Analisis Sentimen Komentar *Youtube* Potensi Gempa *Megathrust* Di Indonesia Menggunakan Algoritma *Support Vector Machine* (SVM)

1st Ade Krisna Subagyo
Sistem Informasi, Fakultas Rekayasa Industri
Universitas Telkom Purwokerto
Purwokerto, Indonesia
adekrisna@student.telkomuniversity.ac.id

2nd Sena Wijayanto
Sistem Informasi, Fakultas Rekayasa Industri
Universitas Telkom Purwokerto
Purwokerto, Indonesia
senawijayanto@telkomuniversity.ac.id

Abstrak — Indonesia berada di kawasan Cincin Api Pasifik, menjadikannya salah satu negara dengan risiko gempa bumi tertinggi di dunia, termasuk potensi gempa *megathrust*. Peringatan resmi dari BMKG tentang potensi gempa besar ini menimbulkan kekhawatiran publik, yang tercermin dalam komentar masyarakat di platform media sosial *YouTube* . Analisis terhadap komentar-komentar ini penting untuk memahami persepsi masyarakat serta mendukung strategi komunikasi risiko bencana. Penelitian ini bertujuan untuk mengidentifikasi dan mengklasifikasikan sentimen masyarakat menjadi dua sentimen, yaitu sentimen positif dan sentimen negatif, menggunakan pendekatan analisis sentimen berbasis machine learning dengan algoritma *Support Vector Machine* (SVM). Data dikumpulkan dari dua video *YouTube* populer, kemudian melalui tahapan pra-pemrosesan, pelabelan menggunakan *SenticNet*, pembobotan TF-IDF, penyeimbangan data dengan SMOTETomek, serta pengujian model menggunakan empat kernel SVM (Linear, RBF, *Polynomial*, dan *Sigmoid*) pada dua skenario pembagian data: 80:20 dan 90:10. Hasil menunjukkan kernel RBF menghasilkan akurasi tertinggi (89%) pada skenario 80:20, namun bias terhadap kelas positif. *Kernel* Linear dan *Sigmoid* lebih seimbang dan stabil. *Kernel* Linear dipilih sebagai model terbaik untuk klasifikasi sentimen dalam penelitian ini.

Kata kunci : gempa *megathrust* , analisis sentimen, komentar *YouTube* , support vector machine, SMOTETomek

I. PENDAHULUAN

Indonesia terletak di Cincin Api Pasifik, wilayah aktif tektonik yang menjadikannya sangat rentan terhadap bencana alam seperti gempa bumi dan tsunami [1]. BMKG pada 30 November 2024 memperingatkan potensi gempa *megathrust* di selatan Jawa dan Sumatra dengan magnitudo hingga 9,1 yang berisiko memicu tsunami besar. Informasi ini menimbulkan berbagai reaksi publik, mulai dari kekhawatiran hingga kritik terhadap kesiapan pemerintah. Tingkat kesadaran masyarakat menjadi elemen penting dalam mitigasi risiko bencana [2].

Di era digital, media sosial seperti *YouTube* menjadi sarana utama penyebaran informasi dan ekspresi opini publik. Kolom komentar pada video-video terkait gempa *megathrust*, seperti milik CNN Indonesia dan BBC News Indonesia, menunjukkan respons beragam yang mencerminkan persepsi masyarakat. Namun, tidak semua komentar akurat, beberapa mengandung hoaks yang dapat memicu kepanikan. Oleh karena itu, analisis sentimen diperlukan untuk memahami persepsi publik secara sistematis [3].

Analisis sentimen merupakan cabang dari *Natural Language Processing* (NLP) yang bertujuan mengklasifikasikan opini masyarakat menjadi kategori positif atau negatif [4]. Salah satu algoritma yang efektif untuk tugas ini adalah metode *Support Vector Machine* (SVM), yang terbukti memiliki akurasi tinggi dalam klasifikasi data teks. Meski demikian, studi yang secara khusus menganalisis sentimen masyarakat Indonesia terkait gempa *megathrust* melalui media sosial masih terbatas.

Penelitian ini memiliki tujuan untuk mengkaji sentimen masyarakat terhadap potensi gempa *megathrust* melalui komentar *YouTube* , serta mengevaluasi kinerja algoritma SVM dalam klasifikasi sentimen. Diharapkan hasil penelitian dapat memberikan wawasan penting bagi pemerintah dan lembaga terkait dalam menyusun strategi komunikasi publik serta mendukung upaya mitigasi bencana yang lebih efektif.

