Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 3, September 2025, pp. 998~1007
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i3.7289

Sentiment Analysis of Reading Difficulties in Grade 7 Secondary School Students Using the Support Vector Machine (SVM) Algorithm

Rasyid Zanur Algipari[1], Shandy Tresnawati[2]
[1] Program Studi Teknik Informatika, Politeknik TEDC Bandung, Cimahi, 40513, Indonesia
[2] Program Studi Teknik Komputer, Politeknik TEDC Bandung, Cimahi, 40513, Indonesia

Informasi Artikel

Diterima : 30 Juni 2025
Revisi : 01 Oktober 2025
Publikasi : 30 September 2025

Kata Kunci:

Analisis Sentimen
Support Vector Machine (SVM)
lexicon-based
TF-IDF
Kesulitan Membaca
YouTube

ABSTRAK

Kemampuan membaca siswa SMP di Indonesia masih tergolong rendah, terbukti dari hasil Asesmen Nasional dan viralnya kasus 29 siswa SMP kelas VII di Pangandaran yang belum lancar membaca. Penelitian ini bertujuan untuk menganalisis persepsi masyarakat terhadap fenomena tersebut melalui pendekatan analisis sentimen berbasis teks komentar YouTube dan menggunakan algoritma Support Vector Machine (SVM). Langkah - langkah metode SEMMA diterapkan, dimulai dengan pengumpulan data komentar, praproses teks menggunakan metode berbasis kamus dan TF-IDF, dan terakhir klasifikasi menggunakan SVM. Dataset yang digunakan meliputi 1.055 komentar. Hasil penelitian menunjukkan bahwa algoritma SVM mampu mengklasifikasikan sentimen ke dalam tiga kategori (positif, negatif, dan netral) dengan akurasi sebesar 87%. Studi ini berkontribusi pada pengenalan pendekatan analisis sentimen berdasarkan komentar YouTube menggunakan metode leksikon dan algoritma SVM untuk menyelidiki kemahiran membaca siswa sekolah menengah. Hal baru dari studi ini adalah penggunaan data opini publik digital sebagai sumber data untuk menilai reaksi publik terhadap isu-isu pendidikan. Model klasifikasi ini dapat berfungsi sebagai referensi untuk pengembangan sistem pemantauan literasi berbasis AI dan pembuatan kebijakan yang didorong oleh opini publik. Hasil ini menunjukkan bahwa mayoritas masyarakat peduli terhadap mutu pendidikan dasar. Penelitian ini memberikan kontribusi terhadap pemahaman komputasional persepsi masyarakat dan dapat dijadikan acuan pedoman literasi berbasis data.

ABSTRACT

The reading ability of junior high school students in Indonesia is still relatively low, as evidenced by the results of the National Assessment and the viral case of 29 grade VII junior high school students in Pangandaran who are not yet fluent in reading. This study aims to analyze public perception of this phenomenon through a sentiment analysis approach based on YouTube comment text and using the Support Vector Machine (SVM) algorithm. The steps of the SEMMA method are applied, starting with collecting comment data, preprocessing the text using dictionary-based methods and TF-IDF, and finally classification using SVM. The dataset used includes 1,055 comments. The results of the study show that the SVM algorithm is able to classify sentiment into three categories (positive, negative, and neutral) with an accuracy of 87%. This study contributes to the introduction of a sentiment analysis approach based on YouTube comments using the lexicon method and the SVM algorithm to investigate the reading proficiency of high school students. What is new about this study is the use of digital public opinion data as a data source to assess public reactions to education issues. This classification model can serve as a reference for the development of an AI-based literacy monitoring system and policy making driven by public opinion. These results indicate that the majority of the public cares about the quality of basic education. This research contributes to the computational understanding of public perception and can be used as a reference for data-based literacy guidelines.

