KOMPARASI METODE SVM DAN LOGISTIC REGRESSION PADA ANALISIS SENTIMEN PROSEDUR SNBP

Muhammad Yusuf, Nur Salwa Fadia Akmar, Muhammad Syaikh Azka, Fathoni
Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang-Prabumulih No.KM. 32, Indralaya Indah, Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia
09031282328089@student.unsri.ac.id

ABSTRAK

Perkembangan *Natural Language Processing* (NLP) mendorong pemanfaatan analisis sentimen untuk memahami pandangan publik terhadap kebijakan pendidikan, termasuk Seleksi Nasional Berdasarkan Prestasi (SNBP) yang menjadi topik hangat setiap tahun ajaran baru. Beragam komentar yang muncul menunjukkan adanya perbedaan opini di masyarakat, sehingga diperlukan analisis untuk mengidentifikasi pola sentimen yang terbentuk. Penelitian ini bertujuan membandingkan kinerja algoritma *Logistic Regression* dan *Support Vector Machine* (SVM) dalam mengklasifikasikan sentimen terhadap prosedur SNBP. Data penelitian diperoleh dari kuesioner daring yang telah dilabeli berdasarkan rating, kemudian diproses melalui tahap preprocessing teks dan ekstraksi fitur menggunakan TF-IDF. Selanjutnya, data digunakan dalam proses klasifikasi dan dievaluasi menggunakan *confusion* matrix serta metrik performa. Hasil pengujian menunjukkan bahwa *Logistic Regression* memperoleh akurasi sebesar 85,96% namun cenderung bias terhadap kelas positif. Sementara itu, SVM memperoleh akurasi sebesar 87,72% dengan hasil klasifikasi yang lebih seimbang pada setiap kelas. Distribusi data yang tidak terlalu timpang antar kategori sentimen memungkinkan SVM bekerja lebih maksimal dalam memetakan variasi opini responden terkait prosedur SNBP dibandingkan *Logistic Regression*.

Kata kunci : *Analisis Sentimen, SNBP, Logistic Regression, Support Vector Machine.*

1. PENDAHULUAN

Perkembangan pesat dalam bidang *Natural Language Processing* (NLP) membuka ruang yang luas bagi pemanfaatan analisis sentimen dalam menelaah opini publik yang bersumber dari data teks digital. Pendekatan ini semakin relevan karena kemampuannya dalam mengolah data tak terstruktur secara otomatis, sehingga dapat digunakan untuk menangkap persepsi masyarakat terhadap kebijakan publik, layanan pendidikan, maupun sistem berbasis daring [1].

Dalam ranah pendidikan tinggi di Indonesia, mekanisme Seleksi Nasional Berdasarkan Prestasi (SNBP) menjadi salah satu isu yang memicu beragam tanggapan dari berbagai kalangan, termasuk siswa, orang tua, serta tenaga pendidik, khususnya melalui platform digital. Kondisi tersebut mendorong perlunya pendekatan berbasis *machine learning* untuk mengekstraksi pola sentimen sebagai bahan pertimbangan dalam evaluasi kebijakan.

Di sisi lain, proses klasifikasi sentimen tidak terlepas dari sejumlah tantangan, terutama terkait dengan keragaman bahasa informal serta distribusi kelas yang tidak seimbang dalam dataset. Permasalahan ini berpotensi menurunkan performa model dalam mengidentifikasi sentimen secara akurat. Oleh karena itu, pemilihan algoritma klasifikasi menjadi aspek krusial dalam menghasilkan analisis yang reliabel. Algoritma konvensional seperti *Support Vector Machine* (SVM) dan *Logistic Regression* (LR) masih banyak digunakan karena memiliki efisiensi komputasi yang baik serta mampu menangani data teks berdimensi tinggi [2].

SVM dikenal efektif dalam membentuk batas pemisah optimal melalui konsep *hyperplane*, terutama ketika dikombinasikan dengan representasi fitur TF-IDF [3].

Sementara itu, Logistic Regression menawarkan keunggulan dari sisi interpretabilitas dan kestabilan probabilistik, sehingga sering dijadikan sebagai model pembanding dalam berbagai penelitian [4]. Sejumlah studi juga menunjukkan bahwa performa kedua algoritma tersebut sangat dipengaruhi oleh karakteristik dataset serta tahapan preprocessing yang diterapkan [5].

Berbagai penelitian terdahulu telah melakukan komparasi antara SVM dan *Logistic Regression* pada beragam domain, seperti e-commerce, ulasan aplikasi, dan media sosial [6].

Namun demikian, kajian yang secara spesifik menelaah sentimen terhadap kebijakan pendidikan nasional, khususnya SNBP, masih relatif terbatas. Selain itu, sebagian besar penelitian sebelumnya belum secara mendalam mengkaji pengaruh ketidakseimbangan kelas serta kompleksitas bahasa Indonesia terhadap kinerja model klasifikasi [7].

Kesenjangan ini menunjukkan perlunya penelitian yang lebih kontekstual dan berfokus pada karakteristik data lokal.

Penelitian ini mengusulkan analisis komparatif SVM dan Logistic Regression dengan evaluasi akurasi, precision, recall, dan F1-score pada dataset sentimen terhadap prosedur SNBP. Data diproses menggunakan preprocessing teks bahasa Indonesia dan representasi fitur TF-IDF agar perlakuan antar model setara. Penelitian ini bertujuan mengidentifikasi

metode yang paling optimal untuk analisis opini kebijakan pendidikan.

Berlandaskan latar belakang tersebut, penelitian ini mengkaji sekaligus melakukan komparasi algoritma *Support Vector Machine* (SVM) dan *Logistic Regression* dalam proses klasifikasi sentimen terhadap prosedur SNBP. Dalam upaya mencapai tujuan tersebut, data dikumpulkan melalui kuesioner daring yang telah dilabeli ke dalam kategori sentimen, kemudian diproses melalui tahapan *preprocessing* teks berbahasa Indonesia serta ekstraksi fitur menggunakan TF-IDF. Selanjutnya, proses klasifikasi dilakukan menggunakan kedua algoritma tersebut, dengan evaluasi kinerja yang mengacu pada confusion matrix serta metrik akurasi, precision, recall, dan F1-score guna menentukan model yang paling sesuai dalam menganalisis opini terhadap kebijakan pendidikan.

