

IMPLEMENTASI METODE NAIVE BAYES DAN SVM DALAM ANALISIS SENTIMEN MASYARAKAT INDONESIA TERKAIT FENOMENA KABUR AJA DULU PADA MEDIA SOSIAL X

Taufiqu Rahman^{1*}, Sejati Waluyo²

^{1,2} Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: ¹*2111500480@student.budiluhur.ac.id, ²sejati.waluyo@budiluhur.ac.id

(* : corresponding author)

Abstrak-Fenomena "Kabur Aja Dulu" sempat menjadi perbincangan di media sosial khususnya X, fenomena ini banyak diperbincangkan baik di kalangan masyarakat Indonesia di dalam negeri maupun yang sedang berada di luar wilayah Indonesia. Perbincangan terkait fenomena ini merefleksikan ketidakpuasan terkait kondisi ekonomi, sosial, pendidikan, atau kebijakan yang diterapkan pemerintah Indonesia. Penelitian ini berfokus pada analisis sentimen masyarakat Indonesia yang timbul sebagai respons terkait fenomena tersebut. Sumber data yang digunakan berupa unggahan pengguna media sosial X dikumpulkan dengan metode *crawling* berjumlah 200 sampel. Dilakukan *Preprocessing* untuk membersihkan data, kemudian data tersebut dipisahkan dalam tiga kelas sentimen: positif, negatif, dan netral melalui leksikon serta melalui dampingan pakar bahasa, dihasilkan 200 data label otomatis dan 200 data label pakar. Penelitian ini melakukan tiga rasio pengujian dengan rasio 20%, 40%, dan 60%, ekstraksi fitur TF-IDF digunakan dengan normalisasi L2, kemudian menerapkan dua algoritma, yaitu *Naive Bayes* dan *Support Vector Machine* (SVM). Pengujian memperoleh hasil dengan tingkat akurasi tertinggi untuk klasifikasi *Naive Bayes* dengan akurasi 73,2% pada label otomatis dan 60% pada label pakar pada rasio uji 20%. Untuk label otomatis, presisi tercatat 68,1%, *recall* 73,17%, dan *f1-score* 70,47%, sedangkan label pakar menunjukkan presisi 61,67%, *recall* 60,0%, dan *f1-score* 60,59%. SVM mencapai akurasi tertinggi 56,8% pada label otomatis dengan rasio uji 60%, disertai presisi 70,79%, *recall* 56,79%, dan *f1-score* 62,94%, dan akurasi sebesar 52,5% pada label pakar dengan rasio 20%, dengan presisi 53,39%, *recall* 52,5%, dan *f1-score* 52,59%. Meskipun demikian, keduanya sulit memprediksi sentimen netral pada label otomatis akibat ketidakseimbangan jumlah sentimen. Menunjukkan bahwa jumlah dataset dan variasi data sangat berpengaruh terhadap akurasi yang dihasilkan. Penelitian ini menyimpulkan bahwa sentimen masyarakat Indonesia terkait fenomena "Kabur Aja Dulu" cenderung negatif, diharapkan hasil penelitian ini dapat memberikan wawasan mendalam terkait dinamika sentimen masyarakat di ruang digital, serta dapat menjadi referensi kepada pembuat kebijakan guna meningkatkan kepuasan masyarakat Indonesia.

Kata Kunci: Kabur Aja Dulu, Analisis Sentimen, TF-IDF Normalisasi L2, Naive Bayes, Support Vector Machine

IMPLEMENTATION OF NAIVE BAYES AND SVM METHODS FOR SENTIMENT ANALYSIS OF THE INDONESIAN PUBLIC REGARDING THE "KABUR AJA DULU" PHENOMENON ON SOCIAL MEDIA X

Abstract - The "Kabur Aja Dulu" phenomenon has become a significant topic of discussion on social media, particularly X, engaging both Indonesians at home and abroad. This widespread conversation reflects a deep dissatisfaction with the country's economic, social, and educational conditions, as well as with government policies. This study focuses on analyzing the Indonesian public sentiment that has emerged in response to this phenomenon. The research utilized 200 data samples collected by crawling social media posts from X users. Following a data cleaning process, the samples were categorized into three sentiment classes: positive, negative, and neutral. A key aspect of the methodology was the use of two distinct labeling methods—one automated and lexicon-based, and the other manual and conducted by language experts—resulting in 200 automatically labeled and 200 expert-labeled data points. The study employed Naive Bayes and Support Vector Machine (SVM) algorithms, with TF-IDF feature extraction and L2 normalization. Tests were conducted across three different data ratios (20%, 40%, and 60%). The Naive Bayes model performed best, achieving a highest accuracy of 73.2% on automatic labels with a 20% test ratio. For this result, precision was 68.1%, recall was 73.17%, and the f1-score was 70.47%. In contrast, SVM reached its highest accuracy of 56.8% on automatic labels with a 60% ratio. A notable challenge for both algorithms was the difficulty in accurately predicting neutral sentiment in the automatic labels, primarily due to a class imbalance. The study concludes that Indonesian public sentiment regarding the "Kabur Aja Dulu" phenomenon is predominantly negative. This research demonstrates that data volume and variation significantly influence accuracy and is intended to provide policymakers with valuable insight into digital public sentiment.

Keywords: KaburAjaDulu, Sentiment Analysis, TF-IDF L2 Normalization, Naive Bayes, Support Vector Machine

1. PENDAHULUAN

Dengan berkembangnya internet dan teknologi, masyarakat semakin mudah mengakses media sosial dan memanfaatkannya, misalnya untuk memberikan kritik terkait suatu kebijakan atau sekedar menyampaikan pendapat serta opini mereka. Banyak masyarakat yang seringkali memberikan pendapatnya terhadap sebuah fenomena yang muncul di ruang digital melalui unggahannya, dan dari opini atau pendapat tersebut banyak mengandung informasi tentang pandangan manusia terhadap sebuah fenomena yang muncul, untuk menganalisa teks yang terkumpul dari seluruh unggahan tersebut dan memahami kandungan yang ada pada kumpulan teks tersebut dapat dilakukan dengan menggunakan suatu metode yaitu Analisis Sentimen.

