

# ANALISIS SENTIMEN MASYARAKAT DI MEDIA SOSIAL X TERHADAP MASALAH ROHINGYA MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE* DAN TEKNIK *RESAMPLING*

Tri Krama<sup>\*1)</sup>, Nouval Trezandy Lapatta<sup>2)</sup>, Chairunnisa Ar. Lamasitudju<sup>3)</sup>, Syahrullah<sup>4)</sup>

1. Teknik Informatika, Fakultas Teknik, Universitas Tadulako, Indonesia
2. Teknik Informatika, Fakultas Teknik, Universitas Tadulako, Indonesia
3. Teknik Informatika, Fakultas Teknik, Universitas Tadulako, Indonesia
4. Sistem Informasi, Fakultas Teknik, Universitas Tadulako, Indonesia

## Article Info

**Kata Kunci:** Analisis Sentimen; *Chi-Square*; *Resampling*; Rohingya; *Support Vector Machine*.

**Keywords:** *Chi-Square*; *Sentiment Analysis*; *Resampling*; Rohingya; *Support Vector Machine*.

## Article history:

Received 18 July 2024

Revised 12 August 2024

Accepted 30 August 2024

Available online 1 September 2025

## DOI :

<https://doi.org/10.29100/jupi.v10i3.6193>

\* Corresponding author.

Tri Krama

E-mail address:

[trikramaabd.haling@gmail.com](mailto:trikramaabd.haling@gmail.com)

## ABSTRAK

Di era *modern* ini, penggunaan media sosial terus meningkat karena kemajuan dalam teknologi informasi dan komunikasi. X, sebagai salah satu *platform* media sosial populer, memungkinkan pengguna untuk berinteraksi dan berdiskusi mengenai berbagai isu, termasuk masalah pengungsi Rohingya. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap isu Rohingya di X menggunakan algoritma *Support Vector Machine* (SVM) dan teknik *resampling* untuk mengatasi ketidakseimbangan data. Metode yang digunakan meliputi *oversampling* dengan *Synthetic Minority Over-sampling Technique* (SMOTE) dan *random undersampling* untuk menyeimbangkan *dataset*. Hasil penelitian menunjukkan bahwa setelah penerapan SMOTE dan *random undersampling*, model SVM mencapai akurasi 89% dengan *precision* 0,9, *recall* 0,89, dan *f1-score* 0,89. Selain itu, penggunaan seleksi fitur dengan *chi-square* juga terbukti efektif dalam meningkatkan akurasi model, meskipun peningkatannya sedikit, yaitu dari 89% menjadi 90%. Temuan ini menekankan pentingnya penggunaan teknik *resampling* dan seleksi fitur yang tepat dalam analisis sentimen sosial yang kompleks, khususnya dalam konteks isu pengungsi Rohingya di *platform* X.

## ABSTRACT

In this modern era, the use of social media continues to increase due to advances in information and communication technology. X, as one of the popular social media platforms, allows users to interact and discuss various issues, including the Rohingya refugee issue. This research aims to analyze public sentiment towards the Rohingya issue on X using the Support Vector Machine (SVM) algorithm and resampling techniques to overcome data imbalance. The methods used include oversampling with Synthetic Minority Over-sampling Technique (SMOTE) and random undersampling to balance the dataset. The results showed that after applying SMOTE and random undersampling, the SVM model achieved 89% accuracy with precision of 0,9 recall of 0,89, and f1-score of 0,89. In addition, the use of feature selection with chi-square also proved effective in improving the accuracy of the model, although the increase was slight, from 89% to 90%. These findings emphasize the importance of using appropriate resampling and feature selection techniques in complex social sentiment analysis, especially in the context of the Rohingya refugee issue on platform X.

## I. PENDAHULUAN

Di era *modern* ini, penggunaan media sosial terus meningkat karena kemajuan dalam teknologi informasi dan komunikasi. Pengguna dapat memanfaatkan media sosial sebagai *platform online* untuk mengakses informasi, berbagi dan menciptakan konten, serta berpartisipasi dalam berbagai kegiatan [1]. Hal ini menjadikan media sosial sebagai elemen yang hampir wajib dimiliki oleh seluruh masyarakat. Pada awal tahun 2023, sekitar 167 juta orang di Indonesia menggunakan media sosial, atau 60,4% dari populasi, menurut data dari Data-reportal.com [2]. X (sebelumnya dikenal sebagai Twitter) adalah salah satu *platform* media sosial terpopuler yang

memungkinkan pengguna mengekspresikan diri dan berinteraksi melalui tweet. *Platform* ini memiliki daya tarik global yang besar dan menjadi tempat diskusi utama untuk berbagai isu populer dan terkini [3]. Pada awal tahun 2023, tercatat bahwa pengguna X di Indonesia berjumlah sekitar 24 juta orang, yang setara dengan 8,7% dari total populasi [2]. Salah satu topik yang sering dibahas di X adalah masalah pengungsi Rohingya.

Rohingya merupakan kelompok etnis yang telah bermukim di wilayah Rakhine, Myanmar, mulai dari abad ke-7 Masehi. Meski demikian, mereka tidak dianggap sebagai warga negara Myanmar, melainkan menganggap mereka sebagai etnis Bengali [4]. Sebagai akibatnya, mereka sering kali menghadapi kekerasan dan pengusiran, sehingga terpaksa mengungsi ke negara-negara tetangga salah satunya adalah Indonesia. Masalah pengungsi Rohingya ini memberikan beban dan tantangan baru bagi negara-negara penerima, baik dari aspek ekonomi, sosial, serta budaya [5].

Kontroversi mengenai isu pengungsi Rohingya juga sangat terasa di media sosial, termasuk X. Banyak cuitan dan diskusi yang membahas masalah ini secara mendalam, mencerminkan perdebatan yang tajam antara pro dan kontra. Dari beberapa metode yang ada, salah satu metode untuk menganalisis respons masyarakat terhadap isu ini di media sosial adalah analisis sentimen. Metode ini melibatkan otomatisasi penambangan sikap, opini, pandangan, dan emosi dari teks melalui *Natural Language Processing* (NLP) [6]. *Support Vector Machine* (SVM) sering kali diterapkan pada analisis sentimen karena kemampuannya yang tinggi dalam mengolah data berdimensi besar dengan tingkat akurasi yang tinggi [7].

