

## ANALISIS SENTIMEN ORANG TUA MURID BARU TERHADAP SMPN 40 SAMARINDA PADA SPMB 2025 MENGGUNAKAN ALGORITMA NAÏVE BAYES

Resifa Ananta Putra✉, Heny Pratiwi, Ahmad Abul Khair

Teknik Informatika, STMIK Widya Cipta Dharma, Samarinda, Indonesia

Email: [2243072@wicida.ac.id](mailto:2243072@wicida.ac.id)

DOI: <https://doi.org/10.46880/jmika.Vol9No2.pp292-299>

### ABSTRACT

*The New Student Admission Selection (SPMB) plays an essential role in ensuring equal educational access in Indonesia. However, during SPMB 2025 at SMPN 40 Samarinda, many candidates living nearby did not choose the school as their first preference, suggesting that perceptions and school image significantly influenced their choices. This study aims to analyze new student parents sentiments toward SMPN 40 Samarinda using the Naïve Bayes algorithm combined with the Term Frequency–Inverse Document Frequency (TF-IDF) technique. Data were collected from 42 respondents and categorized into positive, neutral, and negative sentiments. The model achieved an accuracy of 86%, precision of 56%, and recall of 63%, showing that Naïve Bayes performs effectively on limited data, though less sensitive to minority classes. The analysis revealed that most parents expressed positive perceptions, indicating growing trust that SMPN 40 Samarinda can support students' character development. These findings emphasize the importance of strengthening school image and service quality while highlighting the potential of machine learning–based sentiment analysis as a data-driven approach to understanding educational perceptions.*

**Keyword:** Naïve Bayes, Sentiment Analysis, SMPN 40 Samarinda, SPMB, TF-IDF.

### ABSTRAK

*Sistem Penerimaan Murid Baru (SPMB) memainkan peran penting dalam memastikan akses pendidikan yang setara di Indonesia. Namun, pada SPMB 2025 di SMPN 40 Samarinda, banyak calon murid yang tinggal di sekitar sekolah tidak memilih sekolah tersebut sebagai pilihan pertama mereka, menunjukkan bahwa persepsi dan citra sekolah secara signifikan mempengaruhi pilihan mereka. Penelitian ini bertujuan untuk menganalisis sentimen orang tua murid baru terhadap SMPN 40 Samarinda menggunakan algoritma Naïve Bayes yang dikombinasikan dengan teknik Term Frequency–Inverse Document Frequency (TF-IDF). Data dikumpulkan dari 42 responden dan dikategorikan menjadi sentimen positif, netral, dan negatif. Model ini mencapai akurasi 86%, presisi 56%, dan recall 63%, menunjukkan bahwa Naïve Bayes berfungsi efektif pada data terbatas, meskipun kurang sensitif terhadap kelas minoritas. Analisis menunjukkan bahwa sebagian besar orang tua mengekspresikan persepsi positif, menandakan kepercayaan yang semakin tumbuh bahwa SMPN 40 Samarinda dapat mendukung pengembangan karakter murid. Temuan ini menekankan pentingnya memperkuat citra sekolah dan kualitas layanan, sekaligus menyoroti potensi analisis sentimen berbasis machine learning sebagai pendekatan berbasis data untuk memahami persepsi pendidikan.*

**Kata Kunci:** Naïve Bayes, Analisis Sentimen, SMPN 40 Samarinda, SPMB, TF-IDF.

### PENDAHULUAN

Sistem Penerimaan Murid Baru (SPMB) merupakan acara tahunan yang sangat penting dalam sistem pendidikan Indonesia, termasuk di tingkat sekolah menengah pertama. Pemerintah melalui kebijakan jalur domisili, mengatur bahwa calon murid yang berdomisili di dekat sekolah tersebut diberi prioritas utama untuk diterima dan dapat memilih hingga tiga sekolah terdekat. Sistem ini bertujuan untuk memastikan akses pendidikan yang merata dan memperkuat peran sekolah negeri di setiap daerah.

Sebuah fenomena menarik muncul selama SPMB 2025 di SMPN 40 Samarinda. Berdasarkan pengamatan penulis yang terlibat sebagai anggota panitia penerimaan murid baru, ditemukan bahwa beberapa calon murid yang berdomisili di dekat SMPN 40 Samarinda tidak memilih sekolah tersebut sebagai pilihan pertama mereka selama proses pendaftaran. Meskipun sekolah tersebut dekat dengan tempat tinggal mereka, banyak dari calon murid ini memilih sekolah lain sebagai pilihan pertama mereka, menempatkan SMPN 40 Samarinda sebagai pilihan kedua atau ketiga

mereka. Akibatnya, sejumlah murid baru akhirnya ditempatkan di SMPN 40 Samarinda setelah tidak diterima di sekolah pilihan pertama mereka, sesuai dengan kebijakan domisili yang mengutamakan jarak antara tempat tinggal murid dan sekolah. Fenomena ini menunjukkan kecenderungan rendahnya prioritas SMPN 40 Samarinda dalam pemilihan sekolah, bukan karena keterbatasan akses atau kesempatan, melainkan diduga dipengaruhi oleh persepsi dan citra sekolah di mata masyarakat setempat.

