TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

Analisis Sentimen terhadap Jalan Rusak di Palembang Pada Media Sosial Menggunakan Algoritma Naïve Bayes

Muhammad Rifqi Virgiansyah, Stephanie, Muhammad Rizky Pribadi*

Fakultas Ilmu Komputer dan Rekayasa, Program Studi Informatika, Universitas Multi Data Palembang, Kota Palembang, Indonesia
Email: [1]mrifqivirgiansyah_2226250075p@mhs.mdp.ac.id, [2]stephanietjhai@mhs.mdp.ac.id, [3,*]rizky@mdp.ac.id
Email Penulis Korespondensi: rizky@mdp.ac.id

Abstrak−Pada negara-negara yang maju, pembangunan infrastruktur jalan memainkan peran vital untuk menggerakkan peningkatan ekonomi. Keterlibatan otoritas publik sangat diperlukan dalam penyediaan dana untuk pembangunan infrastruktur tersebut, mengingat jalan adalah salah satu bentuk infrastruktur publik yang sangat penting. Namun, masih terdapat beberapa daerah, seperti Palembang, di mana kondisi jalan rusak masih menjadi masalah serius. Pemerintah daerah Palembang diduga kurang proaktif dalam memperbaiki infrastruktur jalan di wilayah mereka. Dalam upaya memperjuangkan perhatian pemerintah terhadap kondisi jalan di Palembang, masyarakat menggunakan media sosial sebagai sarana untuk menyuarakan keluhan. Sejumlah *tweet*, postingan, dan kiriman yang menggunakan kata kunci "Jalan Rusak Palembang" di Twitter, Instagram, dan TikTok mengungkapkan bahwa sentimen didapati bahwa hanya 1,8% menunjukkan sentimen positif, 13,3% sentimen netral, dan 84,9% sentimen negatif terhadap kondisi jalan yang rusak mendominasi. Dalam rangka mengumpulkan data lebih lanjut mengenai persepsi masyarakat terhadap kondisi jalan di Palembang, dilakukan penelitian menggunakan algoritma Naïve Bayes. Hasil pengujian menunjukkan bahwa model Naïve Bayes memberikan akurasi sebesar sebesar 91,20%, presisi sebesar 92,32%, recall sebesar 91,20%, dan F1 score sebesar 91,26%, menunjukkan performa yang sangat baik dalam mengklasifikasikan sentimen masyarakat terhadap jalan rusak di Palembang.

Kata Kunci: Analisis Sentimen; Jalan Rusak; Palembang; Media Sosial; Naïve Bayes

Abstract−In developed countries, the development of road infrastructure plays a vital role in driving economic development. The involvement of public authorities is very necessary in providing funds for infrastructure development, considering that roads are a very important form of public infrastructure. However, there are still some areas, such as Palembang, where damaged roads are still a serious problem. The Palembang regional government is suspected of being less proactive in improving road infrastructure in their area. In an effort to fight for the government's attention to road conditions in Palembang, the public uses social media as a means to voice complaints. A number of tweets, posts, and submissions using the keyword "Jalan Rusak Palembang" on Twitter, Instagram, and TikTok revealed that sentiment was found to be only 1.8% showing positive sentiment, 13.3% neutral sentiment, and 84.9% sentiment negativity towards damaged road conditions dominates. In order to collect further data regarding public perceptions of road conditions in Palembang, research was conducted using the Naïve Bayes algorithm. The test results show that the Naïve Bayes model provides an accuracy of 91,20, precision of 92,32%, recall of 91,20%, and F1 score of 91,26%, showing excellent performance in classifying public sentiment towards damaged roads in Palembang.

Keywords: Sentiment Analysis; Damaged Roads; Palembang; Social Media; Naïve Bayes

1. PENDAHULUAN

Infrastruktur jalan memiliki peran yang sangat signifikan dalam memfasilitasi progres pembangunan, terutama di negara-negara yang tengah mengalami perkembangan dan peningkatan ekonomi. Data dari (Statistik, 2022), jumlah jalan rusak berat di Sumatera Selatan mencapai 225,29%. Pemerintah memegang peran penting dalam mendukung investasi guna pembangunan infrastruktur. Secara umum, jalan raya diakui sebagai salah satu infrastruktur publik yang membutuhkan keterlibatan pemerintah untuk mencapai hasil terbaik yang memajukan ekonomi. Oleh sebab itu, penelitian ini bertujuan mengidentifikasi faktor jalan rusak di provinsi Sumatera Selatan dan memberikan rekomendasi solusi yang tepat (Arjunanto & Waluyo, 2023).

Text Mining adalah proses penambangan data yang melibatkan teks, dimana sumber datanya berasal dari dokumen dan bertujuan untuk menemukan kata-kata yang merepresentasikan isi dari dokumen tersebut, sehingga memungkinkan untuk menganalisis keterhubungan antar dokumen. Tujuan dari *text mining* adalah mengekstrak informasi yang berguna dari sumber data. Dengan demikian, sumber data yang digunakan dalam text mining adalah kumpulan dokumen yang memiliki format yang tidak terstruktur, dan proses ini melibatkan identifikasi dan eksplorasi pola yang menarik (Anwar, 2022).

Analisis sentimen, atau yang dikenal sebagai analisis sentimen dalam bahasa Indonesia, adalah metodologi yang digunakan untuk menilai bagaimana emosi disampaikan melalui teks tertulis, lalu mengelompokkan emosi tersebut menjadi kategori positif atau negatif (Muhammad Afdal & Elita, 2022). Perspektif serupa diungkapkan dalam referensi (Furqan et al., 2022), di mana analisis sentimen dimanfaatkan untuk memahami komentar pengguna di internet dan menginterpretasikan persepsi mereka terhadap suatu produk atau merek. Berdasarkan hasil (Alrajak et al., 2020), analisis sentimen adalah metode komputasi yang diterapkan untuk mengenali dan mengklasifikasikan pandangan, emosi, serta sikap yang terhadap dalam data tekstual, umumnya memisahkan sentimen negatif dan positif. Seperti yang telah dibahas sebelumnya, banyak pengguna internet terlibat dalam kegiatan mendokumentasikan pengalaman pribadi mereka, menyampaikan pendapat, dan membagikan berbagai aspek hidup mereka. Merinci daftar lengkap perasaan yang biasa dialami individu, yang dikategorikan sebagai menyenangkan, netral, atau negatif, sering kali disampaikan dengan cara yang kompleks.

