusi kelas yang proporsional pada kedua subset.

3.5. Evaluasi Model

Evaluasi performa model dilakukan menggunakan confusion matrix dan metrik evaluasi berupa akurasi, precision, recall, dan F1-score. Metrik ini digunakan untuk menilai kemampuan model dalam mengklasifikasikan tweet penipuan dan non-penipuan secara tepat. Selain itu, dilakukan analisis fitur untuk mengidentifikasi kata-kata yang memiliki pengaruh paling kuat dalam menentukan klasifikasi penipuan dan non-penipuan.

4. Hasil dan Diskusi

Bagian ini menyajikan hasil eksperimen klasifikasi tweet penipuan menggunakan algoritma Multinomial Naive Bayes dengan fitur TF-IDF, serta pembahasan terhadap performa model yang dihasilkan.

4.1. Hasil Praproses dan Statistik Data

Dataset yang digunakan dalam penelitian ini berjumlah 2.593 tweet berbahasa Indonesia yang diperoleh dari Twitter pada periode 2022 hingga 2025. Data telah melalui proses pembersihan teks, stemming, dan penghapusan duplikasi sehingga setiap tweet yang digunakan bersifat unik dan relevan. Hasil pelabelan manual menunjukkan bahwa distribusi data tidak seimbang, dengan jumlah tweet non-penipuan sebanyak 1.829 tweet (70,5%) dan tweet penipuan sebanyak 764 tweet (29,5%). Kondisi ini mencerminkan situasi nyata di media sosial, di mana konten percakapan umum lebih dominan dibandingkan laporan penipuan. Untuk mengatasi ketidakseimbangan ini, diterapkan teknik pembobotan kelas (class weight balanced) pada saat pelatihan model.

4.2. Hasil Klasifikasi Menggunakan Naive Bayes

Model klasifikasi dibangun menggunakan algoritma Multinomial Naive Bayes dengan representasi fitur TF-IDF dan penerapan sample weight balanced untuk menangani ketidakseimbangan kelas. Evaluasi performa model dilakukan menggunakan data uji sebanyak 519 tweet, yang merupakan 20% dari total 2.593 dataset, sesuai dengan skema pembagian data 80:20.

Hasil pengujian menunjukkan bahwa model Naive Bayes mencapai akurasi sebesar 87,09%. Meskipun akurasi sedikit lebih rendah dibandingkan versi tanpa pembobotan, penerapan class weight berhasil meningkatkan recall kelas penipuan secara signifikan. Rincian hasil evaluasi kinerja model disajikan pada Tabel 1, yang mencakup metrik precision, recall, F1-score, dan support untuk masing-masing kelas.

Tabel 1. Hasil Evaluasi Kinerja Model

Metrik	Global / Macro Avg	Kelas 0 (Non-Penipuan)	Kelas 1 (Penipuan)
Accuracy	0.87 (87,09%)	N/A	N/A
Precision	0.84	0.97	0.90
Recall	0.89	0.84	0.71
F1-Score	0.86	0.94	0.81
Support	519	366	153

Berdasarkan Tabel 1, penerapan stemming dan class weight menghasilkan perubahan signifikan pada performa model. Recall kelas penipuan meningkat drastis menjadi 94%, yang berarti model kini berhasil mendeteksi sebagian besar tweet penipuan pada data uji. Hal ini merupakan peningkatan yang sangat berarti dibandingkan eksperimen awal tanpa stemming dan class weight. Nilai precision kelas non-penipuan sebesar 97% dan F1-score sebesar 90% menunjukkan bahwa model tetap stabil dalam mengklasifikasikan kelas mayoritas.

Pada kelas penipuan, model memperoleh recall sebesar 94%, yang menunjukkan bahwa model kini mampu mendeteksi sebagian besar tweet penipuan secara efektif. Nilai precision sebesar 71% mengindikasikan masih adanya sejumlah false positive, di mana beberapa tweet non-penipuan diprediksi sebagai penipuan. Hal ini merupakan trade-off yang wajar dalam sistem deteksi penipuan, di mana recall yang tinggi lebih diprioritaskan

DOI: https://doi.org/10.31004/riggs.v5i1.4991

untuk memastikan tidak ada penipuan yang terlewat. Penerapan class weight balanced terbukti berhasil meningkatkan sensitivitas model terhadap kelas minoritas (penipuan).

Secara keseluruhan, hasil ini menunjukkan bahwa kombinasi stemming, TF-IDF, dan class weight balanced berhasil meningkatkan performa model dalam mendeteksi penipuan secara signifikan. Recall kelas penipuan yang mencapai 94% menjadikan model ini lebih andal untuk diterapkan dalam sistem deteksi penipuan digital secara nyata, di mana kemampuan mendeteksi penipuan (meminimalkan false negative) merupakan aspek yang paling krusial.

