Sistemasi: Jurnal Sistem Informasi
Volume 13, Nomor 5, 2024: 2209-2221
ISSN:2302-8149
e-ISSN:2540-9719

Optimasi Ekstraksi Fitur untuk Meningkatkan Akurasi *Naïve Bayes* dalam Analisis Sentimen Objek Wisata Bulukumba

Feature Extraction Optimization to Improve Naïve Bayes Accuracy in Sentiment Analysis of Bulukumba Tourism Objects

[1]Darmawan Setiawan*, [2]Najirah Umar, [3]M. Adnan Nur
[1,2,3]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Handayani Makassar
[1,2,3]Jl. Adyaksa Baru No. 1, Kota Makassar, Sulawesi Selatan, Indonesia
*e-mail: dermawans1442@gmail.com

(***received*:** 6 September 2024, ***revised*:** 7 September 2024, ***accepted*:** 11 September 2024)

Abstrak

Penelitian ini menggunakan media sosial (*Twitter*) dalam penerapan analisis sentimen untuk menentukan tingkat kepuasan masyarakat terhadap objek wisata Bulukumba. Data teks yang tidak terstruktur menjadi tantangan utama dalam analisis sentimen. Untuk itu, implementasi algoritma *Naïve Bayes* merupakan pendekatan efektif dalam mengatasi tantangan ini, karena kemampuannya dalam menangani data teks dengan baik. Penelitian ini bertujuan untuk mengevaluasi kinerja *Multinomial Naïve Bayes* melalui pengujian kombinasi nilai parameter *Minimum Document Frequency* (*min-df*) dan *Maximum Document Frequency* (*max-df*) dalam menentukan tingkat akurasi. Tahapan analisis ini mencakup pengumpulan data dari *Twitter* terkait objek wisata Bulukumba. Prapemrosesan yang dilakukan meliputi pembersihan data, *casefolding*, normalisasi teks, tokenisasi, penghapusan *stopword*, dan *stemming*. Ekstraksi fitur menggunakan *Count Vectorizer* dan pembobotan *TF-IDF*. Proses diakhiri dengan *10-Fold Cross-Validation* dengan memecah data menjadi data latih dan data uji untuk klasifikasi analisis sentimen, serta evaluasi dengan menggunakan *Confusion Matrix*. Dalam penelitian ini terdapat 10 skenario pengujian yang memiliki kombinasi *min-df* dan *max-df* yang berbeda. Nilai *min-df* yang digunakan terdiri dari 0.001, 0.002, 0.005, 0.01, 0.02 dan untuk *max-df* terdiri dari 0.5, dan 0.8. Hasil dari implementasi *Multinomial Naïve Bayes* pada pengujian tersebut menunjukkan bahwa peningkatan akurasi klasifikasi pada pengaturan parameter *min-df* dan *max-df* yang efektif. Akurasi tertinggi sebesar 0.7910 pada pengujian kombinasi nilai parameter *min-df* 0.001 dan *max-df* 0.8. Sementara itu, rata-rata akurasi setiap pengujian didapatkan nilai tertinggi sebesar 0.7272 dengan *min-df* 0.002 serta *max-df* masing-masing 0.5 dan 0.8.

Kata kunci: *naïve bayes*, analisis sentimen, wisata bulukumba, *min-df*, *max-df*

Abstract

This research employs social media (Twitter) to apply sentiment analysis ascertain the degree of public satisfaction with the Bulukumba tourist attraction. Unstructured text data is a major challenge in sentiment analysis. For this reason, implementing the Naïve Bayes algorithm is an effective approach for conquering this challenge because of its ability to handle text data well. This study aims to evaluate the performance of multinomial Naïve Bayes by testing a combination of minimum document frequency (min-df) and maximum document frequency (max-df) parameter values in determining the level of accuracy. This analysis stage includes collecting data from Twitter related to the Bulukumba tourist attraction. Preprocessing carried out includes data cleaning, casefolding, text normalization, tokenization, stopword removal, and stemming. Feature extraction using Count Vectorizer and TF-IDF weighting. The process ends with 10-Fold Cross-Validation by separating the data into training data and test data for sentiment analysis classification, as well as evaluation using the Confusion Matrix. In this research, there are 10 test scenarios with various combinations of min-df and max-df. The values of employed min-df consists of 0.001, 0.002, 0.005, 0.01, 0.02 and max-df consists of 0.5 and 0.8. The results of implementing Multinomial Naïve Bayes in this test show that classification accuracy increases with effective min-df and max-df parameter settings. The greatest accuracy was 0.7910 in testing a combination of min-df parameter values of 0.001 and max-df 0.8.

 ISSN:2302-8149
e-ISSN:2540-9719

Meanwhile, the average accuracy for each test was obtained the highest value of 0.7272 with min-df of 0.002 and max-df of 0.5 and 0.8 respectively.

Keywords: *naïve bayes, sentiment analysis, bulukumba tourism, min-df, max-df*

1 Pendahuluan

Perkembangan teknologi informasi telah mengubah cara pandang masyarakat dalam menyampaikan opini dan pengalaman mereka, terutama melalui media sosial [1]. Media sosial kini menjadi *platform* utama untuk berbagi pengalaman, pendapat, dan kepuasan terhadap produk atau layanan [2]. sehingga menjadi sumber data yang kaya untuk analisis sentimen guna meningkatkan kualitas produk atau layanan. Penerapan ini efektif dalam berbagai bidang termasuk pariwisata. Pariwisata merupakan salah satu sektor utama yang berperan penting dalam mengembangkan perekonomian di Indonesia [3], termasuk Kabupaten Bulukumba yang memiliki potensi wisata alam, budaya, dan sejarah yang besar, sehingga berhasil menarik minat wisatawan lokal maupun wisatawan mancanegara [4]. Kepuasan wisatawan sendiri merupakan faktor kunci dalam pengembangan sektor pariwisata yang tidak hanya mempengaruhi kunjungan ulang tetapi juga reputasi destinasi wisata [5]. Oleh karena itu, pemahaman yang tepat mengenai sentimen wisatawan sangat penting untuk meningkatkan daya tarik dan kualitas layanan wisata, yang pada akhirnya dapat berdampak positif pada jumlah kunjungan wisatawan dan pendapatan daerah.