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

Platform media sosial berbasis internet memungkinkan pengguna membuat profil yang bisa dilihat oleh publik atau kelompok tertentu. Pengguna bisa menghubungkan diri dengan orang lain yang mereka pilih, berbagi konten, dan melihat profil serta jaringan orang lain. Media sosial juga menjadi tempat bagi pengguna untuk aktif membuat, berbagi, dan berinteraksi dengan konten, serta membentuk komunitas online (Arjunanto & Waluyo, 2023).

Penelitian sebelumnya tentang analisis sentimen mencakup analisis terhadap 350 *tweet* menunjukkan sentimen positif sebesar 30,29% dan sentimen negatif sebesar 69,71% dari tanggal 7 April hingga 19 Juni 2023. Hasil terbaik diperoleh menggunakan K=5, dengan akurasi 60%, presisi 36%, dan penarikan 50% (Arjunanto & Waluyo, 2023). Dalam studi terpisah, analisis sentimen terkait presiden dilakukan. Penelitian ini menggunakan algoritma Naïve Bayes dengan 300 data berlabel manual. Pengujian menggunakan *Confusion Matrix* dan 10 folds *Cross Validation* dengan membagi data menjadi beberapa himpunan. Setelah pengujian pada 300 data dilakukan 10 kali dengan 30 data uji yang berbeda, hasil menunjukkan akurasi 87%, presisi 77%, dan penarikan 73% (Salsabila & Wibowo, 2023). Algoritma Naïve Bayes digunakan dalam penelitian bernama "Analisis Sentimen Kinerja Kepemimpinan Bupati dari Data Komentar Menggunakan Metode Naïve Bayes Classifier". Hasil pada penelitian ini yang menggunakan sejumlah 200 data, yang terdiri dari 78 data yang bernilai positif serta 122 data yang bernilai negatif menunjukkan bahwa metode Naïve Bayes mampu mengklasifikasikan sentiment dengan nilai akurasi yang didapatkan sebesar 82%. Hasil perhitungan rata-rata precision sebesar 89%, perhitungan rata-rata Recall 77% serta perhitungan rata-rata F1-Score sejumlah 79% (Seminar et al., 2024). Pada penelitian "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes" menunjukkan bahwa pembelajaran daring memiliki 30% sentimen positif, 69% sentimen negatif, dan 1% netral pada periode tersebut. Tingginya sentimen negatif dihasilkan karena ketidakpuasaan masyarakat terhadap pembelajaran daring. Beberapa twit menunjukkan kekecewaan dengan kata 'stres' dan 'malas' merupakan kata yang memiliki frekuensi tinggi dalam percakapan (Prabowo & Wiguna, 2021). Dalam penelitian lain, terkait analisis jasa transportasi menggunakan algoritma Naïve Bayes didapatkan hasil sentimen positif sebesar 88.60% dan sentimen negatif sebesar 11.40% dengan akurasi sebesar 86.80%. Hasil menunjukkan tingkat sentimen positif dari *tweet* masyarakat lebih besar dibandingkan dengan tingkat sentimen negatif (Mas Pintoko & Muslim, 2018).

Meskipun penelitian sebelumnya telah banyak mengkaji analisis sentimen di berbagai bidang, termasuk sosial media dan kinerja pemerintahan, belum ada penelitian yang secara khusus mengkaji sentimen publik terkait kondisi infrastruktur jalan di Sumatera Selatan menggunakan teknik *text mining* dan analisis sentimen. Penelitian sebelumnya cenderung berfokus pada analisis sentimen produk atau merek, serta isu politik, tanpa mengaitkannya dengan kondisi infrastruktur yang memiliki dampak langsung terhadap kehidupan sehari-hari masyarakat.

Penelitian ini bertujuan untuk mengisi perbedaan yang ada dengan mengidentifikasi faktor-faktor yang menyebabkan jalan rusak di provinsi Sumatera Selatan melalui pendekatan *text mining* dan analisis sentimen. Selain itu, penelitian ini akan memberikan rekomendasi solusi yang tepat berdasarkan analisis sentimen publik terhadap kondisi jalan rusak, sehingga dapat membantu pemerintah dalam merencanakan dan mengimplementasikan kebijakan yang lebih efektif untuk memperbaiki infrastruktur jalan di Sumatera Selatan.

2. METODOLOGI PENELITIAN

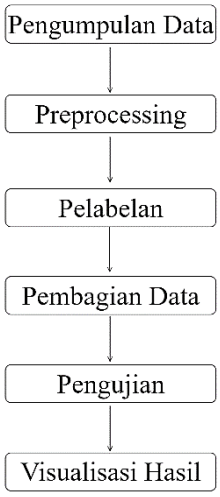
2.1 Data Penelitian

Penelitian ini memanfaatkan data teks dari Twitter berupa *tweet*, Instagram berupa postingan, dan TikTok berupa kiriman yang diperoleh mulai tanggal 16 Maret 2019 hingga 14 Mei 2024 sejumlah 668 data. Penggunaan teknik *crawling* sangat penting dalam hal pengumpulan data serta memantau perkembangan cepat di internet (Rakhmawati et al., 2021). Data itu didapatkan dengan menggunakan pustaka Apify melalui proses *scraping*. Untuk pengumpulan data, penelitian ini berfokus pada satu parameter kata kunci yang berkaitan dengan Palembang. Setelah data diperoleh, data tersebut disimpan dalam format CSV untuk memudahkan pengolahan lebih lanjut. Setiap data kemudian diberi *label* secara manual berdasarkan sentimen, yang dikategorikan menjadi sentimen positif dan negatif. Pelabelan manual bertujuan guna memastikan akurasi yang akan dilakukan menggunakan algoritma Naïve Bayes. Dengan pelabelan yang akurat, diharapkan model Naïve Bayes dapat melakukan klasifikasi sentimen dengan lebih tepat dan dapat diandalkan. Proses ini tidak hanya membantu dalam menguji efektivitas algoritma Naïve Bayes tetapi juga memberikan wawasan yang lebih mendalam mengenai sentimen publik terhadap topik yang berkaitan dengan Palembang selama periode tersebut.