Analisis sentimen yang juga disebut sebagai penggalian opini, adalah proses memeriksa sejumlah besar teks untuk mengetahui apakah dalam teks tersebut terdapat suatu makna yang penting dan juga mengetahui apakah teks tersebut menyampaikan suatu perasaan seperti positif, negatif, atau netral, yang bertujuan untuk memahami opini ataupun emosi seseorang yang dicurahkan dalam teks [1], misalnya yang dicurahkan dalam unggahan pada media sosial salah satunya X. Dengan banyaknya akun politisi, aktivis, artis, dan influencer yang aktif di X menjadikan platform sosial media ini seperti menciptakan ruang debat raksasa, dan fitur yang tersedia seperti *ReTweet* yang membuat opini dapat menyebar secara cepat dan menjadi bahan obrolan nasional maupun global dalam hitungan jam seperti contohnya Fenomena “Kabur Aja Dulu” yang menjadi pembahasan masyarakat secara luas.

Fenomena “Kabur Aja Dulu” pada awalnya muncul di media sosial pada tahun 2023 dan kembali ramai dibicarakan pada bulan Februari 2025 disetiap *tweet* para pengguna dengan menyisipkan #KaburAjaDulu dalam isi *tweet* nya. Fenomena ini biasanya merupakan ungkapan yang sering digunakan masyarakat sebagai bentuk respon terhadap situasi yang dirasa sulit, membingungkan, atau melelahkan baik dalam hal pendidikan, pekerjaan, hingga tekanan hidup sehari-hari, fenomena ini menarik karena dapat merepresentasikan kondisi masyarakat, khususnya bagi generasi muda dalam menyikapi kompleksitas hidup yang dirasakan sehingga mengungkapkan rasa kekecewaannya melalui media sosial [2].

Fenomena “KaburAjaDulu” menggambarkan rasa kecewa masyarakat, terutama kaum muda yang memilih mencari peluang di luar negeri demi memperoleh kehidupan yang lebih layak [3]. Fenomena ini juga menarik perhatian berbagai pihak salah satunya dikalangan akademisi. Dosen Pembangunan Sosial dan Kesejahteraan dari Fakultas Ilmu Sosial dan Ilmu Politik Universitas Gajah Mada, Dr. Hempri Suyatna berpendapat bahwa Fenomena “KaburAjaDulu” menggambarkan kritik sekaligus sindiran generasi muda terhadap kondisi sosial politik di Indonesia saat ini yang dinilai tidak menguntungkan, serta persepsi bahwa negara “kurang hadir” dalam menyelesaikan masalah rakyat. Misalnya dalam ranah pendidikan, kekhawatiran terhadap efisiensi anggaran yang berpotensi mengancam masa depan pendidikan mendorong banyak anak muda memilih pergi ke luar negeri, baik untuk bekerja maupun melanjutkan studi [4].

Beberapa penelitian sebelumnya di Indonesia telah menjadikan analisis sentimen di platform X sebagai pembahasan utama didalamnya, sebagai contoh penelitian yang terkait Program Kampus Merdeka[5]. Mengangkat topik pro dan kontra program Kampus Merdeka yang dirilis pada akhir Januari 2020 oleh Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi periode 2021–2024, penelitian ini memanfaatkan algoritma *Naive Bayes* dan *Support Vector Machine*. Hasil pengujian menunjukkan akurasi 86% untuk *Naive Bayes* dan 93% untuk SVM, menegaskan kemampuan kedua metode tersebut dalam klasifikasi sentimen terkait pro dan kontra program kampus merdeka. Kemudian penelitian yang dilakukan terkait kebijakan TAPERA [6] dengan pembahasan utama yaitu kebijakan iuran tabungan perumahan rakyat (TAPERA) yang menyebabkan perdebatan dikalangan masyarakat pada platform X, Penelitian ini memanfaatkan algoritma *Naive Bayes* dan *Support Vector Machine* dalam proses klasifikasi sentimen. Hasil pengujian menunjukkan bahwa *Support Vector Machine* mencapai akurasi 84%, sedangkan akurasi *Naive Bayes* sebesar 81%.

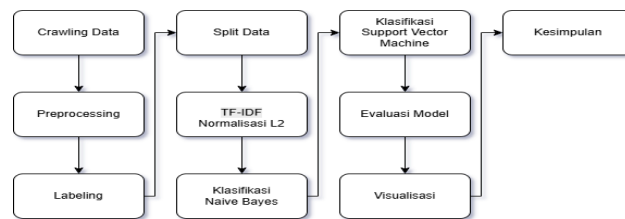
Namun, kedua penelitian tersebut masih memiliki keterbatasan, seperti hanya menggunakan dua kelas sentimen (positif–negatif), tidak melibatkan pakar dalam pelabelan manual, serta evaluasi terbatas pada akurasi. Ada pula penelitian yang menggunakan tiga kelas dengan bantuan lexicon dan SMOTE, tetapi belum divalidasi pakar bahasa serta sangat bergantung pada pustaka tertentu.

Untuk mengisi gap tersebut, penelitian ini dilakukan untuk menganalisis fenomena “KaburAjaDulu” dengan membandingkan hasil klasifikasi berdasarkan data komentar yang diberi label melalui metode lexicon dan validasi pakar, lalu dilakukan pembagian data menggunakan *stratified sampling* tanpa menggunakan SMOTE agar proporsi kelas tetap seimbang secara alami. Pembobotan kata dilakukan dengan TF-IDF dan normalisasi L2 tanpa ketergantungan pustaka scikit-learn. Dari sisi metode, digunakan *Naive Bayes* dan SVM dengan kernel linear, soft margin, SGD, serta pendekatan One-vs-Rest guna mendukung klasifikasi multikelas. Evaluasi mencakup precision, recall, f1-score, macro, dan weighted average, sehingga hasil penelitian lebih kredibel dan komprehensif dibandingkan studi sebelumnya.