Dalam pengamatan yang dilakukan oleh peneliti di Dinas Pariwisata Bulukumba, mereka telah memanfaatkan fitur komentar pada media sosial *Instagram* untuk mengevaluasi sentimen wisatawan. Namun, metode ini kurang efisien karena belum menggunakan sistem cerdas untuk mengidentifikasi dan mengklasifikasikan ratusan bahkan ribuan komentar yang sering kali tidak terstruktur. Selain itu, penilaian kepuasan wisatawan secara konvensional melalui survei langsung juga memerlukan waktu dan biaya besar. Hal ini menyebabkan keterlambatan dalam respons dan pengambilan keputusan yang tepat untuk meningkatkan kualitas pengalaman wisatawan. Analisis sentimen dapat dilakukan untuk menentukan sikap atau emosi penulis terhadap topik tertentu, dengan mengidentifikasi dan mengelompokkan opini dalam teks menjadi positif, negatif, atau netral. Proses ini signifikan dalam berbagai aplikasi, mulai dari penilaian produk hingga pemantauan sentimen publik terhadap kebijakan pemerintah [6].

Twitter, sebagai situs mikroblog terkenal, sering digunakan dalam penelitian *text mining*, khususnya analisis sentimen, karena penggunaannya yang luas untuk memposting pembaruan terkait berbagai topik dalam bentuk *tweet* [7][8].

Analisis sentimen dari *tweet* wisatawan di *Twitter* dapat menggantikan atau melengkapi metode survei konvensional [9]. Tantangan utama dalam penelitian seperti ini biasanya terdapat pada proses mengolah dan menganalisis data teks tidak terstruktur secara efisien dan akurat. Implementasi algoritma *Naïve Bayes* dalam analisis sentimen merupakan pendekatan efektif untuk mengatasi tantangan ini karena kemampuannya dalam menangani data teks dengan baik [10]. Kemudian, algoritma ini dapat mengidentifikasi sentimen positif, negatif, atau netral yang diekspresikan oleh masyarakat maupun wisatawan, serta faktor-faktor yang mempengaruhi kepuasan mereka. Sehingga memungkinkan respons yang lebih cepat dan tepat sasaran untuk perbaikan layanan.

Naïve Bayes memiliki beberapa varian, salah satunya adalah *Multinomial Naïve Bayes* yang sering diterapkan dalam analisis sentimen karena kesederhanaan dan efektivitasnya. Metode ini mengasumsikan bahwa fitur-fitur (kata-kata) dalam dokumen bersifat independen dan menghitung probabilitas kelas berdasarkan distribusi kata [11]. Tezgider et al (2022) menjelaskan bahwa *Multinomial Naïve Bayes* sangat efektif dalam mengklasifikasikan teks dan sering digunakan dalam analisis sentimen karena kemampuannya dalam menangani data teks yang besar dan tidak seimbang dengan efisien [12]. Hasil penelitian oleh Umar & M. Adnan Nur, (2022) juga menjelaskan bahwa *Multinomial Naïve Bayes* memiliki akurasi lebih tinggi di antara tiga variasi *Naïve Bayes* lainnya dengan akurasi sebesar 63,74% [13]. Namun, nilai akurasi yang dihasilkan masih belum optimal untuk aplikasi yang membutuhkan tingkat akurasi yang lebih tinggi. Salah satu cara untuk meningkatkan akurasi tersebut adalah dengan mengoptimalkan proses ekstraksi fitur, yaitu tahapan untuk menentukan fitur-fitur teks yang paling relevan dalam analisis sentimen.

Proses ekstraksi fitur yang efektif sangat penting karena kualitas fitur yang diekstraksi secara langsung mempengaruhi performa algoritma analisis sentimen. Parameter seperti *Minimum Document*

Frequency (*min-df*) dan *Maximum Document Frequency* (*max-df*) dalam algoritma *Multinomial Naïve Bayes* berperan penting dalam menentukan seberapa sering suatu kata muncul di dalam dokumen dan seberapa relevan kata tersebut sebagai fitur untuk klasifikasi sentimen [14]. Optimalisasi parameter ini dapat meningkatkan kemampuan algoritma dalam mengidentifikasi sentimen secara lebih akurat dan efisien.

Penelitian ini bertujuan untuk mengoptimalkan kinerja algoritma *Multinomial Naïve Bayes* dalam analisis sentimen wisatawan terhadap objek wisata di Bulukumba dengan melakukan optimasi pada tahap ekstraksi fitur. Fokus utama penelitian ini adalah mengevaluasi kombinasi nilai parameter *min-df* dan *max-df* untuk mencapai akurasi yang lebih tinggi. Dengan optimasi ini, diharapkan hasil analisis sentimen menjadi lebih akurat dan dapat digunakan oleh pemangku kepentingan di Dinas Pariwisata Bulukumba untuk meningkatkan strategi pemasaran, kebijakan, dan kualitas layanan wisata secara lebih efektif. Dengan demikian, penelitian ini tidak hanya berkontribusi pada pengembangan metode analisis sentimen yang lebih baik untuk sektor pariwisata, tetapi juga menawarkan pendekatan praktis yang dapat diadopsi oleh berbagai destinasi wisata lainnya untuk meningkatkan pengalaman wisatawan berdasarkan data sentimen yang dihasilkan dari media sosial.