2.2 *Preprocessing*

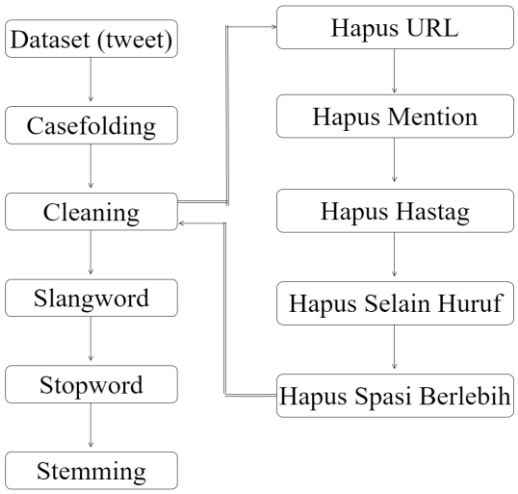
Pada tahap pengumpulan data, data dikumpulkan dari berbagai platform media sosial seperti Twitter, Instagram, dan Tik Tok. Data yang dikumpulkan berhubungan dengan topik jalan rusak di Palembang. Selanjutnya, pada tahap preprocessing, proses ini bertujuan untuk membersihkan dan mempersiapkan data teks agar siap untuk analisis lebih lanjut. Setelah itu, data yang telah dipreprocessing dilabeli untuk menentukan sentimen, seperti positif, netral, dan negatif. Kemudian, dataset dibagi menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data). Algoritma Naïve Bayes dilatih menggunakan data latih, dan kinerjanya diuji menggunakan data uji. Tahap terakhir adalah visualisasi hasil, di mana hasil analisis sentimen ditampilkan dalam bentuk visual untuk memudahkan interpretasi yang dapat dilihat pada Gambar 1.

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239



Gambar 1. Tahap Metode

Tahap pertama dalam analisis sentimen menggunakan data Twitter adalah pra-pemrosesan, yang sangat penting karena secara langsung mempengaruhi hasil performa klasifikasi (Khairunnisa et al., 2021). Memroses data teks yang tidak terstruktur menjadi teks yang terstruktur membutuhkan langkah-langkah pra-pemrosesan data. Karena itu, penelitian ini akan mengeksplorasi beberapa tahapan pra-pemrosesan teks sebagaimana yang ditunjukkan dalamm Gambar 2.



Gambar 2. Tahap *Processing*

a. *Preprocessing* adalah tahapan kritis yang tak terhindarkan dalam evalusi sentimen melalui data Twitter. Langkah ini secara langsung mempengaruhi mutu dan efisiensi klasifikasi yang dihasilkan (Khairunnisa et al., 2021).
b. *Cleaning* adalah proses untuk menghilangkan tanda baca, angka, simbol, *link* URL, dan username dalam teks (Khairunnisa et al., 2021).
c. *Slang word* adalah merupakan prosedur untuk mengubah kata-kata informal menjadi formal dengan menggunakan referensi kamus istilah informal. Kata-kata tersebut kemudian diubah menjadi bentuk yang lebih resmi. Contoh kata *slang slang* yang sering digunakan adalah baper, sotoy, mager, dan lain sebagainya (Nofiyani & Wulandari, 2022).
d. *Stopword removal* adalah mengeliminasi kata-kata penghalang yang bertujuan untuk membuang kata-kata yang dianggap tidak relevan dalam teks, seperti ‘bisa’, ‘tegas’, ‘mampu’, dan ‘tidak’(Khairunnisa et al., 2021).
e. *Stemming* merupakan langkah merubah kata dalam sebuah *tweet* ke bentuk dasarnya guna mengurangi variasi fitur yang memiliki arti serupa. Ini dilaksanakan untuk menangani variasi yang muncul akibat adanya imbuhan pada kata-kata tersebut (Khairunnisa et al., 2021).

2.3 *Labeling*

Labeling merupakan fase di mana dokumen (*tweet)* diberi *label* yang berasal dari hasil *crawling* Twitter dan telah melewati seleksi data sebelumnya (Juniarsih et al., 2020). Dalam penelitian ini, proses memberikan *label* mengkategorikan setiap *tweet* ke dalam kategori positif dan negatif. *Labeling* dilaksanakan secara manual oleh dua individu, sementara satu orang lagi bertanggung jawab untuk memverifikasi data tersebut.

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

2.4 Pembagian Data

Pemisahan data merujuk pada implementasi membagi kumpulan data menjadi dua bagian terpisah: data pelatihan dan data pengujian. Data pelatihan dipakai untuk mengembangkan dan melatih model klasifikasi, sementara data pengujian digunakan untuk mengevaluasi performa model yang telah dilatih. Proses ini melibatkan pembentukan model klasifikasi dengan menggunakan data pelatihan yang telah ditetapkan, yang kemudian dievaluasi dengan menggunakan data pengujian yang terpisah. Biasanya, perbandingan antara data pelatihan dan data pengujian adalah 80:20, di mana 80% data dialokasikan untuk pelatihan dan 20% sisanya untuk pengujian. Keberhasilan klasifikasi bergantung pada komposisi yang sesuai dari data latih, dan jika data latih mencerminkan mayoritas kasus yang mungkin terjadi dalam pengujian, maka hasilnya akan lebih maksimal.

