PENERAPAN DATA MINING UNTUK ANALISIS SENTIMEN PUBLIK TERHADAP WHIPPINK PADA PLATFORM X MENGGUNAKAN RANDOM FOREST

Cherliana, Rahma Ardhia Cahyani, Muhammad Hafiz Al Zaky, Ken Ditha Tania, Allsela Meiriza
Universitas Sriwijaya Kampus Bukit Besar
Jl. Padang Selasa No. 524, Bukit Lama, Kota Palembang, Sumatera Selatan, Indonesia
chcherliana@gmail.com

ABSTRAK

Media sosial menghasilkan volume opini publik yang sangat besar dan tidak terstruktur sehingga memerlukan pendekatan komputasional untuk dianalisis secara sistematis. Salah satu isu yang ramai diperbincangkan di platform X adalah fenomena Whippink yang berkaitan dengan penggunaan nitrous oxide (N_2O) dan memunculkan beragam tanggapan dari masyarakat. Permasalahan penelitian ini adalah bagaimana mengidentifikasi serta mengklasifikasikan kecenderungan sentimen publik terhadap fenomena tersebut berdasarkan data percakapan pada media sosial. Penelitian ini guna menganalisis sentimen masyarakat terhadap fenomena Whippink pada platform X menggunakan algoritma Random Forest. Data penelitian diperoleh melalui teknik web scraping dengan kata kunci #WHIPPINK pada periode Januari 2024 hingga Februari 2026 dan menghasilkan 500 tweet berbahasa Indonesia. Tahapan penelitian meliputi pelabelan sentimen, preprocessing teks, pembobotan fitur diimplementasikan melalui TF-IDF, data dibagi dengan rasio 80:20, serta evaluasi model diimplementasikan dengan confusion matrix dengan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian membuktikan sentimen negatif mendominasi sebesar 71,4%, sedangkan sentimen positif sebesar 28,6%. Model Random Forest dengan 400 pohon keputusan menghasilkan akurasi 85,86% dan weighted F1-score 0,85, yang menunjukkan performa klasifikasi yang baik dalam memetakan persepsi publik.

Kata kunci: analisis sentimen, Random Forest, data mining, media sosial, TF-IDF, Whippink

1. PENDAHULUAN

Big data yang bersumber dari berbagai platform digital merupakan fenomena yang muncul akibat meningkatnya aktivitas masyarakat di ruang digital sehingga mendorong pertumbuhan data dalam jumlah yang sangat besar [1]. Salah satu kontributor utama big data adalah media sosial, yang penggunanya di Indonesia mencapai 143 juta orang berdasarkan laporan Statista yang dimuat pada laman Tempo. Data Similarweb periode 1 Agustus 2025 mengindikasikan bahwa platform media sosial yang memiliki tingkat penggunaan tertinggi di Indonesia antara lain WhatsApp, Instagram, dan X (Twitter). X sebagai platform yang menjadi ruang percakapan publik, termasuk untuk mengikuti berita terbaru, menyampaikan opini, serta berdiskusi mengenai berbagai topik. Karakteristiknya yang bersifat real-time menjadikan platform ini sebagai sumber informasi instan yang banyak diandalkan oleh masyarakat [15].

Tingginya intensitas interaksi pengguna pada media sosial tersebut mencerminkan peran media sosial sebagai ruang ekspresi emosional kolektif, di mana pengguna dapat menyampaikan berbagai tanggapan seperti dukungan, kritik, maupun opini terhadap suatu peristiwa atau isu tertentu. Interaksi yang berlangsung secara masif tersebut menghasilkan kumpulan data opini publik yang sangat besar dan beragam. Namun demikian, data dari media sosial umumnya bersifat tidak terorganisir dan terus bertambah secara dinamis, sehingga proses pengolahan dan analisis data secara manual menjadi tidak efisien serta memakan waktu yang lama. Oleh sebab itu, dibutuhkan pendekatan analisis yang mampu mengelola dan mengklasifikasikan opini publik secara sistematis, salah satunya melalui metode analisis sentimen [2].

Salah satu isu yang memicu berbagai tanggapan publik di media sosial adalah fenomena penggunaan whipping atau yang dikenal sebagai Whippink. Perbincangan mengenai topik ini menjadi semakin meningkat setelah munculnya kasus meninggalnya salah satu influencer di Indonesia yang diduga berkaitan dengan penggunaan produk tersebut, sebagaimana diberitakan oleh Divisi Humas Polri melalui laman TribraTANews Polri. Peristiwa tersebut memicu diskusi yang luas pada berbagai platform media sosial, khususnya X, yang menjadi media penampung untuk masyarakat dalam menyampaikan beragam respons seperti kekhawatiran, kritik, dukungan, maupun perdebatan terkait dampak penggunaan nitrous oxide (N_2O) [14]. Kondisi ini menunjukkan bahwa media sosial dapat menjadi sumber data yang potensial untuk mengidentifikasi kecenderungan opini masyarakat terhadap fenomena yang sedang berkembang.