Penulis Koresponden

Email: nasir@binadarma.ac.id

Cara sitasi IEEE:

R. Z. Algipari, S. Tresnawati, "Sentiment Analysis of Reading Difficulties in Grade 7 Secondary School Students Using the Support Vector Machine (SVM) Algorithm," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 3, pp. 998-1008, September 2025. doi: 10.30811/jaise.v5i3.7289

1. PENDAHULUAN

Literasi membaca di Indonesia masih menjadi tantangan besar, terutama di kalangan remaja usia sekolah. Menurut laporan PISA [1] Indonesia berada di peringkat ke-74 dari 79 negara dalam literasi membaca. Hanya sekitar 30% siswa usia 15 tahun yang mencapai tingkat kemahiran membaca minimum. Data ini menunjukkan bahwa mayoritas remaja Indonesia belum memiliki keterampilan membaca yang memadai untuk memahami teks yang panjang dan kompleks.

Berdasarkan Laporan [2] berdasarkan hasil Asesmen Nasional 2022, hanya 59% siswa SMP di Indonesia yang memiliki kemampuan membaca dan menulis di atas standar minimal. Artinya, masih ada 41% siswa yang belum mencapai ambang batas membaca optimal. Data ini menunjukkan bahwa tantangan literasi sangat nyata, bahkan di SMP. Meskipun siswa memiliki kemampuan membaca dasar, pemahaman teks informatif, analitis, dan panjang masih menjadi tantangan utama.

Masalah ini menjadi perhatian publik pada tahun 2024 ketika sebuah video viral di YouTube dan media sosial lainnya menunjukkan 29 siswa kelas tujuh di sebuah sekolah menengah pertama SMPN 1 Mangunjaya di daerah Pangandaran, yang belum dapat membaca dengan lancar [3]. Video tersebut menarik perhatian luas karena menyoroti kekurangan sistemik dalam pendidikan dasar yang berdampak langsung pada siswa sekolah menengah. Reaksi publik beragam, mulai dari kekhawatiran hingga kritik terhadap sistem pendidikan, yang dianggap tidak efektif dalam menjamin keterampilan dasar siswa.

Selain itu, bacaan yang dianalisis dalam penelitian ini berupa komentar publik di platform YouTube yang membahas kesulitan membaca siswa SMP. Komentar-komentar tersebut dipilih karena merepresentasikan opini masyarakat secara langsung dan spontan terhadap isu tersebut. Analisis dilakukan pada teks berbahasa Indonesia, sebab siswa yang menjadi subjek pembahasan bersekolah di Indonesia dengan bahasa Indonesia sebagai bahasa pengantar utama dalam pembelajaran. Oleh karena itu, kemampuan memahami dan menguasai bahasa Indonesia menjadi indikator penting dalam menilai keterampilan literasi membaca siswa.

Fenomena ini menimbulkan pertanyaan serius tentang penyebab rendahnya kemampuan membaca di sekolah menengah, meskipun kurikulum mengharuskan siswa untuk menguasai keterampilan ini di sekolah dasar. Apakah penyebabnya karena kualitas pengajaran di prasekolah, keterlibatan orang tua yang rendah, konteks sosial, atau keterbatasan akses ke bahan bacaan.

Media sosial telah menjadi ruang terbuka di mana masyarakat dapat menyampaikan pendapat dan keluhan mereka tentang isu pendidikan. Pernyataan publik dalam komentar di YouTube, Twitter, dan platform lainnya dapat dianalisis lebih lanjut untuk mengetahui opini publik tentang isu disleksia.

Untuk mendapatkan pemahaman yang lebih mendalam tentang persepsi publik, penelitian ini menggunakan pendekatan analisis sentimen dengan algoritma Support Vector Machine (SVM). Algoritma ini terbukti efektif dalam mengklasifikasikan opini ke dalam kategori positif, negatif, dan netral. Penelitian sebelumnya oleh Hidayat [4] menunjukkan bahwa SVM dapat mengklasifikasikan sentimen dengan akurasi tinggi meskipun terjadi ketidakseimbangan data. Penelitian oleh Harnelia [5] menunjukkan bahwa SVM dapat mengklasifikasikan ulasan produk dengan akurasi tinggi menggunakan kernel linier.