2.5 Pemodelan

Pemodelan mengacu pada proses sistematis untuk memperoleh pemahaman atau wawasan dari data pelatihan yang tersedia. Seleksi data latih model dilakukan dengan menggunakan teknik sampling kuota. Sampling kuota adalah strategi pengambilan sampel yang dipakai untuk menjamin bahwa jumlah sampel dari suatu populasi memiliki karakteristik atau kriteria tertentu sampai memenuhi kuota yang ditentukan (Priandi & Painem, 2021).

2.6 Naïve Bayes

Naive Bayes Classifier adalah metode klasifikasi yang didasarkan pada prinsip Teorema Bayes. Metode ini menggunakan pendekatan probabilistik dan statistika yang dikembangkan oleh ilmuwan Inggris Thomas Bayes. Metode ini memprediksi probabilitas kejadian di masa depan dengan mengandalkan data historis, sehingga dikenal sebagai Teorema Bayes. Naive Bayes Classifier sering digunakan karena kemampuannya untuk menangani data dengan cepat dan efisien, meskipun dengan asumsi independensi yang sering kali tidak sepenuhnya akurat dalam kasus nyata. Ciri utama dari Naive Bayes Classifier adalah asumsi yang sangat kuat (naif) bahwa setiap fitur dalam data bersifat independen satu sama lain. (Rayuwati et al., 2022).

Persamaan dalam Algoritma Naïve Bayes digunakan sebagai panduan untuk menghitung probabilitas dalam pengambilan keputusan. Dengan demikian, memberikan kerangka kerja yang berguna dalam mengevaluasi probabilitas pada persamaan (1)

$$P(X|H) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Di mana :
X : Data dengan *class* yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X
P(H) : Probabilitas hipotesis H
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

Dalam rumus di atas, P(H|X) adalah probabilitas bahwa hipotesis H benar jika kita memiliki data X, yang dihitung dengan mengalikan probabilitas kemunculan data X dalam hipotesis H dengan probabilitas awal hipotesis H dan kemudian dibagi dengan probabilitas kemunculan data X secara keseluruhan.

2.7 Pengujian

Pengujian model dilakukan dengan menggunakan data uji (Nikmatun et al., 2019). Langkah ini melibatkan pengujian model yang telah dibuat dengan menggunakan data eksperimental yang ada. Hasil pengujian akan dievaluasi menggunakan matriks kontingensi seperti yang ditunjukkan pada Tabel 1 untuk menilai tingkat akurasi, presisi, dan recall. Dalam matriks kontingensi, TP (True Positive) adalah jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1, TN (True Negative) adalah jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0, FP (False Positive) adalah jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1, dan FN (False Negative) adalah jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0 (Normawati & Prayogi, 2021).

Tabel 1. *Confusion Matrix*

		Nilai Aktual	
		TRUE (*positive*)	TRUE (*positive*)
Nilai Prediksi	TRUE (Positive)	TP (True Positive)	TP (True Positive)
	False (Negative)	FN (False Negative	FN (False Negative)

Percobaan dijalankan untuk mengevaluasi efektivitas dan kinerja model latihan dengan menerapkan algoritma yang direkomendasikan, dengan menekankan pada evaluasi nilai, akurasi, presisi, recall, dan F1 score. Penelitian ini

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

melibatkan pengujian dengan membandingkan data prediksi (terutama, data hasil klasifikasi) dengan kumpulan data aktual yang sudah diberi *label*.

3. HASIL DAN PEMBAHASAN

Temuan dari eksperimen menunjukkan tingkat ketepatan algoritma dalam klasifikasi prediksi menggunakan data uji. Akurasi ini didapatkan setelah proses gabungan *scraping* data dari Twitter, Instagram, dan TikTok yang menghasilkan informasi seperti *full_text* atau teks komentar, *username* atau pengguna, *created_at* atau waktu unggah komentar pada TikTok sebagaimana dapat dilihat pada Tabel 2.

Tabel 2. Dataset

full_text	username	created_at
@Zethaan @tagarabak Lagian jgn bicara ketinggian bikin terowongan dana tol ini aja kurang. Akan bertambah lagi beban APBN. Sukur2 kl tidak jd bancakan. Jalan tol Palembang Lampung br 3 tahun aja udah byk rusak. Apakah ada yg bertanggung jawab?? Masih belu	MikirSok	Mon May 13 05:01:55 +0000 2024
@WGreborn @gibran_tweet Gibran itu ga lulus dr UTS australia kata Bima yg anak palembang viral gegara kritik gubernur mslh jalan rusak, bima skrg masih kuliah di australia.	oetomo_ian	Thu May 02 09:05:10 +0000 2024
Marlina : Perbaikan Jalan Rusak akibat pembangunan IPAL oleh project Kementerian akan di Perbaiki Dalam Waktu Dekat Oleh Pemerintah: Palembang,MA-Sosialisasi Pembangunan Jaringan Perpipaan Air Limbah PCSP Paket C1 di tingkat Kelurahan.yang di gelar diâ€¦ https://t.co/WIue1nva6q	ansshajo	Thu Apr 25 14:42:40 +0000 2024
@goraici Sebenernya lagi dirujak gubernur nya disini. Masalah Palembang bukan cuma ini aja sebenernya. Dari mulai jalan rusak sampai banjir. Cuma ga tau ga seviral kota kota lain.	Nandarizzky123	Thu Apr 25 13:43:12 +0000 2024
Viral Aksi Protes Pria Berjas dan Berdasi di Palembang Mandi di Kubangan Jalan Rusak #trending #shorts #indoviral #bbtvi #videoviral #beritaterkini #FYP #viraltwitter #fcklive #bbrightvc #deprem #meme https://t.co/QiIzCtDc8z	boedakjad_l	Thu Apr 25 12:01:26 +0000 2024

