JOISIE (Journal Of Information Systems And Informatics Engineering)
Vol. 9, No.2, Desember 2025, Hlm 403-411
Received: 9 September 2025 Revised: 20 November 2025, Accepted: 16 Desember 2025
p- ISSN: 2503-5304
e- ISSN: 2527-3116

Analisis Sentimen Komentar YouTube terhadap Rumor Peluncuran iPhone 17 Menggunakan Web Scraping dan Studi Komparatif Algoritma Klasifikasi

Nickel Modami[1)], Ephraim Eleazar Reva Manopo[2)], Danendra Rafi Enditama[3)], Afifah Trista Ayunda[4)]
[1,2,3,4] Fakultas Sains dan Teknologi, Universitas Pradita, Kabupaten Tangerang, Banten, Indonesia
email: [1]nickel.modami@student.pradita.ac.id, [2]ephraim.eleazar@student.pradita.ac.id,
[3]danendra.rafi@student.pradita.ac.id, [4]afifah.trista@pradita.ac.id

Abstract

The launch of technology products such as the iPhone always triggers massive discussions that reflect public perceptions of innovation. This study aims to analyze Indonesian public sentiment towards rumors of the iPhone 17 launch using social media data. A total of 1,077 YouTube comments were processed using a text mining approach and TF-IDF weighting to be classified using the Naive Bayes, Random Forest, and K-Nearest Neighbor (KNN) algorithms. The analysis results show that positive sentiment dominates at 43%, driven by enthusiasm for camera features, followed by negative sentiment at 38.8%, highlighting price and design issues. Model evaluation shows Random Forest as the best algorithm with a test accuracy of 69.2% and cross-validation of 65.68%, outperforming other algorithms. This study contributes to mapping Indonesian market perceptions, concluding that, despite strong Apple brand loyalty, price factors and functional innovation are the main determinants of product acceptance.

Keyword: *iPhone, YouTube, Random Forest, Sentiment Analysis, classification*

Abstrak

Peluncuran produk teknologi seperti iPhone senantiasa memicu diskusi masif yang merefleksikan persepsi publik terhadap inovasi. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap rumor peluncuran iPhone 17 memanfaatkan data media sosial. Sebanyak 1.077 komentar YouTube diproses menggunakan pendekatan Text Mining dan pembobotan TF-IDF untuk diklasifikasikan menggunakan algoritma Naive Bayes, Random Forest, dan K-Nearest Neighbor (KNN). Hasil analisis menunjukkan sentimen positif mendominasi sebesar 43% yang didorong antusiasme fitur kamera, diikuti sentimen negatif sebesar 38,8% yang menyoroti isu harga dan desain. Evaluasi model menunjukkan Random Forest sebagai algoritma terbaik dengan akurasi uji 69,2% dan validasi silang 65,68%, mengungguli algoritma lainnya. Penelitian ini memberikan kontribusi dalam memetakan persepsi pasar Indonesia, menyimpulkan bahwa meskipun loyalitas merek Apple kuat, faktor harga dan inovasi fungsional menjadi penentu utama penerimaan produk.

Keywords: *iPhone, YouTube, Random Forest, Sentiment Analysis, classification*

1. PENDAHULUAN

Perkembangan teknologi telekomunikasi, khususnya pasar smartphone, berlangsung sangat pesat di Indonesia. Pengguna smartphone di Indonesia meningkat signifikan dan diprediksi mencapai 89,2% populasi pada akhir 2025 (Fadillah & Batu, 2025). Di tengah persaingan pasar yang ketat, peluncuran seri iPhone terbaru oleh Apple Inc. secara rutin memicu diskusi intens di media sosial, termasuk platform berbagi video YouTube. YouTube, dengan 143 juta pengguna yang menjadikannya salah satu media sosial terbesar di Indonesia, telah menjadi wadah utama bagi masyarakat untuk mengekspresikan opini secara bebas (Kemp, 2025). Data opini yang masif dan tidak terstruktur ini dapat diolah menjadi informasi berharga berupa sentimen (positif, negatif, atau netral). Mengingat volume data yang sangat besar, pendekatan otomatis menggunakan *Machine Learning* menjadi krusial untuk mengekstrak *Voice of Customer* (VoC) ini secara efektif (Bilinski, 2024).