Dalam konteks penelitian sebelumnya, analisis sentimen telah banyak dimanfaatkan untuk mengidentifikasi kecenderungan opini masyarakat terhadap berbagai isu kesehatan yang berkembang di media sosial. Misalnya, penelitian mengenai sentimen kesehatan mental pemuda di platform X menunjukkan bila media sosial berperan sebagai media ekspresi emosional sekaligus sumber informasi yang mempengaruhi pembentukan opini publik [5]. Selain pendekatan machine learning konvensional, sejumlah

penelitian juga memanfaatkan metode deep learning untuk menaikkan kemampuan model dalam memahami pokok bahasan bahasa pada data teks media sosial. Pendekatan deep learning dinilai mampu mempelajari representasi fitur yang lebih majemuk sehingga akurasi dapat ditingkatkan dalam klasifikasi sentimen pada data teks yang tidak terstruktur [16].

Meskipun demikian, penelitian yang secara khusus membahas fenomena penyalahgunaan nitrous oxide (N_2O) yang dikenal sebagai Whippink dalam konteks opini publik di Indonesia masih relatif terbatas. Padahal, fenomena tersebut termasuk salah satu topik yang sedang diperbincangkan di media sosial. Keterbatasan kajian tersebut menunjukkan adanya kebutuhan untuk melakukan penelitian yang mampu memetakan sentimen masyarakat terhadap fenomena Whippink secara lebih sistematis.

Berdasarkan permasalahan tersebut, studi ini berfokus dalam menganalisis dan mengklasifikasikan sentimen opini publik terhadap fenomena Whippink di media sosial X dengan pendekatan machine learning algoritma Random Forest. Berpatokan dengan kemampuannya dalam mencpitakan performa klasifikasi yang stabil melalui pendekatan ensemble learning. Algoritma ini mampu mengurangi risiko overfitting, menangani noise pada data, serta memberikan estimasi tingkat kepentingan fitur yang berkontribusi dalam proses klasifikasi [3]. Selain itu, kinerja model Random Forest juga dipengaruhi oleh konfigurasi parameter seperti jumlah tree dan kedalaman tree, di mana pengaturan parameter yang optimal dapat meningkatkan stabilitas serta kemampuan model dalam menangkap pola data tanpa menyebabkan overfitting [4]. Melalui pendekatan tersebut, penelitian ini diharapkan mampu memberikan gambaran distribusi sentimen publik terhadap fenomena Whippink di media sosial secara lebih sistematis dan terukur, sekaligus memberikan andil dalam pemanfaatan metode machine learning dalam menganalisis data teks tidak terstruktur pada media sosial.

2. TINJAUAN PUSTAKA

2.1. Preprocessing

Dilakukan untuk mentransformasikan data mentah menjadi data yang lebih tertata dan digunakan pada proses pemodelan berbasis *machine learning* [4].

a. *Data Cleaning*
   Pengeleminasian-unsur yang tidak tepat dalam data teks, di antaranya *URL*, emoji, angka, simbol, tanda baca yang berlebihan, *mention*, dan *hashtag*.
b. Case Folding
   Pengkonversian segala karakter huruf dalam teks menjadi huruf kecil (*lowercase*) guna menjaga konsistensi dan keseragaman penulisan.
c. Tokenizing
   Pemecahan unit kata menjadi lebih kecil sehingga setiap kata dapat diperlakukan sebagai elemen tersendiri dalam analisis teks.
d. Stopword Removal
   Penghapusan kata-kata umum dan kurang berkaitan terhadap analisis sentimen, contohnya kata penghubung serta kata depan.
e. Stemming
   Kata-kata yang memiliki imbuhan dinormalisasi dengan menghilangkan imbuhan tersebut sehingga diperoleh bentuk kata dasar.

2.2. Pembobotan Kata (TF-IDF)

Pengukuran tingkat relevansi antara kueri dengan dokumen, sekaligus mengevaluasi hubungan antar dokumen. Pembobotan vektor TF-IDF pada kata kunci kueri terhadap kueri itu sendiri dilakukan untuk menghasilkan nilai kesamaan (similarity score) [8].

2.3. Algoritma Random Forest

Bersifat *ensemble* yang membentuk kumpulan pohon keputusan melalui mekanisme teknik *bootstrap sampling* dan pemilihan subset fitur secara acak pada setiap pohon, bertujuan untuk meningkatkan akurasi serta mengurangi *overfitting* dalam klasifikasi sentimen teks. Penelitian ini menerapkan algoritma Random Forest untuk mengidentifikasi dan mengkategorikan sentimen masyarakat terhadap tagar #WHIPPINK berdasarkan representasi fitur yang diperoleh dari pembobotan TF-IDF [9].