2. METODE PENELITIAN

2.1 Penerapan Metode

Penelitian ini melalui sejumlah tahapan dalam mengembangkan aplikasi analisis sentimen berbasis *web* yang memanfaatkan algoritma *Naive Bayes* dan *Support Vector Machine*. Prosesnya dimulai dari tahap inisiasi hingga implementasi, sedangkan penerapan metode yang digunakan ditampilkan pada Gambar 1.



Gambar 1. Penerapan Metode

2.2 Crawling data

Penelitian ini memanfaatkan data penelitian yang diperoleh melalui metode *crawling* dengan *tools tweet harvest* melalui *Google Colab* dalam rentang waktu 01 Februari 2025 sampai 31 Maret 2025 dengan kata kunci #KaburAjaDulu berbahasa Indonesia, data yang diperoleh yaitu sebanyak 1594 data, kemudian diseleksi menjadi 200 data dan disimpan kedalam format file .csv.

2.3 Preprocessing

Preprocessing adalah langkah awal dalam analisis sentimen yang bertujuan mengubah teks tidak terstruktur menjadi dataset dengan format yang baik agar dapat dimanfaatkan sebagai data masukan terhadap suatu penelitian ataupun analisis untuk mendapatkan hasil prediksi yang tepat, ketika menggunakan data sosial media X untuk melakukan analisis sentimen, akan terjadi masalah yang disebabkan data *tweet* masih mengandung elemen, simbol, atau penulisan dengan singkatan, maka dari itu perlu dilakukan *Preprocessing* untuk mengurangi turunnya performa analisis sentimen yang disebabkan oleh bentuk data yang tidak baik [7].

Berikut penjelasan tahapan *Preprocessing*:

- a. *Cleaning*
Cleaning berperan sebagai proses yang bertujuan untuk menghilangkan simbol, tanda baca, link URL angka, dan elemen yang tidak diperlukan. Proses ini bertujuan menghilangkan noise pada teks agar data menjadi terstruktur, mudah diproses oleh mesin, dan meningkatkan kualitas analisis [7]
- b. *Normalisasi Slangword*
Normalisasi *Slangword* adalah proses konversi kata-kata tidak baku ke dalam bentuk yang baku melalui penggunaan kamus *Slangword* dan menjadikan teks dengan bentuk standar agar lebih akurat untuk proses analisis dan menggunakan kamus yang berisikan kata-kata tidak baku
- c. *Tokenizing*
Proses *tokenizing* bertujuan memisahkan teks kalimat menjadi potongan-potongan kata kecil yang dikenal sebagai token agar lebih mudah diproses dalam proses *Preprocessing* selanjutnya
- d. *Stopwords*
Stopwords merupakan proses untuk kata yang dianggap tidak bermakna yang tidak memiliki pengaruh untuk proses analisis agar mengurangi dimensi data yang tidak diperlukan
- e. *Stemming*
Stemming merupakan proses pengubahan kata berimbuhan menjadi bentuk kata dasar dengan tujuan menghilangkan variasi kata yang memiliki makna sama [8]

2.4 Labeling

Labeling atau pelabelan data *tweet* yang telah kumpulkan dan telah di lakukan *Preprocessing* ini menggunakan pendekatan otomatis berbasis leksikon dengan kamus inset serta menggunakan pendekatan melalui pakar, yang merupakan seorang dosen yang ahli dalam bidang bahasa Indonesia, kemudian data diberi label dalam tiga kelas, yaitu positif, negatif, dan netral. Pelabelan ini dilakukan sebagai dasar pembelajaran mesin yang berbasis *supervised learning* yang membutuhkan data berlabel agar model dapat dilatih secara efektif.

2.5 Split Data

Proses pembagian data atau *Split Data* melibatkan pemisahan dataset menjadi data latih atau sekumpulan data yang dipakai untuk membangun serta melatih model dan data uji yang berfungsi untuk menguji performa model yang telah dibangun tersebut. Dalam pembagian ini, metode *stratified sampling* diterapkan dengan cara mengacak data secara terstruktur berdasarkan tiga kelas sentimen yang telah dilabeli, yaitu positif, negatif, dan netral [9].

2.6 TF-IDF dan Normalisasi L2

Proses TF-IDF mengubah data teks menjadi angka dengan menghitung frekuensi kemunculan kata (tf) dan juga kata-kata yang jarang muncul dalam dokumen (idf). Hasil perhitungan tersebut menghasilkan sebuah vektor dengan bobot yang lebih rendah untuk kata yang sering muncul dan bobot lebih tinggi untuk kata yang jarang muncul. Normalisasi L2 kemudian digunakan untuk menghilangkan bias dengan menyamakan panjang vektor menjadi satu. Persamaan 1 dan persamaan 2 menunjukkan perhitungan TF-IDF dan normalisasi L2 sebagai berikut:

Formula perhitungan TF-IDF pada persamaan 1 sebagai berikut:

$$TFIDF(t, d, D) = (1 + \log(f_{t,d})) \times \left(\log \left(\frac{1 + N}{1 + n_t} \right) + 1 \right) \quad (1)$$

Keterangan:

- t = Kata yang sedang dihitung
- d = Dokumen yang sedang dianalisis
- D = Melambangkan keseluruhan korpus dokumen
- N = Total jumlah dokumen dalam korpus D
- n_t = Menunjukkan banyaknya dokumen yang memuat kata t
- $f_{t,d}$ = Frekuensi kemunculan kata t dalam dokumen d

Formula normalisasi L2 pada persamaan 2 sebagai berikut:

$$TFIDF_{norm}(t, d) = \frac{TFIDF(t, d)}{\sqrt{\sum_{t_i \in d} TFIDF(t_i, d)^2}} \quad (2)$$