II. KAJIAN TEORI

2.1 Analisis Sentimen

Analisis sentimen merupakan metode otomatis yang digunakan untuk mengekstraksi, mengelola, dan memahami data teks tidak terstruktur guna mengidentifikasi sentimen dalam sebuah kalimat, opini, atau pandangan. Teknik ini berfungsi untuk menilai sikap atau kecenderungan opini terhadap suatu topik, apakah bernuansa positif, negatif, atau netral [5].

2.2 *YouTube*

YouTube merupakan platform media sosial berbasis web yang menyajikan informasi dalam bentuk video, di mana pengguna dapat menonton, mengunggah, serta mengekspresikan pandangan, ide, dan kreativitas mereka kapan saja tanpa batasan waktu [6].

2.3 Gempa *Megathrust*

Gempa megathrust merupakan gempa bumi bermagnitudo besar yang terjadi di zona subduksi, yaitu area di mana dua lempeng tektonik bertemu dan salah satunya bergerak ke bawah lempeng lainnya. Fenomena ini umum dijumpai di kawasan Cincin Api Pasifik, yang menyebabkan Indonesia menjadi wilayah dengan tingkat kerentanan tinggi

terhadap jenis gempa tersebut. Gempa *megathrust* dapat memicu tsunami yang besar serta membawa dampak sosial dan ekonomi yang luas [7].

2.4 TF-IDF

Term Frequency – Inverse Document Frequency (TF-IDF) merupakan salah satu teknik pembobotan yang banyak digunakan dalam klasifikasi teks untuk mengevaluasi relevansi sebuah kata terhadap dokumen tertentu dalam sebuah kumpulan data [8].

2.5 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) adalah metode oversampling yang dikembangkan untuk mengatasi ketidakseimbangan jumlah data antar kelas dalam proses klasifikasi. Ketidakseimbangan ini dapat menyebabkan algoritma machine learning memberikan hasil yang bias, terutama ketika kelas minoritas memiliki jumlah data yang jauh lebih sedikit dibandingkan kelas mayoritas. SMOTE bekerja dengan membuat sampel sintetis pada kelas minoritas agar distribusi data menjadi lebih seimbang [9].

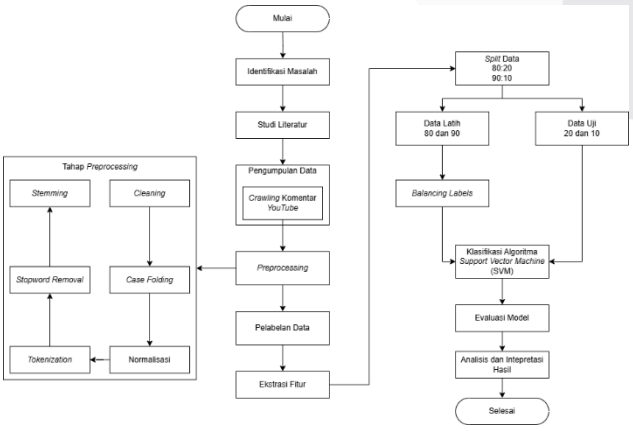
2.6 TOMEKS Link

Tomeks Links merupakan metode *undersampling* yang digunakan untuk mengatasi ketidakseimbangan data dengan cara menghapus data dari kelas mayoritas yang memiliki kesamaan karakteristik dengan kelas minoritas sehingga data menjadi seimbang. Penghapusan ini bertujuan untuk mengurangi ambiguitas antar kelas dan meningkatkan kejelasan batas pemisah antar kelas [10].

2.7 Support Vector Machine

Algoritma *Support Vector Machine (SVM)* merupakan salah satu metode *supervised learning* dalam bidang *machine learning* yang banyak digunakan untuk tugas klasifikasi maupun regresi. SVM bekerja dengan menentukan *hyperplane* optimal yang mampu memisahkan data dari kelas yang berbeda dengan jarak maksimum. *Hyperplane* ini berperan sebagai batas atau pemisah antara satu kelas dengan kelas lainnya. Dengan memanfaatkan berbagai jenis fungsi *kernel* seperti linear, *polynomial*, *Radial Basis Function* (RBF), dan *sigmoid*, algoritma SVM mampu menangani data dengan pola yang tidak linear secara efektif [11].