Analisis Sentimen atau Opinion Mining dalam ranah *Natural Language Processing* (NLP) menjadi pendekatan krusial untuk mengolah data teks tersebut secara otomatis (Bing, 2012). Tantangan utama dalam analisis sentimen berbahasa Indonesia adalah memilih algoritma klasifikasi yang paling akurat dan efisien. Beberapa studi sebelumnya menunjukkan performa beragam dari

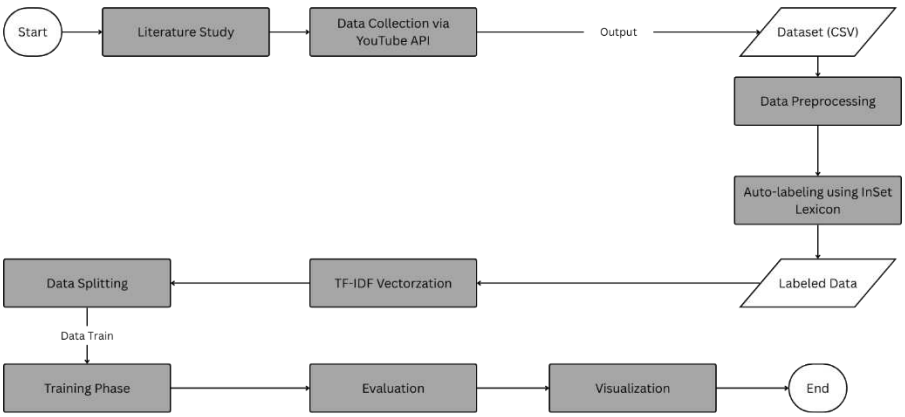
berbagai algoritma: Naïve Bayes dikenal efisien dengan akurasi mencapai 74% pada dataset dari penelitian terdahulu (Wulandari et al., 2023), Random Forest menawarkan stabilitas melalui teknik ensemble dengan akurasi hingga 86,23% pada studi sebelumnya (Tantyoko et al., 2023), sedangkan K-Nearest Neighbor (KNN) bekerja berdasarkan kedekatan fitur data lewat data latih (Data Train) (Kang, 2021). Komparasi ketiga algoritma ini diperlukan untuk menentukan model yang paling optimal dalam menangani karakteristik komentar YouTube Indonesia yang penuh dengan *noise* dan bahasa *slang* (non-baku).

Penelitian mengenai analisis sentimen produk smartphone telah banyak dilakukan, salah satunya studi pada peluncuran iPhone 16 oleh Manurung dan Mayatopani (2025). Meskipun demikian, terdapat kesenjangan (gap) penelitian dalam analisis sentimen terhadap produk yang belum dirilis atau masih berupa rumor, khususnya iPhone 17. Analisis pada fase pra-peluncuran ini menjadi krusial untuk memetakan ekspektasi dan resistensi pasar lebih awal sebelum produk resmi dipasarkan. Analisis pada fase pra-peluncuran ini krusial untuk memetakan ekspektasi dan resistensi pasar lebih awal. Berbeda dengan penelitian sebelumnya, studi ini memberikan kontribusi ilmiah melalui evaluasi komparatif kinerja tiga algoritma klasifikasi (Naïve Bayes, Random Forest, dan KNN) pada dataset komentar YouTube berbahasa Indonesia. Tujuan penelitian ini adalah mengimplementasikan teknik *Text Mining end-to-end*, mulai dari web scraping, pembobotan TF-IDF, hingga evaluasi model, untuk menemukan algoritma dengan akurasi terbaik sekaligus mengungkap sentimen dominan masyarakat sebagai referensi strategis bagi pelaku industri maupun referensi akademis.