Proses dimulai dengan pengumpulan data dari Twitter, Instagram, dan TikTok dan pelabelan sentimen secara manual. Data mentah dari Twitter, Instagram, dan TikTok masih mengandung banyak ikon dan tautan. Oleh karena itu, diperlukan tahap *cleaning* data untuk pembersihan data yang mengacu pada proses mendeteksi dan memperbaiki (atau menghapus) kesalahan dan inkonsistensi dalam sebuah kumpulan data untuk meningkatkan kualitasnya. Hal ini mungkin melibatkan penanganan nilai yang hilang, koreksi entri yang salah, penanganan pencilan (outliers), dan standarisasi format yang ditunjukkan pada Gambar 3.

```
def clean_twitter_text(text):
    text = re.sub(r'@[A-Za-z0-9_]+', '', text)  # Menghapus mentions
    text = re.sub(r'#\w+', '', text)            # Menghapus hashtags
    text = re.sub(r'RT[\s]+', '', text)         # Menghapus RT
    text = re.sub(r'https?://\S+', '', text)    # Menghapus URL

    # Memperbaiki regex untuk mempertahankan spasi
    text = re.sub(r'[^A-Za-z0-9 ]', '', text)
    text = re.sub(r'\s+', ' ', text).strip()

    return text

df['full_text'] = df['full_text'].apply(clean_twitter_text)
```

Gambar 3. *Cleaning*

Preprocessing memiliki beberapa tahapan yaitu *cleaning*, *normalisasi*, *stopword removal*, *tokenization*, dan *stemming*. *Normalisasi* adalah proses untuk mengubah teks menjadi format standar yang konsisten dengan melibatkan penghilangan karakter yang tidak diinginkan seperti tanda baca, konversi seluruh teks menjadi huruf kecil untuk menghindari perbedaan besar antara kata yang ditulis dalam huruf besar dan huruf kecil, serta penggantian kata-kata yang memiliki makna serupa atau sinonim dengan kata yang seragam. Misalnya, mengubah "alangke" menjadi “alangkah" atau mengganti singkatan "uts." menjadi "ujian tengah semester". *Stopword removal* adalah kata-kata umum yang sering muncul dalam teks dan tidak memberikan makna khusus dalam proses analisis teks, seperti "di", "dan", "ke", "yang", "tidak", "atau", dan sebagainya. Proses penghapusan *stopword* melibatkan penghapusan kata-kata tersebut dari teks karena mereka tidak membawa informasi penting dalam analisis. Ini membantu mengurangi dimensi data dan meningkatkan efisiensi dalam pemrosesan dan analisis teks. *Tokenization* adalah proses memecah teks menjadi unit-unit

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

yang lebih kecil, yang biasanya berupa kata atau frasa. Ini adalah langkah penting dalam pemrosesan teks karena memungkinkan komputer untuk memahami struktur teks dan menganalisisnya secara lebih efisien. Pada tahap ini, teks dipecah menjadi token-token berdasarkan spasi atau tanda baca, sehingga setiap kata atau frasa dianggap sebagai token terpisah. Misalnya “km”, “12”, “arah”, “mau”, “kantor”, “alfamart”. *Stemming* adalah proses menghilangkan infleksi dari kata-kata untuk menghasilkan bentuk dasar atau akar kata. Ini membantu dalam konsolidasi variasi kata yang memiliki akar yang sama ke dalam satu bentuk, sehingga meningkatkan konsistensi dan efisiensi dalam analisis teks. Misalnya, kata-kata seperti "sebenarnya" akan diubah menjadi bentuk dasarnya "benar", "perbaikan" akan diubah menjadi bentuk dasarnya "baik". Tabel 3 menunjukkan proses konversi kalimat yang terjadi pada media sosial (Twitter, Instagram, dan TikTok) dengan sentimen setelah tahap pra-pemrosesan.

Tabel 3. Data *Preprocessing*

Original *tweet*, postingan, dan kiriman	hati2 buat yg lewat tol dari dan ke Palembang lagi banyak perbaikan.. rada mendingan dibanding Januari tapi tetep masih ada lubang dan bagian jalan yg rusak.. kalo hujan deras bagian pinggir kanan deket pembatas juga banyak titik genangan waspada potensi aquaplaning
Cleaning	hati2 buat yg lewat tol dari dan ke palembang lagi banyak perbaikan rada mendingan dibanding januari tapi tetep masih ada lubang dan bagian jalan yg rusak kalo hujan deras bagian pinggir kanan deket pembatas juga banyak titik genangan waspada potensi aquaplaning
Normalisasi	hati hati buat yang lewat tol palembang banyak perbaikan agak lebih baik dibanding januari tetap ada lubang dan bagian jalan yang rusak kalau hujan deras bagian pinggir kanan dekat pembatas banyak titik genangan waspada potensi aquaplaning
Stopword removal	hati hati buat lewat tol palembang banyak perbaikan lebih baik dibanding januari tetap lubang bagian jalan rusak kalau hujan deras bagian pinggir kanan dekat pembatas banyak titik genangan waspada potensi aquaplaning
Tokenization	['hati', 'hati', 'buat', 'lewat', 'tol', 'palembang', 'banyak', 'perbaikan', 'lebih', 'baik', 'dibanding', 'januari', 'tetap', 'lubang', 'bagian', 'jalan', 'rusak', 'kalau', 'hujan', 'deras', 'bagian', 'pinggir', 'kanan', 'dekat', 'pembatas', 'banyak', 'titik', 'genangan', 'waspada', 'potensi', 'aquaplaning']
Stemming	hati hati buat lewat tol palembang banyak lebih baik banding januari tetap ada lubang dan bagi jalan rusak kalau hujan deras bagi pinggir kanan dekat batas banyak titik genang waspada potensi aquaplaning