Salah satu tantangan utama yang dihadapi sekolah adalah mempertahankan eksistensinya dan meningkatkan kualitas layanannya agar tetap menjadi pilihan utama masyarakat. Dalam konteks ini, proses penerimaan murid baru merupakan aspek strategis yang menentukan keberlanjutan operasional sekolah. Kemampuan sekolah dalam menarik calon murid tidak hanya mencerminkan reputasi lembaga tetapi juga menunjukkan efektivitas manajemen strategisnya dalam membaca kebutuhan pasar pendidikan yang semakin kompetitif (Suryani et al., 2025). Oleh karena itu, diperlukan pendekatan yang mampu menangkap persepsi masyarakat secara terstruktur dan sistematis. Sebagaimana penelitian sebelumnya oleh (Santoso et al., 2022), *Naïve Bayes* dapat mencapai akurasi yang tinggi dalam klasifikasi sentimen, sehingga cocok untuk data teks pendek. Oleh karena itu, penelitian ini mengadopsi *Naïve Bayes* yang dikombinasikan dengan TF-IDF untuk meningkatkan kinerja klasifikasi. Penelitian ini berupaya untuk menjawab pertanyaan-pertanyaan berikut: (1) Bagaimana analisis sentimen dapat mengungkapkan persepsi orang tua murid baru terhadap SMPN 40 Samarinda selama proses SPMB 2025? (2) Seberapa efektifkah algoritma *Naïve Bayes* yang dikombinasikan dengan TF-IDF dalam mengklasifikasikan umpan balik tekstual dengan data yang terbatas? Signifikansi penelitian ini terletak pada demonstrasi bagaimana klasifikasi sentimen berbasis data dapat membantu sekolah dalam mengevaluasi kepercayaan masyarakat dan meningkatkan citra institusi.

## KAJIAN LITERATUR

Penelitian oleh (Minardi et al., 2024) menerapkan algoritma *Naïve Bayes* untuk mengklasifikasikan sentimen terkait layanan pendidikan di pesantren modern dengan memanfaatkan 387 ulasan yang diperoleh dari media sosial dan forum daring. Hasil penelitian menunjukkan tingkat *accuracy* sebesar 88%, yang menegaskan efektivitas algoritma *Naïve Bayes* dalam mengolah teks tidak terstruktur dari umpan balik di bidang pendidikan. Penulis juga

menyoroti bahwa *Naïve Bayes* menjadi semakin optimal ketika diterapkan bersama kerangka kerja CRISP-DM dalam proses penambangan data.

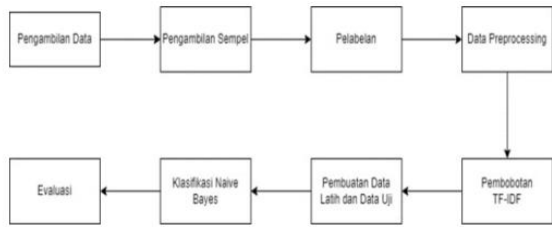
Penelitian lain oleh (Putra et al., 2023) membandingkan kinerja *Naïve Bayes Classifier* dan *Support Vector Machine* (SVM) untuk mengklasifikasikan data kesehatan mental murid, menggunakan *dataset* yang dikumpulkan dari konteks pendidikan. Penelitian ini mencapai *accuracy* sebesar 94,37% dengan *Naïve Bayes* dan 86,87% dengan SVM, yang menyoroti keunggulan *Naïve Bayes* dalam menangani tugas klasifikasi.

Penelitian oleh (Triyono et al., 2025) menerapkan *Naïve Bayes* untuk analisis sentiment terhadap ulasan aplikasi Spotify, memanfaatkan *dataset* sebanyak 1.000 ulasan yang diperoleh dari Google Play Store. Dengan menggabungkan TF-IDF, penelitian ini memperoleh tingkat *accuracy* yang tinggi dalam mengelompokkan sentimen ke dalam 3 kategori (positif, negatif, dan netral), sehingga menunjukkan efektivitas algoritma *Naïve Bayes* dalam memahami ulasan pengguna terhadap aplikasi Spotify.

Penelitian oleh (Lesmana et al., 2020) melakukan analisis sentimen terhadap kebijakan pendidikan, khususnya pada sistem zonasi PPDB (yang kini dikenal sebagai jalur domisili SPMB), dengan menggunakan algoritma *Multinomial Naïve Bayes*. Hasil penelitian tersebut menunjukkan bahwa analisis sentimen dapat dimanfaatkan untuk menilai kebijakan pendidikan berdasarkan opini publik. Keterkaitan penelitian ini memperkuat temuan bahwa metode *Naïve Bayes* efektif dalam mengidentifikasi persepsi masyarakat terhadap kebijakan pendidikan, termasuk dalam konteks proses penerimaan murid baru di SMPN 40 Samarinda.