2.4. Evaluasi Model Klasifikasi

Bekerja dengan mendayagunakan confusion matrix. Berdasarkan *confusion matrix* yang diperoleh, sejumlah metrik untuk menilai kinerja model dapat dihitung, melingkupi *accuracy*, *precision*, *recall*, serta *F1-score*. *Accuracy* merepresentasikan proporsi prakiraan yang tepat dari keseluruhan data pengujian, sedangkan *precision* menunjukkan tingkat keabsahan model dalam mengklasifikasikan suatu kelas. *Recall* menggambarkan kemampuan model dalam mengidentifikasi seluruh instance yang benar terhadap suatu kelas. Sementara itu, *F1-score*, rata-rata antara *precision* dan *recall* yang dimanfaatkan untuk mengevaluasi kinerja model secara lebih proporsional, khususnya pada kondisi ketidakseimbangan data [11].

2.5. Visualisasi

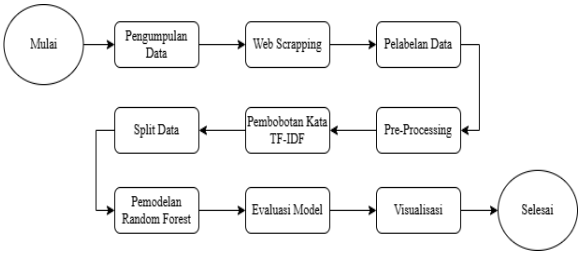
Distribusi sentimen divisualisasikan melalui grafik batang (bar chart) yang menunjukkan jumlah keseluruhan tweet dengan kategori sentimen positif dan negatif. Lebih lanjut, hasil evaluasi model ditampilkan dalam bentuk *heatmap confusion matrix* guna memperjelas perbandingan antara label aktual dan hasil prediksi model. Dengan demikian, kesalahan klasifikasi serta kemampuan prediksi model dapat dikenali secara visual [11].

3. METODE PENELITIAN

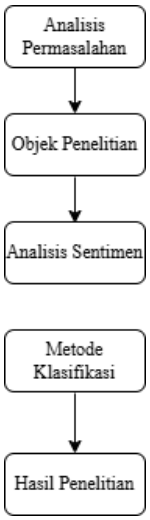
Studi ini memanfaatkan pendekatan deskriptif kuantitatif yang dikombinasikan dengan kajian literatur guna menguraikan hasil penelitian secara objektif.

3.1. Alur dan Kerangka Pemikiran Penelitian

Pendekatan klasifikasi dalam sentimen publik terhadap tagar #WHIPPINK pada platform X diterapkan dengan algoritma Random Forest. Tahapan penelitian disusun secara sistematis, meliputi proses pengumpulan data melalui teknik web scraping, pemberian label sentimen, tahap pra-pemrosesan data, pembobotan fitur menggunakan TF-IDF, dataset dipecah menjadi set pelatihan dan pengujian, pengembangan model klasifikasi, evaluasi kinerja model, serta penyajian hasil analisis dalam bentuk visualisasi ditampilkan pada gambar berikut.



Gambar 1. Alur penelitian



Gambar 2. Kerangka pemikiran

3.2. Teknik Pengumpulan Data

Platform X dimanfaatkan sebagai sumber utama dalam memperoleh data penelitian. Data yang dikumpulkan terdiri dari komentar atau *tweet* pengguna yang berisi opini dan tanggapan terkait fenomena #WHIPPINK. Seluruh komentar yang dipublikasikan oleh pengguna Platform X menjadi populasi dalam penelitian ini, dengan pengambilan data dilakukan menggunakan tagar #WHIPPINK sebagai kata kunci utama. Rentang waktu pengambilan data ditetapkan mulai 01 Januari 2024 hingga Februari 2026 dengan tujuan menangkap dinamika opini publik selama periode berlangsungnya fenomena tersebut. Berdasarkan batasan waktu tersebut, diperoleh sebanyak 500 tweet yang relevan sebagai sampel penelitian. Semua data yang terkumpul disimpan dalam format file .csv agar mempermudah proses *pre-processing* dan analisis lebih lanjut.

3.3. Pelabelan Data

Tahap pelabelan data memiliki peran krusial karena kualitas label sangat memengaruhi kinerja model. Sehingga data teks diberi label sentimen untuk pelatihan dan pengujian yang diperoleh dari platform X terkait #WHIPPINK diberi label berdasarkan polaritas sentimen, yaitu positif dan negatif, sehingga dapat diproses oleh algoritma Random Forest melalui pendekatan *supervised learning*.