2 Tinjauan Literatur

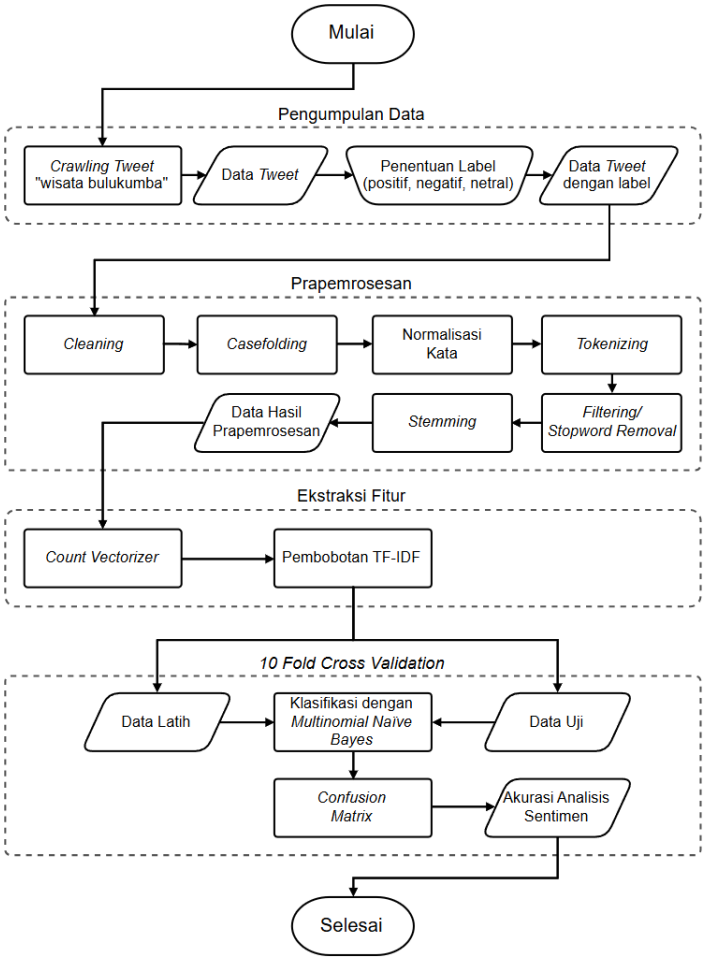
Berbagai penelitian telah dilakukan dalam menilai sentimen publik melalui berbagai *platform*. Penelitian [15] melakukan pengujian pada kombinasi *Multinomial Naïve Bayes* (MNB) dan *Synthetic Minority Oversampling Technique* (SMOTE) untuk menganalisis emosi cuitan di *Twitter* dengan rata-rata akurasi 65%, lebih tinggi 1% jika dibanding dengan tidak menggunakan SMOTE. Demikian pula pada penelitian [16] yang menganalisis sentimen masyarakat di *Twitter* terhadap calon presiden Indonesia 2024 menggunakan algoritma *Naïve Bayes Classifier* dan *Lexicon Based* pada empat skenario pengujian dengan hasil akurasi berturut-turut sebesar 68%, 67%, 70%, dan 71%. Penelitian lain yang berfokus pada kepuasan wisatawan seperti penelitian [17] membandingkan sentimen pada aplikasi *Traveloka* dan *Tiket.com* ditinjau dari harga dan layanan. Hasilnya, *Traveloka* mendapatkan lebih banyak sentimen positif mencapai 97,2% sedangkan *Tiket.com* hanya 46,9%. Penelitian [18] mengenai kepuasan wisatawan Pintu Kota Ambon menggunakan *tools RapidMiner* menunjukkan akurasi negatif 85,42% dan akurasi positif 97,22%, dengan nilai akurasi total 90,65%. Serta penelitian [19] menggunakan *Naïve Bayes Classifier* (NBC) dan SMOTE *Upsampling* terhadap ulasan pengunjung Candi Borobudur di *TripAdvisor* menunjukkan nilai akurasi sebesar 96,36%.

Berdasarkan dari penjelasan diatas, penelitian terdahulu berfokus pada kombinasi pengujian algoritma *Naïve Bayes* dengan metode lain dalam meningkatkan akurasi dan/atau berfokus pada perbandingan tingkat kepuasan wisatawan terhadap layanan *travel online*, *platform TripAdvisor*, serta *RapidMiner* dengan memperhatikan tingkat akurasi sentimen positif. Berbeda dengan penelitian terdahulu, dalam penelitian ini dilakukan evaluasi kinerja *Multinomial Naïve Bayes* melalui pengujian kombinasi nilai parameter *minimum document frequency* (*min-df*) dan *maximum document frequency* (*max-df*) dalam menentukan tingkat akurasi. Validasi akurasi dilakukan menggunakan *10-fold cross-validation* pada data latih dan data uji. Hasil penelitian ini diharapkan memberikan referensi tambahan dalam menentukan nilai *min-df* dan *max-df* yang optimal untuk mencapai akurasi tinggi dalam analisis sentimen.

3 Metode Penelitian

Dalam penelitian yang dilakukan melewati beberapa tahapan mulai dari pengumpulan data, prapemrosesan hingga implementasi algoritma *Multinomial Naïve Bayes*. Adapun tahapan-tahapan tersebut dapat dijabarkan dalam diagram alir pada metode penelitian berikut.

ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. Diagram alir metode penelitian