## METODE PENELITIAN

Dalam penelitian ini, terdapat delapan langkah yang harus dilakukan. Langkah-langkah tersebut meliputi pengambilan data, pengambilan sampel, pelabelan manual, *data preprocessing*, pembobotan TF-IDF, pemisahan data untuk membagi data menjadi data latih dan data uji yang akan diklasifikasikan dengan *Naïve Bayes*, dan terakhir tahap evaluasi untuk menentukan kinerja model.



Gambar 1. Proses Penelitian

### Pengambilan Data

Data yang digunakan dalam penelitian ini dikumpulkan melalui survei menggunakan Google Forms yang disebarakan kepada orang tua murid baru di dua kelas, 7A dan 7B, melalui wali kelas dalam grup WhatsApp. Sebanyak 71 orang tua murid berpartisipasi dalam survei ini.

### Pengambilan Sampel

Dalam pengambilan sampel, *simple random sampling* digunakan untuk memilih sebagian dari populasi. *Simple random sampling* memastikan bahwa setiap anggota populasi memiliki kesempatan yang sama untuk dipilih menjadi sampel. Ukuran sampel keseluruhan dihitung menggunakan metode Slovin.

$$n = \frac{N}{1 + N \cdot e^2} \quad (1)$$

Persamaan (1) adalah rumus Slovin, dengan  $n$  adalah ukuran sampel yang dibutuhkan,  $N$  adalah ukuran populasi, dan  $e$  adalah *margin of error* atau tingkat kesalahan yang ditoleransi.

### Pelabelan

Data analisis yang terkumpul akan diberi label secara manual sebagai positif, negatif, dan netral. Suatu nilai akan diberi label positif jika lebih besar dari 3, negatif jika kurang dari 3, dan netral jika sama dengan 3, seperti yang ditunjukkan pada Tabel 1 berikut ini:

Tabel 1. Tinjauan Pelabelan

Label	Tinjauan
Positif	Lebih dari >3
Negatif	Kurang dari < 3
Netral	Sama dengan 3

Proses pelabelan dalam penelitian ini dilakukan secara manual oleh peneliti berdasarkan skor rata-rata umpan balik orang tua murid baru terhadap 10 pertanyaan kuesioner. Kategori sentimen ditetapkan positif jika skor rata-rata lebih besar dari 3, netral jika skornya sama dengan 3, dan negatif jika skornya kurang dari 3, seperti yang ditunjukkan pada Tabel 1.

Untuk memastikan bahwa pelabelan mencerminkan sentimen yang sebenarnya, umpan balik tertulis setiap responden di bagian kritik dan saran ditinjau untuk memastikan kesesuaian umpan balik tekstual dengan umpan balik skor numerik. Perbandingan ini menunjukkan keselarasan yang kuat antara skor rata-rata kepuasan dan ungkapan tekstual dalam kritik dan saran, memastikan bahwa label sentimen valid dan konsisten untuk digunakan dalam proses analisis sentimen.

### Data Preprocessing

Untuk mengoptimalkan hasil analisis sentimen, perlu dilakukan tahap *data preprocessing*. Tahap ini bertujuan untuk membersihkan *dataset* dan membuatnya lebih sederhana (Syam et al., 2024). Tahap ini terdiri dari empat langkah berikut:

- 1) *Case folding and Cleaning*, yaitu langkah *data preprocessing* yang mengubah semua huruf menjadi huruf kecil dan menghapus semua tanda baca, angka, atau tautan dari respons pengguna.
- 2) *Stopwords removal*, yaitu tahap di mana kata-kata yang sering muncul perlu dihapus dari dokumen tanpa mengubah maknanya. Pada langkah ini, *stopwords* seperti “itu,” “untuk,” “di,” “dan,” “atau” dihapus.
- 3) *Tokenizing*, yaitu tahap di mana kalimat dipisahkan menjadi beberapa kata atau karakter yang disebut token.
- 4) *Stemming*, yaitu tahap akhir dari *data preprocessing*. Untuk mengubah kata-kata yang berafiks menjadi kata dasar, dilakukan proses stemming. Pada langkah ini, kata-kata tambahan seperti awalan, sisipan, dan akhiran akan dihilangkan (Wibowo & Pratama, 2024).

### Pembobotan TF-IDF

TF-IDF memungkinkan transformasi data kata menjadi data numerik dengan memberikan bobot pada kata-kata dalam suatu dokumen (Fahrezi & Verdikha, 2024). Selain itu, metode TF-IDF dapat meningkatkan akurasi proses analisis dengan menggabungkan dua konsep: *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

$$TF.IDF_{std}(t) = tf_a^t \times \log \frac{n}{df^t} \quad (2)$$

TF adalah jumlah kata yang muncul dalam sebuah dokumen, dan jumlah kata dalam sebuah dokumen setara dengan bobot kata-kata tersebut karena kata-kata yang muncul berulang kali memiliki bobot

yang lebih besar. Sebaliknya, konsep IDF menganggap kata-kata yang muncul berulang kali memiliki bobot yang lebih rendah (Yutika et al., 2021).