Keterangan:

- t = term yang sedang dihitung
- d = dokumen yang sedang dianalisis
- t_i = term ke i dalam dokumen d
- $TFIDF(t, d)$ = bobot TF-IDF sebelum di normalisasi
- $TFIDF_{norm}(t, d)$ = bobot term setelah di normalisasi
- $\sum_{t_i \in d} TFIDF(t_i, d)^2$ = jumlah kuadrat bobot TF-IDF

2.7 Klasifikasi Naïve Bayes

Naïve Bayes adalah suatu metode klasifikasi dalam pembelajaran mesin yang mengandalkan teorema Bayes untuk melakukan prediksi peristiwa di masa depan berdasarkan data historis [10], rumus bayes digunakan sebagai dasar pemrograman pada *Naïve Bayes* dapat dilihat pada persamaan 3 sebagai berikut:

$$P(T_i | D) = \frac{P(D | T_i) \times P(T_i)}{P(D)} \quad (3)$$

Keterangan:

- $P(T_i|D)$ = probabilitas dokumen D termasuk kelas T_i
- $P(D|T_i)$ = probabilitas data D muncul kelas T_i
- $P(T_i)$ = probabilitas awal kelas T_i
- $P(D)$ = probabilitas keseluruhan data D

2.8 Klasifikasi Support Vector Machine

Support Vector Machine (SVM) adalah metode untuk mengklasifikasikan data dengan menggunakan *kernel* untuk mencari *hyperplane* terbaik yang bisa membedakan berbagai kelas. Dengan memetakan titik data ke ruang berdimensi lebih tinggi, SVM mampu menentukan garis batas *linear* yang memisahkan dua kelas. Untuk kasus klasifikasi multi-kelas, algoritma ini biasanya menggunakan pendekatan *one-vs-one* yang membagi masalah

menjadi beberapa klasifikasi biner, atau pendekatan *one-vs-rest* yang melakukan klasifikasi satu kelas terhadap kelas lainnya secara terpisah [11].

Formula mencari *hyperplane* terbaik untuk fungsi keputusan pada persamaan 4 sebagai berikut:

$$f(x) = w^T x + b \quad (4)$$

Keterangan:

- w = *Vector* bobot (*weights*) yang menentukan orientasi *hyperplane*
- b = Parameter bias yang menentukan posisi *hyperplane*
- x = *Vector* fitur dari sampel data
- Nilai dari $f(x)$ akan menentukan kelas prediksi dari sampel data (x)

Formula untuk memaksimalkan *margin* pemisah dan mencegah *overfitting* pada persamaan 5 sebagai berikut:

$$J(w, b) = (1/N) \sum_{i=1}^N c_i \cdot \max(0, 1 - y_i(w^T x_i + b)) + (\lambda/2) \|w\|^2 \quad (5)$$

Keterangan:

- $(0, 1 - y_i(w^T x_i + b))$ = *Hinge Loss* yang memberikan penalti pada sampel yang salah klasifikasi dalam *margin*
- $(\lambda/2) \|w\|^2$ = *Term* regularisasi L2. Parameter λ mengontrol tingkat regularisasi, yang berfungsi untuk mencegah *overfitting* dengan memberikan penalti pada bobot model yang terlalu besar
- c_i = Bobot untuk menangani ketidakseimbangan kelas dengan memberi penalti pada kesalahan kelas dengan jumlah data yang lebih sedikit.

Formula meminimalkan fungsi objektif $J(w, b)$ dengan *Stochastic Gradient Descent* (SGD) pada persamaan 6 dan 7 sebagai berikut:

$$w \leftarrow w - \eta \nabla_w J \quad (6)$$

$$b \leftarrow b - \eta \nabla_b J \quad (7)$$

Keterangan:

- η = laju pembelajaran (*learning rate*) yang mengontrol ukuran langkah pada setiap iterasi.
- $\nabla_w J$ dan $\nabla_b J$ = *Gradien* dari fungsi objektif terhadap w dan b

Formula strategi *one vs rest* untuk menangani masalah multi kelas pada persamaan 8, 9, dan 10 sebagai berikut:

$$(w_m)^T \phi(x_m) + b_m \geq 1 - \xi_{im}, \rightarrow y_i = m \quad (8)$$

$$(w_m)^T \phi(x_m) + b_m \leq 1 - \xi_{im}, \rightarrow y_i \neq m \quad (9)$$

$$\xi_{im} \geq 0, \quad i = 1, \dots, l \quad (10)$$

Keterangan:

- $(w_m)^T \phi(x_m) + b_m \geq 1 - \xi_{im}, \rightarrow y_i = m$ = Titik dari kelas m harus berada di sisi benar dari *hyperplane*.
- $(w_m)^T \phi(x_m) + b_m \leq 1 - \xi_{im}, \rightarrow y_i \neq m$ = Titik dari kelas lain harus berada di sisi sebaliknya
- $\xi_{im} \geq 0, \quad i = 1, \dots, l$ = Toleransi kesalahan atau *slack*, diizinkan jika data melanggar *margin*

2.9 Pengujian dan Evaluasi

Pengujian serta evaluasi dalam studi ini memanfaatkan *Confusion Matrix* guna menunjukkan kinerja model dalam melakukan klasifikasi dan prediksi terhadap data aktual yang dimiliki [12] *Confusion Matrix* berisikan detail prediksi seperti *TruePositif*, *FalseNegatifPos*, *FalseNetralPos*, *FalsePositifNeg*, *TrueNegatif*, *FalseNetralNeg*, *FalsePositifNet*, *FalseNegatifNet*, *TrueNetral*. Tabel 1 menunjukkan tabel *Confusion Matrix* secara detail.

