Analisis Sentimen Komentar YouTube Tentang Bencana Banjir Di Sumatera Menggunakan *Support Vector Machine*

Dimas Aditya Putranto[1,*], Sufajar Butsianto[2], Andini Putri Riandani[3]

Fakultas Teknik, Teknik Informatika, Universitas Pelita Bangsa, Kab.Bekasi, Indonesia
Email: [1,*] dimas.312210489@mhs.pelitabangsa.ac.id, [2]sufajar@pelitabangsa.ac.id,
[3]andiniriandani@pelitabangsa.ac.id
[*)] *Email* Penulis Utama

Abstrak− Bencana banjir di Sumatera tahun 2025 memicu berbagai reaksi masyarakat di *YouTube* sebagai salah satu platform media sosial utama di Indonesia. Penelitian ini bertujuan menganalisis sentimen publik dengan mengklasifikasikan komentar menjadi positif, negatif, dan netral menggunakan algoritma *Support Vector Machine* (SVM). Metode yang digunakan mengacu pada kerangka *Knowledge Discovery in Databases* (KDD), dimulai dari pengumpulan 22.776 komentar melalui *YouTube* Data API v3, yang kemudian diproses melalui tahapan *preprocessing* seperti *cleansing*, *case folding*, normalisasi, *tokenizing*, dan *stopword removal* hingga diperoleh 11.021 data valid. Data tersebut dibagi menjadi data latih dan uji dengan rasio 80:20. Hasil evaluasi menggunakan confusion matrix menunjukkan akurasi model sebesar 87%. Distribusi sentimen didominasi oleh positif (46,47%), diikuti netral (28,95%), dan negatif (24,58%). Hasil ini menunjukkan bahwa SVM efektif dalam menganalisis data teks media sosial dan dapat menjadi referensi bagi pihak terkait dalam pengambilan keputusan penanganan bencana.

Kata Kunci: Analisis Sentimen, Banjir Sumatera, *YouTube*, *Support Vector Machine. Knowledge Discovery in Database*

Abstract−*The 2025 floods in Sumatra triggered widespread public reactions on YouTube as one of the main social media platforms in Indonesia. This study aims to analyze public sentiment by classifying comments into positive, negative, and neutral categories using the Support Vector Machine (SVM) algorithm. The methodology follows the Knowledge Discovery in Databases (KDD) framework, starting with the collection of 22,776 comments via the YouTube Data API v3. The data then underwent preprocessing stages, including cleansing, case folding, normalization, tokenizing, and stopword removal, resulting in 11,021 valid data points. The dataset was divided into training and testing sets with an 80:20 ratio. Evaluation using a confusion matrix showed that the model achieved an accuracy of 87%. The sentiment distribution was dominated by positive sentiment (46.47%), followed by neutral (28.95%) and negative (24.58%). These results indicate that SVM is effective in analyzing complex social media text data and can serve as a reference for stakeholders in making decisions related to disaster management.*

Keywords: *Sentiment Analysis, Sumatra Flood, YouTube, Support Vector Machine. Knowledge Discovery in Database*

1. PENDAHULUAN

Bencana banjir yang terjadi di wilayah Sumatera pada tahun 2025 menimbulkan berbagai reaksi masyarakat yang disampaikan melalui media sosial, khususnya YouTube. Kolom komentar pada video terkait banjir menjadi ruang bagi masyarakat untuk menyampaikan dukungan, kritik, harapan, maupun opini terhadap peristiwa tersebut. Besarnya jumlah komentar yang dihasilkan menjadikan analisis sentimen diperlukan untuk memahami kecenderungan persepsi publik secara otomatis dan efisien.

Di tengah situasi krisis ini, *YouTube* mendominasi sebagai platform utama penyebaran informasi bencana di Sumatera, di mana jangkauan penggunanya telah mencakup 65% dari total populasi penduduk di seluruh provinsi Sumatera. Fenomena ini sejalan dengan data bahwa *YouTube* merupakan media sosial yang paling sering diakses di Indonesia, mencapai 63,02%, di mana kolom komentarnya menjadi wadah bagi masyarakat untuk merefleksikan pengalaman langsung dan opini mereka[1]. Video dari kanal berita lokal maupun liputan *drone* BNPB mengumpulkan jutaan *views* dan ribuan komentar yang mencerminkan sentimen public mulai dari apresiasi terhadap relawan PMI hingga kritik terhadap kecepatan bantuan pemerintah.

Analisis opini publik secara konvensional terhambat oleh hambatan geografis dan volume data komentar yang tumbuh secara eksponensial. Membaca ribuan komentar secara manual untuk memetakan respon masyarakat adalah hal yang tidak efisien[2]. Oleh karena itu, diperlukan otomatisasi melalui *Natural Language Processing* (NLP) untuk mengekstraksi emosi dan persepsi publik yang tersirat dalam data teks tersebut secara cepat dan akurat[3].

Dalam mengolah data komentar yang seringkali menggunakan bahasa daerah (Melayu Riau) atau slang Sumatera, tahap *preprocessing* dan penggunaan metode *Support Vector Machine* (SVM) menjadi solusi yang ideal. Penggunaan teknik *preprocessing* seperti *case folding*, *tokenization*, dan *stemming* sangat krusial untuk menangani ketidakteraturan bahasa Masyarakat[4]. Ditambah dengan ekstraksi fitur TF-IDF, algoritma SVM mampu menangani dataset yang tidak seimbang dengan tingkat akurasi tinggi, umumnya berkisar antara 85% hingga 95% [5][1].

Meskipun analisis sentimen pada media sosial telah banyak dilakukan, kajian literatur menunjukkan bahwa penelitian terdahulu masih berfokus pada konteks yang berbeda. Penelitian oleh Jamil et al. (2024) menganalisis

komentar YouTube terkait simulasi manajemen bencana menggunakan algoritma Naïve Bayes dan Support Vector Machine, namun tidak membahas peristiwa bencana aktual maupun kasus banjir di Sumatera[6]. Penelitian lain mengenai sentimen banjir lebih banyak dilakukan pada wilayah berbeda, seperti banjir Bekasi yang berfokus pada evaluasi pengelolaan banjir menggunakan SVM dan teknik seleksi fitur[7]. Selain itu, penelitian terkait banjir Sumatera 2025 yang ditemukan dalam penelusuran literatur umumnya menggunakan sumber data lain, seperti berita daring atau komentar Instagram, serta tidak secara khusus mengkaji komentar YouTube yang merefleksikan respons spontan masyarakat terhadap peristiwa tersebut[8]. Dengan demikian, masih terdapat celah penelitian berupa belum adanya kajian yang secara khusus menganalisis sentimen komentar YouTube terhadap bencana banjir Sumatera tahun 2025 menggunakan algoritma Support Vector Machine. Penelitian ini diharapkan dapat mengisi kesenjangan tersebut sekaligus memberikan gambaran mengenai persepsi publik terhadap bencana yang dipengaruhi oleh faktor antropogenik, seperti deforestasi lahan gambut dan sedimentasi akibat aktivitas pertambangan.