Dengan menggunakan metode ini, penelitian ini bertujuan untuk mengungkap pola opini publik di media sosial mengenai kasus viral terkait siswa sekolah menengah dengan disleksia. Diharapkan hasil

penelitian akan memberikan wawasan baru bagi dunia pendidikan dan mendorong pengembangan langkah-langkah yang lebih tepat sasaran untuk mengatasi kesulitan membaca di Indonesia..

2. METODE

Penelitian ini mengacu pada metode SEMMA (Sample, Explore, Modify, Model, Assess) dalam proses Data Mining sebagai metodologi penelitian[6]. Fokus utama dalam metode ini adalah pengambilan sampel data, eksplorasi, modifikasi, pemodelan, dan evaluasi dalam menganalisis sentimen terhadap kesulitan membaca siswa kelas 7 SMP menggunakan algoritma Support Vector Machine (SVM). Metode ini membantu penelitian dalam memberikan solusi untuk tujuan maupun masalah bisnis. Maka dari itu Penelitian ini menggunakan metode SEMMA dengan tahapan sebagai berikut [7].

2.1 *Sample*

Pada tahap ini dilakukan pengumpulan *dataset* yang bersumber dari komentar video *YouTube* tvOneNews yang berjudul "*Waduh, Puluhan Siswa SMP di Pangandaran Tidak Lancar Membaca*" dengan teknik scraping komentar. Komentar yang didapat berjumlah 1.055 komentar dengan menggunakan *Google Colab* yang memanfaatkan fungsi *YouTube API* komentar-komentar tersebut disimpan didalam file csv.

Langkah yang akan dilakukan oleh peneliti adalah melakukan login atau masuk akun google untuk mengakses fitur-fitur yang disediakan oleh *google* seperti *YouTube API*, *google colab* serta aplikasi-aplikasi pendukung lainnya. Kemudian, dari fitur-fitur tersebut peneliti dapat menggunakannya untuk proses *crawling* data dengan menggunakan Bahasa pemrograman *python*. Tabel 1. menunjukan *dataset* yang diambil dari komentar *Youtube*.

Tabel 1. Komentar *Youtube*

No	Author	Published at
1	@YohanesDwi-m6k	2024-12-12T07:08:50Z
2	@antoniandicharoli3571	2024-08-05T11:55:02Z
3	@fedisekdes8289	2023-09-20T12:26:13Z
…	…	…
1055	@pandalucu2210	2023-08-07T07:53:27Z

2.2 Explore

Pada tahap ini peneliti mendeskripsikan data yang didapat dari hasil *crawling* / *scraping* data komentar video *YouTube* tvOenNews yang berjudul "Waduh, Puluhan Siswa SMP di Pangandaran Tidak Lancar Membaca". Kemudian dilakukan tahap visualisasi data yaitu data yang didapat dibuat grafik berdasarkan waktu untuk menampilkan informasi dari data secara visual [8].

2.3 Modify

Pada tahap Modify di penelitian ini adalah tahapan untuk persiapan data atau *pre-processing*. Tahapan yang akan dilakukan oleh peneliti terdiri dari *case folding, cleaning, tokenize,* dan *normalize*.

2.3.1 Case folding

Langkah pertama pada tahap modify pada penelitian ini adalah *case folding*. Tahap *case folding* dilakukan untuk merubah semua karakter kata dalam *dataset* menjadi huruf kecil. Pada tahap ini dilakukan *case folding* menggunakan *python*. Setelah dilakukan proses tersebut maka didapatkan hasil data set yang sudah tidak mempunyai lagi huruf kapital [9][10].

Tabel 2. Hasil *Case Folding*

No	Author	Published at
1	@YohanesDwi-m6k	2024-12-12T07:08:50Z
2	@antoniandicharoli3571	2024-08-05T11:55:02Z
3	@fedisekdes8289	2023-09-20T12:26:13Z
…	…	…
1055	@pandalucu2210	2023-08-07T07:53:27Z

2.3.2 Cleaning

Selanjutnya adalah membersihkan data dari komponen yang tidak relevan dan tidak memiliki makna seperti ascii, angka, link, hastag, url, tanda baca dan whitespace [11].