III. METODE



GAMBAR 1
(DIAGRAM ALIR PENELITIAN)

3.1 Identifikasi masalah

Penelitian dimulai dengan mengidentifikasi isu gempa *megathrust* yang memicu kekhawatiran publik. Tahap awal meliputi penyusunan latar belakang, rumusan masalah, dan tujuan penelitian.

3.2 Studi literatur

Studi Literatur dilakukan dengan menelusuri referensi terkait analisis sentimen dan algoritma *Support Vector Machine* (SVM), dibatasi pada publikasi lima tahun terakhir dari Google Scholar.

3.3 Pengumpulan data

Pengumpulan Data dilakukan melalui *YouTube* Data API v3 dengan mengambil komentar dari dua video populer terkait gempa *megathrust*. Teknik yang digunakan adalah observasi tidak langsung dan studi dokumentasi untuk memperoleh komentar yang autentik dan representatif. Data dikumpulkan dalam rentang Agustus 2024 hingga April 2025.

3.4 *Preprocessing*

Preprocessing mencakup tahapan *cleaning*, *case folding*, normalisasi, tokenisasi, *stopword removal*, dan *stemming* untuk mempersiapkan data teks.

3.5 Pelabelan data

Pelabelan Data dilakukan dengan pendekatan *lexicon-based* menggunakan SenticNet. Komentar diklasifikasikan menjadi sentimen positif dan negatif berdasarkan skor polaritas.

3.6 Ekstraksi fitur

Ekstraksi Fitur dilakukan menggunakan metode TF-IDF untuk mengubah teks menjadi representasi numerik yang siap digunakan dalam proses klasifikasi.

Persamaan TF dapat dilihat pada persamaan (1).

$$TF = \frac{(t)}{d} \quad (1)$$

Di mana :

t : jumlah kemunculan kata tertentu dalam dokumen d.

d : total keseluruhan kata pada dokumen.

Persamaan idf dapat dilihat pada persamaan (2).

$$IDF = log\frac{N}{df(t)} \quad (2)$$

Di mana :

N : total dokumen yang ada.

df(t) : jumlah dokumen yang memiliki kata t.

Persamaan TD-IDF dapat dilihat pada persamaan (3).

$$TFIDF = TF.IDF \quad (3)$$

3.7 *Split* data

Split Data menggunakan dua skema proporsi: 80:20 dan 90:10 antara data latih dan data uji, guna mengevaluasi performa model dalam dua skenario pelatihan

3.8 *Balancing* labels

Balancing Labels menggunakan teknik SMOTETomek untuk mengatasi ketidakseimbangan kelas, dengan

menggabungkan oversampling (SMOTE) dan pembersihan data ambigu (Tomek Links).

3.9 SVM

Klasifikasi dilakukan menggunakan algoritma SVM dengan empat jenis kernel: linear, *polynomial*, RBF, dan *sigmoid*. Berikut adalah persamaan setiap kernel SVM

1. Linear

$$\boldsymbol{K(x,y) = (x.y)} \quad (4)$$

2. RBF

$$\boldsymbol{K(x,y) = \exp\left(-\gamma||x-y||\right)^2)} \quad (5)$$

3. *Polynomial*

$$\boldsymbol{K(x,y) = (x.y\ + c\ )^d} \quad (6)$$

4. *Sigmoid*

$$\boldsymbol{K(x_i,x) = tanh\left(\gamma(x_i^T x) + r\right)} \quad (7)$$

3.10Evaluasi model

Evaluasi Model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai performa masing-masing kernel.

3.11Analisis dan intepretasi hasil

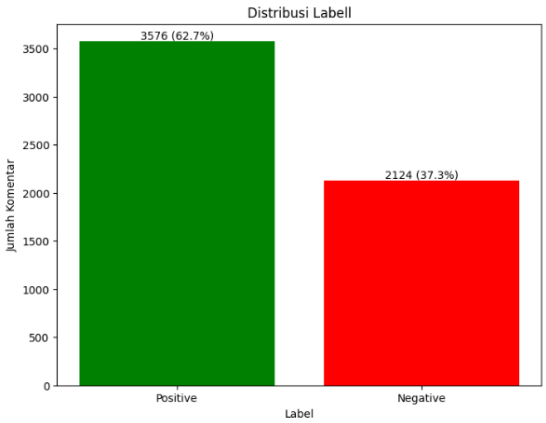
Analisis Hasil dilakukan untuk menafsirkan akurasi terbaik serta memberikan rekomendasi strategi komunikasi publik terkait mitigasi bencana.