Labeling adalah pelabelan yang melibatkan penugasan kategori atau kelas kepada titik data. Dalam pembelajaran mesin berbasis pengawasan, ini adalah proses pemberian keluaran target yang benar kepada data input. Misalnya, dalam tugas klasifikasi, setiap titik data mungkin diberi *label* dengan kelas yang dimilikinya. Pelabelan sangat penting untuk melatih model pembelajaran mesin karena mereka belajar untuk mengasosiasikan fitur input dengan *label* yang sesuai pada Gambar 4. *Labeling manual*, khususnya, adalah metode di mana secara langsung menentukan dan memberikan *label* kepada data. Misalnya, dalam dataset format CSV yang berisi ulasan atau komentar teks, setiap entri mungkin harus dianalisis untuk menentukan sentimen yang tepat, apakah positif, netral, dan negatif. Dalam proses *labeling manual* ini, data yang tidak akurat perlu dikoreksi melalui peninjauan mendalam. Setiap entri dibaca satu per satu, dan konteks serta nuansa emosional dari teks diperhatikan dengan cermat untuk menentukan sentimen yang sebenarnya dapat dilihat pada Gambar 5.

	full_text	translated_text	klasifikasi
0	lagipula jangan bicara tinggi bikin terowongan...	Besides, don't talk high, just make toll tunne...	Negatif
1	gibran tidak lulus uji tengah semester austral...	Gibran did not pass the Australian midterm tes...	Netral
2	marlina baik jalan rusak akibat bangun ipal pr...	Marlina both the road damaged by the building ...	Positif
3	benar rujak gubernur nya sini masalah palemban...	The Governor's Rujak is true, the problem here...	Positif
4	viral aksi protes pria jas das palembang mandi...	Viral Protest Action Men's Suit Palembang Jas ...	Positif
...	...	...	...
663	gila sekali	So crazy	Negatif
664	alangkah macet nyo lurr tengah kota mau gila b...	What a traffic jam nyo lurr in the middle of t...	Negatif
665	ajung tambal mangke dide buhok agi	ajung patch mangke dide buhok agi	Netral
666	apo gwe kamu ngerutuk baed kasih duit kamu cob...	Apo Awe you bluff Based, love your money to vo...	Positif
667	km 12 arah mau kantor alfamart	Km 12 WANT WANT ALFAMART OFFICE	Netral

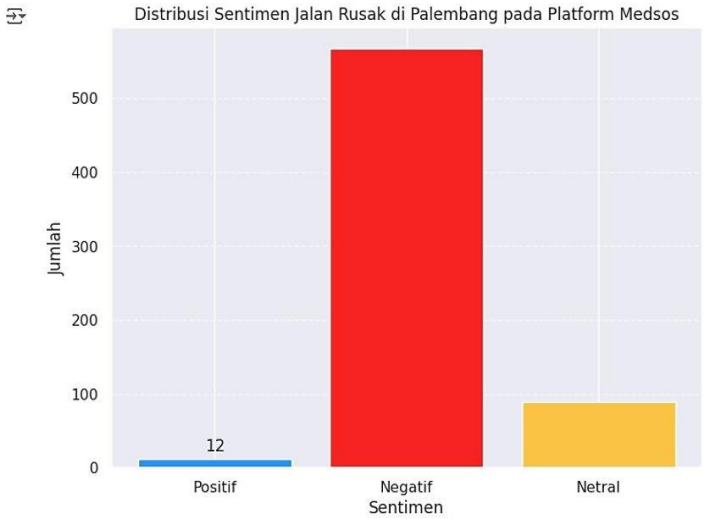
Gambar 4. *Labeling*

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

25	hati hati buat lewat tol palembang banyak lebih baik banding januari tetap ada lubang dan bagi jalan rusak kalau hujan deras bagi pinggir kanan dekat batas banyak titik genang waspada potensi aquaplan	Be careful for those who pass the Palembang toll road, there is a lot of good appeal, there is still a hole and for a damaged road when it rains for the right edge near the limit, a lot	Negatif

Gambar 5. *Labeling Manual*

Visualisasi melibatkan penyajian data secara grafis untuk mendapatkan wawasan atau mengkomunikasikan informasi secara efektif. Ini bisa mencakup pembuatan berbagai jenis plot, grafik, dan diagram untuk menjelajahi pola, hubungan, dan tren dalam data. Visualisasi adalah alat yang sangat kuat untuk memahami kumpulan data yang kompleks dan mengkomunikasikan temuan kepada pemangku kepentingan yang dapat dilihat pada Gambar 6.



Gambar 6. Visualisasi

Data tersebut terbagi menjadi dua bagian: data latih dan data uji. Pengujian data dilakukan untuk setiap algoritma guna memprediksi kelas afektif, menggunakan data latih sebagai referensi. Dengan menggunakan rasio 80:20, 80% dari total data digunakan untuk pelatihan dan 20% untuk pengujian. Data uji ini kemudian diuji dengan algoritma Naïve Bayes untuk memprediksi kelas efektif, menggunakan data latih yang disimpan dengan metode kantong mata. Setelah proses pengujian, nilai akurasi dari algoritma Naïve Bayes pada kelas prediksi menggunakan data uji dianalisis. Akurasi ini mencerminkan seberapa baik model dapat memprediksi kelas afektif berdasarkan data yang telah dilatih. Akurasi yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik untuk mengklasifikasikan sentimen dari teks yang dihasilkan dari Twitter, Instagram, dan TikTok.

Rangkaian proses tersebut, mulai dari pengumpulan data hingga analisis akurasi algoritma, merupakan langkah-langkah penting dalam mengolah dan menganalisis data teks untuk memahami sentimen dari platform sosial seperti Twitter, Instagram, dan TikTok. Dengan demikian, proses ini memberikan pemahaman yang lebih dalam tentang pola, tren, dan sentimen dari data media sosial, yang dapat memberikan wawasan berharga bagi pengambilan keputusan dan pemangku kepentingan.