Tabel 3. Hasil *Cleaning*

Sebelum Cleaning	Sesudah Cleaning
Terimakasih mba nana, kontennya sangat menginspirasi, semoga bisa bermanfaat untuk kita semuaâ•¤	terimakasih mba nana, kontennya sangat menginspirasi, semoga bisa bermanfaat untuk kita semua
Terimakasih mba nana, kontennya sangat menginspirasi, semoga bisa bermanfaat untuk kita semuaâ•¤	terimakasih mba nana, kontennya sangat menginspirasi, semoga bisa bermanfaat untuk kita semua
Terimakasih mba nana, kontennya sangat menginspirasi, semoga bisa bermanfaat untuk kita semuaâ•¤	terimakasih mba nana, kontennya sangat menginspirasi, semoga bisa bermanfaat untuk kita semua

2.3.3 Tokenize

Dilakukan untuk memecah sekumpulan kata (kalimat) menjadi kata yang memiliki arti tertentu (token). Hal ini bertujuan agar kata perkata dapat diolah lebih lanjut dalam proses-proses selanjutnya[12][13].

2.3.4 Normalize

Tahap berikutnya berikutnya adalah normalisasi atau *normalize*. Didalam penelitian ini, tahap normalisasi dilakukan untuk menstandarisasi kata yang memiliki makna yang sama dengan melakukan perubahan penulisan kata yang disingkat dan atau tidak baku. Penelitian ini menggunakan sebuah kamus normalisasi yang didapatkan dari kamus NLP (*Neuro Linguistic Programming*) bahasa Indonesia Resource [14] Tabel 3.5 berikut ini adalah hasil *normalize* pada *dataset* dalam penelitian.

Tabel 4. Hasil Normalize

Sebelum Tokenize	Sesudah Tokenize
[‘siapapun’,’yang’,’baca’,’komen’, ’saya’,’tanamlah’,’pohon’,’tolong’]	[‘siapa’,’pun’,’yang’,’baca’,’komentar’, ’saya’,’tanam’,’lah’,’pohon’,’tolong’]

2.3.5 Stopword Removal

Stopword Removal bertujuan untuk menghapus kata-kata umum yang banyak digunakan namun tidak memberikan pengaruh sentimen pada suatu kalimat. Proses stopword yang digunakan pada penelitian ini adalah dengan memanfatkan library dari Sastrawi yang di dalamnya terdapat corpus stopword bahasa Indonesia[12].

Tabel 5. Hasil *Stopword Removal*

Sebelum Stopword Removal	Sesudah Stopword Removal
[‘siapa’,’pun’,’yang’,’baca’,’komentar, ’saya’,’tanam’,’lah’,’pohon’,’tolong’]	[‘siapa’,’yang’,’baca’,’komentar’, ’saya’,’tanam’,’pohon’,’tolong’]

2.3.6 Lexion Based

Metode lexicon based pada penelitian ini digunakan untuk melakukan klasifikasi kelas pada *dataset*. Suatu komentar atau dokumen pada *dataset* diklasifikasikan ke dalam dua kelas, apakah komentar tersebut termasuk pada kelas positif atau negatif. Kata-kata yang ada dalam komentar terdapat di dalam kamus lexicon,maka kata tersebut akan diberikan nilai score. Jumlah nilai score pada suatu komentar yang menentukan label positif atau negatif. berikut ini adalah hasil dari *dataset* yang sudah melalui pelabelan metode lexicon based[15].Untuk melihat struktur aplikasi ReHome yang digambarkan melalui Class Diagram, yaitu berupa gambaran koneksi atau keterhubungan antar class atau objek yang ada pada alur sistem suatu aplikasi. Class diagram menunjukkan sistem class, metode, atribut, dan hubungan antar objek [15] . Pada gambar dibawah ini menunjukkan bagaimana tiap komponen saling terhubung serta atribut utama yang dimiliki oleh masing-masing class, serta aktifitas apa saja yang bisa dilakukan pada masing-masing fitur aplikasi.