2. METODE PENELITIAN

2.1 Tinjauan Umum

Penelitian ini dirancang secara sistematis mengikuti kerangka *Knowledge Discovery in Databases* (KDD) untuk membedah sentimen publik pada kolom komentar *YouTube* mengenai bencana banjir[9]. Alurnya dimulai dengan mengumpulkan data melalui *YouTube* Data API, lalu mengolah teks tersebut menggunakan teknik *Natural Language Processing* (NLP)[10]. Tujuan utamanya adalah mengubah tumpukan komentar mentah menjadi informasi yang bermanfaat, mulai dari mengelompokkan opini ke dalam kategori positif, negatif, atau netral, hingga melatih algoritma *Support Vector Machine* (SVM) untuk mengenali pola respons masyarakat terhadap situasi banjir.

Mengingat gaya bahasa di media sosial sering kali berantakan, dilakukan pembersihan data secara intensif, seperti menghapus simbol yang tidak perlu, mengubah singkatan menjadi kata baku, hingga membuang kata sambung yang tidak bermakna. Alur penelitian dimulai dari seleksi data (22.776 komentar dari video Najwa Shihab 2025), Semua proses ini dikerjakan di *Google Colab* menggunakan pustaka populer seperti *pandas* dan *scikit-learn*[11]. Untuk menjamin hasil klasifikasinya akurat, model dievaluasi menggunakan *confusion matrix* guna melihat sejauh mana sistem mampu membedakan antara keluhan warga dan informasi bantuan dengan tepat[12].

Melalui pendekatan kuantitatif ini, penelitian berfokus pada keandalan model SVM dalam menangani bahasa "gaul" atau komentar yang tidak beraturan khas netizen Indonesia. Hasil akhirnya adalah sebuah pemetaan sentimen yang divisualisasikan melalui grafik dan *heatmap*, yang menunjukkan kecenderungan emosi publik apakah lebih didominasi oleh kekecewaan terhadap infrastruktur atau dukungan moral bagi para korban. Harapannya, temuan ini bisa memberikan gambaran nyata bagi pihak terkait mengenai dinamika opini masyarakat sehingga penanganan banjir di masa depan bisa lebih sesuai dengan kebutuhan publik[13].

2.2 Sumber Data dan Objek

Sumber data primer dalam penelitian ini berupa komentar pengguna pada video *YouTube* yang membahas dampak banjir, khususnya video liputan pasca-bencana, dokumentasi kerugian warga, dan analisis kegagalan infrastruktur pencegahan banjir. Data dikumpulkan dari platform *YouTube* karena memiliki karakteristik komentar yang kaya akan opini publik, ekspresi emosional warga yang terdampak, serta diskusi *real-time* terkait kritik sosial dan keluhan masyarakat, sehingga sangat cocok untuk digunakan dalam analisis sentimen mengenai dampak bencana tersebut.

2.2.1 Kriteria Pemilihan Video YouTube

a. Video dipilih berdasarkan keterkaitan isi dengan dampak banjir, meliputi laporan kerusakan infrastruktur, dokumentasi kerugian material warga, serta video analisis mengenai penyebab kegagalan sistem *drainase* atau penanganan bencana.
b. Waktu unggahan video dibatasi pada rentang waktu terjadinya bencana banjir dan masa pemulihannya, yakni disesuaikan dengan *timeline* musim penghujan tahun 2025 (misalnya: November - Desember 2025).
c. Menggunakan video dengan ID hiP76ZBZnw4 (sebagai contoh video dari Najwa Shihab yang memiliki tingkat keterlibatan pengguna/ *engagement* yang tinggi serta diskusi publik yang aktif).
d. Menargetkan total 14.000 komentar untuk memastikan ketersediaan ukuran dataset yang memadai dan representatif dalam proses pelatihan model SVM (*Support Vector Machine*).

2.2.2 Atribut Data Yang Dikumpulkan

Setiap komentar yang berhasil di-crawl menyimpan 4 atribut utama dalam format CSV (datadampakbanjir.csv) contoh pada Tabel 1 berikut.

Tabel 1 Atribut Data Komentar Youtube Dampak Banjir

Atribut	Deskripsi	Tipe Data	Contoh
author	Nama pengguna/pengunggah komentar	*String*	"@SuperGinga"
comment	Isi komentar lengkap (teks mentah untuk analisis sentimen)	*String*	"Sumpah air mataku menetes ketika melihat sang !"
likes	Jumlah *like* yang diterima komentar (indikator popularitas opini)	*Integer*	247
published_at	*Timestamp* publikasi komentar (untuk analisis temporal pasca-match)	*Datetime*	"2025-06-15T20:45:00Z"

2.2.3 Karakteristik *Dataset*

a. Dataset mentah: ~22.776 baris × 4 kolom
b. Nama *File*: datadampakbanjir.csv
c. Fokus analisis: Kolom comment sebagai input utama untuk ekstraksi sentimen
d. Data dikumpulkan pada periode Desember 2025 – Januari 2026, bertepatan dengan puncak banjir akibat intensitas hujan ekstrem dan luapan sungai besar (seperti Sungai Musi atau Batanghari).
e. Bahasa Dominan: Bahasa Indonesia dengan intensitas istilah lokal dan bahasa slang masyarakat Sumatera Contoh data mentah ditunjukkan pada tabel 2.

Tabel 2 Data Mentah

Index	*Author*	*Comment*	*Likes*	*Published_at*
0	@godingslow	semoga semua pemimpin punya jiwa kayak bapa 😇	0	2026-01-05T15:43:21Z
1	@srianawati-cf5fl	Semoga Aceh Sumatra cepat pulih kembali	0	2026-01-05T15:01:06Z
2	@BarjoJo-q1v	Jika bapaini main keusubang coba bisa hati sehat baik ijin Alloh benar ,amin	0	2026-01-05T14:34:51Z
3	@BarjoJo-q1v	Per siden nyah Deng KDM baik ,gamerusak orang gabenar ,sama baik bersama makmur ,amin	0	2026-01-05T14:32:10Z
4	@BarjoJo-q1v	Jika bapaini gapercaya coba main keusubang jadi bisa baik bersama baik ok mudah mudahan bapa ini kasih sehat ,amin	0	2026-01-05T14:30:16Z

2.2.4 Batasan Objek Penelitian

a. Hanya komentar dari satu video utama (hiP76ZBZnw4) untuk menjaga konsistensi topik.
b. Fokus pada komentar *top-level* (termasuk *reply* atau *nested comments*).
c. Bahasa utama Indonesia dengan slang suporter lokal (akan ditangani di *preprocessing*).