Ketidakseimbangan data adalah masalah umum dalam analisis sentimen media sosial, di mana model klasifikasi cenderung lebih baik dalam mengenali kelas mayoritas daripada kelas minoritas. Hal ini terjadi karena model lebih sering terpapar pada contoh kelas mayoritas, sehingga dapat mengabaikan kelas minoritas dan mengklasifikasikan semua data ke kelas mayoritas yang lebih dominan [8]. Untuk mengatasi masalah ketidakseimbangan ini, penelitian ini menggunakan teknik *oversampling* seperti *Synthetic Minority Over-sampling Technique* (SMOTE) dan *undersampling* seperti *Random Undersampling*. *Oversampling* dengan SMOTE bertujuan untuk menambah jumlah data pada kelas minoritas dengan menghasilkan contoh baru berdasarkan fitur yang ada, sehingga meningkatkan representasi kelas minoritas. Namun, metode ini dapat berisiko menyebabkan *overfitting* jika tidak dilakukan dengan hati-hati. Sebaliknya, *Random Undersampling* mengurangi jumlah data pada kelas mayoritas untuk menyeimbangkan distribusi data, tetapi dapat menghilangkan informasi penting yang mempengaruhi performa model [9]. Untuk mengatasi kekurangan masing-masing metode, penelitian ini mengimplementasikan kombinasi SMOTE dan *Random Undersampling*. Pendekatan ini diharapkan dapat memperbaiki kualitas klasifikasi sentimen dengan mengurangi bias terhadap kelas mayoritas dan meningkatkan akurasi model dalam mengidentifikasi kelas minoritas. Dengan kombinasi ini, diharapkan model dapat mencapai keseimbangan yang lebih baik dan performa klasifikasi yang lebih akurat pada seluruh kelas dalam *dataset*.

Penelitian oleh Darwis menggunakan algoritma SVM untuk analisis sentimen dan fitur-fitur yang diekstraksi melalui TF-IDF pada data Twitter terkait Komisi Pemberantasan Korupsi (KPK) Indonesia. Dari 2000 data yang diperoleh melalui proses *crawling*, penelitian ini mendapatkan 1890 data dan 3846 *term* setelah *preprocessing*. Data tersebut kemudian dikategorikan menjadi sentimen positif, negatif, dan netral. Hasil dari pengujian menunjukkan bahwa metode SVM berhasil mencapai akurasi sebesar 82%, dengan distribusi sentimen negatif 77%, positif 8%, dan netral 25% [3].

Penelitian yang berjudul "Analisis Sentimen Terkait Layanan Gofood dan Grabfood pada Media Sosial Twitter Menggunakan Algoritma *Support Vector Machine* (SVM)" bertujuan untuk mengevaluasi opini publik tentang layanan Gofood dan Grabfood di Twitter. Data dikumpulkan menggunakan kata kunci "gofood" dan "grabfood" selama periode 23 Februari hingga 27 Februari 2021, serta pada tanggal 5 Maret 2021. Tweet yang dikumpulkan kemudian diberi label secara manual sebagai kelas positif dan negatif, dan diklasifikasikan menggunakan algoritma *Support Vector Machine* (SVM). Pada model pertama, hasil evaluasi menunjukkan akurasi sebesar 80,18%, *recall* sebesar 100%, dan *specificity* sebesar 14%. Karena adanya ketidakseimbangan data, model kedua menggunakan metode *undersampling* pada data latih. Hasil evaluasi model kedua menunjukkan akurasi sebesar 79,26%, *recall* sebesar 86,23%, dan *specificity* sebesar 56% [10].

Chairunnisa dan Ernawati pada tahun 2022 [11] melakukan penelitian tentang bagaimana pengguna menilai aplikasi PeduliLindungi di *Google Play* dengan menggunakan algoritma *Support Vector Machine* dengan fitur *chi-square selection*. Dengan menggunakan *Python* untuk *scraping* di *Google Play*, data ulasan dikumpulkan dan dikategorikan menjadi ulasan positif dan negatif. Hasil klasifikasi menunjukkan performa yang baik dengan akurasi sebesar 93%, *recall* 86%, *precision* 98%, *specificity* 98%, dan *f1-score* 92%.

Penelitian oleh Sabilla dan Vista tahun 2021 [12] yang bertujuan untuk mengembangkan model klasifikasi yang dapat diandalkan untuk memprediksi kebangkrutan bisnis dengan menggunakan kombinasi SMOTE dan *undersampling*. Penelitian ini menggunakan klasifikasi *multilayer perceptron* dan *complement naïve bayes*. Hasil pengujian menunjukkan bahwa SMOTE dan *undersampling* cukup signifikan dalam meningkatkan model

klasifikasi *multilayer perceptron*. Nilai *recall* dan PR AUC juga meningkat pada prediksi menggunakan *complement naïve bayes*. Hasil pengujian menunjukkan peningkatan signifikan dalam kinerja model menggunakan *complement naïve bayes*. *Recall* terbaik sebesar 95,45% dengan metode ini juga diperoleh dengan ROC AUC tertinggi dengan *resampling* sebesar 87,80%.

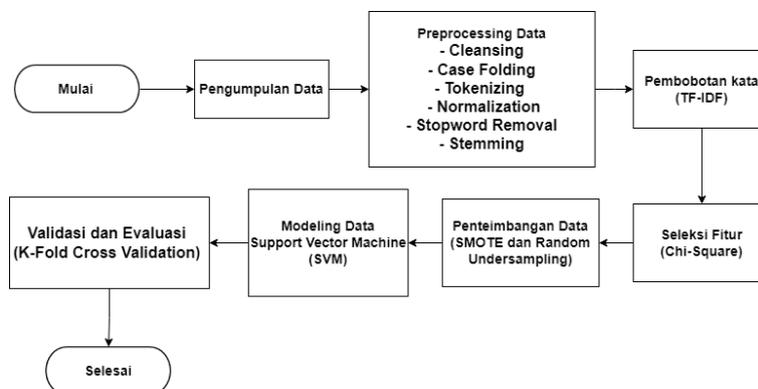
Ananda dan Suryono tahun 2024 [13] membandingkan kinerja algoritma *Naive Bayes* dan SVM dengan menggunakan 3350 tweet untuk mengevaluasi opini publik tentang pengungsi Rohingya di Indonesia. Data dibagi dengan rasio 70:30 untuk pelatihan dan pengujian model. Penelitian ini menerapkan metode *oversampling* menggunakan SMOTE untuk menangani *imbalance* data. Hasil menunjukkan bahwa, dengan tingkat kesalahan yang lebih rendah, model SVM lebih baik dalam memprediksi sentimen dari pada model *Naive Bayes*, dengan akurasi sebesar 76%, sedangkan *Naive Bayes* hanya 70%.