3.1 Pengumpulan Data

Studi ini menggunakan dataset *Twitter* pada kata kunci wisata Bulukumba dan beberapa nama objek wisata seperti pantai Bira, tebing Apparalang, hingga adat Ammatoa Kajang untuk dataset yang lebih spesifik. Dataset yang telah dikumpulkan akan digunakan untuk implementasi algoritma *Naïve Bayes* dalam melakukan analisis sentimen terhadap objek wisata Bulukumba. Untuk mendapatkan data yang diinginkan dilakukan beberapa tahapan, dimulai dengan membuat kode program *crawling tweet* pada *Google Colab* menggunakan bahasa pemrograman *python* dengan memanfaatkan *tweet harvest* versi 2.6.1. *Tweet-harvest* adalah *command-line tool* yang dijalankan menggunakan *Node.js* sehingga efisien dan mudah digunakan untuk melakukan *crawling* data *Twitter*. Alat ini memanfaatkan API *Twitter* untuk mengumpulkan data *tweet* berdasarkan kata kunci tertentu dan menyimpannya dalam format yang mudah diolah untuk analisis lebih lanjut [20]. Perhatikan Gambar 2 dan Gambar 3 berikut.

```
# Crawling Data Tweet

filename = 'wisata-bulukumba.csv'
search_keyword = "wisata bulukumba since:2020-01-01 lang:id"

!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

Gambar 2. Kode program *crawling tweet* pada *google colab*

 ISSN:2302-8149
e-ISSN:2540-9719

	date	time	username	tweet
0	"Sat Dec 03, 2022"	00:02:41	ramawahyudhan	Terdapat surga kecil yang terletak di Desa Kin...
1	"Mon Nov 14, 2022"	10:46:22	Rahman80425328	@sugihidetoshi2 Air terjun bravo 45 bukan?
2	"Thu Jul 28, 2022"	10:21:44	BulukumbaPunya	Air Terjun Bravo 45 Bulukumba https://t.co/SIK...
3	"Sun Aug 23, 2020"	06:13:37	RizkyNs	Kemaren sempet jalan ke Pemandian Bravo 45 di ...
4	"Tue Jan 21, 2020"	17:03:16	cikaliyamato	Air terjun bravo 45 Jngn mi di subreker Cukup ...
...	...	...	...	...
1337	"Wed Jun 08, 2022"	03:09:31	mediumID	Cekcok di Dalam Cafe Wisata Bira Bulukumba War...
1338	"Wed Jun 08, 2022"	00:24:06	Ponggolwatu	Cafe Sawah Baturapa Bulukumba Obyek Wisata Mul...
1339	"Wed Jun 08, 2022"	00:00:32	IniPastiDotCom	Cafe Sawah Baturapa Bulukumba Obyek Wisata Mul...
1340	"Sat May 28, 2022"	03:17:09	Turisiancom	6 Tempat Wisata Bulukumba Wajib Masuk Daftar L...
1341	"Sun May 22, 2022"	10:13:49	Henik_yuli	@m3lodyku @hemaviton99 @Muhayya__ @1yudhiasmar...

1342 rows × 5 columns

Gambar 3. Hasil *crawling tweet* dari *API twitter*

Selanjutnya, kode program *crawling* ini digunakan dalam mengumpulkan *dataset tweet* dengan menggunakan kata kunci “wisata bulukumba”. Dari tahun 2020 hingga 2024 terdapat 1.342 *tweet* yang diambil. Untuk klasifikasi analisis sentimen, setiap *tweet* dalam dataset ini dilabeli positif sebanyak 871, netral 382, dan negatif 89, sesuai pada Gambar 4 berikut.

	date	time	username	tweet	label
0	Sat Dec 03, 2022	00:02:41	ramawahyudhan	Terdapat surga kecil yang terletak di Desa Kin...	1
1	Mon Nov 14, 2022	10:46:22	Rahman80425328	@sugihidetoshi2 Air terjun bravo 45 bukan?	0
2	Thu Jul 28, 2022	10:21:44	BulukumbaPunya	Air Terjun Bravo 45 Bulukumba https://t.co/SIK...	0
3	Sun Aug 23, 2020	06:13:37	RizkyNs	Kemaren sempet jalan ke Pemandian Bravo 45 di ...	1
4	Tue Jan 21, 2020	17:03:16	cikaliyamato	Air terjun bravo 45 Jngn mi di subreker Cukup ...	1
...	...	...	...	...	...
1337	Wed Jun 08, 2022	03:09:31	mediumID	Cekcok di Dalam Cafe Wisata Bira Bulukumba War...	-1
1338	Wed Jun 08, 2022	00:24:06	Ponggolwatu	Cafe Sawah Baturapa Bulukumba Obyek Wisata Mul...	1
1339	Wed Jun 08, 2022	00:00:32	IniPastiDotCom	Cafe Sawah Baturapa Bulukumba Obyek Wisata Mul...	1
1340	Sat May 28, 2022	03:17:09	Turisiancom	6 Tempat Wisata Bulukumba Wajib Masuk Daftar L...	1
1341	Sun May 22, 2022	10:13:49	Henik_yuli	@m3lodyku @hemaviton99 @Muhayya__ @1yudhiasmar...	0

1342 rows × 5 columns

Gambar 4. Data *tweet* dengan label positif, netral, dan negatif

3.2 Prapemrosesan Data

Prapemrosesan merupakan tahapan awal krusial dalam analisis sentimen yang bertujuan untuk meningkatkan kualitas data dan hasil analisis. Tahapan ini mencakup pembersihan data, tokenisasi, penghapusan *stopwords*, hingga *stemming* atau *lemmatisasi*. Prapemrosesan dapat mengurangi *noise* dalam data teks sehingga model dapat fokus pada informasi yang relevan [21]. Prapemrosesan yang efektif secara signifikan dapat meningkatkan akurasi model analisis sentimen [22].

3.2.1 *Cleaning*

Cleaning adalah langkah awal dalam prapemrosesan data teks yang bertujuan menghapus atau memperbaiki teks yang tidak relevan atau bising. Tindakan ini meliputi penghapusan alamat *website* dengan awalan *http://* atau *https://* atau *www*, karakter khusus seperti *tag HTML*, emoji, dan simbol, *tweet* dengan angka, serta *username* @ dan tagar # [23]. Proses ini untuk memastikan bahwa data yang akan dianalisis bersih dan konsisten serta mengoptimalkan waktu dalam proses klasifikasi. Dengan memanfaatkan *library re* (*regular expressions*) untuk mencocokkan pola dalam teks dan memodifikasi *string* berdasarkan pola tersebut.

3.2.2 *Case Folding*

Data *tweet* sering kali tidak konsisten dalam penggunaan huruf besar dan kecil sehingga kata seperti "Wisata" dan "wisata" dianggap berbeda. Untuk itu dilakukan *Casefolding* agar mengganti seluruh karakter dalam teks menjadi huruf-huruf kecil (*lowercase*) dengan mengaplikasikan fungsi *lower()* yang telah ada dalam bahasa pemrograman *python* [24]. Langkah ini penting untuk mengurangi keragaman data dan membantu menyederhanakan proses analisis dengan memastikan bahwa semua teks diperlakukan secara konsisten.

 ISSN:2302-8149
e-ISSN:2540-9719

3.2.3 Normalisasi Kata