Data ini menjadi dasar utama untuk tahap *preprocessing* selanjutnya, di mana kolom *comment* akan diolah menjadi fitur numerik untuk klasifikasi SVM, sementara atribut lain (*likes, published_at*) dapat digunakan untuk analisis tambahan seperti korelasi popularitas komentar dengan sentimen.

2.3 Metode Pengumpulan Data

Pengumpulan data dilakukan melalui teknik *web crawling* menggunakan *YouTube* Data API v3 yang disediakan oleh *Google*, memastikan akses legal dan efisien terhadap komentar publik tanpa melanggar *Terms of Service YouTube.* Proses ini menghasilkan dataset mentah dalam format CSV (datadampakbanjir.csv) yang siap untuk tahap *preprocessing* selanjutnya.

2.3.1 Persiapan API

a. API *Key*: Diperoleh dari Google *Cloud Console* dengan mengaktifkan *YouTube* Data API v3.
b. *Library Python*: *googleapiclient.discovery* untuk membangun service client (*build*('*youtube*', 'v3', *developerKey=api_key*)).

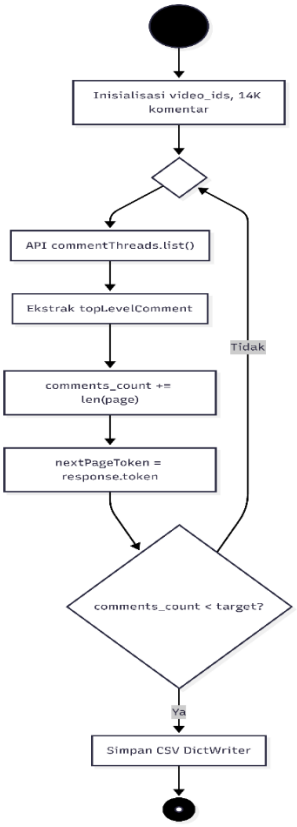
Contoh parameter utama API YouTube terdapat pada tabel 3.

Tabel 3 Parameter Utama API YouTube V3

Parameter	Nilai	Fungsi
part	'snippet'	Mengambil metadata komentar
videoId	' hiP76ZBZnw4 '	ID video target
maxResults	23000	Maksimal komentar per request
textFormat	'plainText'	Format teks bersih tanpa HTML

2.3.2 Algoritma Pengumpulan Komentar

Proses *crawling* diimplementasikan melalui fungsi *get_video_comments*() dan *scrape_comments_from_videos*() dengan logika ditunjukan pada gambar 1.



Gambar 1 *Flowchart* Pengumpulan Komentar YouTube

2.4 Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahap krusial untuk mengubah teks komentar *YouTube* mentah menjadi format yang siap untuk ekstraksi fitur dan klasifikasi SVM. Tahap ini mengikuti *pipeline* NLP standar yang diterapkan secara berurutan pada kolom *comment*, menghasilkan kolom baru bertahap *cleaning*, *case folding*, *normalisasi*, t*okenize, stopword removal*.

2.4.1 Penghapusan Duplikasi Data

```
df.drop_duplicates(subset ="comment", keep = 'first', inplace = True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 7542 entries, 0 to 7603
Data columns (total 1 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   comment  7542 non-null   object
dtypes: object(1)
memory usage: 117.8+ KB
```

Gambar 2 Hasil Penghapusan Duplikasi Data

a. Tujuan: Menghilangkan komentar identik untuk mencegah bias *overfitting* pada model SVM.
b. Hasil: Mengurangi ukuran dataset dari ~11.021 menjadi jumlah unik (dicek via df.info() sebelum dan sesudah).

2.4.2 Text Cleaning

Tahap *Cleaning* pada penelitian ini merupakan tahap pembersihan data komentar *YouTube* yang bertujuan untuk menghilangkan berbagai karakter dan elemen teks yang tidak relevan dengan analisis sentimen terhadap fenomena bencana banjir di Sumatera. Tahap ini dilakukan setelah proses data *crawling* dan sebelum *case folding* maupun tahapan *preprocessing* lainnya, sehingga teks yang masuk ke proses berikutnya sudah dalam kondisi lebih terstruktur dan bebas *noise*, contohnya terdapat pada tabel 4 berikut.

Tabel 4 Contoh *Text Cleaning*

Fungsi	*Regex/Pattern*	Contoh *Input* → *Output*
remove_URL()	`r'https?://\S+	www.\S+'`
remove_usernames()	r'@\w+'	"@Mursid bilang..." → "bilang..."
remove_html()	r'<.*?> '	" Banjir bandang " → "Banjir bandang"
remove_emoji()	Unicode emoji ranges	"Pemerintah 💔😠" → "Pemerintah"
remove_symbols()	r'[^a-zA-Z0-9\s]'	"Pemerintah!!!???" → "Pemerintah"
remove_numbers()	r'\d+'	"Sumbangan 10 Miliar" → "Sumbangan Miliar"

2.4.3 Case Folding

Case Folding dalam penelitian ini merupakan tahap standarisasi penulisan teks dengan mengubah seluruh huruf pada komentar *YouTube* menjadi huruf kecil (*lowercase*). Tahap ini penting agar kata yang sama tetapi berbeda penulisan huruf kapital misalnya "Banjir", "BANJIR", dan "banjir" diperlakukan sebagai satu token yang sama oleh sistem ketika masuk ke proses ekstraksi fitur dan klasifikasi SVM, berikut merupakan contoh *case folding* dari dataset. Contoh case folding ditunjukkan pada tabel 5.

Tabel 5 Contoh *Case Folding*

Index	*Comment (Raw)*	*Cleaning*	*Case Folding*
0	Negara tentu bantu.....ke Aceh...😊	Negara tentu bantuke Aceh	negara tentu bantuke aceh
1	Tulalit	Tulalit	tulalit
2	Ya Allah...😢😢😢 Kami sabar kami kuat...makasih ya Allah kami telah mempunyai sosok yg begitu peduli terhadap rakyat nya...salam dari Takengon bapak Mualim....🙏🙏🙏 Kami dari Takengon bangga mempunyai gubernur seperti babak	Ya Allah Kami sabar kami kuatmakasih ya Allah kami telah mempunyai sosok yg begitu peduli terhadap rakyat nyasalam dari Takengon bapak Mualim Kami dari Takengon bangga mempunyai gubernur seperti babak	ya allah kami sabar kami kuatmakasih ya allah kami telah mempunyai sosok yg begitu peduli terhadap rakyat nyasalam dari takengon bapak mualim kami dari takengon bangga mempunyai gubernur seperti babak