3.4. Preprocessing

Berperan ketika mengolah data mentah dari media sosial agar menjadi sistematis dan dapat diimplementasikan dalam proses pemodelan machine learning. [12].

a. *Data Cleaning*
   Penghapusan URL, simbol, angka, tagar, mention, emoji, dan karakter khusus yang muncul pada teks Platform X.
b. *Case Folding*
   Tahapan ini diperlukan karena data teks dari Platform X menunjukkan ketidakkonsistenan penulisan, misalnya “Whippink”, “WHIPPink”, dan “whippink” yang memiliki arti serupa, tetapi diperlakukan berbeda oleh komputer tanpa adanya proses normalisasi.
c. *Stopword Removal*
   Kata-kata seperti “dan”, “yang”, “di”, serta “itu” dalam teks Bahasa Indonesia tergolong tinggi, tetapi tidak memberikan nilai informasi yang signifikan bagi model klasifikasi, sehingga penghapusannya dilakukan guna menyederhanakan fitur dan menekankan pada kata-kata yang lebih bermakna.
d. *Tokenizing*
   Proses ini penting karena teks pada postingan dan komentar di Platform X biasanya berupa data tidak terstruktur yang terdiri dari dominasi kata dan struktur yang kompleks, sehingga perlu disederhanakan menjadi satuan kata yang lebih terorganisir. Token yang dihasilkan akan menjadi dasar dalam ekstraksi fitur numerik.
e. *Stemming*
   Proses ini penting karena Bahasa Indonesia memiliki beragam afiks seperti awalan, sisipan, dan akhiran yang dapat menyebabkan kata yang secara semantik sama menjadi dianggap berbeda oleh algoritma pembelajaran mesin.

4. HASIL DAN PEMBAHASAN

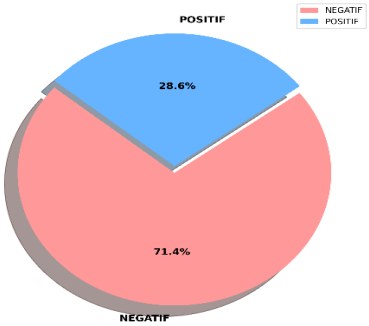
Ringkasan hasil penelitian terhadap analisis sentimen terhadap fenomena *whippink* pada platform X. Pada tahap ini, kumpulan tweet berbahasa Indonesia yang telah diperoleh diolah dan dianalisis untuk mengidentifikasi persepsi masyarakat Indonesia terhadap fenomena tersebut.

4.1. Persiapan Data

Sebelum dilakukan pemodelan menggunakan algoritma Random Forest, data terlebih dahulu melalui beberapa tahap persiapan, yaitu pelabelan sentimen dan proses *pre-processing* teks.

4.2. Pelabelan Data

Pengklasifikasian label pada data diklasifikasian ke dua kategori, yaitu positif dan negatif. Tweet yang mengandung dukungan, pembelaan, atau tanggapan yang bersifat konstruktif terhadap fenomena Whippink diberi label positif. Sebaliknya tweet yang mengandung kritik, kekhawatiran, penolakan, atau sentimen negatif lainnya diberi label negatif. Distribusi hasil pelabelan menunjukkan bahwa dari total 500 tweet, sebanyak 143 termasuk dalam kategori positif dengan persentase 28.6% dan 357 termasuk dalam kategori negatif dengan persentase 71.4%. Adapun gambaran persentase hasil pelabelan sebagai berikut.



Gambar 3. Persentase sentimen publik terhadap Whippink

4.3. Data Cleaning

Hasil penerapan pada Gambar 4 menunjukkan pada baris ke-7 "@AndyHusky666 ketibhan whip pink" menjadi "ketiban whip pink".

full_text	clean_text	label
WASPADA PENYALAHGUNAAN WHIP PINK (N?O) Hai Enc...	WASPADA PENYALAHGUNAAN WHIP PINK NO Hai Encik ...	positif
@gatauahdark @itsquaileggs balik ke ketikan lu...	balik ke ketikan lu sebelumnya bang bukan kore...	positif
@em0_is_n0t_de4d Sedih lihat bg yebe kebangun ...	Sedih lihat bg yebe kebangun mimpiin lula tadi...	positif
Sedih lihat bg yebe kebangun mimpiin lula tadi...	Sedih lihat bg yebe kebangun mimpiin lula tadi...	positif
@ramdaniahmad36 @tanyarffes Lula Lahfah adalah...	Lula Lahfah adalah selebgram berusia tahun ya...	positif
ketiban whip pink	ketiban whip pink	negatif
@AndyHusky666 ketiban whip pink	ketiban whip pink	negatif
ketiban whip pink anjingggggg	ketiban whip pink anjingggggg	negatif
NGAKAK BGT KOCAK KETIBAN WHIP PINK	NGAKAK BGT KOCAK KETIBAN WHIP PINK	negatif
ni org abis ngisep whip pink apa aibon ajg?	ni org abis ngisep whip pink apa aibon ajg	negatif

Gambar 4. *Data cleaning*

4.4. *Case Folding*

Pada gambar 5 ditampilkan transformasi seluruh karakter teks menjadi huruf kecil.