2. METODE PENELITIAN

2.1 ALUR PENELITIAN

Penelitian ini menerapkan metode penelitian deskriptif dengan pendekatan kuantitatif. Metode deskriptif dipilih karena tujuan utama dari penelitian deskriptif adalah untuk menggambarkan secara seakurat mungkin suatu populasi, situasi, atau fenomena, serta karakteristik yang menyertainya. dengan prinsip kuantitatif, penelitian ini melibatkan data yang dikumpulkan dalam bentuk numerik atau kategori data yang dapat diukur (Ghanad, 2023). Pendekatan ini dianggap sebagai teknik terbaik untuk menemukan karakteristik, frekuensi, tren, dan kategori dari data komentar. Secara garis besar, tahapan penelitian yang dilakukan dalam analisis sentimen ini disajikan dalam bentuk diagram alir (flowchart) pada Gambar 1.



Gambar 1. Contoh penggunaan gambar

https://doi.org/10.35145/joisie.v9i2.5684

2.2 PRA-PEMROSESAN TEKS

Data yang digunakan adalah data berupa komentar publik dari platform media sosial YouTube mengenai rumor dan peluncuran iPhone 17. Akuisisi data dilakukan menggunakan metode Web Scraping memanfaatkan YouTube Data API v3. Kata kunci pencarian yang difokuskan pada kanal pengulas teknologi Indonesia adalah "iPhone 17 bocoran" dan "iPhone 17 indonesia".

Data mentah hasil akuisisi kemudian melalui tahapan Pra-pemrosesan yang sistematis untuk mengurangi noise dan menyeragamkan teks. Proses dimulai dengan Cleaning, yaitu pembersihan teks dari karakter yang tidak relevan seperti angka, tanda baca, *URL, hashtag, username* (@), dan *emoji* menggunakan teknik Regular Expression (Regex). Selanjutnya, dilakukan Case Folding untuk mengonversi seluruh teks menjadi huruf kecil (lowercase), memastikan keseragaman data. Untuk mengatasi penggunaan bahasa non-baku dari media sosial, diterapkan Normalization dengan mengubah kata alay atau gaul menjadi kata baku sesuai Kamus Besar Bahasa Indonesia (KBBI) menggunakan kamus normalisasi (colloquial lexicon). Kemudian, dilakukan Stopword Removal untuk menghapus kata-kata umum yang tidak signifikan maknanya (misalnya, "yang," "dan," "di") menggunakan daftar stopword dari pustaka Sastrawi. Tahapan diakhiri dengan Stemming, yaitu pengubahan kata berimbuhan menjadi kata dasar menggunakan algoritma Enhanced Confix Stripping Stemmer yang tersedia dalam pustaka Sastrawi.

Mengingat volume data yang besar, proses pelabelan sentimen dilakukan menggunakan pendekatan otomatis berbasis kamus (Lexicon Based Approach) untuk efisiensi waktu. Penelitian ini memanfaatkan kamus InSet (Indonesian Sentiment Lexicon), yang memuat daftar kata berbahasa Indonesia dengan bobot sentimen positif dan negatif yang telah disesuaikan untuk teks media sosial. Mekanisme penentuan label dilakukan dengan menghitung akumulasi skor polaritas (polarity scoring) dari setiap kata yang terdeteksi dalam kalimat (Koto, 2017). Dalam implementasinya, setiap kata yang cocok dengan entri kamus positif diberi nilai +1, sedangkan kata yang cocok dengan kamus negatif diberi nilai -1. Klasifikasi akhir ditentukan berdasarkan total skor kalimat: kategori Positif jika total skor > 0, Negatif jika total skor < 0, dan Netral jika total skor sama dengan 0.

2.3 LINGKUNGAN PENGEMBANGAN

Implementasi model klasifikasi pada penelitian ini dibangun menggunakan bahasa pemrograman Python di lingkungan (*environment*) Google Colab. Proses manipulasi data tabular memanfaatkan pustaka Pandas, pra-pemrosesan teks bahasa Indonesia menggunakan Sastrawi, serta pembangunan model mesin pembelajaran menggunakan modul Scikit-learn.