Selain algoritma SVM dan teknik *resampling*, terdapat metode lain yang relevan untuk analisis sentimen dalam konteks media sosial, seperti algoritma berbasis *deep learning* dan pendekatan berbasis *neural network*. Metode-metode ini telah menunjukkan hasil yang menjanjikan dalam berbagai penelitian. Model CNN telah memecahkan masalah dalam pemrosesan gambar, para peneliti saat ini mengembangkan CNN untuk proses NLP seperti analisis sentimen, klasifikasi polaritas emosional, kesimpulan teks, dan lainnya. CNN adalah bagian dari *Artificial Neural Network* (ANN) yang dapat mendeteksi informasi dengan sangat akurat. Penelitian yang bertujuan untuk memperoleh persepsi atau gambaran umum masyarakat terhadap penyelenggaraan pilkada 2020 saat pandemi covid-19 telah menunjukkan bahwa CNN dapat menghasilkan performa yang baik dalam klasifikasi sentimen dengan akurasi yang tinggi [14]. Salah satu teknik *deep learning* adalah *Neural Network Recurrent* (RNN). Modifikasi dari metode RNN adalah metode LSTM. Modifikasi ini dilakukan untuk mengatasi masalah RNN, yaitu kurangnya performa RNN dalam akurasi prediksi terhadap ketergantungan yang berlangsung lama. Penelitian yang menganalisis sentimen mahasiswa terhadap evaluasi pengajaran di perguruan tinggi dengan metode LSTM memberikan hasil yang cukup tinggi mencapai 91.08% [15].

Berdasarkan penelitian-penelitian sebelumnya dan kompleksitas permasalahan Rohingya, penelitian ini dilakukan untuk menganalisis sentimen publik terhadap isu Rohingya di X menggunakan algoritma SVM dan mengimplementasikan kombinasi *oversampling* (SMOTE) dan *undersampling* (*Random Undersampling*). Selain itu, penelitian ini juga akan membandingkan performa model dengan dan tanpa menggunakan seleksi *fitur chi-square*, untuk menentukan metode mana yang lebih efektif dalam klasifikasi sentimen. Penelitian ini diharapkan dapat memberikan gambaran mengenai tanggapan masyarakat terhadap masalah Rohingya dan memahami sentimen masyarakat X dalam menanggapi krisis tersebut.

## II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan sistematis untuk menganalisis sentimen masyarakat X terkait isu Rohingya dengan metode *Support Vector Machine* (SVM). Metode penelitian mencakup pengumpulan data melalui *tweet-harvest*, *preprocessing*, pembobotan kata dengan *tf-idf*, seleksi fitur menggunakan *chi-square*, penyeimbangan data dengan SMOTE dan *random undersampling*, pemodelan menggunakan SVM, validasi dan evaluasi dengan *k-fold cross-validation*. Setiap tahap dirancang untuk memastikan data yang bersih, seimbang, dan relevan untuk analisis yang akurat. Tahapan ini divisualisasikan pada Gambar 1 berikut :



Gambar. 1. Tahap Penelitian

### A. Pengumpulan Data

Untuk memeriksa persepsi masyarakat terhadap masalah Rohingya, penelitian ini menggunakan data media sosial X. Alat *tweet-harvest*, yang mengumpulkan setiap postingan berkaitan dengan masalah Rohingya, digunakan

untuk mengumpulkan data melalui teknik *crawling*. Dalam penelitian ini, kata kunci yang digunakan adalah "rohingya," dan periode pengambilan data berlangsung dari November 2023 hingga Desember 2023. Untuk memastikan bahwa label data akurat, setiap data yang dikumpulkan akan diberi label secara manual. Ini juga akan memastikan bahwa hanya data yang relevan dan layak digunakan dalam penelitian. *Dataset* secara keseluruhan berjumlah 2165 tweet, terdiri dari 257 tweet positif, 437 tweet netral, dan 1471 tweet negatif.

TABEL I  
 DATA CRAWLING DAN LABELING DATA

No	Tweet	Label
1.	@nnojaemmm @iam_jellal @tanyakanrl wkwkwk emang 🤔🤔 masa nyamain palestina sama rohingya. meskipun sama" muslim palestina lebih beradab dan rela mati buat mempertahankan tanah mereka. rohingya apaan udah bikin rusuh sana sini malah mau minta tanah orang.	negatif
2	@porhaze @BBCIndonesia Rohingya ini dari mana mereka asalnya? Rohingya ini kasihan juga. Biar bgaimanapun, mreka manusia. Hanya masalahnya mreka terkenal dgn pembangkang & tdk tau terima kasih.	netral
1471.	Ini pemerintah Indonesia harus serius & tdk bisa mreka terus di Indonesia. @jokowi PBB, minimal cari pulau kosong untuk mreka agar mreka tempati.!	positif

### B. Preprocessing Data

Tahap *text preprocessing* merupakan langkah awal yang krusial dalam *text mining*, yang bertujuan untuk mempersiapkan data untuk operasi *knowledge discovery* [16]. Proses ini memastikan bahwa data yang digunakan sudah dalam kondisi optimal untuk analisis lebih lanjut. *Text preprocessing* melibatkan serangkaian rutinitas dan proses yang sistematis. Beberapa tahapan penting dalam *text preprocessing* antara lain:

#### 1) Cleansing

Tujuan dari proses ini adalah untuk menghilangkan elemen teks yang tidak penting seperti *hashtag*, *mention*, dan tanda baca. *Cleansing* membantu menghilangkan gangguan yang bisa mengaburkan hasil analisis. Berikut data *clean* ditampilkan dalam Tabel II.

TABEL II  
 CLEANSING DATA

Tweet	Clean
@nnojaemmm @iam_jellal @tanyakanrl wkwkwk emang 🤔🤔 masa nyamain palestina sama rohingya. meskipun sama" muslim palestina lebih beradab dan rela mati buat mempertahankan tanah mereka. rohingya apaan udah bikin rusuh sana sini malah mau minta tanah orang.	wkwkwk emang masa nyamain palestina sama rohingya meskipun sama muslim palestina lebih beradab dan rela mati buat mempertahankan tanah mereka rohingya apaan udah bikin rusuh sana sini malah mau minta tanah orang

#### 2) Case Folding

Pada tahap ini, semua huruf dalam dokumen diubah menjadi huruf kecil (*lowercase*) untuk memastikan konsistensi dan menghindari perbedaan akibat penggunaan huruf besar dan kecil yang tidak relevan dalam analisis teks. Berikut hasil *case folding* ditampilkan dalam Tabel III.