Index	*Comment (Raw)*	*Cleaning*	*Case Folding*
3	Pemerintah Bareng Warga Aceh Sumatra, kompak banget lawan banjir	Pemerintah Bareng Warga Aceh Sumatra kompak banget lawan banjir	pemerintah bareng warga aceh sumatra kompak banget lawan banjir
4	ALHAMDULILLAH.semoga mualem jadi pemimpin yg sangat amanah.	ALHAMDULILLAHsemoga mualem jadi pemimpin yg sangat amanah	alhamdulillahsemoga mualem jadi pemimpin yg sangat amanah
5	Hadiroh	Hadiroh	hadiroh

2.4.4 Stopword Removal

Stopword Removal pada penelitian ini adalah proses menghapus kata-kata umum yang sering muncul tetapi tidak memberikan makna penting terhadap sentimen, seperti kata sambung, kata depan, kata ganti, dan kata penunjuk. Contohnya kata "yang", "dan", "di", "ke", "ini", "itu", "saja" banyak muncul di komentar *YouTube* tentang bencana banjir di Sumatera, tetapi tidak membedakan apakah komentar tersebut bernada dukungan atau kekecewaan, sehingga kata-kata ini dihilangkan dari daftar token. Langkah ini krusial agar model SVM dapat berfokus pada kata-kata yang memiliki muatan emosi atau informasi penting seperti "banjir", "bantu", "sabar", atau "kecewa". Contohnya di tunjukkan pada tabel 6.

Tabel 6 Contoh *Stopword Removal*

Index	**Normalisasi**	***Tokenizing***	***Stopword Removal***
0	negara tentu bantuke aceh	['negara', 'tentu', 'bantuke', 'aceh']	negara bantuke aceh
1	tulalit	['tulalit']	tulalit
2	ya allah kami sabar kami kuatmakasih ya allah kami telah mempunyai sosok yang begitu peduli terhadap rakyat nyasalam dari takengon bapak mualim kami dari takengon bangga mempunyai gubernur seperti babak	['ya', 'allah', 'kami', 'sabar', 'kami', 'kuatmakasih', 'ya', 'allah', 'kami', 'telah', 'mempunyai', 'sosok', 'yang', 'begitu', 'peduli', 'terhadap', 'rakyat', 'nyasalam', 'dari', 'takengon', 'bapak', 'mualim', 'kami', 'dari', 'takengon', 'bangga', 'mempunyai', 'gubernur', 'seperti', 'babak']	ya allah sabar kuatmakasih ya allah sosok peduli rakyat nyasalam takengon mualim takengon bangga gubernur babak
3	pemerintah bareng warga aceh sumatra kompak banget lawan banjir	['pemerintah', 'bareng', 'warga', 'aceh', 'sumatra', 'kompak', 'banget', 'lawan', 'banjir']	pemerintah bareng warga aceh sumatra kompak banget lawan banjir
4	alhamdulillahsemoga mualem jadi pemimpin yang sangat amanah	['alhamdulillahsemoga', 'mualem', 'jadi', 'pemimpin', 'yang', 'sangat', 'amanah']	alhamdulillahsemoga mualem pemimpin amanah

2.4.5 Frekuensi Kata

Frekuensi kata adalah jumlah kemunculan setiap kata (token) dalam korpus komentar setelah melalui tahapan pembersihan dasar (*cleaning*, *case folding*, normalisasi, tokenisasi, dan biasanya *stopword removal*). Frekuensi ini digunakan untuk melihat kata mana yang paling sering muncul sehingga dianggap mewakili topik atau sentimen dominan dalam data.

Tabel 7 Contoh Frekuensi kata

No	Kata	Frekuensi
1	aceh	1978
2	allah	1594
3	ya	1441
4	pemimpin	1159
5	semoga	1101
6	mualem	953
7	rakyat	869
8	gubernur	733
9	sehat	682
10	beliau	655

Berdasarkan tabel 7, kata "aceh" merupakan kata yang paling sering muncul, menunjukkan bahwa pembahasan berfokus pada Aceh. Tingginya frekuensi kata "allah", "ya", dan "semoga" mencerminkan banyaknya ungkapan doa dan harapan. Sementara itu, kemunculan kata "pemimpin", "mualem", "rakyat", dan "gubernur" menunjukkan adanya pembahasan terkait kepemimpinan dan masyarakat. Secara keseluruhan, teks didominasi oleh topik Aceh, kepemimpinan, serta dukungan dan harapan dari masyarakat.

2.5 Pelabelan Data Sentimen

Pelabelan merupakan tahap krusial untuk mengubah teks komentar *YouTube* yang telah dipra-proses menjadi dataset *supervised learning* dengan tiga kelas sentimen, yaitu Positif, Negatif, dan Netral. Pada penelitian ini, pelabelan tidak dilakukan secara manual oleh manusia, melainkan secara otomatis menggunakan Google Colab dengan bahasa pemrograman Python melalui pendekatan berbasis leksikon (*lexicon-based labeling*). Proses pelabelan memanfaatkan kamus sentimen bahasa Indonesia InSet (Indonesia Sentiment Lexicon) yang terdiri atas daftar kata positif dan kata negatif. Kamus tersebut diimpor dari repositori GitHub dalam bentuk file positive.tsv dan negative.tsv.

Penggunaan pendekatan leksikon dipilih karena mampu memberikan label secara konsisten pada data dalam jumlah besar, mengurangi subjektivitas pelabel manusia, serta banyak digunakan sebagai metode pelabelan awal pada penelitian analisis sentimen berbahasa Indonesia.Contoh pelabelan data sentimen pada tabel 8.

Tabel 8 Pelabelan Data Sentimen

Index	*Stopword Removal*	*Score*	*Sentiment*
0	negara bantuke aceh	0	Netral
1	tulalit	0	Netral
2	ya allah sabar kuatmakasih ya allah sosok peduli rakyat nyasalam takengon mualim takengon bangga gubernur babak	5	Positif
3	pemerintah bareng warga aceh sumatra kompak banget lawan banjir	1	Positif
4	alhamdulillahsemoga mualem pemimpin amanah	1	Positif

2.5.1 Kriteria Pelabelan Sentimen

Kriteria pelabelan adalah seperangkat aturan dan panduan tertulis yang digunakan untuk memutuskan sebuah teks (komentar) harus diberi label sentimen apa: Positif, Negatif, atau Netral, pada Tabel 9.