2.4 MODEL PENGOLAHAN DAN PEMBOBOTAN TEKS

Random Forest adalah algoritma pembelajaran ensemble yang menggabungkan prediksi dari banyak pohon keputusan (decision trees) untuk meningkatkan akurasi dan mencegah overfitting (Halabaku & Bytyçi, 2024). Penerapan teknik bagging (bootstrap aggregating) di mana setiap pohon dilatih pada subset data acak. Keputusan akhir klasifikasi ditentukan berdasarkan sistem pemungutan suara terbanyak (majority voting) dari seluruh pohon yang terbentuk. Secara matematis, prediksi kelas akhir pada Random Forest dapat dirumuskan sebagai berikut (Han et al., 2012):

$$\hat{Y} = \text{mode}\{h_{1(x)}, h_{2(x)}, \dots, h_{B(x)}\} \qquad (1)$$

Keterangan:

1. Yˆ: Prediksi kelas akhir dari algoritma Random Forest.
2. mode: Nilai yang paling sering muncul (*majority vote*).
3. B: Jumlah total pohon (*trees*) dalam hutan.

https://doi.org/10.35145/joisie.v9i2.5684

4. $hb_{(x)}$: Prediksi kelas yang dihasilkan oleh pohon ke-b untuk input x.

Tahap ekstraksi fitur bertujuan untuk mengubah data teks hasil *preprocessing* menjadi format representasi numerik (vektor) agar dapat diproses oleh algoritma pembelajaran mesin. Dalam penelitian ini, metode pembobotan yang digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF). Prinsip utamanya adalah bobot suatu kata akan bernilai tinggi jika kata tersebut sering muncul dalam satu dokumen tertentu (Term Frequency tinggi), namun jarang muncul di dokumen lain dalam korpus (Document Frequency rendah). Hal ini membantu sistem membedakan dokumen satu dengan yang lainnya berdasarkan kata-kata yang unik. Secara matematis, bobot TF-IDF dihitung dengan rumus berikut (Huwaida et al., 2024):

$$W_{i,j} = TF_{i,j} \times \log\left(\frac{N}{DF_i}\right) \qquad (2)$$

Keterangan:

1. $W_{i,j}$: Bobot kata i pada dokumen j.
2. $TF_{i,j}$: Frekuensi kemunculan kata i dalam dokumen j (*Term Frequency*).
3. N: Total jumlah seluruh dokumen dalam korpus data.
4. Tf_i: Jumlah dokumen yang mengandung kata i (*Document Frequency*).
5. log: Fungsi logaritma untuk menormalisasi nilai *Inverse Document Frequency*.

2.5 EVALUASI MODEL

Penelitian ini menerapkan studi komparasi terhadap tiga algoritma Supervised Learning untuk klasifikasi sentimen, yaitu Multinomial Naive Bayes yang efektif menangani data teks berdimensi tinggi, Random Forest Classifier dengan konfigurasi ensemble sebanyak 100 estimators untuk mereduksi overfitting, serta K-Nearest Neighbor (KNN) dengan parameter κ=5 berbasis jarak Euclidean. Untuk skenario pengujian, dataset dibagi dengan proporsi 80% sebagai data latih dan 20% sebagai data uji. Evaluasi performa model diukur menggunakan parameter Accuracy, Precision, Recall, dan F1-Score yang diturunkan dari tabel *Confusion Matrix* dan k-Fold *Cross Validation*.

Confusion Matrix adalah tabel yang menampilkan kinerja algoritma klasifikasi dengan hasil *true positives, true negatives, false positives,* dan *false negatives* (Markoulidakis & Markoulidakis, 2024). 10-fold cross-validation membagi dataset menjadi sepuluh bagian, di mana model dilatih menggunakan sembilan bagian dan divalidasi menggunakan satu bagian yang tersisa, kemudian proses ini dirotasi hingga seluruh fold digunakan. Metode ini memberikan estimasi performa model yang andal (reliable estimate of model performance) serta membantu mencegah adanya overfitting. Berbagai penelitian menunjukkan bahwa 10-fold cross-validation umumnya efektif, dengan k = 10 sebagai pilihan yang paling umum dan k = 5 dapat memadai untuk dataset berukuran besar (Liu & Jody, 2024).