Selanjutnya dilakukan normalisasi kata yang merupakan standarisasi teks untuk menghilangkan variasi yang tidak relevan. Ini termasuk penanganan singkatan, penulisan yang salah, dan bentuk kata yang berbeda menjadi bentuk standar [25]. Berdasarkan data *tweet* pengujian, diambil 592 contoh kata tidak relevan untuk diubah ke dalam kata-kata standar dengan menggunakan fungsi *replace* dari *library os* pada pemrograman *python*.

3.2.4 *Tokenizing*

Tahap selanjutnya yaitu memecah teks menjadi bagian yang lebih kecil yang dikenal sebagai token, biasanya terdiri dari kata atau pun frasa. Sehingga teks menjadi komponen yang dapat dianalisis secara individual [26]. Hasil *tokenizing* dalam bahasa pemrograman *python* dapat diletakkan dalam struktur data *list* untuk memudahkan tahapan prapemrosesan selanjutnya yang berorientasi pada kata.

3.2.5 *Filtering/Stopword Removal*

Filtering atau *Stopword Removal* adalah proses meniadakan kata-kata *general* yang dominan dalam analisis sentimen, tetapi itu tidak memberikan keterangan yang krusial. Menghapus *stopword* dapat meningkatkan akurasi dan efisiensi analisis dengan mengurangi jumlah data yang tidak relevan [27]. Dengan menggunakan modul *stopwords* bahasa Indonesia dari *library NLTK*, serta menerapkan fungsi *remove stopwords* untuk mengimplementasikan penghapusan *stopword* pada tahap prapemrosesan.

3.2.6 *Stemming*

Stemming bertujuan untuk mengubah kata-kata kebentuk dasarnya (*stem*). Teknik ini membantu dalam mengurangi variasi kata yang berbeda menjadi bentuk dasar yang sama, sehingga dapat meningkatkan akurasi dari klasifikasi [28]. Kelas *StemmerFactory* dari *library Sastrawi* digunakan untuk membuat objek *stemmer* yang akan melakukan proses *stemming*.

3.3 Ekstraksi Fitur

Setelah melalui semua tahapan prapemrosesan, data *tweet* menjadi lebih bersih (*cleaned_tweet*), konsisten, dan siap untuk dianalisis. Dalam analisis sentimen, teks yang telah dipraproses akan diekstraksi fitur-fiturnya untuk kemudian diklasifikasikan sebagai sentimen positif, negatif, atau netral. Dua teknik populer untuk ekstraksi fitur dalam analisis sentimen adalah *Count Vectorizer* dan pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) [29].

3.3.1 *Count Vectorizer*

Count Vectorizer adalah teknik untuk mengubah teks menjadi matriks istilah-dokumen (*term-document matrix*). Setiap baris dalam matriks mewakili sebuah dokumen, dan setiap kolom mewakili sebuah istilah (kata) dari seluruh korpus. Nilai di setiap sel adalah jumlah kemunculan kata dalam dokumen tersebut [14]. Parameter minimum dan maksimum *document frequency* (*min-df* dan *max-df*) digunakan pada *count vectorizer* untuk memfilter kata-kata yang muncul terlalu jarang atau terlalu sering dalam korpus, penentuan parameter ini akan mempengaruhi akurasi klasifikasi [13]. Pada penelitian ini dilakukan pengujian dengan beberapa nilai *min-df* dan *max-df* untuk melihat perbedaan akurasi analisis sentimen. Nilai *min-df* yang akan digunakan adalah 0.001 (0.1%) hingga 0.02 (2%) sedangkan *max-df* yaitu 0.5 (50%) dan 0.8 (80%).

3.3.2 *Term Frequency-Inverse Document Frequency (TF-IDF)*

Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode yang bertujuan memberikan nilai pada kata-kata yang sering digunakan. Teknik ini membantu dalam menentukan kepentingan suatu kata dalam dokumen yang relatif pada keseluruhan korpus dokumen. *Term Frequency* (TF) mengukur seberapa sering sebuah terminologi ditemukan pada dokumen. Sedangkan *Inverse Document Frequency* (IDF) mengukur kepentingan suatu kata berdasarkan seberapa jarang kata tersebut ada pada keseluruhan dokumen dalam korpus [30]. Kombinasi ini memberi bobot tertinggi pada kata-kata yang unik pada setiap dokumen dan bobot terendah pada kata-kata yang sangat umum di seluruh korpus. TF-IDF membantu dalam meningkatkan akurasi model analisis sentimen [31]. Pada penelitian ini, baik *count vectorizer* maupun implementasi TF-IDF keduanya menggunakan kelas *CountVectorizer* dan *TfidfTransformer* dari *library python sklearn*.

 ISSN:2302-8149
e-ISSN:2540-9719

3.4 *10-Fold Cross Validation*

Dataset berjumlah 1.342 *tweet* yang telah melalui tahap pra-pemrosesan dan ekstraksi fitur digunakan sebagai sumber data utama penelitian ini. Untuk mengetahui seberapa akurat model klasifikasi yang digunakan, maka diperlukan metode validasi pengujian yaitu *K-fold cross-validation* yang dapat memecah *dataset* ke dalam data *training* dan data *testing* [32]. Penelitian ini menetapkan nilai K=10 untuk mengelompokkan dan melakukan iterasi *training* dan *testing* secara bergantian sebanyak K [33]. Metode ini memberikan estimasi yang lebih stabil dan akurat, sekaligus mengurangi risiko *overfitting*. Untuk setiap iterasi, data kemudian diklasifikasikan dengan menggunakan algoritma *multinomial naïve bayes*, dikarenakan itu adalah salah satu varian *naïve bayes* dengan tingkat akurasi yang tinggi dibanding dengan varian *naïve bayes* lainnya [13]. Kemudian, data latih digunakan dalam membuat model klasifikasi *naïve bayes*. Pelatihan dan pengujian data dilakukan menggunakan *10-fold cross validation* pada beberapa kombinasi TF-IDF (*min-df* dan *max-df*) untuk mendapatkan parameter TF-IDF yang optimal serta mengetahui performa dan akurasi dari *multinomial naïve bayes*.

Setelah model klasifikasi dibuat, data uji tersebut dipergunakan dalam menghitung tingkat akurasinya. Hasil pengujian dari setiap iterasi *10-fold cross validation* pada variasi *min-df* dan *max-df* kemudian dievaluasi. *Confusion matrix* merupakan metode yang dapat diaplikasikan dalam evaluasi model pada penelitian ini untuk mendapatkan nilai akurasi pada setiap *fold* [34].

4 Hasil dan Pembahasan