Tabel 9 Kriteria Pelabelan

Kelas Sentimen	Definisi	Kata Kunci Indikator	Contoh Komentar (*Stopword Removal*)
Positif	Ekspresi optimisme, dukungan, apresiasi, atau pujian terhadap upaya penanggulangan banjir, kinerja pemerintah, relawan, maupun kepedulian tokoh masyarakat.	bangga, dukung, salut, terima kasih, bersyukur, peduli, kuat, amanah, sehat	['alhamdulillahsemoga', 'mualem', 'pemimpin', 'amanah'] ; ['sosok', 'peduli', 'rakyat', 'bangga']
Negatif	Ekspresi kemarahan, kekecewaan, kritik keras, atau hujatan terkait dampak buruk banjir, lambatnya bantuan, maupun kegagalan infrastruktur.	gagal, memalukan, hancur, buruk, lambat, kecewa, rugi, lumpuh, sengsara	['pemerintah', 'lambat', 'bantu', 'masyarakat', 'rugi'] ; ['banjir', 'hancur', 'rumah', 'kecewa']
Netral	Fakta atau informasi deskriptif tanpa emosi kuat, laporan kondisi cuaca/ketinggian air, atau opini yang bersifat informatif tanpa polaritas jelas.	jadwal, lokasi, data, update, informasi, hujan, ketinggian, cuaca, fakta	['ketinggian', 'air', 'sungai', 'musis', 'naik'] ; ['info', 'lokasi', 'posko', 'banjir', 'aceh']

2.6 Pembagian *Dataset*

Pembagian dataset dilakukan menggunakan metode *train–test split* dengan rasio 80:20 untuk menjaga keseimbangan antara ukuran data latih yang cukup besar dan data uji yang representatif. Pendekatan ini memungkinkan model SVM mempelajari pola sentimen dari 80% komentar *YouTube* terkait bencana banjir di Sumatera, kemudian menguji kemampuan generalisasi pada 20% komentar yang tidak pernah dilihat model sebelumnya (*hold-out set*). Rasio ini umum digunakan pada penelitian analisis sentimen berbasis *machine learning* untuk memastikan model memiliki data yang cukup untuk proses pembelajaran (latih) sekaligus memiliki data pembanding yang valid untuk proses evaluasi (uji).

2.7 Metode Yang Digunakan

Penelitian ini menggunakan metode *Support Vector Machine* (SVM) dengan kernel linear sebagai algoritma klasifikasi untuk mengelompokkan sentimen komentar *YouTube* terkait bencana banjir di Sumatera ke dalam tiga kategori, yaitu Positif, Negatif, dan Netral. Pemilihan algoritma SVM didasarkan pada keunggulannya yang sangat cocok untuk mengolah data teks berdimensi tinggi serta kemampuannya dalam membentuk *maximum margin hyperplane*. Hal ini membuat model tetap *robust* terhadap *noise* yang sering ditemukan pada komentar media sosial, sehingga efektif untuk mengekstraksi emosi dan persepsi publik terkait situasi krisis banjir tersebut secara akurat[14].

Berdasarkan karakteristik tersebut, SVM dengan kernel linear diterapkan dalam penelitian ini untuk menganalisis respons masyarakat terhadap bencana banjir di Sumatera berdasarkan komentar YouTube. Pemilihan kernel linear didasarkan pada kemampuannya dalam menangani data teks berdimensi tinggi secara efisien serta membentuk hyperplane pemisah yang optimal. Dengan pendekatan tersebut, model diharapkan mampu memberikan hasil klasifikasi yang baik dalam memetakan sentimen positif, negatif, dan netral dari opini publik yang beragam di media sosial[15].

2.8 Teknik Evaluasi

Evaluasi performa model SVM dalam penelitian ini dilakukan menggunakan metrik klasifikasi multi-kelas standar, yaitu akurasi, *precision*, *recall*, dan *F1-score*, yang dihitung pada data uji independen berisi komentar *YouTube* yang tidak pernah dilihat model saat pelatihan. Selain itu, digunakan *confusion matrix* dan visualisasi heatmap untuk mengidentifikasi pola kesalahan prediksi antar kelas sentimen (Positif, Negatif, Netral), sebagaimana lazim diterapkan pada penelitian analisis sentimen berbasis SVM. Contoh rumus dan interpretasi Confusion Matrix terdapat pada tabel 10.

Tabel 10 Rumus & Interpretasi Confusion Matrix

Metrik	Rumus	Interpretasi
Accuracy	$\frac{TP + TN}{Total}$	Proporsi prediksi benar secara keseluruhan
Precision	$\frac{TP}{TP + FP}$	Dari prediksi kelas benar, berapa % benar-benar kelas itu
Recall	$\frac{TP}{TP + FN}$	Dari data sebenarnya kelas itu, berapa % terdeteksi
F1-Score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	*Harmonic mean, balance precision-recall*

2.8.1 Confusion Matrix

Confusion matrix adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan cara membandingkan label prediksi model dengan label aktual pada data uji. Tabel ini menampilkan jumlah data yang diklasifikasikan benar dan salah untuk setiap kelas, Contoh *Confusion Matrix* pada table 11.

Tabel 11 Contoh Confusion Matrix

Actual \ Prediksi	Negatif	Positif	Netral	Total	*Recall*
Negatif	455 (20.9%)	12	67	534	85.2%
Positif	12	930 (42.8%)	67	1009	92.2%
Netral	66	61	502 (23.1%)	629	79.8%
Total	533	1003	636	2172	-
Precision	85.4%	92.7%	78.9%	-	-

2.9 Instrumen Penelitian

Untuk melaksanakan prosedur penelitian yang telah direncanakan, diperlukan beberapa peralatan atau perlengkapan penting yang dapat memperlancar proses penelitian. Penelitian ini memerlukan peralatan untuk menunjang kemajuan penelitian. Alat yang digunakan dalam penelitian ini:

1. Kebutuhan Perangkat Lunak (*Sofware*)
   Berikut merupakan detail perangkat lunak yang digunakan dalam penelitian ini.
   a. Sistem Operasi: *Windows* 11
   b. Pengolah Kata: *Microsoft Word* 2021
   c. Pengolah Angka: *Microsoft Excel* 2021
   d. *Tools: Google Colab*
2. Kebutuhan Perangkat Keras (*Hardware*)
   Perangkat keras yang digunakan peneliti untuk melakukan implementasi data adalah laptop dengan spesifikasi berikut.
   a. *Processor*: AMD *Ryzen* 57430U *with Radeon Graphics* (2.30 GHz)
   b. *Memory* (RAM): 8 GB
   c. *Hardisk*: 258 GB

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Pengumpulan data dilakukan menggunakan *YouTube* Data API melalui proses *web crawling* pada *platform Google Colab*. Data yang dikumpulkan berupa komentar pengguna pada video *YouTube* yang membahas isu terkait Bencana Banjir yang Menimpa Sumatera. Contoh Crawling Data pada gambar 3.

```
# ID video YouTube yang ingin diambil
video_ids = ['hiP76ZBZnw4']
total_comments = 25000
output_filename = 'datadampakbanjir.csv'

# Ambil komentar dan simpan ke file CSV
comments = scrape_comments_from_videos(video_ids, total_comments, output_filename)

df = pd.DataFrame(comments)
df.info()
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22776 entries, 0 to 22775
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   author        22776 non-null  object
 1   comment       22776 non-null  object
 2   likes         22776 non-null  int64
 3   published_at  22776 non-null  object
dtypes: int64(1), object(3)
memory usage: 711.9+ KB
```

Gambar 3 Crawling Data

3.2 Pra-Pemrosesan Data

Proses pra-pemrosesan dilakukan untuk membersihkan dan menyiapkan data komentar *YouTube* tentang banjir Sumatera agar algoritma klasifikasi dapat bekerja dengan baik.