3. HASIL DAN PEMBAHASAN

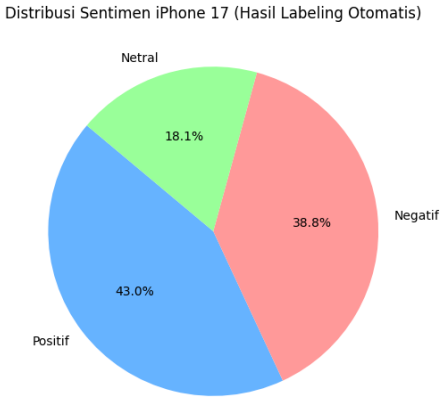
3.1 ANALISIS SENTIMEN PENUTUR BAHASA INDONESIA

Penelitian ini mengumpulkan total 1.077 data komentar mentah dari platform *YouTube* menggunakan teknik crawling melalui YouTube Data API v3. Setelah melalui tahapan pra-pemrosesan (*cleaning, case folding, normalization, stopword removal, stemming*) dan pelabelan otomatis menggunakan kamus InSet, diperoleh distribusi sentimen masyarakat terhadap peluncuran iPhone 17.

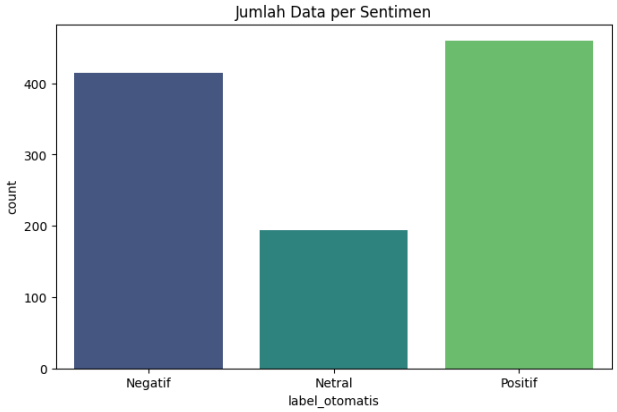
Visualisasi distribusi sentimen pada Gambar 2 menunjukkan bahwa opini masyarakat cenderung positif. Berdasarkan hasil pelabelan, Sentimen Positif dengan persentase 43% (460 data) ini selaras

https://doi.org/10.35145/joisie.v9i2.5684

dengan penelitian (Manurung & Mayatopani, 2025) pada peluncuran iPhone 16, yang juga mencatat sentimen positif sebagai refleksi loyalitas merek (brand loyalty) yang kuat di ekosistem pengguna Apple dan mengindikasikan tingginya antusiasme pasar Indonesia. Sentimen Negatif menempati urutan kedua sebesar 38,8% (415 data), yang menyoroti isu-isu negatif pada produk terkait. Sementara itu, Sentimen Netral merupakan minoritas sebesar 18,1% (194 data), yang berisikan pertanyaan dan pernyataan yang cenderung tidak positif maupun negatif.



Gambar 2. Distribusi Sentimen Komentar



Gambar 3. Chart data per sentiment

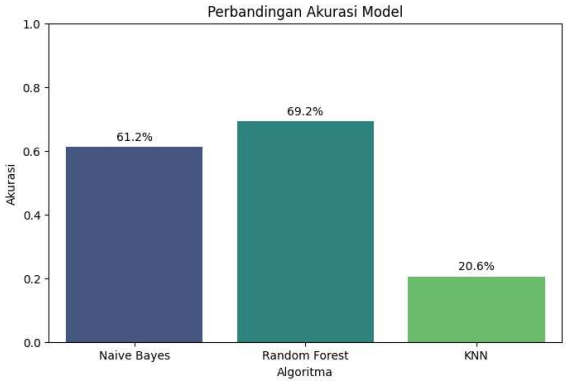
3.2 KOMPARASI KINERJA ALGORITMA KLASIFIKASI

Evaluasi model dilakukan menggunakan data uji (testing set) sebesar 20% dari total dataset. Tiga algoritma dikomparasikan berdasarkan metrik Accuracy,Precision, Recall, dan F1-Score. Ringkasan kinerja ketiga algoritma disajikan pada Tabel 1 dan Gambar 4.