IV. HASIL DAN PEMBAHASAN

Data yang digunakan penelitian ini diperoleh dari dua video *YouTube* yang membahas potensi gempa *megathrust* di Indonesia. Video pertama berasal dari kanal CNN Indonesia dengan judul *"Gempa Besar Megathrust di Indonesia Tinggal Tunggu Waktu"*, sedangkan video kedua dari kanal BBC News Indonesia berjudul *"Gempa Megathrust Ancam Indonesia, Apa Yang Harus Kita Lakukan?"*. Komentar dari kedua video dikumpulkan menggunakan *YouTube* Data API v3 melalui Google Colab dan disimpan dalam format CSV. Hasil crawling menghasilkan 5.135 komentar dari CNN Indonesia dan 793 komentar dari BBC News Indonesia, sehingga total yang dianalisis berjumlah 5.928 komentar.

Tahap selanjutnya adalah preprocessing data komentar, yang meliputi proses cleaning (menghapus tanda baca, URL, dan karakter khusus), *case folding*, normalisasi kata, tokenisasi, penghapusan *stopword*, dan *stemming*. Proses ini bertujuan agar data teks menjadi lebih bersih dan terstruktur untuk digunakan dalam analisis selanjutnya.

Pelabelan dilakukan dengan pendekatan lexicon-based menggunakan kamus SenticNet yang telah diadaptasi ke dalam Bahasa Indonesia. Dari proses pelabelan ini, diperoleh 3.576 data positif dan 2.124 data negatif. Meskipun distribusi tidak sepenuhnya timpang, ketidakseimbangan kelas tetap berisiko menimbulkan bias pada model. Oleh karena itu, diterapkan metode *balancing* menggunakan SMOTETomek untuk menyeimbangkan jumlah data antar kelas dan meningkatkan performa klasifikasi.

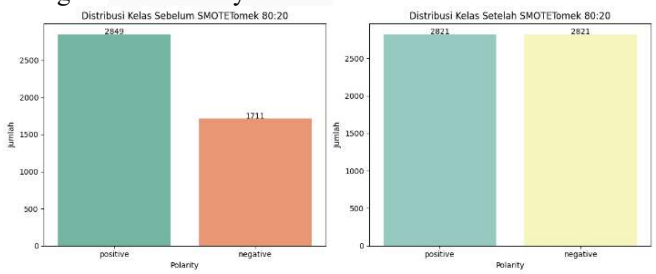


GAMBAR 2
(DISTRIBUSI HASIL LABEL)

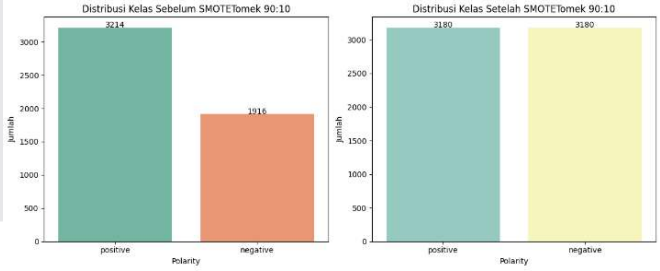
Selanjutnya, ekstraksi fitur dilakukan menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency), yang mengubah data teks menjadi representasi numerik. Hasilnya, diperoleh 6.992 fitur yang digunakan sebagai input pada proses pelatihan model klasifikasi.

Untuk pengujian model, data dibagi ke dalam dua skenario: 80% data latih dan 20% data uji, serta 90% data latih dan 10% data uji. Tujuannya adalah untuk membandingkan performa model berdasarkan jumlah data latih yang digunakan.

Balancing label menggunakan teknik SMOTETomek dengan menggabungkan dua teknik yaitu SMOTE dan Tomek Links. SMOTE digunakan untuk menambah data sintetis, dan Tomek Links digunakan untuk manghapus data ambigu dari kelas mayoritas.



GAMBAR 3
HASIL SMOTETOMEK 80:20



GAMBAR 4
HASIL SMOTETOMEK 90:10

Hasil *balancing* label menggunakan SMOTETomek pada scenario 80:20 menjadi seimbang (2821) dan pada scenario 90:10 menjadi 3180 data. Hasil *balancing* digunakan untuk melatih model. Model dikembangkan menggunakan algoritma *Support Vector Machine* (SVM) dengan empat jenis kernel: linear, RBF, *polynomial*, dan *sigmoid*.