Tahapan pra-pemrosesan yang dilakukan meliputi:

a. *Case folding*, yaitu proses standarisasi dengan mengubah seluruh huruf dalam komentar menjadi huruf kecil (*lowercase*) agar konsisten.
b. *Cleaning*, yaitu tahap pembersihan untuk menghapus elemen teks yang tidak relevan seperti URL, *username* (@), simbol, angka, dan emoji.
c. *Tokenizing*, yaitu memecah kalimat komentar menjadi unit-unit kata (token) individual untuk pemrosesan komputasional.
d. Normalisasi kata, yaitu mengubah kata-kata tidak baku, singkatan (seperti "yg" menjadi "yang"), atau *slang* lokal menjadi bentuk kata baku bahasa Indonesia.
e. *Stopword removal*, yaitu tahap penghapusan kata-kata umum (seperti "dan", "di", "yang") yang tidak memberikan makna penting terhadap muatan sentimen.

Setelah seluruh tahapan dilakukan, jumlah data yang layak digunakan dalam proses klasifikasi adalah sebanyak 11.021 komentar. Contoh jumlah data setelah pra-pemrosesan pada tabel 12.

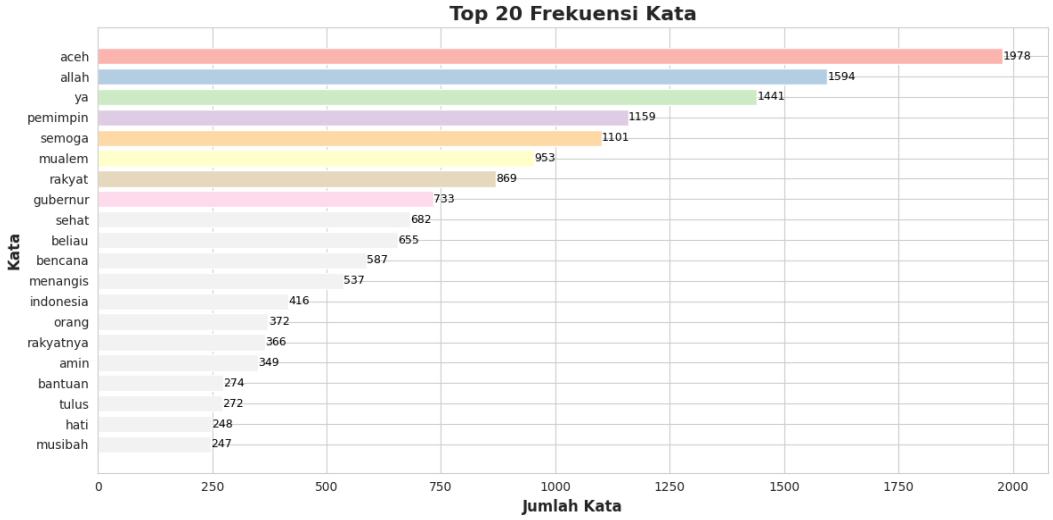
Tabel 12 Jumlah Data Setelah Pra-Pemrosesan

Keterangan	Jumlah Data
Data awal hasil *crawling*	22.776
Data akhir siap klasifikasi	11.021

Terjadi pengurangan jumlah data yang cukup signifikan dari data awal. Hal ini menunjukkan bahwa terdapat komentar yang tidak relevan, duplikat, atau tidak memenuhi kriteria analisis.

3.3 Hasil Frekuensi Kata

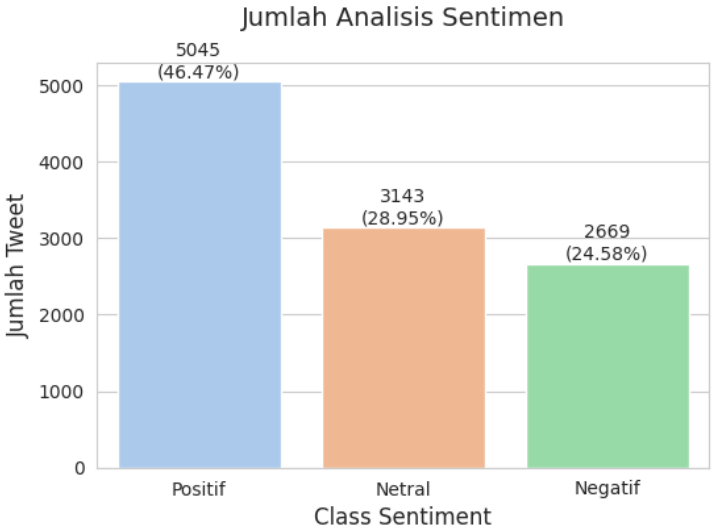
Pada tahap ini frekuensi kata yang sudah terkumpul akan membentuk sebuah grafik yang menunjukkan seberapa sering kata yang sudah digunakan. Contoh Frekuensi kata pada gambar 4.



Gambar 4 Contoh Top 20 Frekuensi Kata

3.4 Hasil Pelabelan Sentimen

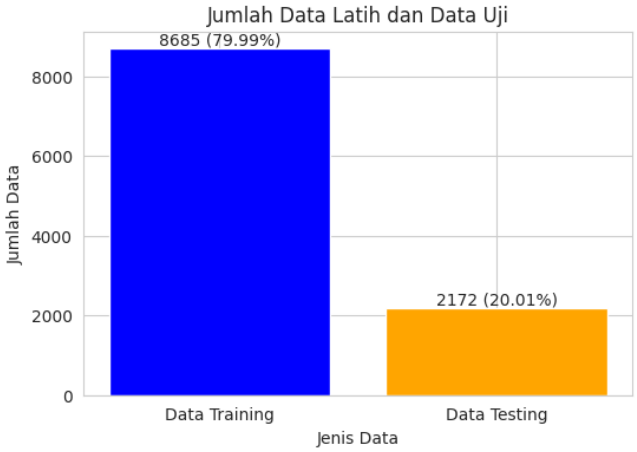
Setelah tahap pra-pemrosesan dan frekuensi kata dilakukan proses pelabelan sentimen dengan mengelompokkan komentar ke dalam tiga kategori, yaitu positif, netral, dan negatif. Contoh Distribusi sentiment pada gambar 5.