3.1 Pembahasan

Berdasarkan struktur tabel *Confusion Matrix* pada Tabel 2.1 yang diberikan, terdapat empat nilai yang diwakili oleh TP (True Positive), TN (True Negative), FP (False Positive), dan FN (False Negative). TP merupakan jumlah positive yang benar, di mana model memprediksi kebenaran positif dari suatu data dan data tersebut memang positif. TN adalah jumlah prediksi yang benar di mana model memprediksi kebenaran negatif dari suatu data dan data tersebut memang negatif. FP adalah jumlah prediksi yang salah, di mana model memprediksi kebenaran positif dari suatu data, namun kenyataannya data tersebut negatif. FN adalah jumlah prediksi yang salah, di mana model memprediksi kebenaran

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

negatif dari suatu data, namun kenyataannya data tersebut positif (Luthfi Bangun Permadi & Gumilang, 2024). Untuk mengukur kinerja model, terdapat beberapa parameter yang digunakan, antara lain akurasi, presisi, recall, dan F1 score. Akurasi merupakan nilai rasio prediksi benar dari total data. Presisi adalah perbandingan nilai rasio prediksi benar dengan total nilai yang diprediksi dengan benar (Nugroho & Religia, 2021). Evaluasi kinerja model dilakukan dengan menggunakan nilai-nilai ini.

Hasil klasifikasi Naïve Bayes pada pengujian model terhadap data uji menunjukkan bahwa dari total 230 data yang dievaluasi, terdapat 110 data yang diklasifikasikan dengan benar sebagai positif (True Positive), 105 data diklasifikasikan dengan benar sebagai netral (True Neutral), dan 96 data diklasifikasikan dengan benar sebagai negatif (True Negative). Namun, terdapat 9 data yang salah diklasifikasikan sebagai positif (False Positive) dan 21 data yang salah diklasifikasikan sebagai netral (False Neutral). Tidak ada data yang salah diklasifikasikan sebagai negatif (False Negative). Hasil ini kemudian dapat digunakan untuk menghitung metrik evaluasi seperti akurasi, presisi, dan recall untuk menilai kinerja model klasifikasi Naïve Bayes yang dapat dilihat pada Tabel 4.

Tabel 4. *Confusion Matrix*

Confusion Matrix	Hasil Klasifikasi Naïve Bayes
True Positive	110
False Positive	9
True Neutral	105
False Neutral	21
True Negative	96
False Negative	0

Hasil evaluasi kinerja model klasifikasi Naïve Bayes menunjukkan tingkat akurasi sebesar 91,20%. Selain itu, nilai presisi model adalah 91,20%, yang mengindikasikan seberapa banyak dari dokumen yang diklasifikasikan sebagai positif yang benar-benar positif. Nilai recall, atau juga dikenal sebagai sensitivitas, adalah 92,32%, yang menunjukkan seberapa banyak dari seluruh dokumen yang benar-benar positif yang berhasil terdeteksi oleh model. Selanjutnya, F1 score, yang menggabungkan presisi dan recall, adalah 91,26%. Hasil-hasil ini memberikan gambaran tentang seberapa baik model Naïve Bayes dalam melakukan klasifikasi sentimen terhadap jalan rusak di Palembang berdasarkan data uji yang digunakan dapat dilihat pada Tabel 5.

Tabel 5. Evaluasi

Akurasi	91,20%
Precall	92,32%
Presisi	91,20%
F1 score	91,26%

Selama fase pengujian dan evaluasi ini, model pelatihan dan data uji digunakan. Berdasarkan model pelatihan yang terdiri dari 668 data pelatihan dengan karakteristik sebagai berikut 12 komentar positif, 85 komentar netral, 564 komentar negatif, serta 332 data uji, berikut pada Tabel 6 adalah sampel hasil prediksi algoritma Naïve Bayes.

Tabel 6. Hasil Pegujian dan Evaluasi

Sebelum	Sesudah	Klasifikasi Positif	Klasifikasi Negatif	Klasifikasi Netral	Total
Nilai Prediksi	Positif	110	9	15	125
	Negatif	0	96	0	96
	Netral	0	6	105	111
Total		110	111	120	332

$$\text{Akurasi} = \frac{True\ Positive + True\ Negative}{Total}$$

$$\text{Presisi} = \frac{True\ Positive}{True\ Positive \times False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Tweet, postingan, dan kiriman dilakukan menggunakan *wordcloud. Wordcloud* digunakan untuk memfasilitasi ilustrasi kata dengan menyediakan visualisasi teks yang menarik dan informatif. *Wordcloud* untuk jalan rusak dapat dilihat pada Gambar 7.

TIN: Terapan Informatika Nusantara
Vol 5, No 1, June 2024, page 1-10
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i1.5239

Gambar 7. *Wordcloud*

Sementara itu, hasil analisis sentimen terhadap 668 komentar menunjukkan bahwa mayoritas pandangan (sentimen) masyarakat Indonesia cenderung negatif 88,90%, dan sentimen positif sebesar 1,8%, sedangkan sentimen netral sebesar 13,30% pada periode 16 Maret 2019 hingga 14 Mei 2024 yang ditunjukkan pada Gambar 8. Diagram lingkaran presentase sentimen pada 668 data 12 komentar positif (1,8%), 564 komentar negatif (88,90%), dan 85 komentar netral (13,30%).

Gambar 8. Diagram Lingkaran

4. KESIMPULAN

Dari analisis dan penilaian terhadap dataset dan algoritma yang digunakan dalam ppenelitian ini, yang terdiri dari 668 *tweet*, postingan, dan kiriman. Kesimpulan dapat diambil bahwa pandangan masyarakat Indonesia terhadap kondisi infrastruktur jalan yang memburuk bisa teramati dengan jelas. Khususnya Palembang, keadaan kerusakan menunjukkan dominasi sentimen negatif, dengan penurunan mencapai 84,9%. Selama rentang 16 Maret 2019 hingga 14 Mei 2024, sentimen positif tercatat sebesar 1,8%, sedangkan sentimen netral sebesar 13,3%. Hal ini menunjukkan bahwa mayoritas masyarakat merasa tidak puas atau prihatin dengan kondisi infrastruktur jalan yang memburuk. Penelitian ini melibatkan beberapa tahapan penting, mulai dari pengumpulan data (*scraping*), pra-pemrosesan data (*preprocessing*), pelabelan, pemodelan, berbagi data, hingga klasifikasi menggunakan algoritma Naïve Bayes. Kualitas tahap pra-pemrosesan memegang peran krusial dalam mencapai hasil optimal pada tahap-tahap berikutnya. Dengan memastikan data yang bersih dan terstruktur, proses analisis dapat dilakukan dengan lebih akurat dan efisien. Hasil penelitian ini diharapkan dapat memberikan wawasan yang berguna bagi pemerintah dan pihak terkait dalam merumuskan kebijakan dan langkah-langkah perbaikan yang lebih tepat guna meningkatkan kondisi infrastruktur jalan yang memadai dan memenuhi kebutuhan masyarakat.