TABEL III  
 CASE FOLDING DATA

Clean	Case Folding
wkwkwk emang masa nyamain palestina sama rohingya meskipun sama muslim palestina lebih beradab dan rela mati buat mempertahankan tanah mereka rohingya apaan udah bikin rusuh sana sini malah mau minta tanah orang	wkwkwk emang masa nyamain palestina sama rohingya meskipun sama muslim palestina lebih beradab dan rela mati buat mempertahankan tanah mereka rohingya apaan udah bikin rusuh sana sini malah mau minta tanah orang

#### 3) Tokenizing

Tahap ini teks dibagi menjadi bagian-bagian yang lebih kecil, seperti kata-kata atau token. Proses ini penting untuk analisis berbasis kata karena memungkinkan pengolahan teks pada tingkat kata per kata. Berikut hasil *tokenizing* ditampilkan dalam Tabel IV.

TABEL IV  
 TOKENIZING DATA

Case Folding	Token
wkwkwk emang masa nyamain palestina sama rohingya meskipun sama muslim palestina lebih beradab dan rela mati buat mempertahankan tanah mereka rohingya apaan udah bikin rusuh sana sini malah mau minta tanah orang	['wkwkwk', 'emang', 'masa', 'nyamain', 'palestina', 'sama', 'rohingya', 'meskipun', 'sama', 'muslim', 'palestina', 'lebih', 'beradab', 'dan', 'rela', 'mati', 'buat', 'mempertahankan', 'tanah', 'mereka', 'rohingya', 'apaan', 'udah', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'minta', 'tanah', 'orang']

#### 4) Normalization

Tahap ini merupakan proses perbaikan kata-kata yang mengalami penyingkatan atau variasi penulisan, sehingga semua variasi dari kata yang sama diwakili dalam bentuk standar yang konsisten. Berikut hasil *normalization* ditampilkan dalam Tabel V.

TABEL V  
 NORMALIZATION DATA

Token	Normalization
['wkwkwk', 'emang', 'masa', 'nyamain', 'palestina', 'sama', 'rohingya', 'meskipun', 'sama', 'muslim', 'palestina', 'lebih', 'beradab', 'dan', 'rela', 'mati', 'buat', 'mempertahankan', 'tanah', 'mereka', 'rohingya', 'apaan', 'udah', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'minta', 'tanah', 'orang']	['wkwkwk', 'memang', 'masa', 'menyamakan', 'palestina', 'sama', 'rohingya', 'meskipun', 'sama', 'muslim', 'palestina', 'lebih', 'beradab', 'dan', 'rela', 'mati', 'buat', 'mempertahankan', 'tanah', 'mereka', 'rohingya', 'apaan', 'sudah', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'meminta', 'tanah', 'orang']

#### 5) Stopword Removal

Pada tahap ini, kata-kata umum seperti "dan", "atau", dan "yang" dihilangkan. Proses ini membantu dalam menurunkan *noise* dan meningkatkan fokus pada kata-kata yang lebih penting. Berikut hasil *stopword removal* ditampilkan dalam Tabel VI.

TABEL VI  
 STOPWORD REMOVAL DATA

Normalization	Stopword
['wkwkwk', 'memang', 'masa', 'menyamakan', 'palestina', 'sama', 'rohingya', 'meskipun', 'sama', 'muslim', 'palestina', 'lebih', 'beradab', 'dan', 'rela', 'mati', 'buat', 'mempertahankan', 'tanah', 'mereka', 'rohingya', 'apaan', 'sudah', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'meminta', 'tanah', 'orang']	['wkwkwk', 'memang', 'masa', 'menyamakan', 'palestina', 'sama', 'rohingya', 'meskipun', 'sama', 'muslim', 'palestina', 'lebih', 'beradab', 'rela', 'mati', 'buat', 'mempertahankan', 'tanah', 'rohingya', 'apaan', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'meminta', 'tanah', 'orang']

#### 6) Stemming

Tahap ini mengembalikan suatu kata ke kata dasarnya (*root form*). *Stemming* membantu mengkonsolidasikan berbagai bentuk kata yang berbeda ke dalam bentuk dasar yang sama, sehingga analisis dapat lebih terfokus pada konsep utama yang diwakili oleh kata-kata tersebut. Hasil dari tahapan *stemming* ditampilkan dalam tabel VII.

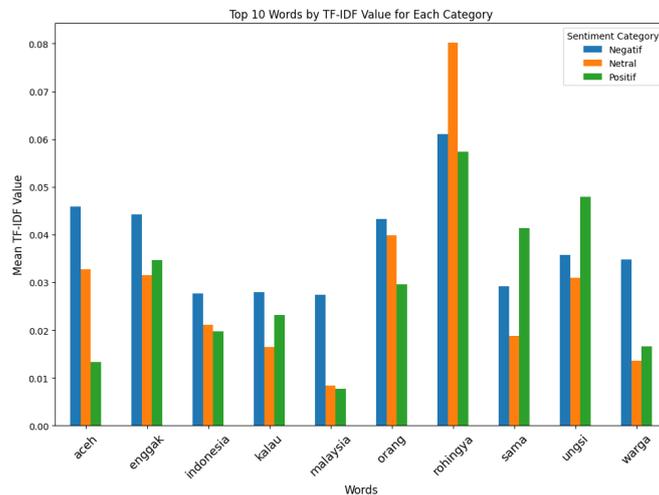
TABEL VII  
 STEMMING DATA

Stopword	Stemming
['wkwkwk', 'memang', 'masa', 'menyamakan', 'palestina', 'sama', 'rohingya', 'meskipun', 'sama', 'muslim', 'palestina', 'lebih', 'beradab', 'rela', 'mati', 'buat', 'mempertahankan', 'tanah', 'rohingya', 'apaan', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'meminta', 'tanah', 'orang']	['wkwkwk', 'memang', 'masa', 'sama', 'palestina', 'sama', 'rohingya', 'meski', 'sama', 'muslim', 'palestina', 'lebih', 'adab', 'rela', 'mati', 'buat', 'tahan', 'tanah', 'rohingya', 'apa', 'bikin', 'rusuh', 'sana', 'sini', 'malah', 'mau', 'minta', 'tanah', 'orang']

### C. Ekstraksi Fitur dengan TF-IDF

Setelah proses *preprocessing*, pembobotan kata adalah tahap penting. Tujuan pembobotan kata adalah untuk membuat data yang tidak terstruktur menjadi terstruktur, yang kemudian dapat digunakan untuk klasifikasi dengan