4.3. Analisis Confusion Matrix

Tabel 2. Confusion Matrix Hasil Klasifikasi Tweet

Kelas Label	Prediksi Kelas 0 (Non-Penipuan)	Prediksi Kelas 1 (Penipuan)
Aktual Kelas 0 (Non-Penipuan)	308 (TN)	58 (FP)
Aktuan Kelas 1 (Penipuan)	9 (FN)	144 (TP)

Keterangan: TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive.

Berdasarkan Tabel 2, dari total 519 data uji, model berhasil mengklasifikasikan 308 tweet non-penipuan (TN) dan 144 tweet penipuan (TP) dengan benar. Terdapat 58 tweet non-penipuan yang diprediksi sebagai penipuan (false positive) dan 9 tweet penipuan yang diprediksi sebagai non-penipuan (false negative). Perbandingan dengan eksperimen sebelumnya menunjukkan penurunan false negative secara drastis dari 44 menjadi 9, yang berarti model kini jauh lebih sensitif dalam mendeteksi penipuan berkat penerapan class weight.

Nilai true positive yang tinggi (144) dan false negative yang sangat rendah (9) menunjukkan bahwa model sangat efektif dalam mendeteksi tweet penipuan. Peningkatan false positive dari 6 menjadi 58 merupakan konsekuensi dari penerapan class weight yang meningkatkan sensitivitas model, namun trade-off ini dapat diterima mengingat tujuan utama sistem adalah mendeteksi penipuan (meminimalkan false negative).

4.4. Analisis Fitur

Analisis fitur dilakukan untuk mengidentifikasi kata-kata yang paling berpengaruh dalam proses klasifikasi setelah penerapan stemming. Hasil analisis menunjukkan bahwa kata-kata seperti tipu, scam, akun, modus, dan deepfake memiliki probabilitas tinggi pada kelas penipuan. Penerapan stemming terbukti efektif karena variasi kata seperti “menipu”, “ditipu”, dan “penipuan” kini direpresentasikan oleh satu token “tipu”, sehingga meningkatkan frekuensi kemunculan fitur kunci. Sebaliknya, kata-kata yang bersifat umum seperti “hati”, “bantu”, dan “udah” lebih dominan pada kelas non-penipuan.

Hasil ini menunjukkan bahwa model mampu mempelajari pola linguistik yang relevan dengan konteks penipuan di Twitter. Keberadaan kata-kata kunci tersebut memperkuat keputusan model dalam mengklasifikasikan tweet sebagai penipuan.

4.5. Perbandingan Naive Bayes dan Decision Tree

Sebagai eksperimen tambahan, dilakukan perbandingan performa antara Multinomial Naive Bayes dan Decision Tree menggunakan konfigurasi yang sama, yaitu fitur TF-IDF dengan class weight balanced. Hasil perbandingan disajikan pada Tabel 3.

Tabel 3. Perbandingan Performa Naive Bayes vs Decision Tree

Model	Accuracy	Precision (Penipuan)	Recall (Penipuan)	F1-Score (Penipuan)
Naive Bayes	87,09%	71,29%	94,12%	81,13%
Decision Tree	96,53%	95,30%	92,81%	94,04%

DOI: https://doi.org/10.31004/riggs.v5i1.4991

Berdasarkan Tabel 3, Decision Tree menghasilkan performa yang lebih tinggi pada hampir semua metrik, dengan akurasi 96,53% dan F1-score penipuan 94,04%. Namun, Naive Bayes dipilih sebagai model utama dalam penelitian ini dengan beberapa pertimbangan. Pertama, Naive Bayes bersifat lebih interpretatif sehingga memudahkan analisis kata-kata yang berpengaruh dalam proses klasifikasi. Kedua, Naive Bayes jauh lebih efisien secara komputasi, terutama untuk dataset berukuran besar. Ketiga, recall kelas penipuan pada Naive Bayes (94,12%) tetap sangat tinggi dan hanya selisih tipis dibandingkan Decision Tree (92,81%), sehingga dari sudut pandang deteksi penipuan, kedua model memiliki kemampuan yang setara. Perbedaan akurasi yang signifikan antara kedua model juga menunjukkan bahwa Decision Tree berpotensi mengalami overfitting pada data latih, yang perlu diverifikasi lebih lanjut melalui cross-validation.