Dataset yang telah melalui tahap prapemrosesan, ekstraksi fitur hingga *10-fold cross validation*, itu kemudian akan diuji dengan menerapkan algoritma *multinomial naïve bayes*. Dalam proses pengujian tersebut terdapat 10 *dataset* yang memiliki kombinasi *min-df* dan *max-df* yang berbeda. Nilai *min-df* yang digunakan terdiri dari: 0.001, 0.002, 0.005, 0.01, 0.02. Sedangkan untuk *max-df* terdiri dari: 0.5, dan 0.8. Untuk penerapan masing-masing nilai *min-df* dan *max-df* tersebut menggunakan *library sklearn python*. Berikut ini adalah hasil dan pembahasan dari Implementasi *multinomial naïve bayes* pada sentimen objek wisata Bulukumba.

4.1 Pengujian dengan *min-df* 0.001 dan *max-df* 0.5

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.001 dan 0.5 pada data *tweet* wisata dapat ditunjukkan pada Tabel 1 berikut.

Tabel 1. Pengujian *min-df* 0.001 dan *max-df* 0.5

	10-Fold Cross Validation				
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.7111	0.7185	0.6716	**0.7835**	**0.6641**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.7164	0.7761	0.7089	0.7238	0.7238

Rata-rata Akurasi: 0.7198

Dalam sepuluh percobaan yang dilakukan, kombinasi pertama menunjukkan akurasi tertinggi mencapai 0.7835 pada *fold-4*, sedangkan akurasi terendah tercatat dalam *fold-5* dengan akurasi 0.6641.

4.2 Pengujian dengan *min-df 0.001* dan *max-df* 0.8

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.001 dan 0.8 pada data *tweet* wisata dapat ditunjukkan pada Tabel 2 berikut.

Tabel 2. Pengujian *min-df* 0.001 dan *max-df* 0.8

	10-Fold Cross Validation				
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.7037	0.7185	0.6865	**0.7910**	**0.6641**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.7164	0.7761	0.7089	0.7238	0.7238

Rata-rata Akurasi: 0.7213

 ISSN:2302-8149
e-ISSN:2540-9719

Pada pengujian kombinasi kedua akurasi tertinggi mencapai 0.7910, sementara akurasi terendah tercatat 0.6641. Hasil ini konsisten dengan skenario sebelumnya, dengan akurasi tertinggi dan terendah masing-masing berada pada *fold-4* dan *fold-5*.

4.3 Pengujian dengan *min-df* 0.002 dan *max-df* 0.5

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.002 dan 0.5 pada data *tweet* wisata dapat ditunjukkan pada Tabel 3 berikut.

Tabel 3. Pengujian *min-df* 0.002 dan *max-df* 0.5

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.7333	0.7185	0.7238	**0.7761**	**0.6567**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.7014	0.7388	0.7164	0.7611	0.7462

Rata-rata Akurasi: 0.7272

Pada pengujian ini, akurasi tertinggi mencapai 0.7761 di *fold-4*, sedangkan akurasi terendah tercatat sebanyak 0.6567 pada *fold-5*. Pola distribusi akurasi ini konsisten dengan hasil pada *fold* di kedua pengujian sebelumnya.

4.4 Pengujian dengan *min-df* 0.002 dan *max-df* 0.8

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.002 dan 0.8 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 4 berikut.

Tabel 4. Pengujian *min-df* 0.002 dan *max-df* 0.8

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.7333	0.7185	0.7238	**0.7761**	**0.6641**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.6940	0.7388	0.7089	0.7686	0.7462

Rata-rata Akurasi: 0.7272

Pada pengujian kombinasi keempat, akurasi tertinggi yang diperoleh adalah 0.7761, sementara akurasi terendah tercatat sebesar 0.6641. Posisi hasil akurasi tertinggi dan terendah tetap konsisten dengan pengujian sebelumnya, yakni pada *fold-4* dan *fold-5* secara berurutan.

4.5 Pengujian dengan *min-df* 0.005 dan *max-df* 0.5

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.005 dan 0.5 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 5 berikut.

Tabel 5. Pengujian *min-df* 0.005 dan *max-df* 0.5

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.7111	0.7037	0.7089	0.7462	**0.6716**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.7089	0.7537	0.7164	0.7313	**0.7611**

Rata-rata Akurasi: 0.7213

Pada kombinasi ini, akurasi tertinggi diperoleh pada *fold-10* dengan nilai akurasi mencapai 0.7611, sementara akurasi terendah dicapai pada *fold-5* dengan nilai 0.6716, konsisten dengan hasil pada skenario lainnya.

4.6 Pengujian dengan *min-df* 0.005 dan *max-df* 0.8

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.005 dan 0.8 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 6 berikut.

Tabel 6. Pengujian *min-df* 0.005 dan *max-df* 0.8

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.7037	0.7037	0.7089	0.7462	**0.6716**

 ISSN:2302-8149
e-ISSN:2540-9719

	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.7014	0.7537	0.7164	0.7313	**0.7686**

Rata-rata Akurasi: 0.7205

Akurasi tertinggi mencapai 0.7686 di *fold-10*, sementara akurasi terendah tercatat sejumlah 0.6716 yang juga terdapat pada *fold-5*.

4.7 Pengujian dengan *min-df* 0.01 dan *max-df* 0.5

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.01 dan 0.5 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 7 berikut.

Tabel 7. Pengujian *min-df* 0.01 dan *max-df* 0.5

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.6814	0.7185	0.7089	**0.7313**	**0.6567**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.6865	0.6865	0.7089	0.7089	0.7164

Rata-rata Akurasi: 0.7004

Pada kombinasi ini, nilai akurasi tertinggi dicapai pada *fold* ke-4 dengan nilai 0.7313. Sebaliknya, nilai akurasi terendah terjadi pada *fold* ke-5, dengan nilai 0.6567, sama seperti pada skenario lainnya.

4.8 Pengujian dengan *min-df* 0.01 dan *max-df* 0.8

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.01 dan 0.8 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 8 berikut.

Tabel 8. Pengujian *min-df* 0.01 dan *max-df* 0.8

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.6888	0.7185	0.7014	**0.7313**	**0.6641**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.6791	0.6865	0.7089	0.7089	0.7238

Rata-rata Akurasi: 0.7011