Tabel 1. Output Rata-rata Makro

Algoritma	Precision	Recall	F1-Score
RF	0.59	0.46	0.44
NB	0.63	0.44	0.42
KNN	0.32	0.33	0.16

https://doi.org/10.35145/joisie.v9i2.5684



Gambar 4. Chart perbandingan akurasi.

Hasil evaluasi komparatif menunjukkan bahwa Random Forest memiliki kinerja tertinggi dengan akurasi 69,2%, hal ini mengkonfirmasi temuan (Tantyoko et al., 2023) yang menyatakan algoritma ini lebih superior dibanding algoritma klasifikasi lainnya, Naive Bayes menunjukkan performa mendekati puncak (61,2%), sedangkan K-Nearest Neighbor (KNN) mencatat akurasi terendah (20,6%).

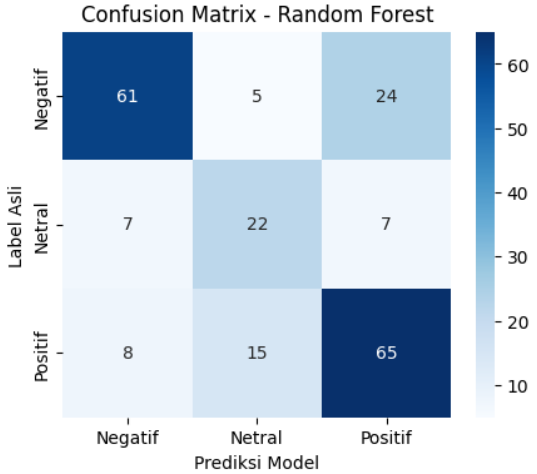
3.3 VALIDASI DAN EVALUASI

Tabel 2. Output Rata-rata 10-fold cross validation

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	65.68	67.04	65.68	65.59
NB	61.55	56.99	61.55	56.72
KNN	25.91	64.48	25.91	21.69

Tabel 2 menunjukkan bahwa Random Forest menjadi algoritma paling stabil jika dibandingkan dengan algoritma klasifikasi lainnya Hasil rata-rata akurasi 65.68% untuk Random Forest, 61.55% untuk NB, dan 25.91% untuk KNN.

Evaluasi Menggunakan Confusion Matrix pada model terakurat dan terstabil yaitu Random Forest untuk mengidentifikasi pola kesalahan klasifikasi. Seperti terlihat pada Gambar 5, model paling berhasil mengklasifikasikan kelas Positif (True Positive: 65 data dan True Negative: 61)



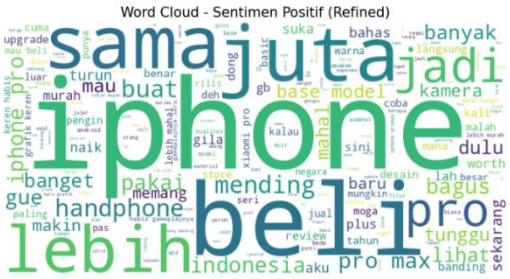
Gambar 5. Confusion Matrix Algoritma Random Forest.

Kesalahan klasifikasi mayoritas terjadi pada sentimen Negatif yang sering diprediksi sebagai Netral. Hal ini kemungkinan besar disebabkan oleh penggunaan gaya bahasa sindiran (sarkasme) atau kalimat implisit (misalnya: "Harganya murah banget ya buat ginjal") yang sulit dideteksi oleh model tanpa pemahaman konteks semantik dan ambiguitas pada kalimat tanya yang terdeteksi sebagai netral.