clean_text	Case_Folding	label
WASPADA PENYALAHGUNAAN WHIP PINK NO Hai Encik ...	waspada penyalahgunaan whip pink no hai encik ...	positif
balik ke ketikan lu sebelumnya bang bukan kore...	balik ke ketikan lu sebelumnya bang bukan kore...	positif
Sedih lihat bg yebe kebangun mimpiin lula tadi...	sedih lihat bg yebe kebangun mimpiin lula tadi...	positif
Sedih lihat bg yebe kebangun mimpiin lula tadi...	sedih lihat bg yebe kebangun mimpiin lula tadi...	positif
Lula Lahfah adalah selebgram berusia tahun ya...	lula lahfah adalah selebgram berusia tahun ya...	positif
ketiban whip pink	ketiban whip pink	negatif
ketiban whip pink	ketiban whip pink	negatif
ketiban whip pink anjingggggg	ketiban whip pink anjingggggg	negatif
NGAKAK BGT KOCAK KETIBAN WHIP PINK	ngakak bgt kocak ketiban whip pink	negatif
ni org abis ngisep whip pink apa aibon ajg	ni org abis ngisep whip pink apa aibon ajg	negatif

Gambar 5. *Case folding*

4.5. *Stopword Removal*

Proses ini menggunakan daftar *stopword* Bahasa Indonesia yang selanjutnya diperluas dengan menambahkan *stopword* yang bersifat kontekstual terhadap media sosial, seperti istilah umum pada Twitter serta variasi bahasa informal. Selain itu, penelitian ini juga mempertahankan kata-kata negasi seperti *tidak, bukan, belum, gak,* dan *jangan* agar tidak terhapus selama tahap *filtering*. Hasil stopword removal disimpan pada kolom *Stopword_Removal* pada gambar 6 dibawah ini.

Case_Folding	Stopword_Removal	label
waspada penyalahgunaan whip pink no hai encik ...	waspada penyalahgunaan no hai encik puan produ...	positif
balik ke ketikan lu sebelumnya bang bukan kore...	ketikan bang bukan korelasi ga bukti kuat kl p...	positif
sedih lihat bg yebe kebangun mimpiin lula tadi...	sedih lihat bg yebe kebangun mimpiin lula mutu...	positif
sedih lihat bg yebe kebangun mimpiin lula tadi...	sedih lihat bg yebe kebangun mimpiin lula mutu...	positif
lula lahfah adalah selebgram berusia tahun ya...	lula lahfah selebgram berusia meninggal mendad...	positif
ketiban whip pink	ketiban	negatif
ketiban whip pink	ketiban	negatif
ketiban whip pink anjingggggg	ketiban anjingggggg	negatif
ngakak bgt kocak ketiban whip pink	ngakak kocak ketiban	negatif
ni org abis ngisep whip pink apa aibon ajg	ni org abis ngisep aibon ajg	negatif

Gambar 6. *Stopword removal*

4.6. *Tokenizing*

Implementasi tokenizing dilakukan menggunakan metode sederhana berbasis fungsi split (), yang memisahkan teks berdasarkan spasi dan menghasilkan daftar kata dalam bentuk list. Hasil tokenisasi disimpan pada kolom *Tokenizing* untuk mempertahankan transparansi proses transformasi data yang ditampilkan pada gambar 7.

Stopword_Removal	Tokenizing_Display	label
waspada penyalahgunaan no hai encik puan produ...	['waspada', 'penyalahgunaan', 'no', 'hai', 'en...	positif
ketikan bang bukan korelasi ga bukti kuat kl p...	['ketikan', 'bang', 'bukan', 'korelasi', 'ga',...	positif
sedih lihat bg yebe kebangun mimpiin lula mutu...	['sedih', 'lihat', 'bg', 'yebe', 'kebangun', '...	positif
sedih lihat bg yebe kebangun mimpiin lula mutu...	['sedih', 'lihat', 'bg', 'yebe', 'kebangun', '...	positif
lula lahfah selebgram berusia meninggal mendad...	['lula', 'lahfah', 'selebgram', 'berusia', 'me...	positif
ketiban	['ketiban']	negatif
ketiban	['ketiban']	negatif
ketiban anjingggggg	['ketiban', 'anjingggggg']	negatif
ngakak kocak ketiban	['ngakak', 'kocak', 'ketiban']	negatif
ni org abis ngisep aibon ajg	['ni', 'org', 'abis', 'ngisep', 'aibon', 'ajg']	negatif

Gambar 7. *Tokenizing*

4.7. Stemming

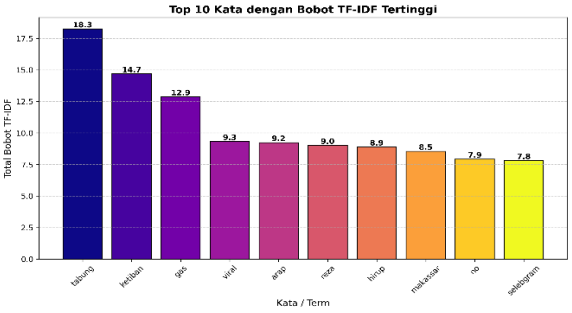
Diterapkan dengan menggunakan pustaka Sastrawi, yang dirancang khusus untuk pemrosesan teks berbahasa Indonesia. Proses dilakukan dengan menggabungkan token menjadi kalimat sementara, kemudian menerapkan algoritma stemming untuk memperoleh bentuk dasar kata, dan selanjutnya mengembalikannya ke dalam bentuk daftar token agar konsisten dengan tahap sebelumnya. Berdasarkan gambar 8, terlihat bahwa kata-kata berimbuhan berhasil direduksi ke bentuk dasarnya tanpa menghilangkan makna utama. Tahap ini membantu mengurangi redundansi fitur dan meningkatkan efektivitas pembobotan kata pada proses TF-IDF.