### Klasifikasi *Naïve Bayes*

*Naïve Bayes* merupakan salah satu algoritma klasifikasi probabilistik yang paling sederhana, dengan asumsi independensi yang kuat antar fitur yang digunakan. Asumsi ini menyatakan bahwa setiap fitur, seperti kata dalam konteks analisis teks, dianggap berdiri sendiri, independen, dan tanpa korelasi langsung satu sama lain, sehingga banyak orang menggunakan algoritma ini untuk melakukan analisis sentimen. Klasifikasi *Naïve Bayes* dapat memprediksi kemungkinan keanggotaan kelas berdasarkan prediksi asumsi independen. Dengan demikian, pengalaman masa lalu dapat digunakan untuk menentukan probabilitas di masa mendatang (Husaini & Lisdiyanto, 2024). Algoritma *Naïve Bayes* memiliki keuntungan karena ringkas dalam analisis sentimen (Herlinawati et al., 2020). Selain itu, algoritma ini seringkali mencapai akurasi yang tinggi bahkan ketika menggunakan data pelatihan dalam jumlah kecil (Zakaria et al., 2024).

Dalam penelitian ini, algoritma *Naïve Bayes* diterapkan untuk mengklasifikasikan komentar orang tua murid baru pada bagian kritik dan saran, ke dalam tiga kategori sentimen (positif, netral, dan negatif). Algoritma ini memperkirakan probabilitas suatu dokumen termasuk dalam kelas tertentu berdasarkan kemunculan kata-kata dalam kelas tersebut. Setiap kata diperlakukan sebagai fitur independen, yang memungkinkan algoritma bekerja secara efisien bahkan dengan *dataset* yang kecil.

Probabilitas kelas sentimen  $C$  untuk dokumen  $X$  ditentukan menggunakan teorema *Naïve Bayes* berikut:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (3)$$

Dimana  $P(C|X)$  merepresentasikan probabilitas bahwa dokumen  $X$  termasuk dalam kelas  $C$ ,  $P(X|C)$  adalah kemungkinan distribusi kata di kelas  $C$ , dan  $P(C)$  adalah probabilitas prioritas setiap sentimen berdasarkan distribusi kelas *dataset*.

Misalnya, dalam komentar “Sekolahnya bagus dan disiplin,” kata “bagus” dan “disiplin” sering muncul dalam data berlabel positif, sehingga menghasilkan probabilitas yang lebih tinggi untuk kelas positif. Sebaliknya, komentar “kurang ramah” berisi istilah yang ditemukan dalam sampel negatif, sehingga menghasilkan probabilitas yang lebih tinggi untuk kelas negatif. Selama pelatihan, 80% data

berlabel digunakan untuk pembelajaran model, dan 20% sisanya digunakan untuk pengujian.

Algoritma *Naïve Bayes* dipilih karena memiliki efisiensi tinggi dalam menganalisis teks berukuran kecil dan menggunakan pendekatan probabilistik yang sederhana, sehingga hasil klasifikasinya mudah dipahami serta relevan untuk konteks penelitian pendidikan.

### Evaluasi Metode

Evaluasi Metode merupakan faktor kunci dalam menilai kinerja klasifikasi dan memandu pemodelan pengklasifikasi (Tharwat, 2021). Metrik yang digunakan adalah *accuracy*, *precision*, dan *recall* yang dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

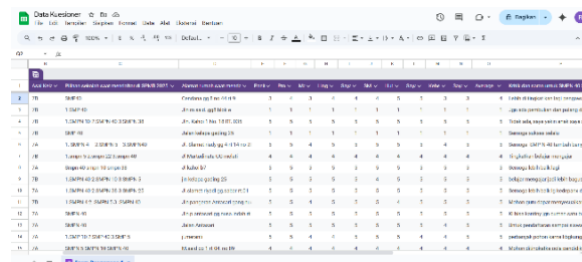
$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Dimana  $TP$  adalah *True Positive*,  $TN$  adalah *True Negative*,  $FP$  adalah *False Positive*, dan  $FN$  adalah *False Negative* (Wibowo & Pratama, 2024).

## HASIL DAN PEMBAHASAN

### Pengambilan Data



Gambar 2. Tangkapan Layar Jawaban di Google Forms

Dalam pengambilan data, kami menggunakan Google Forms. *Forms* didistribusikan melalui grup WhatsApp wali kelas 7A dan 7B. Dari distribusi *forms* tersebut, kami memperoleh data 71 responden dari dua kelas di SMPN 40 Samarinda. Data yang diperoleh dari distribusi tersebut dapat dilihat pada Gambar 2.