2. TINJAUAN PUSTAKA

Bagian ini memaparkan landasan teoretis, konsep-konsep pendukung, serta penelitian terdahulu yang relevan dengan topik penelitian sebagai pijakan dalam membangun kerangka konseptual penelitian. Selain itu, tinjauan pustaka digunakan untuk menemukan celah penelitian yang mendasari pelaksanaan studi ini.

2.1. Penelitian Terdahulu

Sejumlah studi sebelumnya menunjukkan bahwa pendekatan analisis sentimen dengan menggunakan *machine learning* banyak digunakan untuk mengkaji opini pengguna terhadap berbagai layanan digital dan sistem teknologi [8].

Perkembangan teknologi informasi yang pesat menyebabkan meningkatnya jumlah data teks yang dihasilkan pengguna, sehingga diperlukan metode yang efektif untuk mengolah dan menganalisis data tersebut [9].

Berbagai algoritma, seperti *Support Vector Machine*, *Logistic Regression*, hingga model berbasis *deep learning*, telah digunakan untuk melakukan klasifikasi teks dengan hasil yang cukup baik. Di sisi lain, metode klasifikasi konvensional masih sering dipilih karena memiliki keunggulan dalam hal kestabilan dan efisiensi komputasi, terutama pada data dengan jumlah fitur yang besar [9].

Sejumlah studi telah membandingkan kinerja algoritma klasifikasi pada tugas Natural Language Processing (NLP) [10].

Temuan dari beberapa penelitian menunjukkan bahwa algoritma *Support Vector Machine* (SVM) mampu memberikan kinerja yang baik, terutama ketika diterapkan pada data dengan jumlah fitur yang besar dan tingkat kompleksitas yang tinggi. [10].

Sementara penggunaan *Logistic Regression* dipadukan dengan teknik representasi teks seperti TF-IDF juga menunjukkan hasil yang konsisten dengan tingkat akurasi tinggi [10].

Hal ini menegaskan bahwa pemilihan algoritma serta representasi teks menjadi faktor penting dalam keberhasilan klasifikasi [11].

Pada berbagai domain seperti ulasan aplikasi, layanan digital, dan e-commerce, kinerja algoritma klasifikasi dipengaruhi oleh tahapan preprocessing serta karakteristik dataset yang digunakan [9], [10].

Tahapan *preprocessing* seperti tokenisasi, penghapusan *stopword*, dan *stemming* berperan dalam meningkatkan kualitas data teks sebelum klasifikasi. Beberapa studi menunjukkan bahwa penggunaan TF-IDF yang dikombinasikan dengan SVM dan *Logistic Regression* menghasilkan kinerja yang stabil. Penelitian pada teks berbahasa Indonesia juga menunjukkan bahwa pendekatan ini tetap efektif meskipun menghadapi keragaman bahasa yang tinggi [11], [12].

2.2. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan salah satu algoritma machine learning yang cukup dominan diterapkan dalam analisis sentimen karena memiliki kapabilitas yang baik dalam mengelola data berdimensi tinggi dengan jumlah fitur yang kompleks dan berukuran besar. Kemampuan tersebut membuat SVM mampu membentuk pemisahan kelas secara optimal sehingga performanya tetap stabil meskipun dihadapkan pada representasi fitur teks yang sangat banyak. [13].

Pendekatan ini memisahkan data ke dalam kelas yang berbeda dengan membentuk batas keputusan yang optimal, sehingga mampu meningkatkan kemampuan model dalam melakukan generalisasi [14].

Karakteristik tersebut membuat SVM cocok digunakan pada data teks yang umumnya memiliki jumlah fitur besar. Selain itu, SVM dikenal memberikan performa yang konsisten pada berbagai skenario klasifikasi [13].

Dalam penerapannya, SVM menggunakan konsep kernel untuk mengatasi data yang tidak dapat dipisahkan secara linear. Penerapan fitur kernel seperti linear, polynomial, dan radial basis function (RBF) memungkinkan proses pemetaan data ke dalam ruang fitur berdimensi lebih tinggi, sehingga struktur pola yang bersifat kompleks maupun nonlinier dapat terseparasi secara lebih efektif dan optimal.. Fleksibilitas ini membuat SVM digunakan dalam penelitian analisis sentimen berbasis teks [13].

Selain itu, SVM memiliki keunggulan dalam mengurangi risiko overfitting karena hanya bergantung pada support vectors [13].

Alam analisis sentimen, metode ini sering dipadukan dengan ekstraksi fitur TF-IDF untuk meningkatkan kinerja klasifikasi. Berbagai studi menunjukkan kombinasi tersebut menghasilkan akurasi tinggi dan konsisten, sehingga SVM masih banyak digunakan dalam tugas klasifikasi teks sampai sekarang [13].

2.3. *Logistic Regression*

Logistic Regression termasuk metode klasifikasi yang kerap digunakan dalam analisis sentimen karena strukturnya yang sederhana serta mudah untuk diinterpretasikan [15].

Algoritma ini bekerja dengan mengestimasikan probabilitas suatu data untuk termasuk ke dalam kategori kelas tertentu melalui penerapan fungsi logistik yang bersifat sigmoid [15].

Efisiensi komputasi dan stabilitas Logistic Regression pada berbagai dataset membuat Logistic Regression sering dijadikan baseline dalam penelitian klasifikasi teks [14], [15].

Logistic Regression kerap dipadukan dengan ekstraksi fitur TF-IDF untuk meningkatkan representasi data teks [16].

Pendekatan ini membantu model menangkap hubungan kata dan sentimen secara lebih efektif. Selain itu, tingkat interpretabilitas yang tinggi memudahkan analisis dan penjelasan hasil klasifikasi [15].

Logistic Regression diketahui memiliki kinerja yang cukup baik dalam tugas analisis sentimen berdasarkan berbagai studi terdahulu [16].

Meskipun sederhana, metode ini tetap relevan untuk pengolahan teks modern [14].

Selain itu, efisiensi waktu komputasinya membuat Logistic Regression menjadi pilihan efektif dalam berbagai studi klasifikasi teks.

2.4. TF-IDF (*Term Frequency-Inverse Document Frequency*)

TF-IDF digunakan untuk mengukur tingkat kepentingan suatu kata dalam dokumen. [17].

Metode ini mengkombinasikan frekuensi kemunculan kata (*term frequency*) dan tingkat keunikan kata dalam korpus (inverse document frequency). Pendekatan ini mampu menonjolkan kata yang paling informatif sehingga efektif sebagai representasi dasar data teks [17].