Gambar 5 Distribusi Sentimen

3.5 Pembagian *Data Training* dan *Testing*

Dataset kemudian dibagi menjadi dua bagian menggunakan metode *train-test split* dengan rasio 80:20. Contoh ada pada gambar 6.



Gambar 6 Pembagian *Data Training* dan *Testing*

3.6 Distribusi Sentimen Pada *Data Testing*

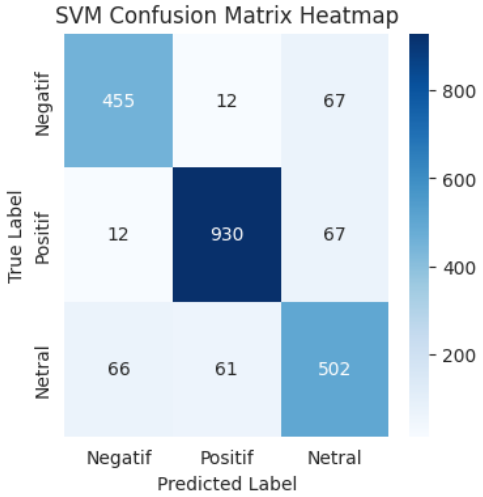
Distribusi sentimen pada data testing menunjukkan bahwa jumlah data tidak tersebar secara merata pada setiap kelas sentimen. Contoh distribusi sentimen data testing pada tabel 13.

Tabel 13 Distribusi Sentimen *Data Testing*

Kelas Sentimen	Jumlah Data
Positif	1009
Netral	629
Negatif	534
Total	2.172

3.7 Hasil Pengujian Model *Confusion Matrix*

Pada tahap ini *confusion matrix* merupakan salah satu metode evaluasi yang digunakan untuk mengetahui kinerja model klasifikasi dengan membandingkan antara label aktual dan label prediksi. Contoh heatmap confusion matrix ada pada gambar 7.



Gambar 7 *Heatmap Confusion Matrix Model* SVM

3.8 Hasil Analisis Klasifikasi SVM

Hasil Kalsifikasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

Tabel 14 Hasil klasifikasi SVM

Kelas Sentimen	*Precision*	*Recall*	*F1-Score*	*Support*
Negatif	0.86	0.85	0.85	534
Netral	0.78	0.80	0.79	629
Positif	0.93	0.92	0.92	1009
Macro Avg	0.86	0.86	0.86	2172
Weighted Avg	0.87	0.87	0.87	2172
Accuracy	-	-	87%	2172

Pada Tabel 14 Model memperoleh akurasi sebesar 87%, yang menunjukkan sebagian besar data uji berhasil diklasifikasikan dengan benar. Nilai *weighted average F1-score* sebesar 0,87 mencerminkan performa yang solid, terutama didorong oleh kelas positif (1009 data) yang memiliki performa tertinggi. Sementara itu, nilai *macro average* 0,86 sedikit di bawah nilai *weighted*, menandakan adanya variasi performa antar kelas. Hal ini terlihat pada sentimen netral dengan *F1-score* terendah (0,79), yang menunjukkan tantangan model dalam membedakan bahasa netral dibandingkan ekspresi positif atau negatif yang lebih kontras. Namun, selisih tipis antara keduanya membuktikan bahwa model tetap stabil meskipun terdapat ketimpangan jumlah sampel.

3.9 Analisis dan Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Support Vector Machine (SVM) mampu mengklasifikasikan sentimen komentar YouTube terkait bencana banjir di Sumatera dengan performa yang sangat baik, ditunjukkan oleh tingkat akurasi sebesar 87%. Nilai akurasi tersebut menunjukkan bahwa dari seluruh data pengujian, sebanyak 87% komentar berhasil diklasifikasikan ke dalam kelas sentimen yang sesuai. Tingginya akurasi yang diperoleh dipengaruhi oleh beberapa faktor. Pertama, proses *preprocessing* yang dilakukan secara bertahap, meliputi *cleansing*, *case folding*, *tokenization*, normalisasi kata, dan *stopword removal*, mampu mengurangi *noise* pada data sehingga fitur yang digunakan lebih representatif. Kedua, penggunaan pembobotan TF-IDF berhasil menonjolkan kata-kata yang memiliki kontribusi penting dalam membedakan setiap kelas sentimen. Ketiga, algoritma SVM dengan kernel linear memiliki kemampuan yang baik dalam menangani data teks yang berdimensi tinggi dan bersifat *sparse*, sehingga dapat membentuk batas pemisah (*hyperplane*) yang optimal antar kelas sentimen.

Proses klasifikasi dilakukan melalui beberapa tahapan sistematis, dimulai dari pengumpulan data komentar YouTube, dilanjutkan dengan *preprocessing* teks, pelabelan sentimen, hingga pelatihan model menggunakan algoritma SVM. Implementasi SVM memungkinkan sistem mengenali pola linguistik dan karakteristik teks spesifik pada setiap kategori sehingga opini publik dapat teridentifikasi secara otomatis.

Kinerja model dievaluasi menggunakan *confusion matrix* yang menghasilkan nilai *precision*, *recall*, dan *F1-score*. Berdasarkan hasil pengujian, model SVM terbukti sangat efektif dalam mengklasifikasikan data, terutama pada kelas sentimen positif yang memperoleh F1-score sebesar 0,92. Hal ini mengindikasikan bahwa komentar positif cenderung memiliki kata-kata yang lebih konsisten dan mudah dikenali oleh model. Sebaliknya, kelas sentimen netral memperoleh F1-score sebesar 0,79, yang menunjukkan masih adanya kesalahan klasifikasi. Kondisi tersebut kemungkinan disebabkan oleh penggunaan bahasa yang ambigu, komentar yang bersifat informatif tanpa ekspresi emosi yang jelas, serta adanya kemiripan kosakata antara sentimen netral dengan sentimen positif maupun negatif. Selain itu, ketidakseimbangan distribusi data antar kelas juga dapat memengaruhi kemampuan model dalam mengenali kelas tertentu.

Nilai *weighted average* sebesar 0,87 menunjukkan bahwa model memiliki keseimbangan yang baik antara *precision* dan *recall* pada seluruh kategori sentimen. Dengan demikian, akurasi sebesar 87% tidak hanya mencerminkan kemampuan SVM dalam melakukan klasifikasi secara tepat, tetapi juga menunjukkan bahwa kombinasi tahapan *preprocessing*, pembobotan TF-IDF, dan karakteristik SVM yang efektif untuk data teks berdimensi tinggi berkontribusi terhadap performa model yang diperoleh.