“Asal Kelas” adalah identitas orang tua murid di kelas 7A atau 7B, “Pilihan Sekolah...” adalah sekolah yang dipilih oleh murid baru pada SPMB 2025, “Alamat Rumah...” adalah tempat tinggal murid baru menurut dokumen resmi, dan kolom berikutnya adalah 10 pertanyaan dengan skala 1 sampai 5, yang hasilnya

akan dirata-ratakan. 10 pertanyaan yang telah diuji validitas dan reliabilitasnya diberikan kepada orang tua murid baru yang mendaftar SPMB 2025 di SMPN 40 Samarinda, dan kolom terakhir adalah untuk kritik dan saran di mana orang tua murid baru dapat menyampaikan pandangan mereka tentang SMPN 40 Samarinda.

### Pengambilan Sampel

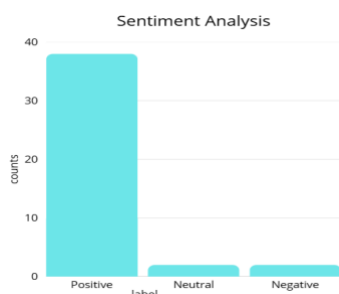
Pengambilan sampel dilakukan menggunakan metode Slovin. Jumlah total murid kelas 7A dan 7B berdasarkan absensi kelas adalah 71. Pengambilan sampel dilakukan seperti yang dijelaskan pada persamaan 7.

$$n = \frac{71}{1 + 71 \cdot 0,01^2} = \frac{71}{1 + 71 \cdot (0,01)} = \frac{71}{1 + 0,71} = \frac{71}{1,71} = 41,52$$

Dari hasil perhitungan metode Slovin, diperoleh sampel sebanyak 41,52. Untuk mempermudah pengumpulan data, hasilnya dibulatkan menjadi 42 responden sentimen.

### Pelabelan

Setelah pengambilan sampel, proses pelabelan dilakukan. Pelabelan dilakukan dengan menggunakan skor rata-rata dari 10 pertanyaan yang diberikan kepada orang tua murid baru. Skor berkisar antara 1 hingga 5, di mana 1 berarti sangat tidak setuju, 2 berarti tidak setuju, 3 berarti netral, 4 berarti setuju, dan 5 berarti sangat setuju. Label sentimen dibagi menjadi tiga kategori: Positif, Netral, dan Negatif, sesuai dengan aturan pada Tabel 1. Hasil pelabelan menunjukkan 38 sentimen positif, 2 sentimen netral, dan 2 sentimen negatif. Hal ini menunjukkan bahwa mayoritas sentimen bersifat positif. Hasil pelabelan dapat dilihat pada Gambar 3.



Gambar 3. Hasil Pelabelan

### Data Preprocessing

1. *Case folding and Cleaning*, yaitu langkah yang melibatkan konversi respons atau komentar pengguna menjadi huruf kecil dan penghapusan karakter yang tidak diinginkan dari komentar agar data lebih mudah dianalisis. Hasil dari proses ini dapat dilihat pada Gambar 4.

komentar	cleaned
Lebih di tingkat kan lagi pengawasan bagi anak2 yg mungkin ketakuannya agak nakal agar bisa menjaga nama baik sekolah	lebih di tingkat kan lagi pengawasan bagi anak yg mungkin ketakuannya agak nakal agar bisa menjaga nama baik sekolah
Jgn ada pembulan dan pulang dan masuk sekolah sesuai jam biar yg jemput enak	jgn ada pembulan dan pulang dan masuk sekolah sesuai jam biar yg jemput enak
Tidak ada, saya yakin anak saya akan berkembang dengan baik pendidiknya di SMPN 40	tidak ada saya yakin anak saya akan berkembang dengan baik pendidiknya di smpn
Semoga sukses selalu	semoga sukses selalu
Semoga SMP N 40 tambah banyak kelasnya Sukses	semoga smp n tambah banyak kelasnya sukses untuk smpn
Tingkatkan belajar mengajar	tingkatkan belajar mengajar
Semoga lebih baik lagi	semoga lebih baik lagi
belajar mengajar jadi lebih bagus	belajar mengajar jadi lebih bagus
Semoga lebih baik lg kedepannya dr segi keamanan	semoga lebih baik lg kedepannya dr segi keamanan

Gambar 4. Case folding and Cleaning

2. *Stopwords removal*, yaitu tahap dimana semua hasil dari proses pembersihan akan diproses kembali untuk menghilangkan kata-kata umum yang sering muncul. Hasil dari tahap ini dapat dilihat pada Gambar 5.

cleaned	no_stopwords
lebih di tingkat kan lagi pengawasan bagi anak yg mungkin ketakuannya agak nakal agar bisa menjaga nama baik sekolah	tingkat pengawasan anak yg mungkin ketakuannya nakal menjaga nama sekolah
Jgn ada pembulan dan pulang dan masuk sekolah sesuai jam biar yg jemput enak	jgn pembulan pulang masuk sekolah sesuai jam biar yg jemput enak
tidak ada saya yakin anak saya akan berkembang dengan baik pendidiknya di smpn	anak berkembang pendidiknya smpn
semoga sukses selalu	semoga sukses
Semoga smp n tambah banyak kelasnya sukses untuk smpn	semoga smp n kelasnya sukses smpn
tingkatkan belajar mengajar	tingkatkan belajar mengajar
semoga lebih baik lagi	semoga
belajar mengajar jadi lebih bagus	belajar mengajar bagus
Semoga lebih baik lg kedepannya dr segi keamanan	semoga lg kedepannya dr segi keamanan
mohon guru dapat menyesuaikan pemahaman materi tidak terlalu cepat dan memberikan kesempatan siswa untuk bertanya atau mengulangi materi yang sulit di pahami	mohon guru menyesuaikan pemahaman materi tidak terlalu cepat kesempatan siswa mengulangi materi sulit pahami