Tokenizing_Str	Stemming_Str	label
waspada, penyalahgunaan, no, hai, encik, puan,...	waspada, penyalahgunaan, no, hai, encik, puan,...	positif
ketikan, bang, bukan, korelasi, ga, bukti, kua...	keti, bang, bukan, korelasi, ga, bukti, kuat, ...	positif
sedih, lihat, bg, yebe, kebangun, mimpiin, lul...	sedih, lihat, bg, yebe, bangun, mimpiin, lula,...	positif
sedih, lihat, bg, yebe, kebangun, mimpiin, lul...	sedih, lihat, bg, yebe, bangun, mimpiin, lula,...	positif
lula, lahfah, selebgram, berusia, meninggal, m...	lula, lahfah, selebgram, usia, tinggal, dadak,...	positif
ketiban	ketiban	negatif
ketiban	ketiban	negatif
ketiban, anjingggggg	ketiban, anjingggggg	negatif
ngakak, kocak, ketiban	ngakak, kocak, ketiban	negatif
ni, org, abis, ngisep, aibon, ajg	ni, org, abis, ngisep, aibon, ajg	negatif

Gambar 8. *Stemming*

4.8. Hasil Pembobotan Kata (TF-IDF)

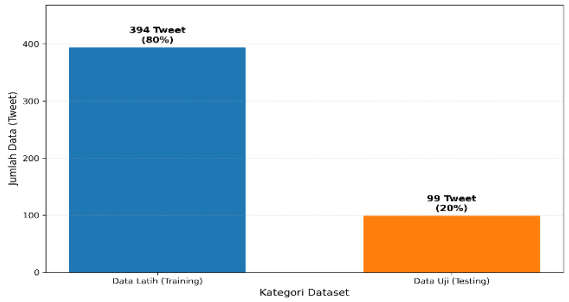
Berdasarkan hasil perankingan, diperoleh 10 kata dengan bobot TF-IDF tertinggi, yaitu *tabung* (18,3), *ketiban* (14,7), *gas* (12,9), diikuti oleh *viral*, *arap*, *reza*, *hirup*, *makassar*, *no*, dan *selebgram* dengan rentang skor 7,8–9,3. Tingginya bobot kata seperti *tabung*, *gas*, dan *hirup* menunjukkan bahwa percakapan publik berfokus pada aspek penggunaan nitrous oxide, sedangkan kemunculan kata *viral*, *selebgram*, serta nama individu tertentu mengindikasikan pengaruh pemberitaan dan figur publik dalam membentuk opini masyarakat. Hasil pembobotan TF-IDF ini tidak hanya berfungsi sebagai representasi fitur numerik untuk proses klasifikasi menggunakan Random Forest, tetapi juga memberikan gambaran tematik mengenai isu-isu utama yang dominan dalam diskusi publik di platform X.



Gambar 9. TF-IDF

4.9. Hasil Split Data

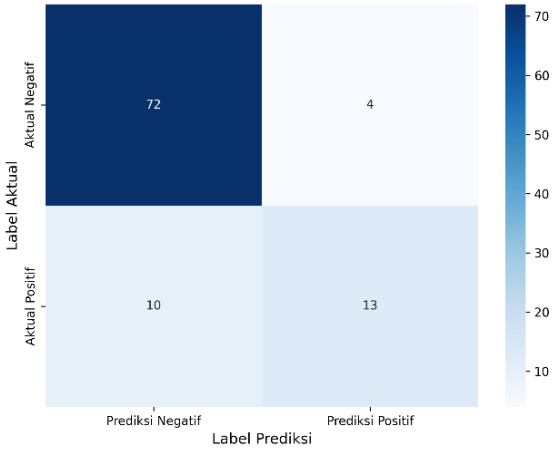
Dataset yang telah melewati *pre-processing* dipartisi sebagai data latih dan data uji dengan rasio 80:20. Dari total 493 *tweet* dalam pemodelan, sebanyak 394 dijadikan data latih, sedangkan 99 sebagai data uji, sebagaimana ditampilkan pada Gambar 10.



Gambar 10. Proporsi pembagian dataset

4.10. Pemodelan Random Forest

Hasil evaluasi model menunjukkan bahwa algoritma Random Forest dengan tepat mendapatkan akurasi sebesar 85,86. Hasil confusion matrix pada Gambar 11 menunjukkan bahwa dari 76 data aktual negatif, sebanyak 72 data berhasil diprediksi dan 4 data salah diklasifikasikan sebagai positif. Sementara itu, dari 23 data aktual positif, sebanyak 13 data berhasil diprediksi dan 10 data salah diklasifikasikan sebagai negatif.



