Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

Analisis Sentimen Opini Publik Tentang Gempa Megathrust di Indonesia Menggunakan Metode Support Vector Machine dan Naïve Bayes

Dicky Kurniawan, Muhammad Najib Dwi Satria*

Fakultas Teknik dan Ilmu Komputer, Program Studi Sistem Informasi , Universitas Teknokrat Indonesia, Lampung, Indonesia
Email: [1]dicky_kurniawan@teknokrat.ac.id, [2,*]najibmuhammad@teknokrat.ac.id
Email Penulis Korespondensi: najibmuhammad@teknokrat.ac.id
Submitted: **06/11/2024**; Accepted: **01/12/2024**; Published**: 03/12/2024**

Abstrak−Indonesia merupakan negara yang rawan terhadap bencana alam, terutama gempa bumi. Salah satu ancaman terbesar adalah potensi gempa megathrust yang dapat menyebabkan kerusakan parah dan korban jiwa, terutama di wilayah Jakarta. Di era digital, informasi mengenai potensi gempa megathrust banyak disebarkan melalui platform media sosial seperti YouTube, yang kemudian memicu beragam opini dan komentar dari publik. Video berjudul "GEMPA MEGATHRUST!! WARGA JAKARTA HARUS BERSIAP DENGAN RUNTUHNYA GEDUNG-GEDUNG?" yang diunggah oleh channel YouTube Kamar Jeri telah menarik perhatian publik dan memicu diskusi mengenai kesiapsiagaan masyarakat terhadap potensi bencana ini. Penelitian ini bertujuan untuk menganalisis sentimen opini publik terhadap video tersebut dengan menggunakan dua metode machine learning, yaitu Naive Bayes dan Support Vector Machine (SVM). Untuk mengatasi ketidakseimbangan data antar kelas sentimen, diterapkan teknik Synthetic Minority Over-sampling Technique (SMOTE). Hasil penelitian menunjukkan bahwa SMOTE efektif dalam meningkatkan kinerja kedua model, namun peningkatan pada SVM lebih signifikan. SVM menunjukkan performa yang lebih baik dibandingkan Naive Bayes dalam mengklasifikasikan sentimen opini publik.

Kata Kunci: Analisis sentiment, Gempa megathrust, Naïve Bayes, Support Vector Machine, YouTube

Abstract−Indonesia is a country prone to natural disasters, especially earthquakes. One of the biggest threats is the potential for megathrust earthquakes that can cause severe damage and loss of life, especially in the Jakarta area. In the digital era, information about the potential for megathrust earthquakes is widely disseminated through social media platforms such as YouTube, which then triggers various opinions and comments from the public. A video titled "MEGATHRUST EARTHQUAKE!! JAKARTA RESIDENTS MUST BE PREPARED FOR THE COLLAPSE OF BUILDINGS?" uploaded by the Kamar Jeri YouTube channel has attracted public attention and sparked discussions about community preparedness for this potential disaster. This study aims to analyze public opinion sentiment towards the video using two machine learning methods, namely Naive Bayes and Support Vector Machine (SVM). To overcome the imbalance of data between sentiment classes, the Synthetic Minority Over-sampling Technique (SMOTE) technique was applied. The results showed that SMOTE was effective in improving the performance of both models, but the improvement in SVM was more significant. SVM performed better than Naive Bayes in classifying public opinion sentiment

Keywords: Megathrust earthquake, Naïve Bayes, Sentiment analysis, Support Vector Machine, YouTube

1. PENDAHULUAN

Indonesia, sebagai negara kepulauan yang membentang luas di garis khatulistiwa, dianugerahi keindahan alam yang luar biasa [1]. Namun, di balik panorama memukau tersebut, tersimpan potensi bencana alam yang cukup mengancam. Letak geografis Indonesia yang berada di jalur "Ring of Fire" Pasifik menjadikannya zona dengan aktivitas seismik dan vulkanik yang tinggi [2]. Kondisi ini menempatkan Indonesia pada risiko terhadap berbagai bencana alam, termasuk letusan gunung berapi, tsunami, dan yang paling sering terjadi, gempa bumi. Salah satu ancaman gempa bumi terbesar yang dihadapi Indonesia adalah potensi gempa megathrust [3].

Gempa megathrust terjadi di zona subduksi, yaitu area di mana lempeng tektonik Indo-Australia menunjam ke bawah lempeng Eurasia [4]. Pergerakan lempeng yang terus berlangsung menyebabkan akumulasi energi yang sangat besar di zona pertemuan lempeng [5]. Ketika friksi antar lempeng mencapai titik kritis, energi yang terakumulasi tersebut akan dilepaskan secara tiba-tiba, menghasilkan gempa bumi dengan kekuatan yang dahsyat, yang dikenal sebagai gempa megathrust. Gempa megathrust tidak hanya berpotensi menimbulkan kerusakan infrastruktur yang masif, tetapi juga mengancam jiwa dan kelangsungan hidup masyarakat, terutama jika disertai oleh tsunami [6].

Menghadapi ancaman gempa megathrust yang nyata, pemerintah Indonesia, bersama para ilmuwan dan berbagai lembaga terkait, terus berupaya meningkatkan pemahaman mengenai karakteristik dan potensi dampaknya [7]. Penelitian dan pemantauan berkelanjutan dilakukan untuk mengidentifikasi zona-zona rawan gempa, mempelajari pola dan siklus gempa, serta mengembangkan strategi mitigasi bencana yang efektif. Informasi mengenai potensi gempa megathrust kemudian disebarluaskan kepada publik melalui berbagai kanal, baik melalui media massa konvensional maupun platform digital seperti video YouTube Channel Kamar Jeri [8]. Di era digital saat ini, platform seperti YouTube memiliki peran penting dalam menyebarkan informasi ke masyarakat luas, mengingat kemudahan akses dan jangkauannya yang luas [9].