Gambar 5. Stopwords removal

3. *Tokenizing*, yaitu setelah data dibersihkan dan kata-kata umum yang sering muncul di tahap selanjutnya telah dihapus, setiap teks akan dipecah menjadi kata-kata individual dari kolom *no\_stopwords* yang telah dibersihkan. Hasil *tokenizing* dapat dilihat pada Gambar 6.

tokens
['tingkat', 'awas', 'anak', 'yg', 'laku', 'nakal', 'manjaga', 'nama', 'sekolah']
['jgn', 'bulian', 'pulang', 'masuk', 'sekolah', 'sesuai', 'jam', 'biar', 'yg', 'jemput', 'enak']
['anak', 'kembang', 'didik', 'nya', 'smp']
['moga', 'sukses']
['moga', 'smp', 'kelas', 'sukses', 'smpn']
['tingkat', 'ajar', 'ajar']
['moga']
['ajar', 'ajar', 'bagus']
['moga', 'lg', 'kedepannya', 'dr', 'segi', 'aman']

Gambar 6. Tokenizing

4. *Stemming*, yaitu proses terakhir, dimana semua afiks dihilangkan, hanya menyisakan kata dasar. Hasil proses stemming dapat dilihat pada Gambar 7.

stemmed
tingkat awas anak yg laku nakal menjaga nama sekolah
jgn bulian pulang masuk sekolah sesuai jam biar yg jemput enak
anak kembang didik smpn
moga sukses
moga smp n kelas sukses smpn
tingkat ajar ajar
moga
ajar ajar bagus
moga lg kedepan dr segi aman

Gambar 7. *Stemming*

#### Pembobotan TF-IDF

Setelah umpan balik dari orang tua murid baru diproses terlebih dahulu, langkah selanjutnya adalah memberi bobot pada kata-kata. TF-IDF menyederhanakan teks menjadi nilai numerik sehingga lebih ringkas. Kata-kata dengan frekuensi tinggi tetapi hanya muncul di beberapa dokumen lebih informatif dan akan diberi bobot lebih tinggi.

ajar	alumni	aman	anak	awas	bagus	baik	berang	bawa	belah	belanja
0.0	0.0	0.0	0.2619452911506205	0.35774939742110957	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.4389948404897019	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.001715140418436	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.7834190344124292	0.0	0.0	0.0	0.0	0.62149385879532	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.43819802853265069	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 8. *Stopwords removal*

Gambar 8 menunjukkan hasil pembobotan kata. Nilai 0 menunjukkan kata tersebut tidak muncul atau tidak memiliki bobot dalam dokumen terkait, sementara nilai yang besar menunjukkan kata tersebut memiliki nilai informatif dalam dokumen. Misalnya, kata “bagus” memiliki nilai 0 yang rendah dalam 10 dokumen, kecuali pada dokumen 8 yang memiliki nilai 0,62149385879532. Hal ini menunjukkan bahwa kata “bagus” sering muncul di dokumen 8 tetapi tidak di dokumen lain. Hal ini menjadikannya informatif dalam dokumen tersebut.

#### Klasifikasi *Naïve Bayes*

Setelah seluruh kata diberi bobot menggunakan metode TF-IDF, tahap selanjutnya adalah melatih model klasifikasi menggunakan algoritma *Naïve Bayes*. *Dataset* dibagi menjadi dua bagian, yaitu 80% untuk

data pelatihan (*training*) dan 20% untuk data pengujian (*testing*). Pada tahap pelatihan, model mempelajari distribusi probabilitas kemunculan setiap kata terhadap tiga kategori sentiment positif, netral, dan negatif. Selanjutnya, pada tahap pengujian, model menghitung kemungkinan sebuah komentar baru termasuk ke dalam salah satu kategori tersebut dan menetapkan kelas dengan nilai probabilitas tertinggi sebagai hasil prediksi akhir.

Proses ini memungkinkan model untuk mengenali kata-kata yang paling mewakili setiap sentimen. Misalnya, kata “bagus” cenderung muncul pada sentimen positif, sedangkan kata “kurang” lebih dominan pada sentimen negatif. Sementara itu, komentar netral umumnya berisi kalimat deskriptif tanpa ekspresi emosional yang jelas. Melalui pendekatan ini, *Naïve Bayes* menghitung peluang suatu sentimen berdasarkan kemunculan kata dan menormalkannya terhadap distribusi kelas secara keseluruhan.