Gambar 11. Confusion matrix

Mengacu pada Tabel 1, performa klasifikasi pada kelas negatif menghasilkan *precision* 0,88, *recall* 0,95, dan *F1-score* 0,91. Di sisi lain, kelas positif memperoleh nilai precision 0,76, recall 0,57, serta F1-score 0,65. Nilai weighted average F1-score sebesar 0,85 mengindikasikan bahwa model memiliki performa yang cukup baik secara keseluruhan, meskipun kemampuan dalam mendeteksi sentimen positif masih lebih rendah disebabkan oleh dominasi sentimen negatif dalam distribusi dataset, sehingga performa model dalam mendeteksi pola pada kelas mayoritas menjadi lebih baik.

Tabel 1. performa klasifikasi

Sentiment	***Precision***	***Recall***	***F1-score***	***Support***
Negatif	0,88	0,95	0,91	76
Positif	0,76	0,57	0,65	23
Accuracy			0,86	99
Macro avg	0,82	0,76	0,78	99
Weighted avg	0,85	0,86	0,85	99
Total Accuracy	**0,8586 = 85,86%**			

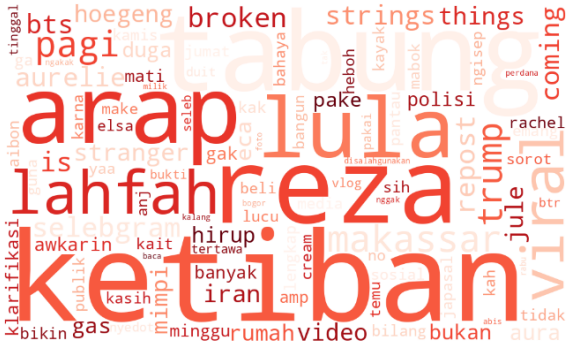
4.11. Visualisasi *Word Cloud*

Kata yang mendominasi dalam masing-masing kategori sentimen berdasarkan pembobotan TF-IDF ditampilkan melalui *word cloud*. Proses ini bertujuan memberikan representasi visual mengenai tema dan persepsi publik terhadap fenomena #WHIPPINK. Hasil visualisasi sentimen positif terdapat pada Gambar 12, kata-kata yang dominan antara lain *sah*,

milik, *utama*, *oksigen*, *kuliner*, *ringan*, dan *kenal*. Kemunculan kata-kata tersebut menunjukkan adanya persepsi bahwa sebagian masyarakat memandang penggunaan nitrous oxide dalam konteks tertentu, seperti kuliner, sebagai sesuatu yang legal atau wajar selama digunakan sesuai fungsi utamanya. Hal ini mengindikasikan adanya opini yang bersifat defensif atau pembenaran terhadap penggunaan produk tersebut.

Gambar 12. *Word cloud* sentimen positif

Sementara itu, sentimen negatif diterpkan juga *word cloud* sebagaimana ditampilkan pada gambar 13. Kata-kata didominasi mengarah pada dampak kesehatan, risiko, serta kekhawatiran publik. Perbedaan distribusi kata antara sentimen positif dan negatif menunjukkan adanya polarisasi opini masyarakat terhadap fenomena Whippink. Secara umum, visualisasi ini memperkuat hasil analisis sebelumnya bahwa sentimen negatif lebih dominan, namun tetap terdapat kelompok masyarakat yang memberikan respons positif atau netral terhadap isu tersebut.



Gambar 13. *Word cloud* sentimen negatif

5. KESIMPULAN DAN SARAN

Berdasarkan temuan penelitian, penerapan data mining dengan pendekatan analisis sentimen menggunakan algoritma Random Forest mampu secara efektif dan andal mengklasifikasikan opini publik terkait #WHIPPINK di Platform X. Dari 500 tweet yang dikumpulkan pada periode Januari 2024 hingga Februari 2026, distribusi sentimen menunjukkan dominasi opini negatif sebesar 71,4%, sedangkan sentimen positif sebesar 28,6%, yang mengindikasikan bahwa mayoritas masyarakat cenderung menyoroti aspek risiko dan dampak kesehatan dari penggunaan nitrous oxide. Rangkaian proses preprocessing seperti data cleaning, case folding, penghapusan stopword, tokenisasi, dan stemming terbukti efektif dalam meningkatkan kualitas fitur sebelum tahap pembobotan menggunakan TF-IDF. Model Random Forest dengan 400 pohon keputusan menghasilkan akurasi sebesar 85,86% serta weighted F1-score sebesar 0,85, yang merepresentasikan bahwa performa klasifikasi yang baik, meskipun kemampuan deteksi sentimen positif masih lebih rendah akibat ketidakseimbangan data. Visualisasi word cloud turut memperkuat temuan bahwa percakapan negatif lebih menonjol dibandingkan positif. Dengan demikian, studi ini mengindikasikan bahwa penerapan *machine learning* berbasis algoritma Random Forest dapat dimanfaatkan sebagai metode analisis opini publik yang andal untuk mendukung proses pengambilan keputusan serta perencanaan strategi komunikasi kesehatan. Penelitian selanjutnya, disarankan penggunaan dataset yang lebih besar dan seimbang, penambahan kategori sentimen seperti netral, serta perbandingan dengan algoritma lain agar diperoleh model dengan performa lebih optimal dan kemampuan generalisasi yang lebih baik.