Penelitian ini akan menganalisis sentimen opini publik terhadap video YouTube Channel Kamar Jeri tentang gempa megathrust di Indonesia, dengan fokus pada video berjudul "GEMPA MEGATHRUST!! WARGA JAKARTA HARUS BERSIAP DENGAN RUNTUHNYA GEDUNG-GEDUNG?"[10]. Video ini dipilih karena menarik perhatian publik dan memicu banyak diskusi *online* mengenai potensi gempa megathrust di Jakarta.[11] Analisis sentimen akan dilakukan dengan menggunakan dua metode *machine learning*, yaitu Naïve Bayes dan Support Vector

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

Machine (SVM)[12]. Metode SMOTE akan diterapkan untuk menyeimbangkan data dan meningkatkan akurasi klasifikasi. Naïve Bayes adalah algoritma klasifikasi probabilistik yang relatif sederhana namun efektif, sementara Support Vector Machine dikenal dengan kemampuannya dalam menangani data berdimensi tinggi dan kompleks. Penelitian ini bertujuan untuk mengetahui bagaimana masyarakat merespon informasi mengenai gempa megathrust yang disampaikan melalui video YouTube dan seberapa siap mereka dalam menghadapi potensi bencana tersebut [13].

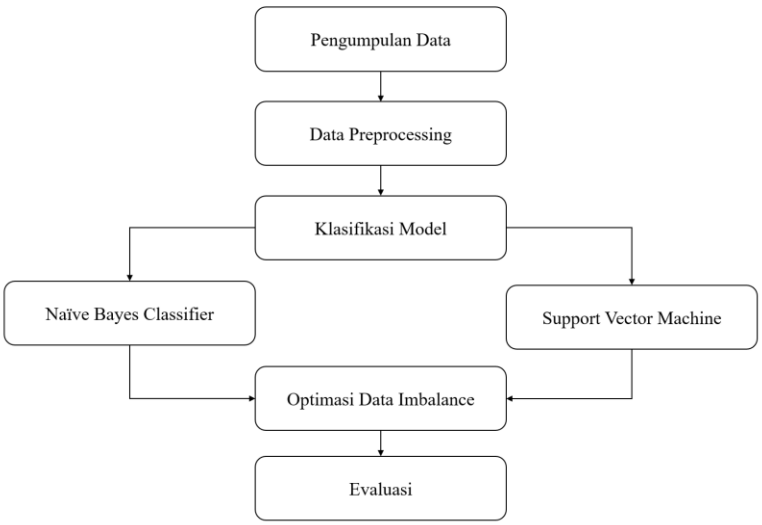
Penelitian tentang analisis sentimen dengan *machine learning* telah dilakukan oleh beberapa peneliti sebelumnya, salah satunya adalah penelitian yang dilakukan oleh Dhea Ananda dkk yang berjudul "Analisis Sentimen Publik Terhadap Pengungsi Rohingya di Indonesia dengan Metode Support Vector Machine dan Naïve Bayes" [14]. Penelitian tersebut menganalisis sentimen publik terhadap pengungsi Rohingya dengan menggunakan metode *machine learning* SVM dan Naive Bayes serta teknik SMOTE untuk menangani ketidakseimbangan data. Hasil penelitian menunjukkan bahwa optimasi dengan metode SMOTE meningkatkan performa model dalam memprediksi sentimen pada dataset yang digunakan. Meskipun terjadi peningkatan performa pada kedua model setelah optimasi SMOTE, model SVM menunjukkan hasil yang lebih baik dibandingkan dengan model Naive Bayes, dengan nilai akurasi mencapai 76% [15].

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis sentimen publik terhadap video YouTube berjudul "GEMPA MEGATHRUST !! WARGA JAKARTA HARUS BERSIAP DENGAN RUNTUHNYA GEDUNG-GEDUNG?" dari channel Kamar Jeri. Data yang digunakan berupa komentar-komentar yang diberikan oleh penonton video tersebut [16].

Untuk menganalisis sentimen yang terkandung dalam komentar, penelitian ini memanfaatkan dua metode machine learning, yaitu Naïve Bayes dan Support Vector Machine (SVM) [17]. Kedua metode ini dipilih karena memiliki karakteristik yang berbeda dan diharapkan dapat memberikan hasil analisis yang lebih komprehensif. Naïve Bayes merupakan algoritma yang sederhana namun cukup powerful, sementara SVM dikenal dengan kemampuannya dalam menangani data yang kompleks [18].

Setelah dilakukan analisis dengan kedua metode tersebut, hasilnya akan dibandingkan untuk mengetahui performa masing-masing dalam mengklasifikasikan sentimen pada data komentar YouTube. [19] Proses ini meliputi pengumpulan data komentar, pembersihan dan persiapan data, pelabelan data dengan sentimen (positif, negatif), pemodelan dengan Naïve Bayes dan SVM, serta evaluasi dan perbandingan hasil klasifikasi. Tujuan akhir dari penelitian ini adalah untuk memahami sentimen dan reaksi masyarakat terhadap informasi mengenai potensi gempa megathrust yang disampaikan melalui video tersebut. Untuk tahapan penelitian bisa dilihat pada Gambar 1 di bawah ini [20].



Gambar 1. Tahapan Penelitian

2.1 State Of The Art

Isu potensi gempa megathrust di Indonesia telah mengemuka sejak awal abad ke-21, dipicu oleh gempa dan tsunami Aceh 2004 dan diperkuat oleh penelitian ilmiah seperti yang dipublikasikan oleh Kerry Sieh dkk. pada tahun 2010. Meskipun BMKG dan berbagai lembaga telah menyebarluaskan informasi dan melakukan edukasi, ketidakpastian waktu kejadian dan potensi dampak yang besar membuat isu ini terus menjadi perhatian, apalagi dengan peran media sosial yang dapat memperkuat atau melemahkan persepsi publik. Oleh karena itu, mitigasi yang

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

komprehensif tetap menjadi kunci dalam mengurangi risiko bencana megathrust di Indonesia. Untuk State Of The Art Perkembangan isu Gempa Megathrust di Indonesia bisa dilihat pada Gambar 2 di bawah ini.

Gambar 2. State of The Art Perkembangan isu Gempa Megathrust di Indonesia

2.2 Data Collecting

Penelitian ini mengambil data dengan cara scraping komentar di *YouTube* pada tayangan " GEMPA MEGATHRUST!! WARGA JAKARTA HARUS BERSIAP DENGAN RUNTUHNYA GEDUNG-GEDUNG" melalui channel *YouTube* Kamar Jeri. Scraping dilakukan dengan memanfaatkan YouTube Data API menggunakan bahasa pemrograman Python. Data yang diambil dari proses scraping berupa komentar dan balasan, yang kemudian disimpan dalam bentuk CSV untuk diproses lebih lanjut.