Tabel 1. *Confusion Matrix*

Aktual	Prediksi		
	Positif	Negatif	Netral
Positif	<i>TruePositif</i> (TP)	<i>FalsePositifNeg</i> (FPN)	<i>FalsePositifNet</i> (FPNt)
Negatif	<i>FalseNegatifPos</i> (FNP)	<i>TrueNegatif</i> (TN)	<i>FalseNegatifNet</i> (FNNt)
Netral	<i>FalseNetralPos</i> (FNtP)	<i>FalseNetralNeg</i> (FNtN)	<i>TrueNetral</i> (TNt)

Keterangan:

TruePositif (TP) = Data positif benar diidentifikasi positif.

TrueNegatif (TN) = Data negatif benar diidentifikasi negatif.

TrueNetral (TNt) = Data netral benar diidentifikasi netral.

FalsePositifNeg (FPN) = Data positif, diidentifikasi negatif.

FalseNegatifPos (FNP) = Data negatif, diidentifikasi positif.

FalseNetralPos (FNtP) = Data netral, diidentifikasi positif.

FalsePositifNet (FPNt) = Data positif, diidentifikasi netral.

FalseNetralNeg (FNtN) = Data netral, diidentifikasi negatif.

FalseNegatifNet (FNNt) = Data negatif, diidentifikasi netral.

Untuk menghitung skor akurasi, presisi, *recall*, *f1-score*, *macro average*, *weighted average* dapat digunakan formula sebagai berikut:

Formula menghitung skor akurasi dapat dilihat pada persamaan 11.

$$Akurasi = \frac{TP + TN + TNt}{Total} \quad (11)$$

Formula menghitung skor presisi dapat dilihat pada persamaan 12 sebagai berikut:

$$Precision_P = \frac{TP}{TP + FNP + FNtP}$$

$$Precision_N = \frac{TN}{TN + FPN + FNtN} \quad (12)$$

$$Precision_T = \frac{TNt}{TNt + FPNt + FNNt}$$

Formula menghitung skor *recall* dapat dilihat pada persamaan 13 sebagai berikut:

$$Recall_P = \frac{TP}{TP + FPN + FPNt}$$

$$Recall_N = \frac{TN}{TN + FNP + FNNt} \quad (13)$$

$$Recall_T = \frac{TNt}{TNt + FNtP + FNtN}$$

Formula menghitung *f1-score* dapat dilihat pada persamaan 14 sebagai berikut:

$$F1_P = 2 \times \frac{Precision_P \times Recall_P}{Precision_P + Recall_P}$$

$$F1_N = 2 \times \frac{Precision_N \times Recall_N}{Precision_N + Recall_N} \quad (14)$$

$$F1_T = 2 \times \frac{Precision_T \times Recall_T}{Precision_T + Recall_T}$$

Formula menghitung *macro average* dapat dilihat pada persamaan 15 sebagai berikut:

$$Precision_{macro} = \frac{Precision_P + Precision_N + Precision_T}{3}$$

$$Recall_{macro} = \frac{Recall_P + Recall_N + Recall_T}{3} \quad (15)$$

$$F1_{macro} = \frac{F1_P + F1_N + F1_T}{3}$$

Formula menghitung *weighted average* dapat dilihat pada persamaan 16 sebagai berikut:

$$Precision_{weighted} = \frac{Precision_P \times N_P + Precision_N \times N_N + Precision_T \times N_T}{N_P + N_N + N_T} \quad (16)$$

$$Recall_{weighted} = \frac{Recall_P \times N_P + Recall_N \times N_N + Recall_T \times N_T}{N_P + N_N + N_T}$$

$$F1_{weighted} = \frac{F1_P \times N_P + F1_N \times N_N + F1_T \times N_T}{N_P + N_N + N_T}$$

3. HASIL DAN PEMBAHASAN

3.1 Crawling data

Proses pengumpulan data melalui metode *crawling* dalam penelitian ini berhasil mengumpulkan data sebanyak 1594 data yang kemudian diseleksi menjadi 200 data dan disimpan dalam file dengan format .csv yang kemudian disimpan kedalam *database*.

3.2 Preprocessing

Preprocessing dilakukan terhadap dataset yang telah di *import* sebelumnya, proses ini akan melakukan tahapan-tahapan merujuk pada penjelasan di sub bab 2.3, pada gambar 2 merupakan hasil proses *Preprocessing*.