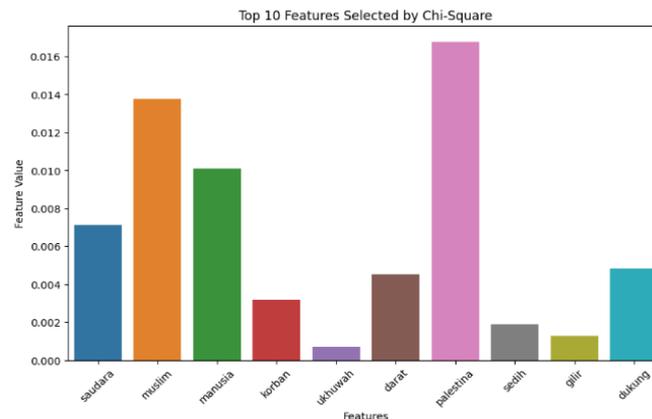
menggunakan metode *classifier* [17]. Metode TF-IDF adalah cara yang bagus untuk mengetahui seberapa penting hubungan antara kata-kata (*terms*) dan dokumen. Metode ini menggabungkan dua ide utama yaitu frekuensi kata dalam sebuah dokumen (*term frequency*) dan frekuensi inversi dokumen atau seberapa jarang kata muncul diseluruh dokumen. Tujuannya adalah untuk meningkatkan kemampuan untuk mencari dokumen yang relevan dan penting. Oleh karena itu, standar ideal adalah kata-kata yang sering muncul tetapi jarang muncul di satu dokumen [18]. Pada gambar 2 menunjukkan distribusi rata-rata nilai TF-IDF dari 10 kata teratas untuk setiap kategori sentimen positif, netral, maupun negatif.



Gambar. 2. 10 Nilai TF-IDF Teratas

#### D. Seleksi Fitur dengan Chi-Square

Menemukan *subset* fitur yang ideal dalam data berdimensi tinggi merupakan tantangan dalam klasifikasi. Seleksi fitur adalah proses untuk memilih fitur yang relevan dengan tujuan memperoleh fitur yang optimal. Salah satu teknik penyeleksian fitur adalah *Chi-Square*, yang menggunakan teori statistika untuk mengevaluasi independensi sebuah kata dengan kategorinya [19].



Gambar. 3. 10 Nilai *Chi-Square* Teratas

Gambar 3 tersebut menampilkan sepuluh fitur terbaik dengan skor *chi-square* tertinggi menunjukkan hubungan yang kuat antara fitur-fitur ini dan kategori sentimen. Dari 3820 fitur yang ada, dipilih sebanyak 2000 fitur terbaik yang akan digunakan. Penentuan nilai k atau jumlah fitur ini didasarkan pada eksperimen awal dengan berbagai nilai k, di mana k=2000 menunjukkan hasil yang optimal dalam hal akurasi dan efisiensi.

#### E. Penyeimbangan Data

Masalah ketidakseimbangan data umum terjadi dalam analisis sentimen, dimana beberapa kelas sentimen memiliki jumlah yang signifikan lebih banyak dari pada yang lain. Model pembelajaran mesin seringkali memprediksi kelas mayoritas dan mengabaikan kelas minoritas karena ketidakseimbangan ini. Akibatnya, mereka lebih sulit untuk mengklasifikasikan kelas yang lebih jarang. Untuk mengatasi masalah ini, teknik penyeimbangan data berikut diterapkan pada dataset ini:

1) SMOTE

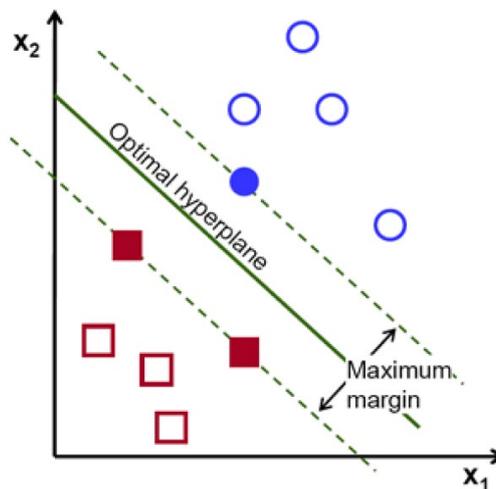
Oversampling adalah teknik untuk menyeimbangkan jumlah data minoritas sehingga jumlahnya setara dengan data mayoritas [20]. SMOTE akan membuat sampel sintetik dari kelas minoritas dengan menggabungkan sampel minoritas dengan sampel terdekat [21].

2) Random Undersampling

Metode random undersampling akan memilih data dari kelas mayoritas untuk dihapus dari kumpulan data latih dalam proses undersampling. Dengan melakukan random undersampling, jumlah data latih dari kelas mayoritas akan berkurang. Proses undersampling dapat diulang hingga diperoleh distribusi kelas yang diinginkan pada data latih [10].

F. Modeling Data

Model pembelajaran mesin Support Vector Machine (SVM) beroperasi di bawah pengawasan algoritma pembelajaran terarah. Algoritma ini menganalisis data yang digunakan dalam analisis regresi dan klasifikasi [18]. Pada awalnya, Support Vector Machine (SVM) diterapkan untuk klasifikasi dalam dua kelas, tetapi telah berkembang dan diperluas untuk klasifikasi multi kelas. Jika pemisah tersebut berupa garis dalam dua dimensi, pemisah tersebut berupa plane dalam tiga dimensi, dan pemisah yang memiliki lebih dari tiga dimensi disebut hyperplane [19].



Gambar. 4. Hyperplane Terbaik Memisahkan 2 Kelas

Gambar 2 merupakan ilustrasi dari hyperplane yang ditemukan, berada diantara dua kelas yaitu bulat biru dan kotak merah, hyperplane ini dihasilkan dengan maksimalisasi margin antara kedua kelas, yang berarti garis tersebut ditempatkan sedemikian rupa sehingga memberikan jarak maksimum antara titik-titik terdekat dari masing-masing kelas, yang dikenal sebagai support vectors. SVM biasanya diterapkan untuk pengklasifikasian secara linier dengan rumus :

$$K(x, y) = x \cdot y \tag{1}$$

Dengan menggunakan kernel trick, Support Vector Machine dapat bekerja pada dataset yang sangat besar. Untuk membangun model klasifikasi, Support Vector Machine hanya memanfaatkan beberapa titik data yang berperan sebagai support vector [22]. Kernel yang bisa digunakan antara lain adalah fungsi linear, polynomial, gaussian RBF dan sigmoid. Berikut rumus kernel trick yang umum digunakan adalah sebagai berikut :