REFERENCES

Alrajak, M. S., Ernawati, I., & Nurlaili, I. (2020). Analisis sentimen terhadap Pelayanan PT PLN di Jakarta pada Twitter dengan Algoritma K- Nearest Neighbor (K-NN). *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *1*(2), 110–122.

Anwar, K. (2022). Analisa sentimen Pengguna Instagram Di Indonesia Pada Review Smartphone Menggunakan Naive Bayes. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *2*(4), 148–155. https://doi.org/10.30865/klik.v2i4.315

Arjunanto, M. R., & Waluyo, S. (2023). Analisis Sentimen Pada Media Sosial Twitter Terhadap Infrastruktur Jalan Di

Lampung Menggunakan Algoritma K-Nearest Neighbor. *Prosiding Seminar Nasional …*, *2*(September), 810–817. http://senafti.budiluhur.ac.id/index.php/senafti/article/view/914%0Ahttps://senafti.budiluhur.ac.id/index.php/senafti/article/download/914/557

Furqan, M., Sriani, S., & Sari, S. M. (2022). Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia. *Techno.Com*, *21*(1), 51–60. https://doi.org/10.33633/tc.v21i1.5446

Juniarsih, S., Ripanti, E. F., & Pratama, E. E. (2020). Implementasi Naive Bayes Classifier pada Opinion Mining Berdasarkan Tweets Masyarakat Terkait Kinerja Presiden dalam Aspek Ekonomi. *Jurnal Sistem Dan Teknologi Informasi (Justin)*, *8*(3), 239. https://doi.org/10.26418/justin.v8i3.39118

Khairunnisa, S., Adiwijaya, A., & Faraby, S. Al. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). *Jurnal Media Informatika Budidarma*, *5*(2), 406. https://doi.org/10.30865/mib.v5i2.2835

Luthfi Bangun Permadi, M., & Gumilang, R. (2024). Penerapan Algoritma CNN (Convolutional Neural Network) Untuk Deteksi Dan Klasifikasi Target Militer Berdasarkan Citra Satelit. *Jurnal Sosial Teknologi*, *4*(2), 134–143. https://doi.org/10.59188/jurnalsostech.v4i2.1138

Mas Pintoko, B., & Muslim, K. (2018). Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naïve Bayes Classifier. *E-Proceeding of Engineering*, *5*(3), 8121–8130.

Muhammad Afdal, & Elita, L. R. (2022). *Penerapan Text Mining Pada Aplikasi Tokopedia Menggunakan Algoritma K-Nearest Neighbor*. https://ejournal.uin-suska.ac.id/index.php/RMSI/article/view/16595

Nikmatun, Alvi, I., Waspada, & Indra. (2019). Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Jurnal SIMETRIS*, *10*(2), 421–432.

Nofiyani, N., & Wulandari, W. (2022). Implementasi Electronic Data Processing Untuk meningkatkan Efektifitas dan Efisiensi Pada Text Mining. *Jurnal Media Informatika Budidarma*, *6*(3), 1621. https://doi.org/10.30865/mib.v6i3.4332

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 697–711.

Nugroho, A., & Religia, Y. (2021). Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(3), 504–510. https://doi.org/10.29207/resti.v5i3.3067

Prabowo, W. A., & Wiguna, C. (2021). Sistem Informasi UMKM Bengkel Berbasis Web Menggunakan Metode SCRUM. *Jurnal Media Informatika Budidarma*, *5*(1), 149. https://doi.org/10.30865/mib.v5i1.2604

Priandi, M., & Painem. (2021). Analisis Sentimen Masyarakat Terhadap Pembelajaran Daring di Era Pandemi Covid-19 pada Media Sosial Twitter Menggunakan Ekstraksi Fitur Countvectorizer dan Algoritma K-Nearest Neighbor. *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *2*(2), 311–319.

Rakhmawati, N. A., Waskitho, R. B., Rahman, D. A., & Nuha, M. F. A. U. (2021). Klasterisasi Topik Konten Channel Youtube Gaming Indonesia Menggunakan Latent Dirichlet Allocation. *Journal of Information Engineering and Educational Technology*, *5*(2), 78–83. https://doi.org/10.26740/jieet.v5n2.p78-83

Rayuwati, Husna Gemasih, & Irma Nizar. (2022). IMPLEMENTASI AlGORITMA NAIVE BAYES UNTUK MEMPREDIKSI TINGKAT PENYEBARAN COVID. *Jural Riset Rumpun Ilmu Teknik*, *1*(1), 38–46. https://doi.org/10.55606/jurritek.v1i1.127

Salsabila, F., & Wibowo, A. (2023). *Analisis Sentiment Terhadap Presiden Pada Facebook Dengan Menggunakan Metode Naive Bayes*. *2*(September), 818–825.

Seminar, P., Sains, N., Pada, A., Bayes, T., Na, M., Sentimen, A., Teknik, F., & Wahid, U. (2024). *253 | Fakultas Teknik Universitas Wahid Hasyim*. *April 2008*, 253–258.

Statistik, B. P. (2022). *Badan Pusat Statistik Provinsi Sumatera Selatan*.