Proses	Sebelum	Setelah
Cleaning	@dhanis : @suryafis: Buatkan bhw negara gagal membuat lingkungan pekerjaan bagi rakyatnya? Bkn bemaakud membuat nilai burak ipod org? yg #KahurAjaDulu toh aku nndng jg pdlaku #KahurAjaDulu Mnd d' indo kita trvada lpgangan pekerjaan bn memfakha kbg dg layak org & aka prg or mkan d ln	bukti bhw negara gagal membuat lingkungan pekerjaan bagi rakyatnya bkn bemaakud membuat nilai burak ipod org yg kahur_aja_dulu toh aku nndng jg pdlaku kahur_aja_dulu mnd d' indo kita trvada lpgangan pekerjaan bn memfakha kbg dg layak org & aka prg or mkan d ln
Normalisasi	bukti bhw negara gagal membuat lingkungan pekerjaan bagi rakyatnya? Bkn bemaakud membuat nilai burak ipod org? yg kahur_aja_dulu toh aku nndng jg pdlaku kahur_aja_dulu mnd d' indo kita trvada lpgangan pekerjaan bn memfakha kbg dg layak org & aka prg or mkan d ln	bukti bahwa negara gagal membuat lingkungan pekerjaan untuk rakyatnya bkn bemaakud membuat nilai burak ipoda orang yg kahur_aja_dulu toh aku nndng jg pdlaku kahur_aja_dulu mnd d' indo kita trvada lpgangan pekerjaan bn memfakha kbg dg layak org & aka prg or mkan d ln
Tokemizing	bukti bahwa negara gagal membuat lingkungan pekerjaan untuk rakyatnya? Bkn bemaakud membuat nilai burak ipoda orang yg kahur_aja_dulu toh aku nndng jg pdlaku kahur_aja_dulu mnd d' indo kita trvada lpgangan pekerjaan bn memfakha kbg dg layak org & aka prg or mkan d ln	'bukti', 'bahwa', 'negara', 'gagal', 'membuat', 'lingkungan', 'pekerjaan', 'bagi', 'rakyatnya', 'bukan', 'bemaakud', 'membuat', 'nilai', 'burak', 'ipoda', 'orang-orang', 'yang', 'kahur_aja_dulu', 'toh', 'saya', 'sendiri', 'juga', 'pdlaku', 'kahur_aja_dulu', 'mnd', 'indonesia', 'nda', 'trvada', 'lpgangan', 'pekerjaan', 'bn', 'memfakha', 'kbg', 'dg', 'layak', 'org', 'aka', 'prg', 'or', 'mkan', 'd', 'ln', 'ngeni'
Stopwords	'bukti', 'bahwa', 'negara', 'gagal', 'membuat', 'lingkungan', 'pekerjaan', 'bagi', 'rakyatnya', 'bukan', 'bemaakud', 'membuat', 'nilai', 'burak', 'ipoda', 'orang-orang', 'yang', 'kahur_aja_dulu', 'toh', 'saya', 'sendiri', 'juga', 'pdlaku', 'kahur_aja_dulu', 'mnd', 'indonesia', 'nda', 'trvada', 'lpgangan', 'pekerjaan', 'bn', 'memfakha', 'kbg', 'dg', 'layak', 'org', 'aka', 'prg', 'or', 'mkan', 'd', 'ln', 'ngeni'	'bukti', 'negara', 'gagal', 'membuat', 'lingkungan', 'pekerjaan', 'bagi', 'rakyatnya', 'bukan', 'bemaakud', 'membuat', 'nilai', 'burak', 'ipoda', 'orang-orang', 'yang', 'kahur_aja_dulu', 'toh', 'saya', 'sendiri', 'juga', 'pdlaku', 'kahur_aja_dulu', 'mnd', 'indonesia', 'nda', 'trvada', 'lpgangan', 'pekerjaan', 'bn', 'memfakha', 'kbg', 'dg', 'layak', 'org', 'aka', 'prg', 'or', 'mkan', 'd', 'ln', 'ngeni'
Stemming	'bukti', 'negara', 'gagal', 'membuat', 'lingkungan', 'pekerjaan', 'bagi', 'rakyatnya', 'bukan', 'bemaakud', 'membuat', 'nilai', 'burak', 'ipoda', 'orang-orang', 'yang', 'kahur_aja_dulu', 'toh', 'saya', 'sendiri', 'juga', 'pdlaku', 'kahur_aja_dulu', 'mnd', 'indonesia', 'nda', 'trvada', 'lpgangan', 'pekerjaan', 'bn', 'memfakha', 'kbg', 'dg', 'layak', 'org', 'aka', 'prg', 'or', 'mkan', 'd', 'ln', 'ngeni'	'bukti', 'negara', 'gagal', 'beer', 'lwoang', 'lmpa', 'lakvat', 'bukan', 'maksud', 'ben', 'nlla', 'burak', 'orang', 'kahur_aja_dulu', 'aku', 'sendiri', 'baku', 'kahur_aja_dulu', 'mnd', 'indonesia', 'nda', 'lpgang', 'kepa', 'ndakab', 'kbg', 'layak', 'orang', 'prgi', 'car', 'mkan', 'har', 'ngeni'

Gambar 2. Data Hasil *Preprocessing*

3.3 Labeling

Pada tahap ini, teks hasil preprocessing diberi label menggunakan dua pendekatan, yaitu berbasis leksikon dan pakar. Pelabelan berbasis leksikon dilakukan secara otomatis dengan memanfaatkan kamus InSet, di mana setiap kata dalam tweet diberi bobot sentimen positif, negatif, atau netral sesuai dengan nilai yang terdapat dalam kamus. Skor total dari seluruh kata dalam satu teks kemudian dijumlahkan, dan hasilnya menentukan kelas akhir sentimen. Sementara itu, pelabelan pakar dilakukan secara manual oleh seorang dosen bidang bahasa Indonesia yang memiliki keahlian dalam analisis linguistik dan pemaknaan teks. Secara teknis, data tweet disajikan dalam bentuk file Excel yang kemudian diteliti satu per satu oleh pakar untuk memberikan label positif, negatif, atau netral sesuai dengan konteks bahasa. Dari hasil pelabelan, pendekatan leksikon menghasilkan 45 positif, 148 negatif, dan 7 netral, sedangkan pelabelan pakar menghasilkan 70 positif, 80 negatif, dan 50 netral, dengan jumlah masing-masing 200 data

3.4 Split Data

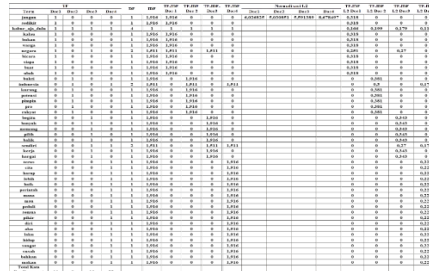
Split Data atau pembagian data dilakukan menggunakan metode *stratified sampling* dengan menggunakan tiga rasio uji, yaitu 80:20 atau 20%, 40:60 atau 40%, dan 60:40 atau 60%. Jumlah pembagian data yang dihasilkan pada label otomatis serta label pakar dengan metode *stratified sampling* ini berbeda, dikarenakan menggunakan pembulatan keatas dan diakumulasikan dengan total label dari setiap kelas sentimen yang dihasilkan dari proses pelabelan, hasil dari pembagian data pada label otomatis dan label pakar ditunjukkan pada tabel 2.