2.3 Data Preprocessing

Data komentar *YouTube* yang tidak terstruktur dan tidak konsisten perlu dibersihkan melalui proses *text preprocessing* sebelum dilakukan analisis sentimen. Setelah data bersih, proses pelabelan akan dilakukan untuk memberikan label sentimen pada setiap komentar. *Preprocessing* dilakukan menggunakan *library NLTK* dan *spaCy* dalam bahasa pemrograman *Python* dengan tahapan sebagai berikut: *Cleansing*, *Case Folding*, *Tokenizing*, *Stopword Removal*, *Stemming*, dan Pelabelan Data. Berikut ini adalah penjelasan dari tahap *preprocessing*:

a. Cleansing: Dataset mentah dibersihkan dari symbol, emoji, tanda baca, link url, tanda baca dan angka. Untuk contoh *Cleansing* bisa dilihat pada Tabel 1 di bawah ini.

Tabel 1. Cleansing

Comment	Cleansing
Cuma bisa doa dan pasrah sama Allah...panik pun gak guna klo musibah Uda dlm skenario Allah...berusaha menjalani hidup sebaik2nya❤	Cuma bisa doa dan pasrah sama Allah panik pun gak guna klo musibah Uda dlm skenario Allah berusaha menjalani hidup sebaik2ny
klo gada uang apa yg mau dibeli? tau ga org lg pada ga punya duit,dagang sepi, urusan gmpa biar jd urusan tuhan, urusan perut lbh pnting	klo gada uang apa yg mau dibeli tau ga org lg pada ga punya duit dagang sepi urusan gmpa biar jd urusan tuhan urusan perut lbh pnting

b. *Case Folding*: Dataset yang memuat huruf besar diubah menjadi huruf kecil. Untuk contoh *Case Folding* bisa dilihat pada Tabel 2 di bawah ini

Tabel 2. Case Folding

Comment	Case Folding
Alhamdulillah Sulawesi tenggara baik baik saja	alhamdulillah sulawesi tenggara baik baik saja
Sejak megathrust rame,orang2 dg anxiety semakin parah,lg tidur kebangun jantung berdebar takut gempa,jd ujung2nya ya sudahlah pasrah aja	sejak megathrust rame orang2 dg anxiety semakin parah lg tidur kebangun jantung berdebar takut gempa jd ujung2nya ya sudahlah pasrah aja

c. *Tokenize*: Kalimat komentar dipecah menjadi kata per kata seperti “dia sedang mewawancarai kursi” menjadi “dia”, “sedang”, “mewawancarai”, “kursi”. Untuk contoh *Tokenize* bisa dilihat pada Tabel 3 di bawah ini.

Tabel 3. Tokenize

Comment	Tokenize
Bukan menyepelekan peringatan bang ,cuma setiap ada peringatan itu gak pernah terjadi ,sedangkan bila benar terjadi tidak ada peringatan apapun	'buka', 'menyepelekan', 'peringatan', 'bang', 'cuma', 'setiap', 'ada', 'peringatan', 'itu', 'gak', 'pernah', 'terjadi', 'sedangkan', 'bila', 'benar', 'terjadi', 'tidak', 'ada', 'peringatan', 'apapun'

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

Knpa pemerintah diam tidak memfasilitasi yg akan terjadi kalo Indonesia akan ada terjadi gempa yg tidak di duga Bng mohon saran nya untuk pemerintah Han kita gimana Bng	'knpa', 'pemerintah', 'diam', 'tidak', 'memfasilitasi', 'yg', 'akan', 'terjadi', 'kalo', 'indonesia', 'akan', 'ada', 'terjadi', 'gempa', 'yg', 'tidak', 'di', 'duga', 'bng', 'mohon', 'saran', 'nya', 'untuk', 'pemerintah', 'han', 'kita', 'gimana', 'bng'

d. *Stopword Removal*: Kata hubung atau kata-kata yang tidak penting dalam kalimat, misalnya “di”, “ke”, “dari” atau “yang”. Untuk contoh *Stopword* bisa dilihat pada Tabel 4 di bawah ini.

Tabel 4. Stopword

Comment	Stopword
jadi keinget kemarin ada ramalan kiamat dari prindapan tp gk jadi dan dulu gempar 2012 kiamat, lah ini ada lagi	jadi inget kemarin ada ramal kiamat dari prindapan tp gk jadi dan dulu gempar 2012 kiamat, lah ini ada lagi
Gak ada 1pun sebuah bangunan yg tahan dari bencana alam yg dari allah hadirkan , sekuat-sekuatnya sebuah bangunan jika sang penguasa sudah berkehendak akan runtuhlah jua apapun ituI	Gak ada 1pun sebuah bangunan yg tahan dari bencana alam yg dari allah hadir , sekuat-sekuatnya sebuah bangunan jika sang penguasa sudah berkehendak akan runtuh jua apapun itu

e. ***Stemming***: Pada proses ini dilakukan pengubahan kata-kata yang memiliki imbuhan seperti “mainan” menjadi “main”. Untuk contoh *Stemming* bisa dilihat pada Tabel 5 di bawah ini.

Tabel 5. Stemming

Comment	Stemming
Percaya apa gk, nanti kalau memang terjadi gempa bumi beneran pasti keluar anggaran bantuan ini itu, Di korupsi gk kira2 itu dananya	'percaya', 'apa', 'gk', ‘nanti', 'kalau', 'memang', 'jadi', 'gempa', 'bumi', 'beneran', 'pasti', 'keluar', 'anggar', 'bantu', 'ini', 'itu', 'di', 'korupsi', 'gk', 'kira2', 'itu', 'dana
Thn serakah bencana akan melanda ini peringkatan para pejabat pejabat Jogja, garut, sukabumi, banten, kota Bandung	'thn', 'serakah', 'bencana', 'akan', 'landa', 'ini', 'peringkat', 'jabat', 'jabat', 'jogja', 'garut', 'sukabumi', 'banten', 'kota', 'bandung'

f. **Pelabelan**: Pada tahap ini, dataset akan diberi label positif, dan negatif. Penentuan label pada dataset dilakukan secara manual dengan memberi sentiment mengenai komentar yang ada di *YouTube*. Untuk contoh Pelabelan bisa dilihat pada Tabel 6 di bawah ini.