Polynomial Kernel :

$$K(x, y) = (x \cdot y + c)^d \tag{2}$$

Sigmoid Kernel :

$$K(x, y) = \tanh(\sigma(x \cdot y) + c) \tag{3}$$

Radial Basis Function Kernel :

$$K(x, y) = \exp \frac{-||x - y||^2}{2 \cdot \sigma^2} \tag{4}$$

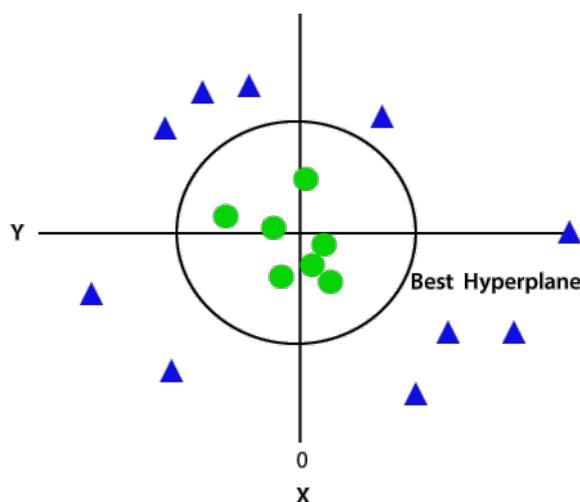
Keterangan :

$c$  = constanta

$d$  = degree

exp = eksponensial

Berikut Gambar 3 yang merupakan prinsip *Support Vector Machine* (SVM) dalam menangani data *non-linear* dengan menggunakan transformasi ke ruang berdimensi lebih tinggi. Lingkaran dalam gambar tersebut merupakan *hyperplane* terbaik yang memisahkan dua kelas data yaitu kelas berbentuk bulat dan kelas berbentuk segitiga.



Gambar. 5. *Hyperplane* Terbaik Non-Linear

### G. Validasi dan Evaluasi

Evaluasi dan validasi model sangat penting untuk mengukur kinerja algoritma yang digunakan dalam analisis sentimen. Pada penelitian ini, evaluasi dan validasi dilakukan menggunakan metode yang sering digunakan untuk memastikan hasil yang akurat dan andal. *Cross Validation* merupakan metode yang dimanfaatkan untuk menelisik keberhasilan suatu model dengan menerapkan perulangan dengan komposisi setiap data pernah menjadi data latih maupun data uji sehingga algoritma yang digunakan teruji validitasnya. Sedangkan *K-fold cross validation* adalah metode validasi silang yang menggabungkan data ke dalam set data  $k$  dengan pembagian jumlah yang seimbang. Dalam proses ini, pelatihan dan pengujian dilakukan sebanyak  $k$  kali, dan data dibagi menjadi  $k$  bagian [23].

## III. HASIL DAN PEMBAHASAN

Dalam penelitian ini, *dataset* yang digunakan terdiri dari 2165 data dengan distribusi yang tidak seimbang, yaitu 1471 data negatif, 437 data netral, dan 257 data positif. Ketidakseimbangan ini dapat mempengaruhi performa model analisis sentimen karena model cenderung lebih baik dalam mengenali kategori yang memiliki lebih banyak contoh (negatif) dibandingkan kategori yang memiliki lebih sedikit contoh (netral dan positif). Oleh karena itu, dilakukan teknik *oversampling* dan *undersampling* untuk menyeimbangkan distribusi data.

*Synthetic Minority Over-sampling Technique* (SMOTE) digunakan untuk meningkatkan jumlah data dalam kategori minoritas (netral dan positif) yang ditingkatkan menjadi dua kali jumlah data netral awal, yaitu 874 data. Setelah melakukan *oversampling*, diterapkan teknik *random undersampling* untuk menurunkan jumlah data dalam kategori mayoritas (negatif) dari 1471 menjadi 874, sehingga distribusi data menjadi seimbang. Berikut merupakan hasil *oversampling* dan *undersampling* yang ditampilkan pada Tabel VIII.

TABEL VIII  
 HASIL OVERSAMPLING DAN UNDERSAMPLING

Sentimen	Sebelum	Sesudah
Negatif	1471	874
Netral	437	874
Positif	257	874
Total	2165	2622

Pemilihan *oversampling* hanya untuk kategori netral dan positif, serta penambahan jumlah data berdasarkan dua kali jumlah data netral, dilakukan untuk menghindari pembuatan data sintetis yang terlalu banyak. Jumlah data netral yang awalnya 437 dipilih karena berada di tengah antara jumlah data positif dan negatif, menjadikannya dasar yang wajar untuk penyeimbangan data. Dengan meningkatkan jumlah data netral dan positif menjadi 874, proses *oversampling* menghasilkan jumlah data sintetis yang moderat dan efisien. Selain itu, dengan melakukan *undersampling* terhadap data negatif dari 1471 menjadi 874, membuat jumlah data yang dihapus tidak terlalu banyak, sehingga tetap mempertahankan informasi penting dalam data negatif.

Setelah proses *resampling*, algoritma *Support Vector Machine* (SVM) dengan *kernel Radial Basis Function* (RBF) dan parameter  $C=100$  diterapkan untuk klasifikasi sentimen. *Kernel RBF* dipilih berdasarkan hasil *Grid Search* dengan parameter  $\{C: [0.1, 1, 100], 'kernel': ['rbf', 'poly', 'sigmoid', 'linear']\}$ . *Grid Search* menunjukkan bahwa kombinasi  $C=100$  dan *kernel RBF* menghasilkan nilai skor rata-rata tertinggi sebesar 0.8999, menjadikannya sebagai pilihan optimal dibandingkan *kernel* lainnya. Tabel IX di bawah ini menampilkan lima kombinasi parameter teratas dari *Grid Search*:

TABEL IX  
HASIL 5 TERATAS GRID SEARCH PARAMETE SVM

C	Kernel	Mean Score
100	'rbf'	0.8999
1	'rbf'	0.8875
1	'poly'	0.8770
100	'linear'	0.8693
1	'linear'	0.8417

*Kernel RBF* dipilih karena kemampuannya dalam menangani data yang tidak terpisahkan secara linier dengan menyesuaikan *margin* keputusan secara fleksibel. Parameter  $C=100$  berfungsi untuk mengontrol *trade-off* antara *margin* maksimal dan kesalahan klasifikasi, dengan penalti yang lebih besar terhadap kesalahan untuk mencapai *margin* pemisahan yang lebih baik. *Dataset* dibagi menjadi data latih dan data uji dengan proporsi 80:20. Proses klasifikasi juga dibandingkan pada saat menggunakan *Chi-Square* dan tidak. Evaluasi menggunakan *10-fold cross validation* menunjukkan hasil akurasi pengujian sebesar 89.0%, rata-rata akurasi *cross-validation* 0.89, rata-rata *precision* 0.90, rata-rata *recall* 0.89, dan rata-rata *F1-score* 0.89. Seleksi fitur menggunakan *chi-square* meningkatkan akurasi pengujian menjadi 90.0%, dengan semua matriks evaluasi meningkat menjadi 0.90. Meskipun peningkatan akurasi dari 89% ke 90% terlihat kecil, peningkatan ini signifikan dalam konteks aplikasi praktis. Peningkatan akurasi sebesar 1% dapat berarti ratusan atau ribuan prediksi tambahan yang benar dalam skala besar, yang dapat sangat bermanfaat dalam analisis sentimen untuk pengambilan keputusan bisnis atau pemantauan opini publik. Berikut adalah hasil klasifikasi yang dirangkum dalam Tabel X :