DAFTAR PUSTAKA

[1] S. W. Sri and S. Sya'roni, "Analisis Sentimen Dalam Big Data: Studi Kasus Pada Media Sosial," *Innov. Manag. Technol. Educ. Bus.*, vol. 1, no. 01, 2025.

[2] N. M. Damayanti, I. D. Ariningtyas, M. I. A. Icham, and A. P. Sari, "Analisis Sentimen Publik Pada Tagar# Btscomeback Di Platform X Menggunakan Indobertweet," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, 2025.

[3] D. C. Novitasari and R. R. Putri, "Analisis Sentimen Cuitan X terhadap Pendaki Asing Gunung Rinjani Menggunakan Algoritma SVM dan Random Forest," *INTEGER J. Inf. Technol.*, vol. 10, no. 2, 2025.

[4] R. Wijanarko, D. E. Ratnawati, and P. P. Adikara, "Analisis Sentimen Dampak Perkembangan Artificial Intelligence (AI) pada Media Sosial X/Twitter Menggunakan Metode Random Forest," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 8, no. 5, 2024.

[5] A. Agustin, J. Junadhi, F. Zoromi, and P. Kudadiri, "Analisis Sentimen Kesehatan Mental Pemuda di Media Sosial Menggunakan Deep Learning," *DEVICE J. Inf. Syst. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 816–830, 2025.

[6] R. Hidayat, I. Imran, and R. Ramadhan, "Peran Media Sosial Dalam Mengkonstruksi Opini Publik Terkait Kebijakan Pemerintah: Studi Kasus Wacana Publik Tahun 2025," *CORE J. Commun. Res.*, pp. 64–74, 2025.

[7] M. P. Pratiwi, E. Y. Natalia, and R. Fauzi, "Desain Sistem Big Data untuk Deteksi Pola Perilaku Konsumen dengan Menggunakan

Algoritma Machine Learning," *J. Desain Dan Anal. Teknol.*, vol. 5, no. 1, pp. 12–19, 2026.

[8] P. P. D. D. Wisata, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 1, p. 171, 2024.

[9] D. Ramadhansyah, A. Asrofiq, and Y. Yunefri, "Analisis sentimen ulasan penumpang maskapai penerbangan di Indonesia dengan algoritma Random Forest dan KNN," *Zo. J. Sist. Inf.*, vol. 6, no. 2, pp. 287–297, 2024.

[10] A. D. Sugiarto and M. S. Utomo, "Analisis Sentimen Ulasan Pengguna Bca Mobile Di Google Play Store Menggunakan Metode Decision Tree," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 5, pp. 8004–8009, 2025.

[11] R. M. Dinata, M. Marhaeni, K. Atmadja, E. Rayhana, V. Hadi, and U. Al Kaf, "Analisis Komprehensif Kinerja Model Klasifikasi Sentimen: Evaluasi Lintas Metrik pada Dataset Tweet Film Bahasa Indonesia: Data Sentimen Analitik dari Tweteer (X) Tentang Film Berbahasa Indoensia," *J. REKAYASA Inf.*, vol. 14, no. 1, pp. 38–47, 2025.

[12] A. Pangestu and R. T. Iswahyudi, "Pengaruh Data Preprocessing Terhadap Performa Regresu Linier Dalam Prediksi Saham," *IKRA-ITH Inform. J. Komput. dan Inform.*, vol. 9, no. 2, pp. 204–210, 2025.

[13] Divisi Humas Polri, "Polisi ungkap hasil pemeriksaan whip pink terkait kematian Lula Lahfah," Tribratanews Polri, Feb. 5, 2026. [Online]. Available: https://tribratanews.polri.go.id/blog/nasional-3/polisi-ungkap-hasil-pemeriksaan-whip-pink-terkait-kematian-lula-lahfah-98483

[14] Tempo.co, "Daftar media sosial yang paling banyak digunakan di Indonesia," Aug. 19, 2025. [Online]. Available: https://www.tempo.co/digital/daftar-media-sosial-yang-paling-banyak-digunakan-di-indonesia-2060584

[15] P. Ayuningtiyas, K. D. Tania, and W. K. Sari, "Sentiment-Based Knowledge Discovery pada Aplikasi iPusnas Menggunakan Metode Machine Learning dan Deep Learning," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2486–2497, 2025.