Tabel 6. Pelabelan

Comment	Pelabelan Data
di berita judulnya jelas hanya potensi potensi artinya kemungkinan yg akan terjadi di masa depan tapi gak pasti dan megatrush artinya bukan gempa tepatnya cincin api yg meliputi asia tenggara maaf bang jeri perbaiki narasi informasi yg lebih tepat sehingga tidak menimbulkan rasa was2 pada masyarakat awam yg belum mengerti arti mega trush	Positif
biarin dah dari pada hidup keree bencana megathrus untuk memusnah kan orang orang udah pada jahat semua	Negatif

2.4 Klasifikasi Model

Model klasifikasi dapat diartikan sebagai suatu cara sistem untuk mengelompokkan data ke dalam berbagai kategori berdasarkan karakteristiknya. Proses ini direpresentasikan dalam bentuk matematis. Di antara banyaknya pendekatan yang ada, algoritma *Naive Bayes* dan *Support Vector Machine* menjadi pilihan populer karena efisiensi dan kemampuannya untuk diandalkan. Tujuan utama dari model ini adalah untuk memberikan prediksi yang akurat dan konsisten ketika dihadapkan dengan data baru.

2.5 Support Vector Machine

Support Vector Machine (SVM) adalah teknik dalam machine learning yang bertujuan untuk memilah data ke dalam kelompok-kelompok berbeda. Caranya adalah dengan menemukan suatu pembatas optimal yang memaksimalkan jarak antar kelompok tersebut. SVM bisa digunakan untuk permasalahan yang melibatkan dua kelompok atau lebih. Intinya, SVM mencari pembatas terbaik yang bisa memisahkan data dengan sebaik-baiknya. Keunggulan SVM adalah kemampuannya untuk bekerja dengan data yang memiliki banyak variabel dan rumit.

2.6 Naïve Bayes

Algoritma *Naive Bayes* adalah metode klasifikasi yang sederhana namun efektif, yang bekerja berdasarkan prinsip teorema *Bayes*. Ide dasarnya adalah mengasumsikan bahwa setiap fitur dalam data berdiri sendiri dan tidak saling terkait. Meskipun terdengar sederhana, algoritma ini banyak digunakan dalam berbagai aplikasi, seperti pemrosesan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

teks (misalnya, mengklasifikasikan email spam), dan sistem rekomendasi. Cara kerja Naive Bayes adalah dengan memprediksi label untuk data baru berdasarkan probabilitas kemunculan label tersebut, dengan mempertimbangkan karakteristik data yang diamati. Rumus $P(X|Y)=(P(Y|X).P(X))/(P(Y))$ merupakan inti dari algoritma ini. Sederhananya, rumus ini menghitung probabilitas suatu kejadian X terjadi jika kejadian Y sudah terjadi. Walaupun *Naive Bayes* mengasumsikan bahwa fitur-fitur data saling *independen* (yang tidak selalu benar dalam dunia nyata), algoritma ini seringkali memberikan hasil yang cukup baik, terutama untuk dataset yang besar. Oleh karena itu, *Naive Bayes* sering dijadikan tolak ukur untuk membandingkan performa algoritma klasifikasi yang lebih kompleks..

2.7 Optimasi Data Imbalance

Optimasi data imbalance adalah proses untuk mengatasi masalah ketidakseimbangan jumlah data antar kelas dalam suatu dataset. Ketidakseimbangan ini terjadi ketika satu kelas memiliki jumlah data yang jauh lebih banyak atau lebih sedikit daripada kelas lainnya. Hal ini dapat menyebabkan model machine learning menjadi bias dan tidak akurat dalam memprediksi kelas minoritas. Tujuan optimasi data imbalance adalah meningkatkan kualitas model machine learning dengan mengurangi dampak dari ketidakseimbangan tersebut. Salah satu metode yang umum digunakan adalah SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE bekerja dengan cara menambah jumlah data pada kelas minoritas secara artifisial. Caranya adalah dengan menciptakan data sintetis baru berdasarkan data yang sudah ada di kelas minoritas. Data sintetis ini dibuat berdasarkan karakteristik tetangga terdekat dari data asli. Dalam kasus ini, SMOTE digunakan untuk meningkatkan jumlah data sentimen negatif agar sebanding dengan jumlah data sentimen positif. Awalnya, data sentimen negatif jauh lebih sedikit daripada data sentimen positif. Setelah SMOTE diterapkan, jumlah data pada kedua kelas menjadi seimbang. Dengan menyeimbangkan jumlah data antar kelas, model machine learning dapat belajar dengan lebih baik dan menghasilkan prediksi yang lebih akurat, karena bias akibat ketidakseimbangan data telah dikurangi.

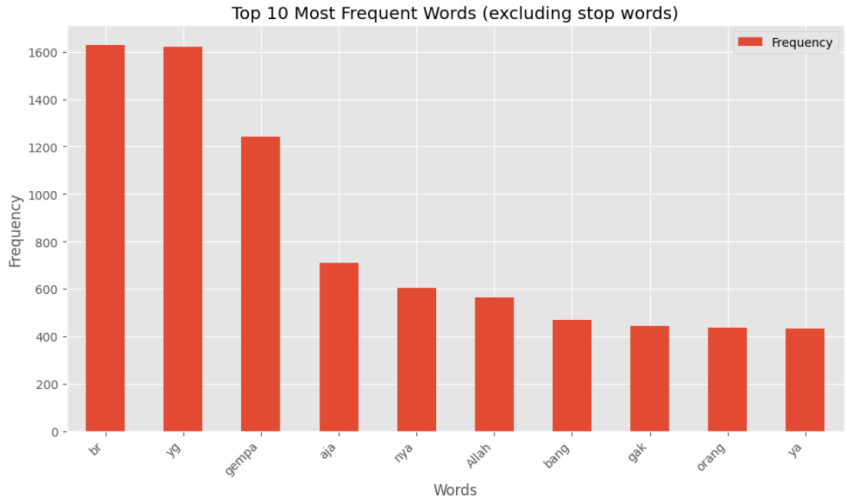
2.8 Evaluasi

Evaluasi model merupakan tahapan penting dalam penelitian analisis sentimen untuk mengukur performa model *machine learning* yang telah dilatih. Pada tahap ini, penulis akan membandingkan akurasi model *Naive Bayes* dan SVM dalam mengklasifikasikan sentimen komentar *YouTube*.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