TABEL X  
HASIL K-FOLD CROSS VALIDATION DENGAN DAN TANPA CHI-SQUARE

Fold	Tanpa <i>Chi-Square</i>				Menggunakan <i>Chi-Square</i>			
	Akurasi	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	Akurasi	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
Fold 1	0,89	0,89	0,88	0,88	0,90	0,9	0,9	0,9
Fold 2	0,91	0,91	0,91	0,91	0,89	0,89	0,89	0,89
Fold 3	0,9	0,9	0,9	0,9	0,9	0,9	0,9	0,9
Fold 4	0,88	0,89	0,88	0,88	0,9	0,9	0,9	0,9
Fold 5	0,89	0,9	0,89	0,89	0,91	0,91	0,91	0,91
Fold 6	0,88	0,88	0,88	0,88	0,9	0,91	0,91	0,91
Fold 7	0,9	0,9	0,9	0,9	0,9	0,9	0,9	0,9
Fold 8	0,9	0,9	0,89	0,89	0,91	0,91	0,91	0,91
Fold 9	0,9	0,9	0,9	0,9	0,89	0,89	0,89	0,89
Fold 10	0,9	0,91	0,9	0,9	0,91	0,92	0,92	0,92
Rata-rata	0.89	0,9	0,89	0,89	0,9	0,9	0,9	0,9

Visualisasi data adalah langkah penting dalam analisis sentimen untuk memahami distribusi dan tren dalam *dataset*. Dalam penelitian ini, digunakan *word cloud* untuk visualisasi. *Word cloud* adalah gambar visual yang menunjukkan frekuensi kata-kata dalam teks, dengan ukuran huruf mencerminkan frekuensi kata, kata yang lebih sering muncul ditampilkan dengan huruf lebih besar [24]. Pada Gambar 6, 7 dan 8 memperlihatkan kata-kata yang paling sering muncul dalam tweet dengan sentimen positif, netral dan negatif.



SVM dibandingkan *multilayer perceptron* dan *complement naïve bayes*, namun keduanya menunjukkan bahwa *resampling* dapat meningkatkan performa model secara signifikan dalam konteks data yang tidak seimbang.

Dalam hal penggunaan seleksi fitur *chi-square*, temuan penelitian ini berbanding dengan studi Chairunnisa dan Ernawati tahun 2022 [11], yang mencapai akurasi 93%, *recall* 86%, *precision* 98%, *specificity* 98%, dan *F1-score* 92% dalam analisis sentimen aplikasi PeduliLindungi menggunakan *chi-square* dengan SVM. Meskipun penelitian ini menunjukkan peningkatan akurasi dari 89% menjadi 90% dengan penggunaan *chi-square*, peningkatan tersebut tidak sebesar yang ditemukan dalam studi tersebut. Hal ini mengindikasikan bahwa meskipun *chi-square* efektif dalam meningkatkan performa model, dampaknya bervariasi tergantung pada konteks dan data yang digunakan.

Temuan ini memperluas pemahaman tentang penerapan teknik *resampling* dan seleksi fitur dalam analisis sentimen, terutama dalam konteks isu sosial yang kompleks seperti pengungsi Rohingya. Meskipun peningkatan akurasi dari 89% ke 90% terbilang kecil, hal ini menunjukkan bahwa *chi-square* dapat memberikan kontribusi positif terhadap performa model SVM dalam analisis sentimen. Penelitian ini menegaskan pentingnya memilih teknik yang tepat untuk menangani ketidakseimbangan data dan menilai bagaimana seleksi fitur dapat mempengaruhi hasil analisis dalam konteks yang berbeda.

Hasil penelitian ini menunjukkan bahwa kombinasi teknik penyeimbangan data dan algoritma SVM dengan seleksi fitur *chi-square* dapat menghasilkan model analisis sentimen yang kuat dan akurat. Teknik-teknik ini dapat membantu model mengenali sentimen dari berbagai kategori secara merata, serta mengurangi bias model terhadap kategori mayoritas. Oleh karena itu, penelitian ini membuka jalan bagi eksplorasi tambahan dalam penerapan teknik-teknik seleksi fitur dan *resampling* pada analisis data serta hasil penelitian ini sejalan dengan studi terdahulu, menunjukkan bahwa teknik-teknik tersebut dapat meningkatkan performa model.

Penelitian selanjutnya dapat mengeksplorasi berbagai teknik penyeimbangan data dan algoritma tambahan untuk meningkatkan performa model analisis sentimen. Algoritma seperti *Long Short-Term Memory* (LSTM), *Random Forest*, dan *Gradient Boosting* memiliki potensi untuk menawarkan hasil yang lebih baik dalam klasifikasi sentimen. Untuk teknik penyeimbangan data, metode *oversampling* dapat dilakukan dengan *Support Vector Machine SMOTE* (SVM-SMOTE) atau *Random Over Sampling* (ROS), sementara metode *undersampling* dapat diterapkan menggunakan *TomekLinks*, *Edited Nearest Neighbors* (ENN), atau *Neighborhood Cleaning Rule* (NCR). Kombinasi antara teknik *oversampling* dan *undersampling*, seperti SMOTE-ENN dan SMOTE-Tomek, juga layak untuk dieksplorasi dalam upaya mengoptimalkan distribusi data. Lebih jauh, penelitian mendatang dapat mengkaji penerapan teknik seleksi fitur yang lebih canggih serta integrasi model-model berbasis *deep learning* seperti LSTM untuk meningkatkan akurasi dan performa keseluruhan model dalam analisis sentimen. Eksplorasi ini diharapkan dapat membuka jalan bagi peningkatan signifikan dalam kemampuan model untuk mengenali dan mengklasifikasikan sentimen secara lebih akurat dan adil.

#### IV. KESIMPULAN

Penelitian ini menyajikan analisis sentimen terhadap masalah Rohingya dengan menggunakan algoritma *Support Vector Machine* (SVM) dan teknik *resampling* data. Hasil penelitian menunjukkan bahwa penggunaan teknik SMOTE dan *random undersampling* efektif dalam mengatasi ketidakseimbangan data, meningkatkan performa model secara signifikan. Setelah menerapkan teknik-teknik tersebut, model mencapai nilai akurasi sebesar 89%, *recall* sebesar 90%, *precision* sebesar 89%, dan *f1-score* sebesar 89%. Selain itu, penggunaan seleksi fitur *chi-square* memberikan peningkatan akurasi dari 89% menjadi 90%. Hal ini menunjukkan bahwa *chi-square* dapat membantu meningkatkan kinerja model, meskipun peningkatan ini kecil. Temuan ini menunjukkan bahwa kombinasi teknik-teknik tersebut dapat digunakan untuk meningkatkan akurasi analisis sentimen, khususnya dalam konteks data yang tidak seimbang. Penelitian ini memberikan kontribusi penting dalam mengembangkan metode analisis sentimen yang lebih efektif dan dapat diterapkan dalam berbagai konteks.