Tabel 2. Hasil Pembagian data

Rasio	Data Latih Label Otomatis	Data Uji Label Otomatis	Data Latih Label Pakar	Data Uji Label Pakar
20%	159	41	160	40
40%	119	81	120	80
60%	79	121	80	120

3.5 TF-IDF dan Normalisasi L2

Proses TF-IDF merupakan metode untuk memberi bobot pada kata dengan mengonversi teks ke format numerik, menghitung frekuensi kata (tf), mengukur pentingnya sebuah kata dalam dokumen (idf), mengalikan tf dengan idf, dan menormalisasi hasilnya menggunakan L2, gambar 3 menunjukkan hasil dari seluruh perhitungan.



Gambar 3. Hasil Perhitungan TF-IDF dan Normalisasi L2

3.6 Pengujian dan Evaluasi

Pengujian dan evaluasi akan mengukur kinerja model yang dihasilkan *Naïve Bayes* dan model yang dihasilkan SVM dengan membandingkan hasil klasifikasi dari dua jenis sumber data berlabel. Kedua sumber data, yaitu data yang dilabeli secara otomatis (leksikon) dan data yang dilabeli oleh pakar, masing-masing berjumlah 200 data. Untuk mengevaluasi performa model secara menyeluruh, pengujian dilakukan dengan beberapa skema yang berbeda, yaitu pada rasio uji 20% atau 80:20, 40% atau 60:40, dan 60% atau 40:60.

Gambar 4 menunjukkan *Confusion Matrix* setiap rasio pengujian yang diperoleh dari proses *Naïve Bayes* dan SVM menggunakan label otomatis dan label pakar pada tiga rasio data yang telah dilakukan yang menunjukkan jumlah prediksi untuk setiap kelas sentimen yaitu positif, negatif, dan netral

Aktual	Prediksi Label Otomatis Naive Bayes			Prediksi Label Pakar Naive Bayes			Prediksi Label Otomatis SVM			Prediksi Label Pakar SVM		
	Positif	Negatif	Netral	Positif	Negatif	Netral	Positif	Negatif	Netral	Positif	Negatif	Netral
Rasio 80:20												
Positif	3	6	0	9	2	3	3	6	0	9	2	3
Negatif	3	27	0	2	10	4	3	27	0	2	10	4
Netral	2	0	0	1	4	5	2	0	0	1	4	5
Rasio 60:40												
Positif	6	10	2	13	9	6	6	10	2	13	9	6
Negatif	11	47	2	7	16	9	11	47	2	7	16	9
Netral	1	2	0	7	6	7	1	2	0	7	6	7
Rasio 40:60												
Positif	11	13	3	11	18	13	11	13	3	11	18	13
Negatif	20	62	7	17	19	12	20	62	7	17	19	12
Netral	3	2	0	9	10	11	3	2	0	9	10	11

Gambar 4. *Confusion Matrix*

Pada gambar 5 merupakan matriks nilai akurasi, presisi, f1-score, support, macro average, dan weighted average yang dihasilkan dari *Naïve Bayes* dan SVM menggunakan label otomatis dan label pakar pada rasio 80:20. Menunjukkan akurasi *Naïve Bayes* label otomatis sebesar 73,2%, *Naïve Bayes* label pakar 60,0%. SVM label otomatis 63,4% dan SVM label pakar sebesar 52,5%. Tingkat akurasi lebih besar dihasilkan label otomatis dikarenakan label otomatis menghasilkan jumlah sentimen yang tidak seimbang disetiap kelas, berbeda dengan label pakar yang menghasilkan jumlah sentimen yang stabil untuk setiap kelas.

Aktual	Naive Bayes Otomatis				Rasio	SVM Otomatis			
	Precisi	Recall	F1-Score	Support		Accuracy	Support	Accuracy	
Positif	37.5%	33.33%	35.29%	9	73.2%	42.86%	66.67%	52.17%	9
Negatif	81.82%	90.0%	85.71%	30		90.91%	66.67%	76.92%	30
Netral	0.0%	0.0%	0.0%	2		0.0%	0.0%	0.0%	2
Rata-rata macro	39.77%	41.11%	40.34%	-	80:20	44.59%	44.44%	43.03%	-
Rata-rata weighted	68.1%	73.17%	70.47%	-		75.93%	63.41%	67.74%	-
					60.0%	SVM Pakar			
Positif	37.5%	33.33%	35.29%	9		56.25%	64.29%	60.0%	14
Negatif	81.82%	90.0%	85.71%	30		61.54%	50.0%	55.17%	16
Netral	0.0%	0.0%	0.0%	2	36.36%	40.0%	38.1%	10	
Rata-rata macro	39.77%	41.11%	40.34%	-	80:20	51.38%	51.43%	51.09%	-
Rata-rata weighted	68.1%	73.17%	70.47%	-		53.39%	52.5%	52.59%	-

Gambar 5. Matriks Evaluasi Pengujian Dengan Rasio 80:20

Pada gambar 6 merupakan matriks nilai akurasi, presisi, f1-score, support, macro average, dan weighted average yang dihasilkan dari *Naïve Bayes* dan SVM menggunakan label otomatis dan label pakar pada rasio 60:40. Menunjukkan akurasi *Naïve Bayes* label otomatis sebesar 65,4%, *Naïve Bayes* label pakar 45,0%. SVM label otomatis 56,8% dan SVM label pakar sebesar 43,8%. Tingkat akurasi lebih besar dihasilkan label otomatis