Tabel 1 Hasil Pengujian Model 80:20

Kernel	Accuracy	Precision		Recall		F1-score	
		Positif	Negatif	Positif	Negatif	Positif	Negatif
Linear	88%	91%	83%	90%	85%	**91%**	**84%**
RBF	**89%**	90%	**87%**	**93%**	81%	**91%**	**84%**
Poly	87%	91%	81%	89%	84%	90%	83%
Sigmoid	88%	**92%**	81%	88%	**86%**	90%	83%

Pada skenario pembagian data 80:20 yang ditampilkan pada Tabel 1, kernel RBF mencatat akurasi tertinggi sebesar 89%, dengan performa precision dan recall yang cukup baik terutama pada kelas positif, yaitu precision 90% dan recall 93%. Namun, nilai recall pada kelas negatif menurun hingga 81%, yang menunjukkan adanya kecenderungan model lebih optimal dalam mengenali sentimen positif dibanding negatif. Kernel Linear menunjukkan performa yang paling seimbang, dengan akurasi 88%, precision 91% (positif) dan 83% (negatif), serta F1-score yang cukup tinggi dan merata pada kedua kelas, yaitu 85% dan 84%. Kernel *Polynomial* dan *Sigmoid* masing-masing mencatat akurasi 87% dan 88%, dengan performa yang cukup stabil namun sedikit di bawah kernel Linear, terutama pada recall dan F1-score kelas negatif. Kernel *Sigmoid* unggul dalam precision kelas positif (92%) namun lemah pada recall negatif (86%), sedangkan *Polynomial* menunjukkan hasil yang konsisten namun tidak menonjol di semua metrik.

Tabel 2 Hasil Pengujian Model 90:10

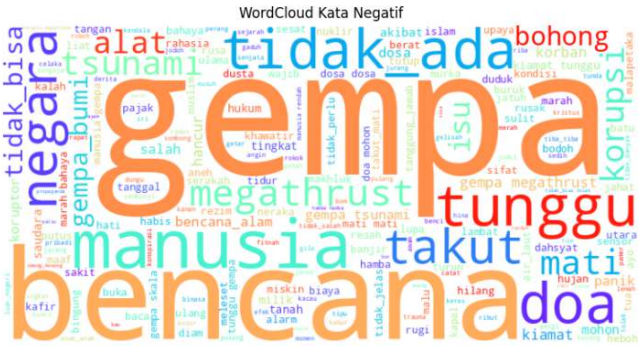
Kernel	Accuracy	Precision		Recall		F1-score	
		Positif	Negatif	Positif	Negatif	Positif	Negatif
Linear	87%	**91%**	81%	89%	**85%**	90%	**83%**
RBF	**88%**	89%	86%	**93%**	80%	**91%**	**83%**
Poly	**88%**	**91%**	**93%**	90%	84%	90%	**83%**
Sigmoid	86%	90%	80%	88%	83%	89%	81%

Pada skenario pembagian data 90:10 yang ditampilkan pada Tabel 2, pola performa model relatif konsisten. Kernel Linear tetap menunjukkan performa yang stabil dan seimbang dengan akurasi 87%, serta nilai precision dan recall yang seimbang pada kedua kelas. Kernel RBF kembali meraih akurasi tertinggi 88%, namun dengan kelemahan serupa seperti pada skenario sebelumnya, yaitu recall kelas negatif yang lebih rendah dibanding positif. Kernel *Polynomial* mengalami peningkatan signifikan pada precision kelas negatif (93%) dan mempertahankan akurasi di angka 88%, menjadikannya lebih kompetitif pada skenario ini. Sementara itu, kernel *Sigmoid* mencatat akurasi terendah (86%) dan performanya cenderung lemah, terutama dalam recall kelas negatif (81%) dan F1-score yang lebih rendah dibanding kernel lainnya.

Berdasarkan kedua skenario, kernel Linear dipilih sebagai kernel terbaik secara keseluruhan, karena menunjukkan performa yang stabil, seimbang, dan tidak bias terhadap salah satu kelas. Meskipun kernel RBF unggul pada akurasi, ketidakseimbangannya terhadap kelas negatif menjadikannya kurang ideal dalam konteks klasifikasi sentimen yang memerlukan sensitivitas yang sama pada kedua kelas.