Akurasi tertinggi yang dicapai dalam pengujian ini adalah 0.7313, yang diperoleh pada *fold* ke-4. Sebaliknya, akurasi terendah tercatat pada *fold* ke-5, yaitu sebesar 0.6641. Distribusi hasil ini konsisten dengan hasil dari empat pengujian awal.

4.9 Pengujian dengan *min-df* 0.02 dan *max-df* 0.5

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.02 dan 0.5 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 9 berikut.

Tabel 9. Pengujian *min-df* 0.02 dan *max-df* 0.5

10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.6518	0.6518	0.6417	**0.7089**	**0.6343**
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.6567	0.6716	0.6865	0.6492	0.7014

Rata-rata Akurasi: 0.6654

Masih sama dengan pengujian sebelumnya dimana nilai akurasi tertinggi diperoleh pada *fold-4* sebesar 0.7089. Sedangkan akurasi terendah sebesar 0.6343 pada *fold-5*.

4.10 Pengujian dengan *min-df* 0.02 dan *max-df* 0.8

Hasil penerapan kombinasi *min-df* dan *max-df* masing-masing dengan nilai 0.02 dan 0.8 pada data *tweet* wisata dapat ditunjukkan dalam Tabel 10 berikut.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 10. Pengujian *min-df* 0.02 dan *max-df* 0.8

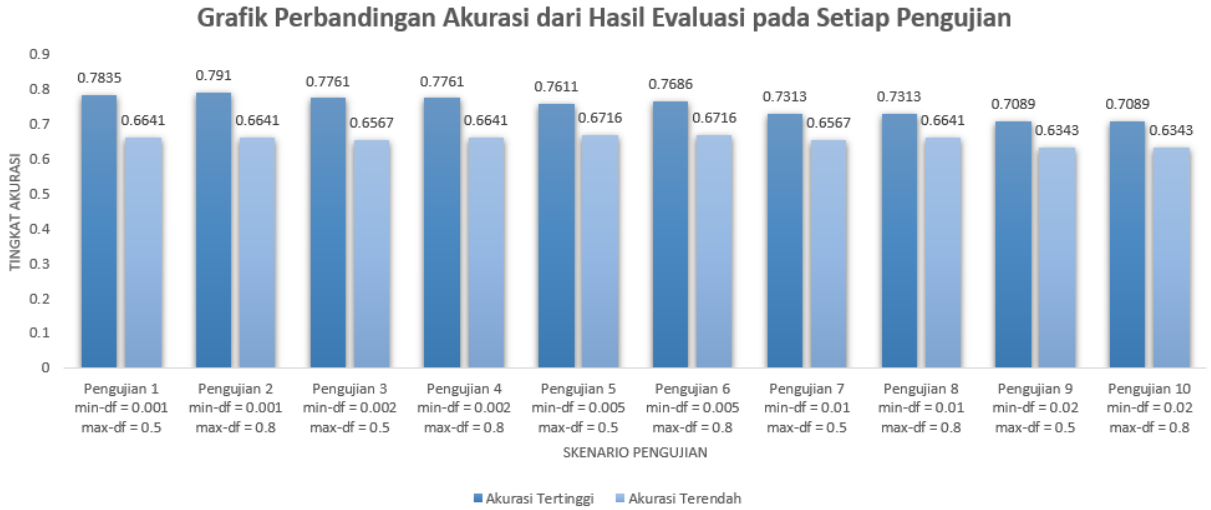
10-Fold Cross Validation					
Akurasi	*fold-1*	*fold-2*	*fold-3*	*fold-4*	*fold-5*
	0.6518	0.6518	0.6492	**0.7089**	0.6417
	fold-6	*fold-7*	*fold-8*	*fold-9*	*fold-10*
	0.6716	0.6791	0.6940	**0.6343**	**0.7089**

Rata-rata Akurasi: 0.6691

Pada pengujian terakhir, terdapat perbedaan signifikan dibandingkan dengan pengujian sebelumnya. Akurasi tertinggi tercatat sebesar 0.7089 pada dua *fold* berbeda, yaitu *fold-4* dan *fold-10*. Sebaliknya, akurasi terendah tercatat sebesar 0.6343 pada *fold-9*, menunjukkan posisi yang cukup berbeda dibandingkan dengan hasil pengujian sebelumnya.

4.11 Perbandingan Akurasi pada Setiap Pengujian

Berdasarkan 10 pengujian kombinasi nilai parameter *min-df* dan *max-df*, itu menunjukkan bahwa akurasi tertinggi secara konsisten dicapai pada *fold-4* dan *fold-10*, dengan delapan pengujian di *fold-4* dan tiga pengujian di *fold-10*, sementara pengujian terakhir menunjukkan akurasi serupa di keduanya. Akurasi terendah secara dominan muncul pada sembilan pengujian di *fold-5*, sementara pengujian terakhir akurasi terendah ada di *fold-9*. Untuk nilai akurasi tertinggi mencapai 0.7910 pada pengujian kedua dengan konfigurasi *min-df* 0.001 dan *max-df* 0.8 sementara akurasi terendah sebesar 0.6343 pada pengujian kesembilan dan kesepuluh dengan *min-df* 0.02 serta *max-df* masing-masing 0.5 dan 0.8. Lebih lanjut, untuk rata-rata akurasi pada tiap pengujian diperoleh nilai tertinggi pada pengujian ke-3 dan ke-4 yang mencapai 0.7272 dengan *min-df* 0.002 dan *max-df* yang masing-masing 0.5 dan 0.8. Hasil tersebut menunjukkan bahwa pengujian pada kombinasi nilai parameter *min-df* dan *max-df* yang berbeda dapat memberikan pengaruh yang signifikan terhadap tingkat akurasi analisis sentimen. Grafik berikut menunjukkan perbandingan akurasi tertinggi dan terendah pada setiap pengujian, yang dihasilkan dari penentuan nilai parameter *min-df* dan *max-df*.



Gambar 5. Grafik perbandingan akurasi pada setiap pengujian

5 Kesimpulan

Dalam implementasi *Multinomial Naïve Bayes* pada data *tweet* wisata kabupaten Bulukumba, kombinasi pengaturan nilai parameter *min-df* dan *max-df* menunjukkan bahwa nilai akurasi tertinggi mencapai 0.7910 didapatkan pada pengujian kedua dengan *min-df* 0.001 dan *max-df* 0.8. Sedangkan, rata-rata akurasi tiap pengujian didapatkan nilai tertinggi sebesar 0.7272 diperoleh pada pengujian ketiga dan keempat dengan *min-df* 0.002 serta *max-df* masing-masing 0.5 dan 0.8. Hal ini menunjukkan bahwa pengaturan *min-df* dan *max-df* yang efektif sangat berpengaruh pada tingkat akurasi klasifikasi analisis sentimen. Sehingga dapat disimpulkan bahwa dalam penentuan *min-df* dan *max-df* yang tepat mampu meningkatkan akurasi klasifikasi.

 ISSN:2302-8149
e-ISSN:2540-9719

Referensi

[1] S. H. Putri and L. O. Maharani, "Penggunaan Media Sosial Twitter @Txtdari Pemerintah Sebagai Saluran Penyebaran Berita Dalam Membentuk Opini Publik," *J. Komun. dan Desain*, vol. 04, no. 02, pp. 79–88, 2021.

[2] S. Kumar and M. S. Shah, "Social Media and Its Impact on Consumers Behaviour," *Int. J. Multidiscip. Res.*, vol. 5, no. 2, pp. 1–9, 2023, doi: 10.36948/ijfmr.2023.v05i02.2252.

[3] N. Aida, G. Atiqasani, and W. A. Palupi, "The Effect of the Tourism Sector on Economic Growth in Indonesia," *Wseas Trans. Bus. Econ.*, vol. 21, pp. 1158–1166, 2024, doi: 10.37394/23207.2024.21.95.

[4] K. RI, "7 Destinasi Wisata Unggulan di Bulukumba yang Wajib Dikunjungi," 2021. https://kemenparekraf.go.id/hasil-pencarian/7-destinasi-wisata-unggulan-di-bulukumba-yang-wajib-dikunjungi

[5] M. A. Rahim, N. A. Bakar, N. A. A. N. Hashim, N. M. M. Nawi, and H. Wee, "Empirical Evidence From the Tourism Industry on the Factors That Affect Tourist Destination Satisfaction," *Geoj. Tour. Geosites* , vol. 44, no. 4, pp. 1209–1215, 2022, doi: 10.30892/gtg.44404-936.

[6] R. Kora and A. Mohammed, "An enhanced approach for sentiment analysis based on meta-ensemble deep learning," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, pp. 1–13, 2023, doi: 10.1007/s13278-023-01043-6.

[7] P. Agarwal, "Developing an Approach to Evaluate and Observe Sentiments of Tweets," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 5, no. 3, pp. 473–479, 2019, doi: 10.32628/cseit1953143.

[8] M. Dagar, A. Kajal, and P. Bhatia, "Twitter Sentiment Analysis using Supervised Machine Learning Techniques," *2021 5th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2021*, no. March, pp. 1–18, 2021, doi: 10.1109/ISCON52037.2021.9702333.