Dalam analisis sentimen, TF-IDF terbukti meningkatkan performa algoritma klasifikasi seperti SVM dan Logistic Regression [18].

Metode ini mengurangi pengaruh kata umum yang kurang relevan sekaligus memberi bobot lebih pada kata yang informatif [17], [18]

Selain sederhana dan mudah diimplementasikan, TF-IDF digunakan dalam berbagai penelitian analisis te [18].

Sejumlah penelitian terbaru menunjukkan bahwa TF-IDF tetap relevan meskipun telah muncul berbagai metode representasi teks berbasis deep learning [19].

Teknik ini masih mampu memberikan hasil yang kompetitif, terutama pada dataset berukuran kecil hingga menengah. Selain itu, kombinasi TF-IDF dengan algoritma klasifikasi tradisional terbukti menghasilkan model yang stabil dan akurat, sehingga tetap menjadi komponen penting dalam analisis sentimen [18], [19].

2.5. Sintesis Tinjauan Pustaka

Berbagai studi menunjukkan bahwa *Support Vector Machine* dan *Logistic Regression* mampu memberikan kinerja yang baik dalam analisis sentimen teks, terutama ketika dipadukan dengan TF-IDF. Meski demikian, penelitian yang ada umumnya masih berfokus pada domain produk, aplikasi, dan layanan digital. Analisis sentimen pada kebijakan pendidikan, khususnya terkait SNBP, masih relatif terbatas sehingga membuka peluang untuk penelitian lebih lanjut.

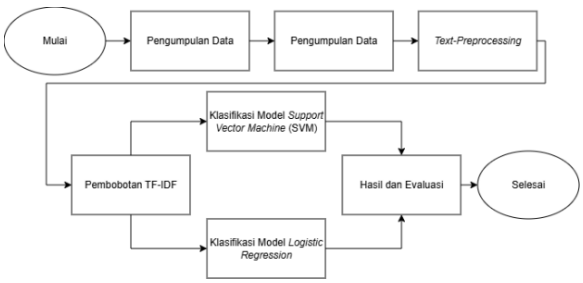
3. METODE PENELITIAN

Metode penelitian menguraikan tahapan yang ditempuh untuk mencapai tujuan studi secara terstruktur. Bagian ini memuat pendekatan penelitian, jenis penelitian, teknik pengumpulan data, serta metode analisis yang digunakan sehingga hasil yang diperoleh dapat dipertanggungjawabkan secara ilmiah. Penyusunan metode yang terstruktur membantu memastikan validitas dan reliabilitas penelitian sebelum memasuki pembahasan kerangka kerja penelitian.

3.1. Kerangka Kerja Penelitian

Dalam penelitian ini diterapkan alur kerja klasifikasi teks berbasis machine learning yang mencakup beberapa tahap, yaitu pengumpulan data, pemberian label sentimen, dan prapemrosesan sebelum model dibangun. Setelah pelabelan, dilakukan pembersihan dan normalisasi teks serta ekstraksi fitur menggunakan TF-IDF yang efektif merepresentasikan bobot kata [20].

Selanjutnya, model klasifikasi dikembangkan dengan menggunakan algoritma *Logistic Regression* dan *Support Vector Machine* (SVM), yang dalam berbagai penelitian komparatif menunjukkan performa kompetitif pada analisis sentimen berbasis teks. Seluruh tahapan tersebut disusun secara terpadu untuk memastikan proses eksperimen berjalan terukur serta menghasilkan model klasifikasi sentimen yang optimal berdasarkan evaluasi kinerja yang terstandarisasi.



Gambar 1. Kerangka kerja penelitian

Kerangka kerja ini bertujuan memberikan gambaran tahapan penelitian yang jelas agar mudah direplikasi. Alur yang sistematis membantu meminimalkan kesalahan prosedural serta mempermudah analisis dan pembahasan hasil.

Pendekatan ini mendukung validitas metodologis penelitian secara keseluruhan.

3.2. Pengumpulan Data

Sumber data pada penelitian ini dikumpulkan melalui kuesioner online berbasis Google Form yang diisi oleh responden. Pendekatan survei online dipilih karena telah banyak digunakan dalam penelitian kuantitatif sebagai metode yang efektif untuk menjangkau responden secara luas serta mempermudah proses distribusi instrumen tanpa batasan geografis [21].

Responden dalam penelitian ini terdiri dari siswa kelas XII SMA Negeri 1 Indralaya yang memiliki pengalaman dan pemahaman terkait prosedur SNBP. Selain itu, Google Form juga mempermudah rekapitulasi data secara otomatis sehingga pengumpulan data dapat dilakukan dengan lebih efektif [22].

Data yang dikumpulkan berupa tanggapan dan opini responden terhadap prosedur SNBP yang digunakan sebagai bahan analisis sentimen. Responden memberikan penilaian dalam bentuk rating serta ulasan tertulis yang mencerminkan persepsi mereka. Data hasil kuesioner kemudian diekspor ke format spreadsheet untuk diproses lebih lanjut, sehingga seluruh data bersumber dari pengalaman langsung responden.

3.3. Pelabelan Data

Pelabelan data dilakukan guna mengelompokkan tanggapan responden ke dalam kategori sentimen positif, netral, dan negatif berdasarkan rating yang diberikan saat pengisian kuesioner. Pada penelitian ini, rating 1 dan 2 diasosiasikan sebagai sentimen negatif karena merefleksikan kecenderungan respon yang bernada keberatan, ketidakpuasan, maupun hambatan terhadap prosedur SNBP. Rating 3 ditempatkan pada kategori netral sebab menunjukkan opini yang bersifat moderat atau belum condong ke salah satu sisi. Sementara itu, rating 4 dan 5 diklasifikasikan sebagai sentimen positif karena menggambarkan apresiasi responden terhadap kejelasan informasi serta alur pelaksanaan SNBP. Skema pelabelan ini diterapkan agar interpretasi sentimen lebih ajek dan dapat meminimalkan distorsi subjektivitas pada proses anotasi data.

Tabel 1. Contoh Pelabelan Data Sentimen

Rentang Rating	Label Sentimen	Keterangan
1–2	Negatif	Menggambarkan tanggapan yang cenderung menunjukkan ketidakpuasan, kebingungan, atau kendala terhadap prosedur SNBP.
3	Netral	Menunjukkan opini yang berada di posisi moderat atau belum condong ke sentimen positif maupun negatif.
4–5	Positif	Merepresentasikan tanggapan yang bernada apresiatif terhadap kejelasan informasi dan alur prosedur SNBP.