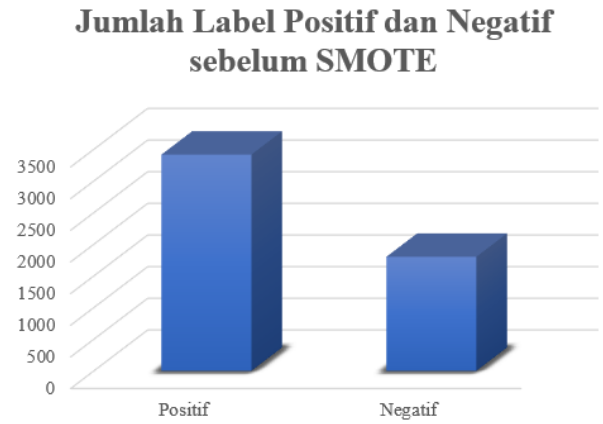
Dataset yang digunakan dalam penelitian ini berasal dari 5.218 komentar pengguna YouTube. Sebelum dianalisis, komentar-komentar tersebut diproses dengan membersihkan dan mempersiapkan teks, seperti menghapus tanda baca, mengubah huruf menjadi huruf kecil, dan menghilangkan kata-kata yang tidak penting. Selanjutnya, setiap komentar diberi label sentimen, yaitu positif atau negatif. Model machine learning (*Naive Bayes* dan SVM) kemudian digunakan untuk mengklasifikasikan sentimen komentar. Dari analisis data, ditemukan 10 kata dengan frekuensi tertinggi, baik pada sentimen positif maupun negatif. Kata-kata tersebut adalah "br", "yg", "gempa", "aja", "nya", "Allah", "bang", "gak", "orang", dan "ya". Untuk memvisualisasikan frekuensi kata-kata ini, digunakan wordcloud, yaitu representasi visual dari kata-kata di mana ukuran setiap kata menunjukkan seberapa sering kata tersebut muncul dalam data. Dalam penelitian ini, dibuat wordcloud yang menampilkan frekuensi kata-kata dari semua komentar, tanpa membedakan sentimen, dan juga wordcloud khusus untuk komentar berlabel positif dan komentar berlabel negatif. Visualisasi wordcloud ini memberikan gambaran sekilas tentang kata-kata yang paling sering muncul dalam setiap sentimen. Untuk contoh Frekuensi kata yang sering muncul bisa dilihat pada Gambar 3 di bawah ini.



Gambar 3. Frekuensi kata yang sering muncul

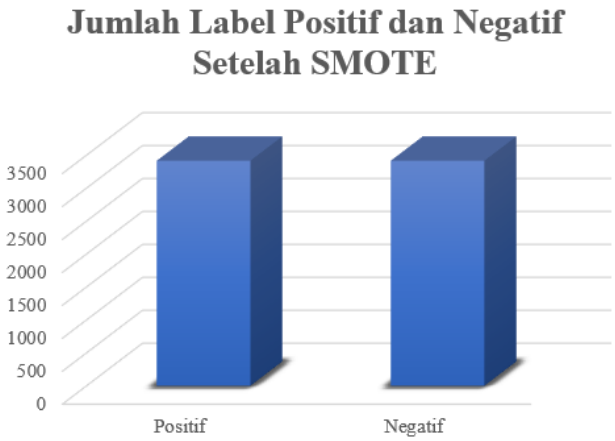
Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

Penelitian ini menemukan adanya ketidakseimbangan jumlah data antara sentimen positif dan negatif, di mana terdapat 3.416 data sentimen positif dan 1.802 data sentimen negatif. Ketidakseimbangan ini dapat mempengaruhi kinerja model machine learning, sehingga perlu diatasi. Untuk itu, penelitian ini menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE berfungsi untuk menyeimbangkan jumlah data antar kelas dengan cara menambah data sintetis pada kelas minoritas (dalam hal ini, sentimen negatif). Data sintetis ini dibuat berdasarkan karakteristik data asli yang sudah ada, sehingga model machine learning dapat mempelajari pola data dengan lebih baik dan menghasilkan prediksi yang lebih akurat. Untuk Jumlah Klasifikasi Sentiment Sebelum Menggunakan SMOTE bisa dilihat pada Gambar 4 di bawah ini.



Gambar 4. Jumlah Klasifikasi Sentiment sebelum Menggunakan SMOTE

Penerapan teknik SMOTE berhasil menyeimbangkan jumlah data antara sentimen positif dan negatif, sehingga masing-masing kelas memiliki 3.416 data. Penyeimbangan ini merupakan langkah krusial dalam mengatasi masalah ketidakseimbangan kelas yang dapat mengganggu kinerja model analisis sentimen. Dengan jumlah data yang seimbang, model machine learning dapat mempelajari pola-pola dari kedua kelas sentimen dengan lebih baik. Hal ini akan meningkatkan kemampuan model dalam mengklasifikasikan sentimen secara akurat, karena model tidak lagi bias terhadap kelas mayoritas. Hasil penyeimbangan data setelah penerapan SMOTE dapat dilihat pada gambar 4. Visualisasi ini menunjukkan distribusi data yang seimbang antara sentimen positif dan negatif, yang diharapkan dapat meningkatkan performa model dalam menganalisis sentimen. Untuk Jumlah Klasifikasi Sentiment Setelah Menggunakan SMOTE bisa dilihat pada Gambar 5 di bawah ini.