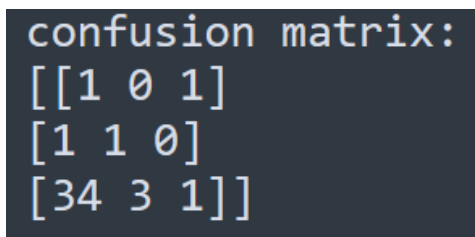
Hasil pelatihan menunjukkan bahwa model berhasil mengidentifikasi pola umum dari umpan balik orang tua murid baru terhadap sekolah. Namun, kinerja algoritma mengalami penurunan ketika data yang digunakan tidak seimbang, karena model cenderung lebih fokus mempelajari kelas dengan jumlah data yang lebih besar. Hal ini juga terlihat dalam penelitian ini, di mana sentimen positif mendominasi sebanyak 38 data, sementara sentimen netral dan negatif masing-masing hanya terdiri atas dua data. Ketimpangan tersebut menyebabkan model lebih efektif dalam mengenali sentimen positif, tetapi kurang tepat dalam mengklasifikasikan sentimen dengan jumlah data yang lebih sedikit.

Meskipun demikian, dengan tingkat *accuracy* sebesar 86%, *precision* 56%, dan *recall* 63%, model masih mampu menunjukkan performa yang cukup baik untuk ukuran *dataset* yang relatif kecil. Hasil ini sejalan dengan temuan penelitian sebelumnya (Fitriani et al., 2021) yang menegaskan bahwa algoritma *Naïve Bayes* unggul dalam hal efisiensi dan tetap mampu menghasilkan klasifikasi yang akurat meskipun menggunakan data terbatas. Pada penelitian mendatang, performa model dapat ditingkatkan melalui penerapan teknik *oversampling* seperti *Random Oversampling* (ROS) guna menyeimbangkan proporsi data antar kelas, sehingga kemampuan model dalam mengenali sentimen minoritas dapat diperbaiki tanpa mengorbankan tingkat akurasi keseluruhan.



### Evaluasi Model

Hasil diperoleh dari *confusion matrix* di mana baris pertama adalah kelas negatif dengan 1 kelas negatif diprediksi negatif, 0 kelas negatif diprediksi netral, dan 1 kelas negatif diprediksi positif. Baris kedua adalah baris netral, di mana 0 sampel kelas netral diprediksi negatif, 1 sampel kelas netral diprediksi netral dengan benar, dan 1 sampel netral diprediksi positif. Baris ketiga adalah baris positif, di mana 1 sampel kelas positif diprediksi negatif, 3 sampel positif diprediksi netral, dan 34 kelas positif diprediksi positif dengan benar. Hasil proses dapat dilihat pada Gambar 9.



confusion matrix:
[1 0 1]
[1 1 0]
[34 3 1]

Gambar 9. *Confusion matrix*

Hasil dari *confusion matrix* kemudian digunakan untuk evaluasi. Hasil menunjukkan bahwa model mencapai *accuracy* sebesar 86%, *precision* sebesar 56%, dan *recall* sebesar 63%. Hasil menunjukkan bahwa *Naïve Bayes* mampu mengklasifikasikan sentimen sesuai dengan label yang telah diberikan. Namun, model tersebut masih memiliki kekurangan yaitu hanya dapat mengenali label positif dengan baik dibandingkan dengan label lainnya, seperti yang ditunjukkan pada *confusion matrix* pada Gambar 9. Hal ini dikarenakan lebih banyak label positif dibandingkan dengan label netral maupun negatif, sehingga permasalahan ini disebabkan oleh ketidakseimbangan data. Hal ini menyebabkan model hanya mengenali label dominan dan tidak mengenali label lainnya, sehingga mengakibatkan klasifikasi yang tidak seimbang (Khair et al., 2024).

### KESIMPULAN

Penelitian ini menunjukkan bahwa proses SPMB di SMPN 40 Samarinda tidak hanya dipengaruhi oleh kebijakan domisili, tetapi juga oleh persepsi dan citra sekolah di mata masyarakat. Melalui penerapan algoritma *Naïve Bayes* dengan metode TF-IDF, penelitian ini berhasil mengidentifikasi sentimen orang tua murid baru yang sebagian besar bersifat positif, khususnya terkait pada proses SPMB di tahun ini. Model mencapai akurasi sebesar 86%, *precision* 56%, dan *recall* 63%, menunjukkan efektivitas *Naïve Bayes* dalam menganalisis teks berukuran kecil meskipun

terdapat ketidakseimbangan data. Hasil tersebut menegaskan bahwa penerapan analisis sentimen dengan pendekatan *machine learning* berpotensi menjadi sarana pendukung dalam menilai citra sekolah serta membantu proses pengambilan keputusan strategis di sektor pendidikan. Namun, aspek komunikasi publik dan strategi promosi sekolah masih perlu ditingkatkan agar kepercayaan masyarakat terhadap sekolah semakin kuat. Penelitian selanjutnya disarankan memperluas sumber data salah satunya dengan opini publik di media sosial, serta menggunakan algoritma pembandingan seperti *Support Vector Machine* atau *Logistic Regression* untuk membantu validasi kinerja *Naïve Bayes*.