3.4. *Text Preprocessing*

Proses preprocessing dimulai dengan case folding, yaitu mengubah segala teks menjadi huruf kecil untuk meminimalkan perbedaan akibat penggunaan kapital [23].

Dilanjutkan dengan pembersihan data dengan menghilangkan angka, tanda baca, serta simbol yang tidak relevan. Tahap berikutnya adalah dengan menghilangkan kata-kata umum yang kurang bermakna [23].

Setelah itu, dilakukan tokenisasi dengan memecah teks menjadi unit kata sebagai dasar dalam pembobotan fitur dan proses klasifikasi [24].

3.5. Pembobotan TF-IDF

Setelah melewati tahapan preprocessing, data teks kemudian ditransformasikan ke dalam representasi numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) guna mengubah data tekstual menjadi vektor fitur yang dapat diproses oleh algoritma klasifikasi. Metode ini berfungsi untuk mengukur tingkat signifikansi suatu istilah dalam sebuah dokumen dengan mempertimbangkan frekuensi kemunculannya serta distribusinya di seluruh korpus [25].

TF-IDF menggabungkan frekuensi kemunculan kata dalam dokumen dan tingkat kelangkaannya pada seluruh dokumen, sehingga kata yang lebih diskriminatif memiliki bobot lebih tinggi [25].

Secara konseptual, TF-IDF merupakan teknik transformasi data teks tidak terstruktur menjadi representasi vektor numerik yang dapat diproses oleh algoritma *machine learning* [26].

Setiap dokumen dinyatakan sebagai vektor yang memuat bobot kata-kata penting. Representasi tersebut membantu model dalam mengenali pola distribusi kata pada tiap kategori sentimen. Hasil pembobotan akan digunakan sebagai fitur masukan pada tahap klasifikasi.

3.6. Klasifikasi *Logistic Regression*

Logistic Regression adalah metode klasifikasi yang menggunakan pendekatan probabilitas untuk membagi sentimen ke dalam kategori berupa positif, netral, dan negatif. Algoritma ini memodelkan probabilitas kelas menggunakan fungsi sigmoid dan banyak digunakan dalam klasifikasi teks [27].

Selain itu, algoritma ini mampu mengolah data berdimensi tinggi dari hasil TF-IDF serta menawarkan

kemudahan dalam interpretasi hasil dibandingkan metode lain [28].

Dalam penelitian ini, Logistic Regression dilatih menggunakan data yang telah melalui preprocessing dan pembobotan TF-IDF, kemudian diuji untuk mengukur akurasi prediksi sentimen. Beberapa parameter seperti jumlah iterasi maksimum dan class weighting digunakan untuk meningkatkan performa model. Hasilnya kemudian dibandingkan dengan Support Vector Machine (SVM) untuk melihat perbedaan kinerja dalam analisis sentimen [29].

3.7. Klasifikasi *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan algoritma klasifikasi yang menavigasikan *hyperplane* terbaik untuk memilah data antar kelas dengan memaksimalkan margin. Metode ini efektif untuk data teks berdimensi tinggi dan bersifat sparse seperti hasil TF-IDF, sehingga digunakan sebagai pembanding Logistic Regression dalam penelitian ini.

SVM dilatih menggunakan data yang sama dengan Logistic Regression untuk menjaga konsistensi, kemudian diuji untuk memperoleh nilai akurasi dan metrik evaluasi lainnya. Hasil evaluasi digunakan untuk menganalisis kemampuan model dalam mengklasifikasikan sentimen responden terhadap prosedur SNBP ke dalam kelas positif, netral, dan negatif.

3.8. *Confusion Matrix*

Proses evaluasi performa model dilakukan menggunakan *Confusion Matrix*, yaitu matriks evaluasi yang digunakan untuk merepresentasikan hubungan antara label aktual dengan hasil prediksi model secara terstruktur dalam bentuk tabel. Melalui pendekatan ini, dapat diidentifikasi jumlah prediksi benar dan salah pada tiap kategori sentimen serta pola kesalahan yang muncul.

Confusion Matrix juga digunakan sebagai acuan dalam menghitung metrik evaluasi seperti akurasi, presisi, dan recall. Akurasi menunjukkan tingkat ketepatan keseluruhan prediksi, sedangkan presisi dan recall memberikan gambaran performa tiap kelas. Dengan demikian, *Confusion Matrix* memungkinkan evaluasi model secara lebih komprehensif dalam menentukan model terbaik.

4. HASIL DAN PEMBAHASAN

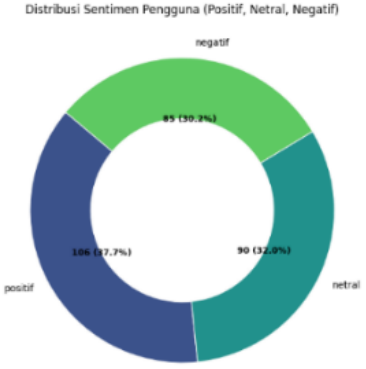
4.1. Pengumpulan Data

qqqData penelitian dikumpulkan melalui kuesioner daring menggunakan Google Form. Responden merupakan siswa kelas XII SMA Negeri 1 Indralaya tahun 2026 untuk memperoleh perspektif berdasarkan pengalaman langsung terhadap prosedur SNBP. Kuesioner disebarkan kepada total 392 siswa, namun sebanyak 281 responden memberikan jawaban dan data tersebut digunakan sebagai bahan analisis dalam penelitian ini.

4.2. Pelabelan Data

Data yang telah terkumpul selanjutnya melalui tahap pelabelan sentimen untuk mengelompokkan opini responden terhadap prosedur SNBP ke dalam tiga kategori, yaitu positif, netral, dan negatif. Proses pelabelan dilakukan berdasarkan rating yang diberikan responden pada saat pengisian kuesioner, dengan tetap mempertimbangkan konteks komentar untuk memastikan kesesuaian makna sentimen yang disampaikan. Pendekatan ini digunakan agar proses pelabelan lebih konsisten serta mampu meminimalkan subjektivitas dalam penentuan label sentimen.

Berdasarkan hasil pelabelan, diperoleh distribusi data sentimen yang terdiri atas 106 data sentimen positif, 90 data sentimen netral, dan 85 data sentimen negatif. Visualisasi distribusi sentimen responden dapat dilihat pada Gambar 2. Pelabelan distribusi sentimen data.