Secara keseluruhan, penelitian ini menunjukkan bahwa metode Support Vector Machine sangat efektif digunakan untuk analisis sentimen berbasis teks pada media sosial. Model ini mampu memberikan gambaran yang akurat mengenai kecenderungan opini publik terhadap bencana banjir di Sumatera yang tercermin melalui distribusi sentimen pada kolom komentar YouTube.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan mengenai analisis sentimen komentar *YouTube* terhadap bencana banjir di Sumatera menggunakan algoritma *Support Vector Machine* (SVM), maka dapat diambil kesimpulan sebagai berikut:Algoritma *Support Vector Machine* (SVM) terbukti sangat efektif dalam mengklasifikasikan volume data komentar yang besar dan tidak terstruktur dari platform *YouTube*. Hal ini ditunjukkan oleh kemampuan model dalam menangani data teks yang mengandung bahasa slang dan istilah lokal melalui serangkaian tahap *preprocessing* yang sistematis. Hasil pemrosesan data menunjukkan bahwa opini publik didominasi oleh Sentimen Positif sebesar 46,47%, yang mayoritas berisi doa, dukungan moral, dan apresiasi terhadap relawan. Sementara itu, Sentimen Netral sebesar 28,95% dan Sentimen Negatif sebesar 24,58% yang berisi keluhan terhadap infrastruktur serta penanganan bencana di lapangan.Model klasifikasi SVM yang dibangun memiliki tingkat performa yang solid dengan Akurasi sebesar 87%. Hasil evaluasi menunjukkan nilai *weighted average F1-score* sebesar 0,87, di mana model paling unggul dalam mengenali kelas positif (*F1-score* 0,92), namun masih memiliki tantangan pada kelas netral karena ambiguitas bahasa masyarakat.

UCAPAN TERIMAKASIH

Terima kasih disampaikan kepada semua pihak yang terlibat dalam penelitian ini, serta kepada program studi Teknik Informatika Universitas Pelita Bangsa yang telah memfasilitasi kelancaran penelitian ini.

DAFTAR PUSTAKA

[1] T. Muhayat, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Komentar Video Youtube Menggunakan Support Vector Machines," pp. 231–240, 2022, doi: 10.35889/progresif.v19i1.1060.

[2] G. W. M. H. Riwanto, P. Ardhiyansyah, R. A. Adiansyah, A. Alfiansyah, "ANALISIS SENTIMEN KOMENTAR YOUTUBE TERKAIT PENERAPAN MAKAN BERGIZI GRATIS MENGGUNAKAN MODEL ALGORITMA SVM," vol. 6, no. 12, pp. 1–14, 2025.

[3] F. Wijaya, R. A. Ramadhani, and A. B. Setiawan, "Analisis Sentimen Komentar Video Youtube Dengan Leksikon Textblob Menggunakan Algoritma Svm Berdasarkan Rasio Data Uji," vol. 9, pp. 1894–1902, 2025, doi: 10.29407/h1z62886.

[4] A. A. Syam, G. H. M, A. Salim, D. F. Surianto, and M. F. B, "Analisis teknik preprocessing pada sentimen masyarakat terkait konflik israel-palestina menggunakan support vector machine," vol. 9, no. 3, pp. 1464–1472, 2024, doi: 10.29100/jipi.v9i3.5527.

[5] S. R. Putri and L. Rosnita, "Sentiment Analysis of Youtube and Gotube Reviews on Google Play Using the Support Vector Machine (SVM) Method in Indonesia," vol. 9, no. 3, pp. 1025–1033, 2025, doi: 10.30871/jaic.v9i3.9461.

[6] M. Jamil, H. Hadiyanto, and R. Sanjaya, "Sentiment Analysis : Classifying Public Comments on YouTube in Disaster Management Simulation in Indonesia Using Naïve Bayes and Support Vector Machine," *Int. Inf. Eng. Technol. Assoc.*, vol. 29, no. 2, pp. 437–446, 2024, doi: 10.18280/isi.290205.

[7] D. Maulana, E. Widodo, A. Firmansyah, and M. Danny, "Optimizing Sentiment Analysis for Bekasi Flood Management Using SVM and Naive Bayes with Advanced Feature Selection," *Briliance*, vol. 4, no. 1, pp. 362–371, 2024, doi: 10.47709/brilliance.v4i1.4268.

[8] N. Alfarizi, P. Lydia, M. S. Novelan, A. Putra, and S. Sinurat, "Comparative Machine Learning Analysis for Sentiment Classification of Sumatra Disaster 2025," *J. Technol. Comput.*, vol. 3, no. 1, pp. 68–76, 2026.

[9] M. S. Alfandi and Z. Fatah, "PENERAPAN DATA MINING MENGGUNAKAN METODE K-MEANS CLUSTERING UNTUK ANALISA PENJUALAN TOKO UMAMA HIJAB KALIWATES JEMBER," *J. Ris. Sist. Inf.*, vol. 1, no. 4, pp. 94–102, 2024, doi: 10.69714/3ty90586.

[10] Nurfiyah N. Ramadhandi, "Penerapan Metode Natural Language Processing (NLP) Pada Question Answering System Untuk Media Informasi Mahasiswa Universitas Bhayangkara Jakarta Raya," *J. Inf. Inf. Secur.*, vol. 4, no. 2, pp. 175–186, 2023.

[11] M. Farid, D. Wahyudi, A. R. Chrisvo, and M. F. Hidayatullah, "ANALISIS DISTRIBUSI SELISIH TINGKAT PENGGANGGURAN TERBUKA (TPT) DAN TINGAT PARTISIPASI ANGKATAN KERJA (TPAK) MENGGUNAKAN GOOGLE COLAB," *J. Tek. Inform. DAN Multimed.*, vol. 4, no. 2, pp. 39–45, 2024.

[12] V. Novalia, I. Saputra, M. Ula, and M. Danil, "Application of Artificial Intelligence Chi-Square Model and Classification Of KNN in Heart Disease Detection," *JITE(Journal Informatics Telecommun. Eng.*, vol. 6, no. July, pp. 180–188, 2022.

[13] R. A. Syahputra, R. Arifin, and M. Iqbal, "Sentiment Analysis on Tabungan Perumahan Rakyat (TAPERA) Program by using Support Vector Machine (SVM)," vol. 8, no. 2, pp. 531–541, 2024.

[14] A. Mareta, D. E. Subroto, L. Aulia, and S. Nuryanah, "Peran Media Sosial Youtube sebagai Media Edukasi dalam Pendidikan Generasi Z," *GURUKU J. Pendidik. dan Sos. Hum.*, vol. 3, no. 1, pp. 99–105, 2025.

[15] D. Irawan, E. B. Perkasa, D. Wahyuningsih, and E. Helmud, "Perbandingan Klassifikasi SMS Berbasis Support Vector Machine , Naive Bayes Classifier , Random Forest dan Bagging Classifier," *J. SISFOKOM (Sistem Inf. dan Komputer)*, vol. 10, no. 3, pp. 432–437, 2021.