https://doi.org/10.35145/joisie.v9i2.5684

3.4 INTERPRETASI TOPIK SENTIMEN

Analisis topik menggunakan Word Cloud memberikan informasi berharga mengenai konteks pembicaraan. Pada visualisasi Gambar 6 dan 7, kelompok sentimen positif didominasi oleh kata kunci "mau", "bagus", dan "lebih", yang menunjukkan bahwa loyalitas konsumen masih sangat didorong oleh kualitas spesifikasi teknis dan fitur fotografi. Di sisi lain, Word Cloud sentimen negatif didominasi oleh kata "cuma", "mahal", "pajak", dan "mirip". Temuan ini memberikan sinyal penting bagi pelaku industri bahwa hambatan utama adopsi produk iPhone 17 di Indonesia adalah faktor harga yang dinilai overpriced dan persepsi kurangnya inovasi desain yang signifikan dibandingkan seri pendahulunya.



Gambar 6. Word Cloud Sentimen Positif

Gambar 7. Word Cloud Sentimen Negatif

Analisis frekuensi kata (word frequency analysis) mengungkap fokus diskusi yang kontras pada kedua polaritas sentimen. Pada klaster sentimen positif, kata kunci dominan meliputi "iPhone" (172 kemunculan), "Pro" (127), dan "beli" (90). Tingginya frekuensi kata "lebih" (69) dan "juta" (65) dalam konteks ini mengindikasikan adanya intensi pembelian yang kuat serta apresiasi komparatif terhadap keunggulan spesifikasi varian Pro, di mana konsumen cenderung mentoleransi nominal harga jutaan rupiah demi peningkatan fitur yang ditawarkan. Sebaliknya, pada sentimen negatif, kata "enggak" (102) dan "harga" (76) menjadi indikator utama resistensi konsumen. Tingginya frekuensi kata negasi "enggak" yang muncul bersamaan dengan "beli" (83) merefleksikan keraguan atau penolakan pasar akibat penetapan harga yang dinilai terlalu tinggi, Fenomena sensitivitas harga ini konsisten dengan perilaku konsumen yang dipaparkan oleh Fadillah et al. (2025), di mana variabel harga terbukti memiliki pengaruh negatif signifikan terhadap keputusan pembelian smartphone di pasar negara berkembang, mengalahkan faktor promosi. Selain itu, kemunculan kata "kayak" (53) sering kali merujuk pada perbandingan skeptis, di mana produk baru dianggap memiliki kemiripan yang signifikan dengan seri sebelumnya atau produk kompetitor, sehingga dinilai minim inovasi.

4. SIMPULAN

4.1 KESIMPULAN

Berdasarkan hasil analisis sentimen terhadap rumor peluncuran iPhone 17, penelitian ini menyimpulkan dua hal utama. Pertama, dari segi persepsi publik, sentimen positif mendominasi sebesar 43% yang didorong oleh loyalitas terhadap spesifikasi teknis dan fitur kamera, namun diimbangi oleh resistensi sentimen negatif sebesar 38,8% yang menyoroti isu harga overpriced dan stagnasi desain. Kedua, dari segi komputasi, evaluasi komparatif membuktikan bahwa Random Forest

https://doi.org/10.35145/joisie.v9i2.5684

adalah algoritma terbaik dengan akurasi 69,2% dan stabilitas validasi silang 65,68%, mengungguli Naïve Bayes dan K-Nearest Neighbor (KNN). Temuan ini mengindikasikan bahwa dominasi pasar Apple di Indonesia mulai menghadapi tantangan sensitivitas harga, sehingga narasi pemasaran perlu bergeser dari sekadar estetika menuju penekanan nilai fungsional yang konkret.