Gambar 2. Pelabelan distribusi sentimen data

Di Pelabelan data dilakukan dengan membagi tanggapan responden ke dalam tiga kategori sentimen berdasarkan rating yang diberikan. Rating 1–2 dikategorikan sebagai sentimen negatif karena umumnya mencerminkan ketidakpuasan terhadap prosedur SNBP, seperti alur yang dianggap rumit atau kendala teknis yang dialami. Rating 3 dimasukkan ke kategori netral karena menunjukkan penilaian yang "di tengah jalan", sedangkan rating 4–5 termasuk sentimen positif karena menggambarkan tanggapan yang cenderung baik terhadap kejelasan informasi dan proses SNBP. Selain rating, isi komentar juga dicermati agar makna opini yang disampaikan tetap sejalan dengan label sentimen yang diberikan Contoh hasil pelabelan data komentar, rating, dan label sentimen dapat diamati pada Tabel 1. Contoh Pelabelan Data Sentimen.

Tabel 2. Contoh Pelabelan Data Sentimen

No	Komentar	Rating	Label Sentimen
1	susah dipahami dan ribet ketika server ngelag	2	negatif
2	lumayan jelas arahannya, tapi server masih sering drop ketika jam 3 sore	3	netral
3	lumayanlah walaupun cukup kelabakan	3	netral
4	Prosedur SNBP mudah dipahami dan informasi yang beredar jelas	4	positif
5	sangat mudah dipahami dan usaha memang tidak mengkhianati hasil	5	positif

4.3. *Text Preprocessing*

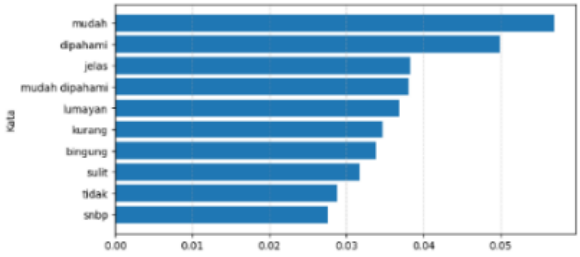
Data teks kuesioner diproses terlebih dahulu melalui tahap preprocessing sebelum dilakukan klasifikasi, yang mencakup case folding untuk mengubah huruf menjadi kecil, pembersihan tanda baca dan karakter yang tidak relevan, serta penghapusan kata-kata yang kurang bermakna. Tahap ini bertujuan menyeragamkan format teks agar lebih mudah diproses oleh algoritma klasifikasi. Hasilnya, data menjadi lebih bersih dan terstruktur sehingga meningkatkan kualitas representasi fitur pada tahap berikutnya.

Tabel 3. Hasil *text preprocessing*

Teks Sebelum Preprocessing	Teks Setelah Preprocessing
Baik namun tolong evaluasi dan ditingkatkan kualitas orang yang turun di lapangan agar tidak terjadi miskonsepsi dan misinformasi serta ...	tolong evaluasi tingkat kualitas orang turun lapang tidak miskonsepsi misinformasi tolong siap matang bijak libat siswa
Sangat sulit dipahami di berbagai penjelasan karena keterbatasan pemahaman tentang media yang diberikan (komputer dan sebagainya)	sulit paham jelas batas paham media komputer
Kejelasan pelaksanaan SNBP terkategori mudah untuk dipahami karena tahapan2 yang ada dalam pengisian data seleksi SNBP sangat runtut dan ...	jelas laksana snbp kategori mudah paham tahapan2 isi data seleksi snbp runtut sedia informasi mudah laksana nilai kurang nilai sistem snbp ...
Prosedur SNBP 2026 terstruktur rapi mulai dari PDSS sekolah → registrasi siswa → pilih prodi → seleksi rapor/prestasi → pengumuman ...	prosedur snbp 2026 struktur rapi pdss sekolah registrasi siswa pilih prodi seleksi raporprestasi umum lebih transparan bas data akademik ...

Dari tabel di atas, dapat terlihat bahwa tahap pra-pemrosesan ini berhasil untuk memfilter serta membuat data yang diolah menjadi satu format yang sama, semua data kini tidak lagi memiliki huruf kapital, tidak ada unsur selain unsur kata, seperti emoji, tanda baca, dan semacamnya, serta kata yang tidak memiliki arti secara tersendiri juga sudah tidak ada lagi.

4.4. Pembobotan TF-IDF



Gambar 3. Pembobotan TF-IDF

Setelah melewati tahap preprocessing, teks selanjutnya diproses menggunakan TF-IDF untuk menghasilkan representasi numerik berbobot. Kata atau istilah yang memiliki frekuensi kemunculan tinggi pada suatu dokumen, namun relatif jarang ditemukan pada dokumen lainnya, akan memperoleh bobot yang lebih besar karena dianggap memiliki tingkat representativitas informasi yang lebih signifikan.. Pendekatan ini memungkinkan model mengenali kata-kata penting yang berkontribusi terhadap sentimen, sehingga data dapat langsung dimanfaatkan dalam proses pelatihan model klasifikasi.

Berdasarkan gambar 3 tersebut, menampilkan bahwa beberapa kata seperti “mudah” memiliki bobot tertinggi dibandingkan kata yang lain. Dalam kasus ini, semakin tinggi angka pembobotan TF-IDF, semakin penting pula kata tersebut dalam konteks dokumen tersebut, namun jarang muncul di dokumen lain.

4.5. Klasifikasi *Logistic Regression*

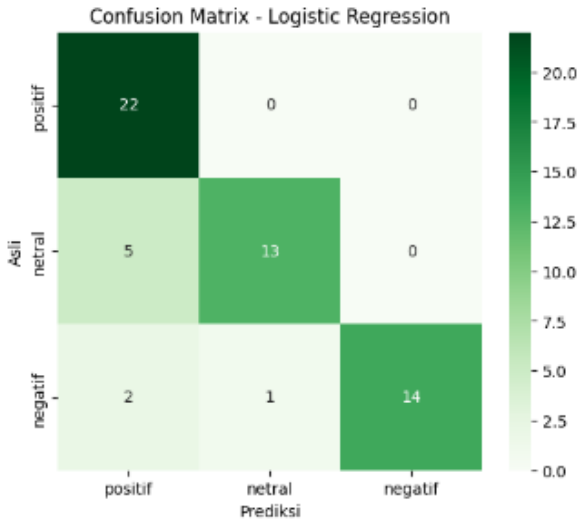
Data yang telah diubah ke bentuk numerik melalui TF-IDF selanjutnya digunakan untuk melatih model Logistic Regression. Model ini memperkirakan peluang suatu teks termasuk ke dalam kelas sentimen tertentu berdasarkan kombinasi linier dari fitur yang tersedia. Logistic Regression dipilih karena mampu mengolah data teks berdimensi tinggi dan menghasilkan kinerja yang relatif stabil. Untuk evaluasi, data dipisah menjadi data latih dan data uji guna menilai performa model.