#### DAFTAR PUSTAKA

- [1] R. Wasono, "Twitter Mengenai Netflix Diblokir Telkom Menggunakan Support," 2022.
- [2] S. Kemp, "Digital 2023: Indonesia - DataReportal - Global Digital Insights," *datareportal.com*, 2023. <https://datareportal.com/reports/digital-2023-indonesia> (accessed Jan. 12, 2024).
- [3] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *Educat - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 1–11, 2020, doi: 10.21107/edutic.v7i1.8779.
- [4] K. Arianta, D. G. S. Mangku, and N. P. R. Yuliantini, "Perlindungan Hukum Bagi Kaum Etnis Rohingya Dalam Perspektif Hak Asasi Manusia Internasional," *J. Komunitas Yustitia Univ. Pendidik. Ganessa Jur. Ilmu Huk.*, vol. 3, no. 2, pp. 166–176, 2020.
- [5] B. Budaya, "Dampak Kewarganegaraan Etnis Rohingya Di Myanmar Terhadap Pelanggaran Hak Asasi Manusia Dan Negara Sekitar," *J. Ilm. Huk.*, vol. 11, no. 1, pp. 106–120, 2017, [Online]. Available: <https://maksigama.wisnuwardhana.ac.id/index.php/maksigama/article/view/44>
- [6] A. Andreyestha and A. Subekti, "Analisa Sentiment Pada Ulasan Film Dengan Optimasi Ensemble Learning," *J. Inform.*, vol. 7, no. 1, pp. 15–23, 2020, doi: 10.31311/ji.v7i1.6171.
- [7] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

- [8] A. Syukron, S. Sardiarto, E. Saputro, and P. Widodo, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," *J. Teknol. Inf. dan Terap.*, vol. 10, no. 1, pp. 47–50, 2023, doi: 10.25047/jtit.v10i1.313.
- [9] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, 2021, doi: 10.33387/jiko.v4i1.2561.
- [10] R. Agsar Dwi Anggoro, R. Al Habsi, M. Arlanda Valio, Y. Widiastiti, and N. Chamidah, "Analisis Sentimen Terkait Layanan Gofood dan Grabfood pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, no. September, pp. 249–259, 2021.
- [11] C. Chairunnisa, I. Ernawati, and M. M. Santoni, "Klasifikasi Sentimen Ulasan Pengguna Aplikasi PeduliLindungi di Google Play Menggunakan Algoritma Support Vector Machine dengan Seleksi Fitur Chi-Square," *Inform. J. Ilmu Komput.*, vol. 18, no. 1, p. 69, 2022, doi: 10.52958/iftk.v17i4.4594.
- [12] W. I. Sabilla and C. Bella Vista, "Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan," *J. Komput. Terap.*, vol. 7, no. 2, pp. 329–339, 2021, doi: 10.35143/jkt.v7i2.5027.
- [13] D. Ananda and R. R. Suryono, "Analisis Sentimen Publik Terhadap Pengungsi Rohingya di Indonesia dengan Metode Support Vector Machine dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 8, no. April, pp. 748–757, 2024, doi: 10.30865/mib.v8i2.7517.
- [14] S. N. Listyarini and D. A. Anggoro, "Analisis Sentimen Pilkada di Tengah Pandemi Covid-19 Menggunakan Convolution Neural Network (CNN)," *J. Pendidik. dan Teknol. Indones.*, vol. 1, no. 7, pp. 261–268, 2021, doi: 10.52436/1.jpti.60.
- [15] M. A. Amrustian, W. Widayat, and A. M. Wirawan, "Analisis Sentimen Evaluasi Terhadap Pengajaran Dosen di Perguruan Tinggi Menggunakan Metode LSTM," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 535, 2022, doi: 10.30865/mib.v6i1.3527.
- [16] A. Z. Amrullah, A. Sofyan Anas, and M. A. J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *Jurnal*, vol. 2, no. 1, pp. 40–44, 2020, doi: 10.30812/bite.v2i1.804.
- [17] R. Ramadhan, Y. A. Sari, and P. P. Adikara, "Perbandingan Pembobotan Term Frequency-Inverse Document Frequency dan Term Frequency-Relevance Frequency terhadap Fitur N-Gram pada Analisis Sentimen," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 11, pp. 5075–5079, 2021, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [18] M. Nurjannah and I. Fitri Astuti, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman," *J. Inform. Mulawarman*, vol. 8, no. 3, pp. 110–113, 2013.
- [19] T. Emayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, "Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naive Bayes Untuk Analisis Sentimen Pelanggan Tokopedia," *J. Gaussian*, vol. 11, no. 4, pp. 562–571, 2023, doi: 10.14710/j.gauss.11.4.562-571.
- [20] S. Diantika, "Penerapan Teknik Random Oversampling Untuk Mengatasi Imbalance Class Dalam Klasifikasi Website Phishing Menggunakan Algoritma Lightgbm," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 1, pp. 19–25, 2023, doi: 10.36040/jati.v7i1.6006.
- [21] M. Rahayu, A. Luthfiarta, L. Cahyaningrum, and A. Nurfaiza Azzahra, "Pengaruh Oversampling dan Cross Validation Pada Model Machine Learning Untuk Sentimen Analisis Kebijakan Luaran Kelulusan Mahasiswa," *J. Media Inform. Budidarma*, vol. 8, no. 1, pp. 163–172, 2024, doi: 10.30865/mib.v8i1.7012.
- [22] Herianto, "Penerapan Text-Mining Untuk Mengidentifikasi," vol. VIII, no. 2, pp. 36–44, 2019.
- [23] D. F. Zhafira, B. Rahayudi, and I. Indriati, "Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF Berdasarkan Komentar pada Youtube," *J. Sist. Informasi, Teknol. Informasi, dan Edukasi Sist. Inf.*, vol. 2, no. 1, pp. 55–63, 2021, doi: 10.25126/justsi.v2i1.24.
- [24] I. T. Julianto, "Analisis Sentimen Terhadap Sistem Informasi Akademik Institut Teknologi Garut," *J. Algoritma*, vol. 19, no. 1, pp. 449–456, 2022, doi: 10.33364/algoritma/v.19-1.1112.