[9] R. Situmorang, U. M. Husni Tamyis, and L. S. Andar Muni, "Analisis Sentimen Destinasi Wisata Di Jawabarat Pada Twitter Menggunakan Algoritma Naive Bayes Classifier," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 8, no. 2, pp. 339–342, 2023, doi: 10.51876/simtek.v8i2.287.

[10] K. H. Chan and S. K. Im, "Sentiment analysis by using Naïve-Bayes classifier with stacked CARU," *Electron. Lett.*, vol. 58, no. 10, pp. 411–413, 2022, doi: 10.1049/ell2.12478.

[11] A. Sabrani, I. G. W. Wedashwara W, and F. Bimantoro, "Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA)*, vol. 2, no. 1, pp. 89–100, 2020, doi: 10.29303/jtika.v2i1.87.

[12] M. Tezgider, B. Yildiz, and G. Aydin, "Text classification using improved bidirectional transformer," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 9, pp. 1–12, 2022, doi: 10.1002/cpe.6486.

[13] N. Umar and M. Adnan Nur, "Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 4, pp. 585–590, 2022, doi: 10.29207/resti.v6i4.4179.

[14] K. M. Suryaningrum, "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech," *Eng. Math. Comput. Sci. J.*, vol. 5, no. 2, pp. 79–83, 2023, doi:

10.21512/emacsjournal.v5i2.9978.

[15] F. A. J. Ayomi and K. E. Dewi, "Analisis Emosi pada Media Sosial Twitter Menggunakan Metode Multinomial Naive Bayes dan Synthetic Minority Oversampling Technique," *Komputa J. Ilm. Komput. dan Inform.*, vol. 12, no. 2, pp. 9–19, 2023, doi: 10.34010/komputa.v12i2.9454.

[16] Ismail, A. M. Nugroho, and R. Sulistiyowati, "Sentiment Analysis Netizens on Social Media Twitter Against Indonesian Presidential Candidates in 2024 Using Naive Bayes Classifier Algorithm," vol. 7, no. 3, pp. 1611–1622, 2023, doi: 10.30865/mib.v7i3.6536.

[17] N. Agustiana, O. N. Pratiwi, and H. Fakhrurroja, "Comparison Of Sentiment Analysis Of Traveloka And Tiket.Com Applications On Twitter Using The Naive Bayes Method," *ITEJ (Information Technol. Eng. Journals)*, vol. 8, no. 2, pp. 73–83, 2023, doi: 10.24235/itej.v8i2.119.

[18] J. C. Aponno, "Penerapan Algoritma Sentimen Analysis dan Naïve Bayes terhadap opini pengunjung di tempat wisata pantai Pintu Kota, Kota Ambon," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 4, pp. 3180–3188, 2022, doi: 10.35957/jatisi.v9i4.2697.

[19] Y. A. Singgalen, "Analisis Sentimen Wisatawan Melalui Data Ulasan Candi Borobudur di Tripadvisor Menggunakan Algoritma Naïve Bayes Classifier," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, p. 1343−1352, 2022, doi: 10.47065/bits.v4i3.2486.

[20] S. A. Putra and A. Wijaya, "Analisis Sentimen Artificial Intelligence (Ai) Pada Media Sosial Twitter Menggunakan Metode Lexicon Based," *JuSiTik J. Sist. dan Teknol. Inf. Komun.*, vol. 7, no. 1, pp. 21–28, 2023, doi: 10.32524/jusitik.v7i1.1042.

[21] I. Harifal *et al.*, "Naive Bayes Optimization with PSO for Predicting ICU Needs for Covid-19 Patients," vol. 11, no. September, pp. 724–734, 2022, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[22] J. P. Munggaran, A. A. Alhafidz, M. Taqy, D. A. R. Agustini, and M. Munawir, "Sentiment Analysis of Twitter Users' Opinion Data Regarding the Use of ChatGPT in Education," *J. Comput. Eng. Electron. Inf. Technol.*, vol. 2, no. 2, pp. 75–88, 2023, doi: 10.17509/coelite.v2i2.59645.

[23] R. Rasenda, H. Lubis, and R. Ridwan, "Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter," *J. Media Inform. Budidarma*, vol. 4, no. 2, pp. 369–376, 2020, doi: 10.30865/mib.v4i2.2051.

[24] R. Rifaldi, J. Indra, A. R. Pratama, and A. R. Juwita, "Analisis Sentimen Pemboikotan Produk dengan Pendekatan Algoritma Naïve Bayes Media Sosial X," vol. 5, no. 4, pp. 940–946, 2024, doi: 10.47065/josh.v5i4.5420.

[25] M. A. Nur and N. Wardhani, "Optimasi Normalisasi Kata Pada Data Twitter Untuk Meningkatkan Akurasi Analisis Sentimen (Studi Kasus Respon Masyarakat Terhadap Layanan Teman Bus)," *J. Fokus Elektroda Energi List. Telekomun. Komputer, Elektron. dan Kendali)*, vol. 7, no. 4, pp. 237–243, 2022.

[26] Maharani and L. Andraini, "Analisis Part of Tagging Bahasa Indonesia tentang Swamedikasi Pada Dialog Interactive Qestion dengan Brill TAGGER," *Teknologipintar.org*, vol. 2, no. 10, pp. 1–11, 2022.

[27] A. P. Wibawa, F. Miftahuddin, and Suyono, "K-Medoids Clustering for the Establishment of

 ISSN:2302-8149
e-ISSN:2540-9719

Javanese Language Stopword Database," *Ranah J. Kaji. Bhs.*, vol. 10, no. 2, pp. 261–269, 2021, [Online]. Available: https://doi.org/10.26499/rnh/v9i2.1490

[28] N. Chafid, L. Mujianto, and I. N. Irmansyah, "Penerapan Filter Kata Menggunakan Metode Stemming Pada Aplikasi Chatting Berbasis Web," vol. 1, no. 1, pp. 1–9, 2020.

[29] S. Mohd Sofi and A. Selamat, "Aspect Based Sentiment Analysis: Feature Extraction using Latent Dirichlet Allocation (LDA) and Term Frequency - Inverse Document Frequency (TF-IDF) in Machine Learning (ML)," *Malaysian J. Inf. Commun. Technol.*, vol. 8, no. 2, pp. 169–179, 2023, doi: 10.53840/myjict8-2-102.

[30] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen," *Pros. Semin. Nas. Sist. Inf. dan Teknol.*, pp. 145–149, 2022.

[31] M. T. Razaq, D. Nurjanah, and H. Nurrahmi, "Analisis Sentimen Review Film Menggunakan Naive Bayes Classifier dengan Fitur TF-IDF," *e-Proceeding Eng.*, vol. 10, no. 2, pp. 1698–1712, 2023.

[32] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 5, pp. 178–185, 2022, doi: 10.30865/klik.v2i5.362.

[33] L. Mardiana, D. Kusnandar, and N. Satyahadewi, "Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak," *Bul. Ilm. Mat. Stat. dan Ter.*, vol. 11, no. 1, pp. 97–102, 2022.

[34] P. A. Hartanto, "Penerapan Algoritma Decision Tree untuk Seleksi Penerima Beasiswa (Studi Kasus: Smpn 1 Soreang)," *Int. J. Technol.*, vol. 47, no. 1, pp. 1294–1302, 2023.