Tabel 4. Akurasi klasifikasi *logistic regression*

Kelas	Precision	Recall	F1-Score	Support
negatif	1.00	0.82	0.90	17.0
netral	0.93	0.72	0.81	18.0
positif	0.76	1.00	0.86	22.0
accuracy	0.86	0.86	0.86	0.86
macro avg	0.90	0.85	0.86	57.0
weighted avg	0.88	0.86	0.86	57.0
Akurasi Logistic Regression :			**85.96%**	

Model Logistic Regression berhasil mencatat tingkat akurasi sebesar 85,96%, yang menunjukkan bahwa performa model sudah berada pada jalur yang cukup baik dalam melakukan klasifikasi sentimen terkait prosedur SNBP. Pada kategori sentimen positif, model memperoleh nilai precision sebesar 0,76, recall sebesar 1,00, dan f1-score sebesar 0,86, sehingga dapat dikatakan bahwa model mampu menangkap hampir seluruh opini bernada dukungan tanpa banyak yang terlewat. Di sisi lain, kelas netral memperoleh precision sebesar 0,93, recall sebesar 0,72, dan f1-score sebesar 0,81, yang mengindikasikan bahwa model masih belum sepenuhnya "pas di tengah jalan" dalam membedakan opini netral dari sentimen lainnya. Sementara itu, untuk sentimen negatif diperoleh precision sebesar 1,00, recall sebesar 0,82, dan f1-score sebesar 0,90, sehingga mayoritas komentar bernada kritik dapat dikenali dengan baik oleh model tanpa banyak kesalahan klasifikasi.

Berdasarkan confusion matrix, model Logistic Regression tampil cukup solid dalam mengenali sentimen positif dengan seluruh 22 data berhasil dipetakan secara tepat. Pada kelas netral, model masih "sedikit meleset arah" karena hanya 13 data yang terklasifikasi benar, sementara 5 data lainnya bergeser ke kelas positif. Adapun pada sentimen negatif, sebanyak 14 data berhasil dikenali dengan benar, sedangkan 2 data diprediksi positif dan 1 data masuk ke kategori netral. Secara umum, performa model sudah tergolong baik dalam membedakan ketiga jenis sentimen.



Gambar 4. *Confusion Matrix – Logistic Regression*

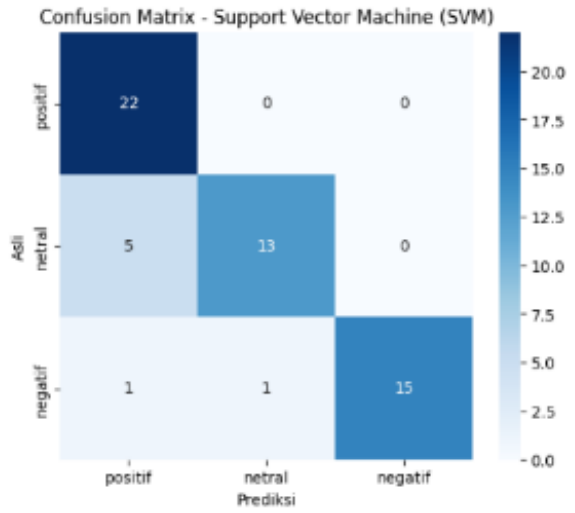
4.6. Klasifikasi *Support Vector Machine* (SVM)

Support Vector Machine (SVM) sebagai metode pembanding dalam klasifikasi sentimen terhadap prosedur SNBP juga dipakai dalam penelitian ini. SVM bekerja dengan membentuk batas pemisah antar kelas yang optimal melalui pemaksimalan margin, sehingga efektif digunakan pada data berdimensi tinggi seperti hasil TF-IDF.

Tabel 5. Akurasi klasifikasi *Support Vector Machine*

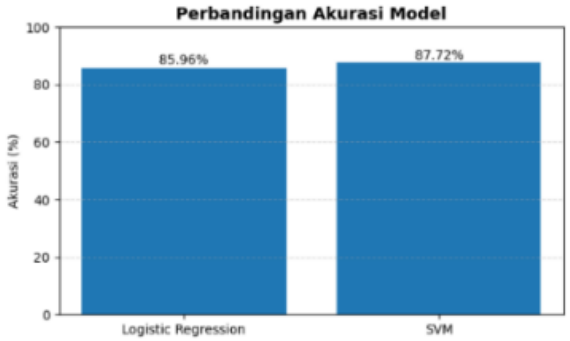
Kelas	Precision	Recall	F1-Score	Support
negatif	1.00	0.88	0.94	17.0
netral	0.93	0.72	0.81	18.0
positif	0.79	1.00	0.88	22.0
accuracy	0.88	0.88	0.88	0.88
macro avg	0.90	0.87	0.88	57.0
weighted avg	0.89	0.88	0.88	57.0
Akurasi SVM :			**87.72%**	

Hasil pengujian memperlihatkan bahwa model SVM mencatat akurasi sebesar 87,72%, yang menandakan performanya sudah cukup bisa diandalkan dalam memetakan sentimen terkait prosedur SNBP. Pada kelas positif, model memperoleh precision 0,79, recall 1,00, dan f1-score 0,88, sehingga hampir seluruh opini bernada dukungan berhasil ditangkap tanpa ada yang luput. Untuk kelas netral, diperoleh precision 0,93, recall 0,72, dan f1-score 0,81, yang menunjukkan bahwa model masih sedikit "abu-abu" ketika membedakan opini netral dengan sentimen lain. Sementara itu, pada kelas negatif, model menghasilkan precision 1,00, recall 0,88, dan f1-score 0,94, sehingga sebagian besar komentar bernada kritik dapat dikenali dengan cukup rapi dan minim salah klasifikasi.