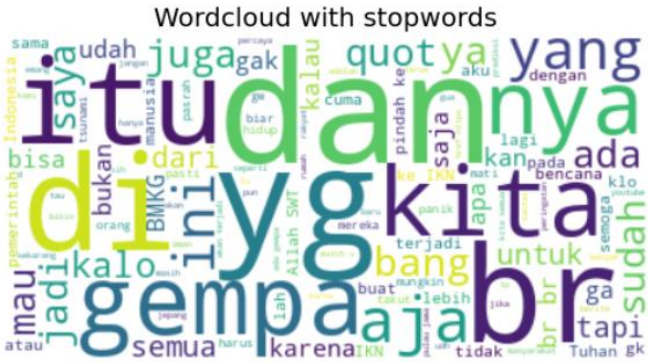


Gambar 5. Jumlah Klasifikasi Sentiment Setelah Menggunakan SMOTE

3.2 Visualisasi Word Cloud

Wordcloud digunakan untuk memvisualisasikan kata-kata yang paling sering muncul dalam suatu dataset. Ukuran kata dalam wordcloud menunjukkan frekuensinya, sehingga kata yang berukuran besar menandakan bahwa kata tersebut sering muncul dalam data. Gambar 6 dalam penelitian ini menunjukkan hasil visualisasi wordcloud dari dataset yang digunakan. Melalui wordcloud ini, kita dapat dengan mudah mengidentifikasi kata-kata yang paling dominan dalam komentar-komentar YouTube yang dianalisis. Wordcloud memberikan cara yang intuitif dan menarik untuk memahami pola dan tren dalam data teks. Dalam konteks analisis sentimen, wordcloud dapat membantu mengidentifikasi kata-kata kunci yang berhubungan dengan sentimen positif dan negatif. Untuk Hasil WordCloud Semua Sentimen bisa dilihat pada Gambar 6 di bawah ini.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213



Gambar 6. Hasil WordCloud Semua Sentimen

3.3 Tahapan Pengujian

Data yang telah melalui tahap preprocessing seperti cleansing dan casefolding dianalisis menggunakan dua algoritma klasifikasi, yaitu Support Vector Machine (SVM) dan Naïve Bayes. SVM mencari hyperplane optimal untuk memisahkan data ke dalam kelas-kelas yang berbeda, sedangkan Naïve Bayes menerapkan teorema Bayes untuk klasifikasi. Data dibagi menjadi 80% data latih dan 20% data uji, dan performa model dievaluasi berdasarkan akurasi, F1-score, presisi, recall, dan confusion matrix. Hasil evaluasi sebelum penerapan SMOTE menunjukkan bahwa SVM mencapai akurasi 84%, presisi 86%, recall 79%, dan F1-score 81%, sedangkan Naïve Bayes mencapai akurasi 71%, presisi 76%, recall 71%, dan F1-score 65%. Setelah penerapan SMOTE untuk mengatasi ketidakseimbangan kelas, terjadi peningkatan performa pada kedua model. SVM menunjukkan peningkatan akurasi menjadi 85%, recall menjadi 81%, dan F1-score menjadi 83%, dengan presisi tetap di angka 86%. Naïve Bayes juga mengalami peningkatan akurasi menjadi 74%, presisi menjadi 78%, recall tetap di angka 74%, dan F1-score meningkat menjadi 75%. Secara keseluruhan, SVM menunjukkan performa yang lebih baik daripada Naïve Bayes sebelum dan sesudah penerapan SMOTE. SMOTE memberikan peningkatan performa pada kedua model, namun peningkatan pada SVM lebih signifikan. Tabel 7 dan 8 menunjukkan hasil laporan klasifikasi sebelum dan sesudah penerapan SMOTE, yang memberikan informasi lebih detail tentang kinerja klasifikasi model untuk setiap kelas sentimen.

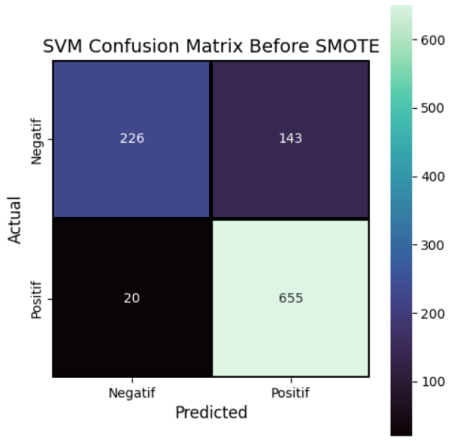
Tabel 7. Hasil laporan klasifikasi sebelum SMOTE

Model	Accuraccy	Precision	Recall	f1-score
SVM	84%	86%	79%	81%
Naïve Bayes	71%	76%	71%	65%

Tabel 8. Hasil laporan klasifikasi Setelah SMOTE

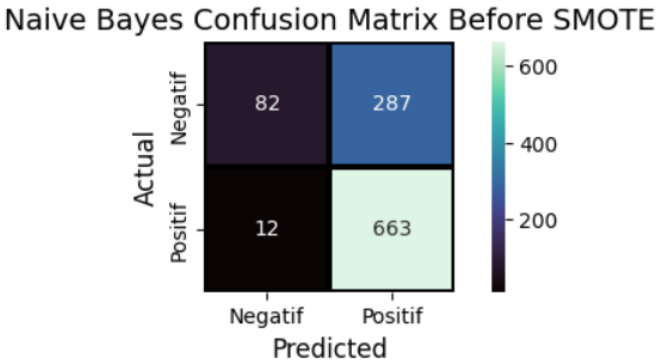
Model	Accuraccy	Precision	Recall	f1-score
SVM	85%	86%	81%	83%
Naïve Bayes	74%	78%	74%	75%

Confusion matrix adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi. Tabel ini membandingkan hasil prediksi dari model dengan label sebenarnya dari data. Dalam penelitian ini, digunakan confusion matrix 2x2, yang berarti model mengklasifikasikan data ke dalam dua kategori (misalnya, positif dan negatif). Untuk melihat perbandingan Hasil Confusion Matrix Algoritma SVM dan Naive Bayes sebelum Optimasi SMOTE bisa dilihat pada Gambar 7 dan 8 di bawah ini.