4.6. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diakui. Pertama, pelabelan data dilakukan secara manual oleh tim peneliti tanpa pengukuran inter-annotator agreement, sehingga terdapat potensi subjektivitas dalam proses labeling. Kedua, dataset terbatas pada tweet berbahasa Indonesia, sehingga model belum tentu generalisasi dengan baik untuk konten berbahasa daerah atau tweet campuran (code-switching). Ketiga, penelitian ini belum menerapkan teknik validasi silang (cross-validation), sehingga evaluasi performa hanya berdasarkan satu kali pembagian data. Keempat, model belum mempertimbangkan konteks percakapan atau thread tweet, yang dapat memberikan informasi tambahan untuk klasifikasi. Keterbatasan-keterbatasan ini dapat menjadi arah pengembangan pada penelitian selanjutnya.

Diskusi

1. Dampak Stemming terhadap Performa Model

Salah satu temuan paling signifikan dalam penelitian ini adalah peningkatan recall kelas penipuan dari 66% pada eksperimen awal menjadi 94,12% setelah penerapan stemming dan class weight balanced. Peningkatan sebesar 28 poin persentase ini tidak dapat sepenuhnya dikaitkan hanya dengan class weight, karena perubahan distribusi bobot saja tidak akan mengubah ruang fitur yang digunakan model. Stemming berperan langsung dalam memperkaya representasi fitur dengan menyatukan variasi morfologi kata ke bentuk dasarnya. Sebagai ilustrasi, sebelum stemming diterapkan, kata “menipu”, “ditipu”, “penipuan”, dan “menipu-nipu” diperlakukan sebagai empat token yang sepenuhnya berbeda oleh TF-IDF, sehingga masing-masing memiliki bobot yang rendah karena frekuensi kemunculannya terpecah. Setelah stemming, keempat bentuk tersebut disatukan menjadi satu token “tipu” yang kemunculannya menjadi lebih dominan dan mendapat bobot TF-IDF yang lebih tinggi. Fenomena ini secara langsung meningkatkan kemampuan model untuk mengenali tweet yang mengandung unsur penipuan meskipun ditulis dengan diksi yang beragam.

Temuan ini konsisten dengan argumen Rafdi et al. (2021) yang menekankan bahwa kualitas representasi fitur, termasuk hasil praproses yang menyeluruh, memberikan kontribusi lebih besar terhadap performa model dibandingkan pemilihan algoritma semata. Dalam konteks bahasa Indonesia yang memiliki sistem afiksasi kompleks, stemming menjadi tahap praproses yang tidak dapat diabaikan, terutama untuk domain penipuan yang kosakatanya sangat bervariasi secara morfologis.

2. Pola Linguistik Penipuan Digital di Twitter

Analisis distribusi fitur yang dihasilkan model memberi gambaran yang menarik mengenai pola linguistik penipuan digital berbahasa Indonesia di Twitter. Kata-kata dengan probabilitas log tertinggi pada kelas penipuan, yaitu “tipu”, “scam”, “modus”, “hati-hati”, dan “deepfake”, mencerminkan dua jenis tweet penipuan yang berbeda secara pragmatis. Pertama, tweet yang berasal dari korban atau saksi yang secara eksplisit melaporkan kejadian penipuan dan menggunakan kata-kata seperti “tipu” dan “scam”. Kedua, tweet peringatan dari pengguna yang mencoba menginformasikan masyarakat tentang modus tertentu, yang ditandai dengan kata “modus”, “hati-hati”, dan “deepfake”. Kemunculan kata “deepfake” sebagai fitur berpengaruh mengindikasikan bahwa dataset yang dikumpulkan pada periode 2022–2025 sudah mencerminkan perkembangan modus penipuan berbasis teknologi kecerdasan buatan, yang semakin marak digunakan untuk membuat konten manipulatif di media sosial.

Temuan ini berbeda dari penelitian Nugroho et al. (2025) yang mendeteksi kecurangan pemilu, di mana fitur dominan lebih bersifat politis. Dalam konteks penipuan transaksi digital, bahasa yang digunakan cenderung lebih

DOI: https://doi.org/10.31004/riggs.v5i1.4991

personal, emosional, dan berorientasi pada tindakan ("tolong viralkan", "hati-hati", "lapor"). Pola ini menunjukkan bahwa model deteksi penipuan yang dilatih pada satu domain tidak dapat langsung ditransfer ke domain lain tanpa proses fine-tuning, karena distribusi fitur linguistiknya berbeda secara fundamental.