4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma Naive Bayes mendapatkan hasil klasifikasi yang lebih baik daripada Support Vector Machine (SVM). Model yang dihasilkan ini berhasil mencapai akurasi tertinggi sebesar 73,2% pada data yang dilabeli secara otomatis, serta akurasi sebesar 60,0% ketika menggunakan data berlabel pakar dengan rasio 80:20 atau 20%. Temuan ini juga memberikan wawasan penting terkait kualitas data, meskipun data berlabel otomatis menghasilkan tingkat akurasi yang lebih baik, hal itu akibat dari tidak seimbangannya distribusi label. Sebaliknya, data berlabel pakar yang lebih merata memungkinkan model untuk belajar dengan lebih andal di seluruh kategori sentimen. Selain itu, tahapan preprocessing yang diterapkan, seperti *cleaning*, normalisasi, *stopwords*, dan *stemming*, terbukti berkontribusi dalam meningkatkan akurasi dengan cara mengurangi *noise* serta menghasilkan representasi kata yang lebih konsisten pada proses pembobotan kata. Dengan demikian, kualitas data, keseimbangan distribusi label, serta tahap *preprocessing* menjadi faktor penentu utama dalam kinerja model pembelajaran mesin. Hasil analisis secara keseluruhan mengarah pada kesimpulan bahwa sentimen masyarakat Indonesia terhadap fenomena "Kabur Aja Dulu" cenderung negatif. Dengan demikian, diharapkan dapat menggunakan jumlah dataset yang lebih besar untuk penelitian berikutnya, serta diharapkan untuk mengeksplorasi lebih lanjut pemilihan metode pembobotan kata selain menggunakan metode TF-IDF, dan melakukan eksplorasi algoritma *supervised learning* lainnya, dengan harapan memperoleh hasil yang lebih akurat untuk prediksi yang dihasilkan.

DAFTAR PUSTAKA

- [1] R. Savira, A. Solichin, And M. Syafrullah, "Analisis Sentimen Pada Twitter Terhadap Kenaikanbbm 2022 Dengan Lexicon Dan Support Vector Machine," *Pros. Semin. Nas. Mhs. Fak. Teknol. Inf.*, Vol. 2, No. 1, Pp. 211–218, 2023, [Online]. Available: <https://senafiti.budiluhur.ac.id/index.php/senafiti/article/view/564>
- [2] N. A. Manurung And Suparman, "Perspektif Publik Terhadap Tagar #Kaburajadulu: Sebuah Analisis Sentimen Komunikasi Digital Di Indonesia," *Nivedana J. Komun. Dan Bhs.*, Vol. 6, No. 2, Pp. 258–269, 2025, [Online]. Available: <https://doi.org/10.53565/Nivedana.V6i1.1613>
- [3] (2025) Cnn Indonesia, "Tren #Kaburajadulu Menyeruak Di Kalangan Anak Muda Indonesia." [Online]. Available: <https://www.cnnindonesia.com/Gaya-Hidup/20250207082558-277-1195561/Tren-Kaburajadulu-Menyeruak-Di-Kalangan-Anak-Muda-Indonesia>
- [4] (2025) Kezia Dwina Nathania, "Ramai Tagar Kabur Aja Dulu, Pakar Ugm: Bentuk Sikap Kritis Dan Sindiran Anak Muda Atas Situasi Di Tanah Air." [Online]. Available: <https://ugm.ac.id/Id/Berita/Ramai-Tagar-Kabur-Aja-Dulu-Pakar-Ugm-Bentuk-Sikap-Kritis-Dan-Sindiran-Anak-Muda-Atas-Situasi-Di-Tanah-Air/>
- [5] I. P. Rahayu, A. Fauzi, And J. Indra, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine," *J. Sist. Komput. Dan Inform.*, Vol. 4, No. 2, P. 296, 2022, Doi: 10.30865/Json.V4i2.5381.
- [6] A. M. Rizqiyah And I. K. D. Nuryana, "Analisis Sentimen Masyarakat Terhadap Kebijakan Iuran Tabungan Perumahan Rakyat (Tapera) Pada Platform X Menggunakan Algoritma Naive Bayes Classifier Dan Support Vector Machine," Vol. 05, No. 03, Pp. 298–306, 2024.
- [7] S. Khairunnisa, A. Adiwijaya, And S. Al Faraby, "Pengaruh Text Preprocessing Terhadap Analisis Sentimen Komentar Masyarakat Pada Media Sosial Twitter (Studi Kasus Pandemi Covid-19)," *J. Media Inform. Budidarma*, Vol. 5, No. 2, P. 406, 2021, Doi: 10.30865/Mib.V5i2.2835.
- [8] F. Hadaina And U. Budiyanto, "Implementasi Metode Multinomial Naive Bayes Untuk Sentiment Analysis Terhadap Data Ulasan Produk Colearn Pada Google Play Store," *Semin. Nas. Mhs. Fak. Teknol. Inf. Jakarta-Indonesia*, No. September, Pp. 660–666, 2022, [Online]. Available: <https://senafiti.budiluhur.ac.id/index.php>
- [9] J. Adiputra And D. Mahdiana, "Analisis Sentimen Dengan Algoritma Support Vector Machine Terhadap Penyakit Hepatitis Akut Misterius," *J. Ilm. Inform.*, Vol. 6, No. April 2022, Pp. 1–8, 2023, [Online]. Available: <https://ejournal.upbatam.ac.id/index.php/jif/article/view/6611%0ahttps://ejournal.upbatam.ac.id/index.php/jif/article/download/6611/3098>
- [10] D. Duei Putri, G. F. Nama, And W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (Dpr) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. Dan Tek. Elektro Terap.*, Vol. 10, No. 1, Pp. 34–40, 2022, Doi: 10.23960/Jitet.V10i1.2262.
- [11] R. W. Pratama, Y. H. Chrisnanto, And G. Abdillah, "Klasifikasi Efek Kerusakan Gempa Bumi Berdasarkan Skala Modified Mercalli Intensity Menggunakan Algoritma Multiclass Support Vector Machine," *Jati (Jurnal Mhs. Tek. Inform.)*, Vol. Vol. 8, No. 2, Pp. 1739–1745, 2024, Doi: 10.36040/Jati.V8i2.9211.
- [12] D. Musfiroh, U. Khaira, P. E. P. Utomo, And T. Suratno, "Analisis Sentimen Terhadap Perkuliahan Daring Di Indonesia Dari Twitter Dataset Menggunakan Inset Lexicon," *Malcom Indones. J. Mach. Learn. Comput. Sci.*, Vol. 1, No. 1, Pp. 24–33, 2021, Doi: 10.57152/Malcom.V1i1.20.