Gambar 5 Wordcloud Positif

Visualisasi dalam bentuk *word cloud* label positif menunjukkan bahwa kata-kata seperti *"tuhan"*, *"Indonesia"*, *"moga"*, *"BMKG"* dan *"waspada"* banyak muncul. Hal ini menunjukkan bahwa respons masyarakat terhadap isu gempa *megathrust* sangat dipengaruhi oleh faktor keagamaan dan harapan masyarakat akan keselamatan.



Gambar 6 Wordcloud Negatif

Sementara itu, *word cloud* untuk label negatif pada Gambar 4.52 memperlihatkan kata-kata *"takut"*, *"bohong"*, *"korupsi"*, dan *"tidak ada"* yang menunjukkan adanya ketidakpercayaan sebagian masyarakat terhadap otoritas dan informasi resmi yang disampaikan.

Word cloud ini dapat menjadi dasar dalam menyusun strategi komunikasi kebencanaan, seperti menekankan pesan religius dan meningkatkan kepercayaan publik terhadap informasi resmi.

V. KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma *Support Vector Machine* (SVM) dalam analisis sentimen komentar *YouTube* terkait opini masyarakat terhadap potensi gempa *megathrust* di Indonesia. Hasil klasifikasi menunjukkan bahwa mayoritas komentar bersentimen positif sebesar 62,7%, sedangkan komentar negatif berjumlah 37,3%, mencerminkan beragam respons emosional publik terhadap isu bencana.

Evaluasi model menunjukkan bahwa algoritma SVM memberikan performa yang baik, dengan kernel RBF mencatat akurasi tertinggi sebesar 89% pada skenario pembagian data 80:20, dan 88% pada skenario 90:10. Meskipun kernel RBF unggul dalam mendeteksi sentimen positif, performanya terhadap kelas negatif kurang optimal. Sebaliknya, kernel Linear dan *Sigmoid* menunjukkan

performa yang lebih seimbang antar kelas. Secara keseluruhan, SVM terbukti efektif dalam klasifikasi sentimen karena mampu menghasilkan akurasi tinggi dan distribusi metrik evaluasi yang stabil.

REFERENSI

[1] M. Ridwan *et al.*, "Pengembangan Peta Bahaya Gempabumi di Batuan Dasar untuk Daerah Cilacap dan Sekitarnya Development of Seismic Hazard Map on Bedrock in Cilacap Area and its Vicinity," *Jurnal Geologi dan Sumberdaya Mineral*, vol. 24, pp. 31–38, 2023.

[2] T. Adventari *et al.*, "Penjalaran Tsunami Menuju Ke Outlet Airlindo Berdasarkan Skenario Gempa *Megathrust* Selatan Jawa," 2021.

[3] N. Amaly, A. Fakultas Dakwah dan Ilmu Komunikasi, and U. Antasari Banjarmasin, "Peran Kompetensi Literasi Digital Terhadap Konten Hoaks dalam Media Sosial," 2021.

[4] V. Puspaning Ramadhan and G. Mareskoti Namung, "Klasterisasi Komentar Cyberbullying Masyarakat di Instagram berdasarkan K-Means Clustering," *Sistem Informasi, Jl. Terusan Dieng*, vol. 65146, pp. 57–59, 2022.

[5] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma *Support Vector Machine* (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, pp. 147–156, 2021.

[6] E. Lutfiyatun, "Optimasi Keterampilan Menyimak Bahasa Arab Dengan Media *YouTube* ," 2022.

[7] G. A. Rifa'i, T. Muliawati, and D. G. Harbowo, "Analisis Probabilitas Gempa Bumi di Pulau Jawa Menggunakan Model Markov Chain," *Seminar Nasional Sains Data*, pp. 602–614, 2024.

[8] F. Rozi, F. Sukmana, and M. N. Adani, "Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Inverse Document Frequency (TF-IDF)," *JIMP: Jurnal Informatika Merdeka Pasuruan*, vol. 6, pp. 1–5, 2021.

[9] A. Farhan and A. Y. Rahman, "Analisis Sentimen Ulasan Aplikasi Identitas Kependudukan Digital Di Google Play Store Dengan," 2025.

[10] H. Utami, "Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network," *Indonesian Journal of Applied Statistics*, vol. 5, no. 1, p. 31, May 2022.

[11] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 17–23, Feb. 2021.