3. Perbandingan dengan Penelitian Terdahulu

Bila dibandingkan dengan penelitian sejenis, performa model dalam penelitian ini tergolong kompetitif. Ghozali et al. (2025) yang menggunakan Naive Bayes untuk analisis sentimen pinjaman online di Twitter memperoleh akurasi di kisaran yang sebanding, meskipun tidak secara spesifik mengukur recall kelas negatif sebagai metrik utama. Hadianti dan Tember (2022) melaporkan bahwa Naive Bayes pada dataset sentimen Covid-19 menghasilkan performa yang lebih rendah dibandingkan SVM pada metrik akurasi, namun keunggulan Naive Bayes dalam efisiensi komputasi dan interpretabilitas tetap menjadi pertimbangan praktis yang relevan. Penelitian Aprilia dan Isnain (2025) menggunakan dataset yang lebih besar namun tidak menerapkan teknik penanganan imbalanced data, sehingga performa pada kelas minoritas tidak dioptimalkan. Perbedaan pendekatan inilah yang membuat perbandingan lintas penelitian perlu dilakukan dengan hati-hati, karena perbedaan dataset, domain, dan konfigurasi praproses dapat menghasilkan perbedaan performa yang signifikan meskipun algoritma yang digunakan sama.

4. Implikasi Praktis dan Arah Pengembangan

Model yang dihasilkan dalam penelitian ini memiliki potensi untuk dikembangkan menjadi sistem deteksi penipuan semi-otomatis yang dapat membantu moderasi konten di platform media sosial. Dengan recall kelas penipuan sebesar 94,12%, sistem ini dapat menjaring hampir seluruh laporan penipuan yang masuk ke dalam antrean pemeriksaan manual, sehingga secara signifikan mengurangi beban kerja moderator. Precision sebesar 71,29% berarti sekitar 29% tweet yang ditandai sebagai penipuan sebenarnya bukan penipuan, sehingga tahap verifikasi manual tetap diperlukan sebelum tindakan diambil. Kombinasi ini mencerminkan arsitektur sistem deteksi dua lapis yang umum diterapkan di industri, di mana model machine learning berfungsi sebagai filter awal berkecepatan tinggi sebelum verifikasi manusia.

Untuk pengembangan lebih lanjut, terdapat beberapa arah yang menjanjikan. Pertama, eksplorasi model berbasis transformer seperti IndoBERT atau IndoNLU yang telah dilatih pada korpus bahasa Indonesia dalam skala besar berpotensi meningkatkan pemahaman konteks secara lebih mendalam, terutama untuk mendeteksi penipuan yang disampaikan secara implisit atau menggunakan eufemisme. Kedua, penggabungan fitur non-teks seperti metadata akun (usia akun, jumlah pengikut, rasio mengikuti/diikuti) dan metadata tweet (waktu unggah, tingkat retweet) dapat memberikan dimensi tambahan yang berguna untuk membedakan akun penipuan dari akun pengguna biasa. Ketiga, implementasi sistem deteksi berbasis aliran data (streaming detection) yang dapat memproses tweet secara real-time akan meningkatkan nilai praktis penelitian ini, mengingat penipuan digital seringkali bersifat sementara dan memanfaatkan momentum tertentu seperti bencana, tren viral, atau momen belanja daring.

5. Kesimpulan

Penelitian ini berhasil menerapkan metode TF-IDF dan algoritma Multinomial Naive Bayes untuk mendeteksi tweet penipuan berbahasa Indonesia pada media sosial Twitter. Dengan penambahan stemming menggunakan Sastrawi Stemmer dan penerapan class weight balanced untuk menangani ketidakseimbangan kelas, model berhasil mencapai akurasi sebesar 87,09% dengan recall kelas penipuan yang meningkat signifikan menjadi 94,12%. Sebagai perbandingan, algoritma Decision Tree menghasilkan akurasi lebih tinggi sebesar 96,53% dengan recall penipuan 92,81%, namun Naive Bayes tetap dipilih sebagai model utama karena sifatnya yang lebih interpretatif dan efisien secara komputasi. Hasil analisis fitur menunjukkan bahwa kata-kata seperti "tipu", "scam", "modus", dan "deepfake" memiliki pengaruh yang kuat dalam mendeteksi penipuan setelah proses stemming, sehingga membuktikan bahwa kombinasi stemming dan pendekatan berbasis teks efektif untuk digunakan dalam konteks deteksi penipuan digital. Penelitian ini dapat menjadi dasar bagi pengembangan sistem deteksi penipuan otomatis pada media sosial. Untuk penelitian selanjutnya, disarankan untuk menambah jumlah data, menerapkan teknik oversampling seperti SMOTE, serta mengeksplorasi model berbasis deep learning seperti LSTM atau BERT berbahasa Indonesia guna meningkatkan performa klasifikasi lebih lanjut.