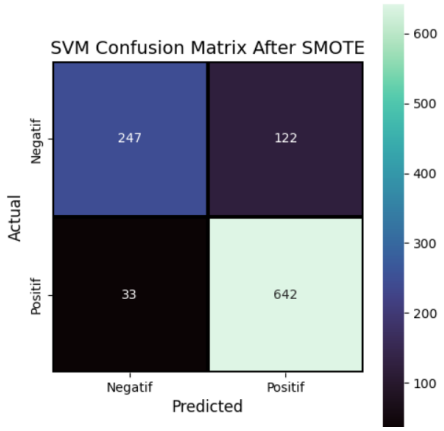
Gambar 7. Confusion Matrix Algoritma SVM Sebelum Optimasi SMOTE

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

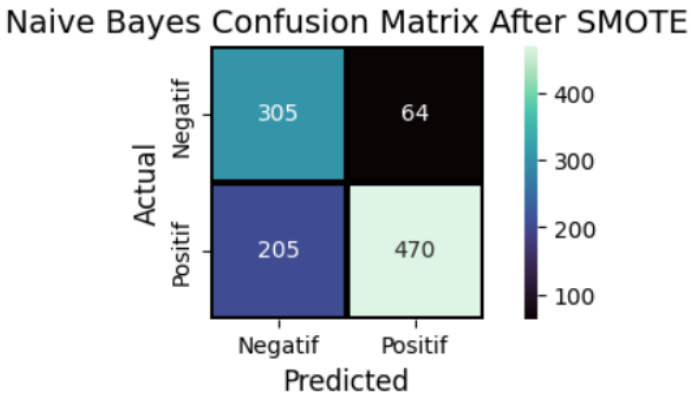


Gambar 8. Confusion Matrix Algoritma Naive Bayes Sebelum Optimasi SMOTE

Sebelum teknik SMOTE diterapkan, *confusion matrix* menunjukkan bahwa baik model *Naive Bayes* maupun SVM cenderung lebih akurat dalam mengklasifikasikan sentimen positif dibandingkan dengan sentimen negatif. Artinya, model lebih sering benar dalam memprediksi sentimen positif, tetapi lebih sering salah dalam memprediksi sentimen negatif. Hal ini bisa terjadi karena jumlah data sentimen positif lebih banyak daripada data sentimen negatif, sehingga model lebih banyak "belajar" tentang sentimen positif. Untuk melihat perbandingan Hasil Confusion Matrix Algoritma SVM dan Naive Bayes setelah Optimasi SMOTE bisa dilihat pada Gambar 9 dan 10 di bawah ini.



Gambar 9. Confusion Matrix Algoritma SVM Setelah Optimasi SMOTE



Gambar 10. Confusion Matrix Algoritma Naive Bayes Setelah Optimasi SMOTE

Setelah penerapan SMOTE, jumlah data sentimen positif dan negatif menjadi seimbang. *Confusion matrix* menunjukkan bahwa akurasi dalam mengklasifikasikan sentimen negatif meningkat pada kedua model. Ini berarti model menjadi lebih baik dalam memprediksi sentimen negatif setelah data diseimbangkan. Namun, terjadi sedikit penurunan dalam akurasi prediksi sentimen positif. Hal ini mungkin terjadi karena penambahan data sintetis pada kelas negatif sedikit "menggeser" fokus model dari sentimen positif. Secara keseluruhan, SMOTE membantu meningkatkan kemampuan model dalam memprediksi sentimen negatif, meskipun ada sedikit penurunan dalam akurasi prediksi sentimen positif. Penyeimbangan data dengan SMOTE bermanfaat untuk meningkatkan kinerja model, terutama dalam mengklasifikasikan kelas minoritas.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

4. KESIMPULAN

Penelitian ini membuktikan efektivitas metode Synthetic Minority Oversampling Technique (SMOTE) dalam meningkatkan kinerja model Support Vector Machine (SVM) dan Naive Bayes untuk prediksi sentimen. SVM awalnya lebih unggul (84%) dibandingkan Naive Bayes (71%) dalam memprediksi sentimen pada dataset yang digunakan. Keunggulan awal SVM ini mengindikasikan bahwa algoritma ini memiliki kemampuan lebih baik dalam memodelkan hubungan kompleks dan non-linear antar fitur pada data, sehingga lebih mampu menangkap pola dan karakteristik data untuk tugas prediksi sentimen. Penerapan SMOTE berhasil meningkatkan akurasi kedua model. SVM mencapai akurasi 85%, sedangkan Naive Bayes mencapai 74%. Peningkatan ini menunjukkan bahwa SMOTE efektif dalam mengatasi ketidakseimbangan data dengan menghasilkan sampel sintetis pada kelas minoritas. Hal ini memungkinkan model untuk belajar dengan lebih baik dan menghindari bias terhadap kelas mayoritas, sehingga meningkatkan kemampuan generalisasi model, khususnya pada kelas minoritas. Meskipun kedua model mengalami peningkatan, SVM tetap menunjukkan performa yang lebih baik setelah penerapan SMOTE. Hal ini menunjukkan SVM lebih sesuai dan efektif dalam memprediksi sentimen pada dataset yang digunakan, terutama setelah data diselaraskan dengan SMOTE. SVM dengan fungsinya yang mampu memetakan data ke dimensi yang lebih tinggi memungkinkan untuk menemukan hyperplane optimal yang memisahkan kelas-kelas sentimen dengan lebih baik. Peningkatan akurasi yang lebih signifikan pada SVM menunjukkan bahwa model ini lebih efektif dalam memanfaatkan data seimbang yang dihasilkan SMOTE. Hasil penelitian ini menegaskan pentingnya penanganan ketidakseimbangan data dan keunggulan SVM dalam prediksi sentimen. Ke depannya, penelitian lanjutan dapat mengeksplorasi parameter dan kernel SVM yang berbeda, serta kombinasi SMOTE dengan teknik lain seperti undersampling untuk mengoptimalkan kinerja prediksi sentimen.

REFERENCES

[1] A. P. Astuti, S. Alam, and I. Jaelani, "Komparasi Algoritma Support Vector Machine Dengan Naive Bayes Untuk Analisis Sentimen Pada Aplikasi BRImo," Bangkit Indonesia, vol. 11, no. 2, 2022. DOI: 10.52771/bangkitindonesia.v11i2.196.

[2] S. H. Ramadhani and M. I. Wahyudin, "Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), vol. 6, no. 4, 2022. DOI: 10.35870/jtik.v6i4.530.