4.2 KETERBATASAN DAN SARAN

Keterbatasan dalam penelitian ini terletak pada kegagalan model menangkap nuansa sarkasme akibat penggunaan pendekatan leksikon dan TF-IDF yang bersifat statis, sehingga komentar bermuatan sindiran sering mengalami misklasifikasi . Untuk mengatasi rigiditas tersebut, penelitian selanjutnya direkomendasikan beralih pada arsitektur Deep Learning berbasis konteks seperti BERT (Bidirectional Encoder Representations from Transformers) atau IndoBERT yang mampu memahami hubungan semantik antar kata secara dua arah. Akurasi deteksi juga perlu diperkuat melalui integrasi teknik human annotation sebagai ground truth dan rekayasa fitur khusus, seperti analisis pola tanda baca dan penggunaan emoji, agar model lebih adaptif mengenali sentimen negatif yang disampaikan secara implisit

5. UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Pradita, yang telah memberikan fasilitas dan dukungan akademik sehingga penelitian ini dapat terlaksana dengan baik. Penulis juga menyampaikan apresiasi kepada seluruh pihak, termasuk rekan diskusi dan dosen pengampu, yang telah memberikan masukan berharga dalam penyusunan artikel ilmiah ini.

6. DAFTAR PUSTAKA

Bilinski, P. (2024). The Content of Tweets and the Usefulness of YouTube and Instagram in Corporate Communication. *European Accounting Review*, *8180*. https://doi.org/10.1080/09638180.2022.2084759

Bing, L. (2012). Sentiment Analysis and Opinion Mining. *Morgan & Claypool Publishers*, *May*.

Fadillah, M. R., & Batu, R. L. (2025). Pengaruh Harga , Kualitas Produk , Promosi terhadap Keputusan Pembelian Smartphone Xiaomi (Studi Kasus pada Followers Instagram Xiaomi). *El-Mal: Jurnal Kajian Ekonomi & Bisnis Islam*, *5*(1), 132–139. https://doi.org/1047467/elmal.v5i1.310

Ghanad, A. (2023). An Overview of Quantitative Research Methods. *International Journal of Multidisciplinary Research and Analysis*, *06*(08), 3794–3803. https://doi.org/10.47191/ijmra/v6-i8-52

Halabaku, E., & Bytyçi, E. (2024). Overfitting in Machine Learning : A Comparative Analysis of Decision Trees and Random Forests. *Intelligent Automation & Soft Computing*, *39*(6), 987–1006. https://doi.org/10.32604/iasc.2024.059429

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. Morgan Kaufmann.

Huwaida, S. F., Kusumawati, R., & Isnaini, B. (2024). Analisis sentimen komentar youtube terhadap pemindahan ibu kota negara menggunakan metode Naïve Bayes. *Jambura Journal of Informatics*, *6*(1), 26–39. https://doi.org/10.37905/jji.v6i1.24718

Kang, S. (2021). k -Nearest Neighbor Learning with Graph Neural Networks. *Mathematics*, *9*(830).

Kemp, S. (2025). *Digital 2025: Indonesia*. DATAREPORTAL.

Koto, F. (2017). InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs. *International Conference on Asian Language Processing*, 391–394.

Liu, S., & Jody, D. (2024). Using cross-validation methods to select time series models : Promises and pitfalls. *The British Journal of Mathematical and Statistical Psychology*, *May 2023*, 337–355. https://doi.org/10.1111/bmsp.12330

Manurung, C. E., & Mayatopani, H. (2025). Sentiment Analysis of Indonesian Society Toward the Launch of iPhone 16 Using Naive Bayes , Random Forest , and KNN Algorithms. *Jurnal Komputer, Informasi Dan Teknologi*, *5*(1), 1–13. https://doi.org/10.53697/jkomitek.v5i1.2219

Markoulidakis, I., & Markoulidakis, G. (2024). Probabilistic Confusion Matrix : A Novel Method for Machine Learning Algorithm Generalized Performance Analysis. *Technologies*, *12*(133).

Tantyoko, H., Sari, D. K., & Wijaya, A. R. (2023). Prediksi Potensial Gempa Bumi Indonesia menggunakan Metode Random Forest dan Feature Selection. *Idealis: Indonesia Journal Information System*, *6*(2), 83–89.

Wulandari, F., Haerani, E., Fikry, M., & Budianita, E. (2023). Analisis sentimen larangan penggunaan obat sirup menggunakan algoritma naive bayes classifier. *Jurnal Computer Science and Information Technology (CoSciTech)*, *4*(1), 88–96.

https://doi.org/10.35145/joisie.v9i2.5684