DOI: https://doi.org/10.31004/riggs.v5i1.4991

Referensi

1. D. Setyo Nugroho, I. F. Hanifuddin, M. A. Hasbi, F. Fredianto, A. M. Saputra, dan R. Zildjian, "Sentiment Analysis of Alleged 2024 Election Fraud Based on Tweets Using the Naïve Bayes Classifier Algorithm," *Malcom: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, 2025, doi: 10.57152/malcom.v4i3.1496.
2. N. W. A. S. Aprilia dan A. R. Isnain, "Analisis Sentimen Terhadap Media Sosial Twitter dengan Kasus Kampanye Anti-Korupsi di Indonesia Menggunakan Naive Bayes," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, 2025, doi: 10.30865/mib.v8i2.7582.
3. M. Imam Ghozali, W. H. Sugiharto, dan A. F. Iskandar, "Analisis Sentimen Pinjaman Online Di Media Sosial Twitter Menggunakan Metode Naive Bayes," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 6, 2025, doi: 10.30865/klik.v3i6.936.
4. R. Sovia, D. A. Valkyrie, R. H. Zain, dan F. Firdaus, "Language Processing for Detecting Fake News on Twitter Using a Long Short-Term Memory Architecture," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 935–943, 2025, doi: 10.29207/resti.v9i4.6570.
5. A. Lestari, A. I. Purnamasari, A. Bahtiar, dan E. Tohidi, "Sentiment Analysis to Classify TikTok Shop Users on Twitter with Naïve Bayes Classifier Algorithm," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 2, 2025, doi: 10.59934/jaiea.v4i2.748.
6. K. Machova, M. Mach, dan M. Vasilko, "Comparison of Machine Learning and Sentiment Analysis in Detection of Suspicious Online Reviewers on Different Type of Data," *Sensors*, vol. 22, no. 1, 2022, doi: 10.3390/s22010155.
7. Apif Supriadi, Fatmasari," Implementasi Metode Klasifikasi Naive Bayes Pada Sistem Analisis Opini Pengguna Twitter Berbasis Web, " February 3, 2021 // DOI: 10.51998/jsi.v10i1.356.
8. Bagus Muhammad Akbar, Ahmad Taufiq Akbar, Rochmat Husaini, "Sinovac Vaccine Sentiment and Emotion Analysis on Twitter Using Naïve Bayes and Valence Shifter, " December 30, 2021 // DOI: 10.54914/jtt.v7i2.433
9. Asrumi Asrumi, Didik Suharijadi, Agustina Dewi Setiari, Diah Putri Wulanda, "Analisis Sentimen dan Penggalian Opini, " November 27, 2023.
10. K. Ponmani, M. Thangaraj, "Clustering Based Sentiment Analysis on Twitter Data for COVID-19 Vaccines in India, " 2022 // DOI: 10.53730/ijhs.v6nS2.6126
11. Ni Kadek Indah Puspaningrum, Putu Diah Sastri Pitanatri, Ni Wayan Chintia Pinaria, "Analisis Sentimen dalam Mengurangi Pembatalan Reservasi di The Westin Resort & Spa Ubud, " July 2025 // DOI: 10.35912/simo.v6i2.4707
12. Sri Hadianti, Firman Yosep Tember, "Analisis Sentiment Covid-19 di Twitter Menggunakan Metode Naive Bayes dan SVM," June 30, 2022 // DOI: 10.36294/jurti.v6i1.2557
13. Rizka Safitri Lutfiyani, Niken Retnowati, "Implementation of Email Spam Detection Using Naïve Bayes Algorithm and Decision Tree J48 Text Mining Method, " October 30, 2021 // DOI: 10.35508/jicon.v9i2.5304
14. Rengga Asmara, Muhammad Febrian Ardiansyah, Muhammad Anshori, "Analisa Sentiment Masyarakat terhadap Pemilu 2019 Berdasarkan Opini di Twitter Menggunakan Metode Naive Bayes Classifier," 2020 // DOI: 10.35314/isi.v5i2.1095
15. Abi Rafdi, Herman Mawengkang, Syahril Efend, "Sentiment Analysis Using Naive Bayes Algorithm with Feature Selection Particle Swarm Optimization (PSO) and Genetic Algorithm," 2021 // DOI: 10.25008/ijadis.v2i2.1224
16. Muhammad Dwison Alizah, Arifin Nugroho, Ummu Radiyah, Windu Gata, "Sentimen Analisis Terkait Lockdown pada Sosial Media Twitter, " 2020 // DOI: 10.31294/ijse.v6i2.8991

DOI: https://doi.org/10.31004/riggs.v5i1.4991