Gambar 5. *Confusion Matrix – SVM*

Berdasarkan confusion matrix, model SVM menunjukkan performa yang cukup matang dalam memetakan tiap kelas sentimen. Pada kategori positif, seluruh 22 data berhasil diprediksi dengan tepat tanpa terjadi salah klasifikasi. Untuk kelas netral, model mampu mengenali 13 data secara benar, meskipun masih ada 5 data yang bergeser ke kelas positif. Sementara itu, pada sentimen negatif terdapat 15 data yang berhasil diklasifikasikan dengan benar, sedangkan 1 data diprediksi sebagai positif dan 1 data lainnya masuk ke kategori netral. Secara keseluruhan, model SVM terlihat lebih konsisten dalam membaca pola sentimen, terutama pada kelas negatif dan positif.



Gambar 6. Perbandingan akurasi Logistic Regression dan *Support Vector Machine* (SVM)

Hasil evaluasi mengindikasikan bahwa SVM memperoleh akurasi sebesar 87.72%, lebih tinggi dibandingkan Logistic Regression sebesar 85.96%. Selain itu, SVM juga menghasilkan kinerja klasifikasi yang lebih seimbang dan representatif pada setiap kelas sentimen, sedangkan Logistic Regression masih cenderung bias terhadap kelas positif. Temuan ini menunjukkan bahwa SVM lebih optimal dalam mengenali variasi sentimen pada data opini prosedur SNBP.

4.7. *Wordcloud*

WordCloud digunakan sebagai visualisasi data teks untuk menampilkan kata berdasarkan frekuensi kemunculannya dalam dokumen. Kata dengan frekuensi kemunculan lebih tinggi sering muncul ditampilkan dengan ukuran lebih besar sehingga memudahkan identifikasi kata dominan pada setiap kategori sentimen. Visualisasi ini membantu memberikan gambaran umum kecenderungan opini responden terhadap prosedur SNBP secara intuitif.

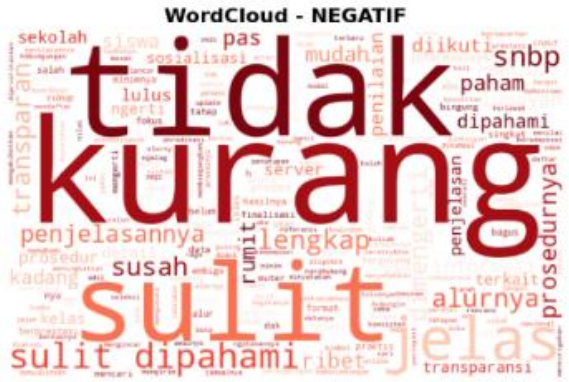
Gambar 7. *Wordcloud* sentimen positif

Berdasarkan WordCloud sentimen positif, kata seperti “mudah”, “paham”, dan “prosedur” tampak paling dominan. Hal ini menunjukkan bahwa sebagian besar responden menilai proses SNBP cukup jelas dan tidak terlalu rumit untuk dipahami. Selain itu, munculnya kata “nilai”, “informasi”, dan “sekolah” mengindikasikan bahwa peserta merasa alur seleksi serta penyampaian informasinya sudah cukup membantu. Secara umum, tanggapan positif yang muncul menggambarkan bahwa sistem SNBP dianggap lebih mudah dipahami dibanding “jalan berliku” yang sering dibayangkan peserta.

Gambar 8. *Wordcloud* Sentimen Netral

Berdasarkan WordCloud sentimen netral, kata seperti “tidak”, “kurang”, dan “bingung” terlihat paling dominan. Hal ini menunjukkan bahwa tanggapan responden lebih banyak berada di posisi “di tengah jalan”, yakni belum sepenuhnya positif maupun negatif. Beberapa peserta masih merasa kurang paham terhadap alur SNBP, meskipun terdapat kata seperti

"mudah" dan "transparan" yang menunjukkan adanya penilaian cukup baik terhadap sistem. Secara umum, sentimen netral menggambarkan adanya campuran pendapat, sehingga respons yang muncul cenderung abu-abu dan saling tumpang tindih.



Gambar 9. *Wordcloud* sentimen negatif

Berdasarkan WordCloud sentimen negatif, kata seperti "tidak", "kurang", dan "sulit" muncul paling dominan. Hal ini menunjukkan bahwa masih banyak responden merasa proses SNBP belum cukup jelas dan mudah dipahami. Kemunculan kata "bingung", "transparansi", serta "penjelasannya" juga mengarah pada keluhan terkait alur dan penyampaian informasi. Selain itu, kata seperti "server" dan "jadwal" menandakan adanya kendala teknis selama proses berlangsung. Secara umum, tanggapan negatif menggambarkan bahwa sebagian peserta merasa proses SNBP masih "jalan terjal" karena informasi dan mekanismenya dianggap belum sepenuhnya terang benderang.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian, metode Support Vector Machine (SVM) menunjukkan performa yang sedikit lebih unggul dibandingkan *Logistic Regression* dalam klasifikasi sentimen prosedur SNBP. SVM memperoleh akurasi 87,72%, sedangkan Logistic Regression mencapai 85,96%. Selain lebih akurat, SVM juga lebih mantap dalam membedakan tiap kelas sentimen, sementara Logistic Regression masih cenderung berat sebelah pada sentimen positif. Hasil analisis memperlihatkan bahwa mayoritas responden memberikan tanggapan positif terhadap kejelasan informasi dan alur SNBP, meskipun masih ditemukan opini netral serta negatif terkait transparansi dan kendala teknis. Temuan ini menunjukkan bahwa SVM lebih adaptif dalam membaca ragam opini responden. Oleh sebab itu, penelitian selanjutnya disarankan menggunakan data yang lebih besar dan distribusi sentimen yang lebih merata agar hasil klasifikasi dapat semakin optimal.

DAFTAR PUSTAKA

[1] V. Umarani, A. Julian, and J. Deepa, "Sentiment analysis using various machine learning and deep learning techniques," *Journal of the Nigerian Society of Physical Sciences*, pp. 385–394, 2021.

[2] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," *Inf. Process. Manag.*, vol. 59, no. 2, p. 102798, 2022.

[3] D. Ogaga and A. Olalere, "Evaluation and comparison of svm, deep learning, and naïve bayes performances for natural language processing text classification task," 2023.