[3] U. Kusnia and F. Kurniawar, "Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes," Explore: Jurnal Keilmuan dan Aplikasi Teknik Informatika, vol. 5, no. 3, 2022. DOI: 10.31004/innovative.v4i3.10769.

[4] R. Fazal and L. Andraini, "Membandingkan Support Vector Machines Dan Naive Bayes Pada Analisis Sentimen Data Twitter," Portaldata.org, vol. 2, no. 10, 2022.

[5] D. Alita and R. B. A. Shodiqin, "Sentimen Analisis Vaksin Covid-19 Menggunakan Naive Bayes Dan Support Vector Machine," Journal of Artificial Intelligence and Technology Information (JAITI), vol. 1, no. 1, pp. 1-12, Mar. 2023. DOI: 10.58602/jaiti.v1i1.20.

[6] F. N. Hidayat and Sugiyono, "Analisis Sentimen Masyarakat Terhadap Perekrutan PPPK Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine," Jurnal Sains dan Teknologi, vol. 5, no. 2, pp. 665-672, Dec. 2023. DOI: 10.55338/saintek.v5i1.1359.

[7] I. Yunanto and S. Yulianto, "Twitter Sentiment Analysis PeduliLindungi Application Using Naïve Bayes and Support Vector Machine," Jurnal Teknik Informatika (JUTIF), vol. 3, no. 4, pp. 807-814, Aug. 2022. DOI: 10.20884/1.jutif.2022.3.4.292.

[8] A. T. Zy and W. Hadikristanto, "Implementasi Algoritma Metode Naive Bayes dan Support Vector Machine Tentang Pembobolan dan Kebocoran Data di Twitter," Bulletin of Information Technology (BIT), vol. 4, no. 1, pp. 49-56, Mar. 2023. DOI: 10.47065/bit.v3i1.493.

[9] F. M. Sarimole and Kudrat, "Analisis Sentimen Terhadap Aplikasi Satu Sehat Pada Twitter Menggunakan Algoritma Naive Bayes Dan Support Vector Machine," Jurnal Sains dan Teknologi, vol. 5, no. 3, pp. 783-790, Feb. 2024. DOI: 10.55338/saintek.v5i1.2702.

[10] R. Noviana dan I. Rasal, "Penerapan Algoritma Naive Bayes dan SVM untuk Analisis Sentimen Boy Band BTS pada Media Sosial Twitter," Jurnal Teknologi dan Sistem (JTS), vol. 2, no. 2, pp. 51-60, Juni 2023. DOI: https://doi.org/10.56127/jts.v2i2.791.

[11] D. Ananda and R. R. Suryono, "Analisis Sentimen Publik Terhadap Pengungsi Rohingya di Indonesia dengan Metode Support Vector Machine dan Naïve Bayes," Jurnal Media Informatika Budidarma, vol. 8, no. 2, pp. 748-757, 2024. DOI: 10.30865/mib.v8i2.7517.

[12] R. M. Arthansa, D. I. Sagita, and A. P. Sari, "Komparasi Analisis Sentimen Ulasan Film Avengers: Endgame di IMDb Menggunakan Metode Naïve Bayes dan SVM," STORAGE – Jurnal Ilmiah Teknik dan Ilmu Komputer, vol. 3, no. 3, pp. 156-166, 2024. DOI: 10.55123.

[13] R. Sulistiawati dan M. K. Sulaeman, "Analisis Sentimen Aplikasi Maskapai Penerbangan Lion Air Menggunakan Metode SVM dan Naïve Bayes," Indonesian Journal of Computer Science, vol. 13, no. 3, 2024. DOI: https://doi.org/10.33022/ijcs.v13i3.3836.

[14] U. R. H. Baba, "Analisa Sentimen Menjelang Pemilihan Umum Presiden 2024 di Indonesia Menggunakan Perbandingan Performa Support Vector Machine (SVM) dan Naïve Bayes," INNOVATIVE: Journal of Social Science Research, vol. 4, no. 3, pp. 11972-11990, 2024. DOI: https://doi.org/10.31004/innovative.v4i3.10761.

[15] R. Rasiban and S. Riyadi, "Analisis Sentimen Opini Masyarakat Terhadap Stadion Jakarta Internasional Stadium (JIS) Pada Twitter Dengan Perbandingan Metode Naive Bayes Dan Support Vector Machine," Jurnal Sains dan Teknologi, vol. 5, no. 3, pp. 1010-1017, Feb. 2024. DOI: 10.55338/saintek.v5i3.2962.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1559−1568
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6213

[16] S. Lestari and S. Berliani, "Analisis Sentimen Masyarakat Terhadap Isu Pecat Sri Mulyani Pada Twitter Menggunakan Metode Naive Bayes Dan Support Vector Machine," Jurnal Sains dan Teknologi, vol. 5, no. 3, pp. 951-960, Feb. 2024. DOI: 10.55338/saintek.v5i3.2746.

[17] A. D. Dayani, Yuhandri, and G. W. Nurcahyo, "Analisis Sentimen Terhadap Opini Publik pada Sosial Media Twitter Menggunakan Metode Support Vector Machine," Jurnal KomtekInfo, vol. 11, no. 1, pp. 1-10, 2024. DOI: 10.35134/komtekinfo.v11i1.439.

[18] R. Rasiban and S. Riyadi, "Analisis Sentimen Opini Masyarakat Terhadap Stadion Jakarta Internasional Stadium (JIS) Pada Twitter Dengan Perbandingan Metode Naive Bayes Dan Support Vector Machine," Jurnal Sains dan Teknologi, vol. 5, no. 3, pp. 1010-1017, Feb. 2024. DOI: 10.55338/saintek.v5i3.2962.

[19] L. A. Hayurian and N. Hendrastuty, "Comparison of Naïve Bayes Algorithm and Support Vector Machine in Sentiment Analysis of Boycott Israeli Products on Twitter," Jurnal Teknik Informatika (JUTIF), vol. 5, no. 3, pp. 731-738, Jun. 2024. DOI: 10.52436/1.jutif.2024.5.3.1813.

[20] U. Kusnia dan F. Kurniawan, "Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play Menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes," Explore: Jurnal Keilmuan dan Aplikasi Teknik Informatika, vol. 12, no. 1, pp. 1-10, Juni 2022. DOI: 10.35891/explorit.