### DAFTAR PUSTAKA

- Fahrezi, I. A., & Verdikha, N. A. (2024). Analisis sentimen Twitter atas isu hak angket menggunakan pembobotan TF-IDF dan algoritma SVM. *Sci-Tech Journal*, 3(2), 179–192.
- Fitriani, R. D., Yasin, H., & Tarno, T. (2021). Penanganan klasifikasi kelas data tidak seimbang dengan Random Oversampling pada Naïve Bayes (Studi kasus: Status peserta KB IUD di Kabupaten Kendal). *Jurnal Gaussian*, 10(1), 11–20.
- Guno, I. A., Wibowo, M., & Pratama, I. (2024). Analisis sentimen ulasan aplikasi identitas kependudukan digital menggunakan metode Support Vector Machine. *Jurnal Teknologi dan Sistem Informasi Bisnis*, 6(4), 715–722. <https://doi.org/10.47233/jteksis.v6i4.1552>
- Herlinawati, N., Yuliani, Y., Faizah, S., Gata, W., & Samudi, S. (2020). Analisis sentimen Zoom Cloud Meetings di Play Store menggunakan Naïve Bayes dan Support Vector Machine. *CESS (Journal of Computer Engineering and System Sciences)*, 5(2), 293. <https://doi.org/10.24114/cess.v5i2.18186>
- Husaini, A. P., & Lisdiyanto, A. (2024). Sistem prediksi penjualan produk APD terlaris di PT A3 Karunia Sidoarjo menggunakan metode Naïve Bayes. *Jurnal Teknologi dan Sistem Informasi Bisnis*, 6(2), 431–437. <https://doi.org/10.47233/jteksis.v6i2.1266>
- Khair, A. A., Pratiwi, H., & Saputra, N. J. (2024). Penerapan algoritma K-Nearest Neighbor untuk klasifikasi penerima beasiswa pada STMIK Widya Cipta Dharma.
- Lesmana, L., Mukrodin, & Nabya, F. (2020). Analisis sentimen pengguna Twitter terhadap kebijakan sistem zonasi PPDB menggunakan algoritma Multinomial Naïve Bayes. *Jurnal Sistem Informasi dan Teknologi Peradaban (JSITP)*, 1(1), 24–28.

- Minardi, J., et al. (2024). Analysis of sentiment towards educational services in modern Islamic boarding schools using the Naïve Bayes method. *Scientific Journal of Informatics*, 11(4).
- Putra, H. D., Khairani, L., & Hastari, D. (2023). Comparison of Naïve Bayes Classifier and Support Vector Machine algorithms for classifying student mental health data. In *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat* (pp. 120–125).
- Santoso, H., Armansyah, A., & Desliani, D. (2022). Analisis sentimen mahasiswa terkait pembelajaran tatap muka menggunakan metode Naïve Bayes Classifier. *Techno.com*, 21(3), 644–654.
- Suryani, L., Afgani, M. W., & Afriantoni, A. (2025). Optimalisasi penerimaan peserta didik baru melalui pendekatan manajemen strategi di Sekolah Dasar Negeri 52 Prabumulih. *Indonesian Research Journal on Education*, 5(2), 1360–1368.
- Syam, A. A., Hardy, G. M., Salim, A., Surianto, D. F., & Fajar, M. B. (2024). Analisis teknik preprocessing pada sentimen masyarakat terkait konflik Israel-Palestina menggunakan Support Vector Machine. *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 9(3), 1464–1472. <https://doi.org/10.29100/jupi.v9i3.5527>
- Tharwat, A. (2021). *Classification assessment methods. Applied Computing and Informatics*, 17(1), 168–192.
- Triyono, A., Faqih, A., Dwilestari, G., & Fathurrohman, F. (2025). *Implementation of the Naïve Bayes method in sentiment analysis of Spotify application reviews. Journal of Artificial Intelligence Engineering and Applications*, 4(2).
- Wibowo, E., & Pratama, I. (2024). Analisis sentimen terhadap ulasan hotel melalui platform Google Review menggunakan metode stacking. *Jurnal Teknologi dan Sistem Informasi Bisnis*, 6(4), 774–784. <https://doi.org/10.47233/jteksis.v6i4.1475>
- Yutika, C. H., Adiwijaya, A., & Al Faraby, S. (2021). Analisis sentimen berbasis aspek pada review Female Daily menggunakan TF-IDF dan Naïve Bayes. *Jurnal Media Informatika Budidarma*, 5(2), 422–430.
- Zakaria, M. A., Pratiwi, H., & Saad, M. I. (2024). Sistem pakar diagnosis penanganan pasca panen kelapa sawit dengan metode Naïve Bayes berbasis web. *Sebatik*, 28(2).