[4] F. Sun and G. Shi, "Study on the application of big data techniques for the third-party logistics using novel support vector machine algorithm," *Journal of Enterprise Information Management*, vol. 35, no. 4–5, pp. 1168–1184, 2022.

[5] R. D. Cahyani and P. T. Prasetyaningrum, "Sentiment Analysis of User Reviews for AI Applications: Evaluating SVM, Logistic Regression, and Random Forest," *Journal of Information Systems and Informatics*, vol. 8, no. 1, pp. 1–27, 2026.

[6] S. Jatmiko and C. Dometian, "Analisis Sentimen Ulasan Google Play Store: Studi Komparatif Algoritma SVM, Naïve Bayes, dan Logistic Regression," *JURNAL FASILKOM*, vol. 15, no. 3, pp. 610–620, 2025.

[7] K. Apipah, M. Martanto, F. Dikananda, A. Bahtiar, and R. Narasati, "PERBANDINGAN KINERJA SUPPORT VECTOR MACHINE ORIGINAL DAN SUPPORT VECTOR MACHINE BERBASIS SMOTE UNTUK ANALISIS SENTIMEN APLIKASI JKN MOBILE," *JURNAL TEKNOLOGI INFORMASI DAN KOMUNIKASI*, vol. 17, no. 1, pp. 20–26, 2026.

[8] P. T. Prasetyaningrum, "Sentiment Analysis and Classification of User Reviews of the'Access by KAI'Application Using Machine Learning Methods to Improve Service Quality," *Journal of Information Systems and Informatics*, vol. 7, no. 2, pp. 1418–1442, 2025.

[9] M. I. Adrian and E. A. Laksana, "Optimalisasi Parameter Support Vector Machine dengan Algoritma PSO untuk Tugas Klasifikasi Sentimen Ulasan IMDb," *Jurnal Algoritma*, vol. 22, no. 1, pp. 288–299, 2025.

[10] R. Rosita and P. T. Prasetyaningrum, "Comparative Analysis of Machine Learning Algorithms for Sentiment Classification of Discord App Reviews," *Journal of Information Systems and Informatics*, vol. 7, no. 4, pp. 4384–4406, 2025.

[11] S. Butsianto and A. M. Rifa'i, "Analisis Sentimen Ulasan Aplikasi Jamsostek dengan SVM, Random Forest, dan Logistic Regression," *Jurnal Informatika Ekonomi Bisnis*, pp. 700–706, 2025.

[12] A. A. Qolbu, N. Fitriyati, and N. Inayah, "Performa Naïve Bayes, SVM, dan IndoBERT

pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang," *Jurnal Fourier*, vol. 14, no. 1, pp. 29–44, 2025.

[13] A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, "Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema," *Sci. Rep.*, vol. 15, no. 1, p. 35038, 2025.

[14] L. P. Mudarakola, R. K. Gatla, A. S. N. Raju, A. Y. Jaffar, A. Alzahrani, and A. A. Dessalegn, "Multi stage sentiment analysis for product reviews on Twitter using optimized machine learning algorithm," *Sci. Rep.*, vol. 15, no. 1, p. 39777, 2025.

[15] D.-M. Cristea, I. Sima, and L. B. Iantovics, "Comparative analysis of optimized logistic regression with state-of-the-art models for complex gastroenterological image analysis," *Front. Med. (Lausanne).*, vol. 12, p. 1655612, 2025.

[16] A. Riswawan, "Implementing TF-IDF and Logistic Regression for sentiment analysis of YouTube comments on the iPhone 16," *Jurnal Teknologi dan Open Source*, vol. 8, no. 2, pp. 604–611, 2025.

[17] R. N. Rathi and A. Mustafi, "The importance of Term Weighting in semantic understanding of text: A review of techniques," *Multimed. Tools Appl.*, vol. 82, no. 7, pp. 9761–9783, 2023.

[18] K. P. Harmandini, "Analysis of TF-IDF and TF-RF feature extraction on product review sentiment," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 8, no. 2, pp. 929–937, 2024.

[19] T. F. ABDILLAH, "Comparison of TF-IDF and GloVe Word Embedding for Sentiment Analysis of 2024 Presidential Candidates-Dalam bentuk buku karya ilmiah," 2024.

[20] V. B. Lestari and C. A. Hutagalung, "Evaluation of TF-IDF Extraction Techniques in Sentiment Analysis of Indonesian-Language Marketplaces Using SVM, Logistic Regression, and Naive Bayes," *J-KOMA: Jurnal Ilmu Komputer dan Aplikasi*, vol. 8, no. 1, pp. 36–44, 2025.

[21] E. Grewenig, P. Lergetporer, L. Simon, K. Werner, and L. Woessmann, "Can internet surveys represent the entire population? A practitioners' analysis," *Eur. J. Polit. Econ.*, vol. 78, p. 102382, 2023.

[22] J. A. Beto, P. Gleason, J. E. Harris, and E. Metallinos-Katsaras, "Electronic survey methodology for data collection and analysis in nutrition and dietetics research," *J. Acad. Nutr. Diet.*, vol. 125, no. 5, pp. 603–614, 2025.

[23] N. M. Diaa, S. S. Ahmed, H. M. Salman, and W. A. Sajid, "Statistical challenges in social media data analysis sentiment tracking and beyond," *Journal of Ecohumanism*, vol. 3, no. 5, pp. 365–384, 2024.

[24] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[25] S. Samidi and D. Fatmawati, "Sentiment classification of IT service feedback via TF-IDF," *CogITo Smart Journal*, vol. 10, no. 2, pp. 403–417, 2024.

[26] K. P. Harmandini, "Analysis of TF-IDF and TF-RF feature extraction on product review sentiment," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 8, no. 2, pp. 929–937, 2024.

[27] K. Ferhati, A. Burlea-Schiopoiu, and A.-G. Nascu, "A Text-Based Project Risk Classification System Using Multi-Model AI: Comparing SVM, Logistic Regression, Random Forests, Naive Bayes, and XGBoost," *Systems*, vol. 13, no. 12, p. 1078, 2025.

[28] A. Bahar, T. Astuti, and P. Arsi, "Performance Comparison Of Svm, Naive Bayes, And Logistic Regression Classification Algorithms In Analyzing Noice App User Reviews," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 469–477, 2024.

[29] S. F. Kadir and A. Fairuzabadi, "Analisis Sentimen Ulasan Shopee di Google Play dengan TF-IDF dan Logistic Regression," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 7940–57945, 2025