Tabel 6. Hasil Lexicon Based

No	Author	Label
1	Kalau lihat ini mengapa ibu kota negara harus pindah? padahal dengan pindah akan menghilangkan hutan dan ekologi. entahlah!	Negatif
2	siapapun yang baca komen saya, tanamlah pohon, tolong...	Positif
3	Sejak dulu kita semua tinggal di planet bumi jika planet ini rusak kita tidak tau lagi mau tinggal di mana.	Negatif
…	…	…
1054	Manusia sudah terlampau melewati batas, usaha apapun yang kita lakukan untuk memperbaikinya itu hanya menunda kehancuran saja	Negatif
1055	bumi makin Tua Pemanasan Global dam efec rumah kaca Wabah penyakit dan perubahan iklim	Negatif

2.4 Model

Pada tahap ini, dilakukan proses pemodelan menggunakan algoritma *Support Vector Machine* (SVM) untuk mengklasifikasikan sentimen dalam komentar *YouTube* terkait kesulitan membaca siswa SMP [16]. Model ini dipilih karena kemampuannya dalam menangani data berukuran besar dan menghasilkan pemisahan kelas yang optimal melalui *Hyperplane* [17][5].Sebelum model dilatih, data yang telah melalui tahap *pre-processing* dikonversi ke dalam bentuk Term Frequency - Inverse Document Frequency (TF-IDF).

Setelah data siap, model SVM dengan kernel linear diterapkan karena sesuai untuk tugas klasifikasi teks. *Dataset* kemudian dibagi menjadi 80% data latih dan 20% data uji untuk mengukur performa model.

Model dilatih menggunakan data latih, di mana setiap komentar dikategorikan sebagai positif atau negatif berdasarkan metode lexicon yang telah digunakan sebelumnya.

Setelah pelatihan selesai, model diuji menggunakan data uji untuk melihat sejauh mana kemampuannya dalam mengklasifikasikan sentimen dengan benar. Evaluasi dilakukan untuk mengetahui apakah model sudah cukup baik atau perlu dilakukan penyempurnaan lebih lanjut, seperti penyesuaian parameter atau penambahan data latih.

2.5 Asses

Evaluasi model dilakukan menggunakan akurasi, *precision*, *recall*, dan *F1-score* [16]. Jika performa model kurang optimal, dilakukan penyesuaian parameter atau teknik *cross-validation* untuk menghindari overfitting. Hasil evaluasi menjadi dasar dalam penyempurnaan model agar dapat mengklasifikasikan sentimen dengan lebih akurat [18].

2.6 Proses dan Hasil Penelitian

Pada Proses penelitian ini dimulai dengan mengumpulkan data komentar *YouTube* menggunakan *YouTube API*, dilanjutkan dengan eksplorasi dan pembersihan data melalui pra-pemrosesan seperti sensitivitas huruf besar- kecil, sanitasi, tokenisasi, *stemming*, dan penghapusan stopword. Sentimen diklasifikasi menggunakan metode berbasis leksikon, kemudian direpresentasikan menggunakan TF-IDF dan diklasifikasi menggunakan *Support Vector Machine* (SVM). Hasil penelitian menunjukkan bahwa model SVM memiliki akurasi yang tinggi dalam membedakan komentar positif dan negatif, ditinjau dari presisi, *recall* dan skor F1[19]. Analisis lebih lanjut menggunakan awan kata mengungkapkan pola kata yang dominan dan memberikan wawasan mengenai persepsi masyarakat terhadap kesulitan membaca siswa sekolah menengah.

2.7 Analisis

Analisis Hasil analisis sentimen menggunakan *Support Vector Machine* (SVM) menunjukkan bahwa model mampu membedakan komentar positif, netral dan negatif dengan akurasi tinggi [20][21]. Evaluasi berdasarkan presisi, *recall* dan skor F1 menegaskan bahwa model bekerja maksimal dalam mengidentifikasi pola mood.

Analisis lebih lanjut menggunakan cloud word menunjukkan kata-kata yang paling sering muncul dalam komentar dan menggambarkan sentimen umum yang berkembang di masyarakat mengenai kesulitan membaca siswa sekolah menengah. Komentar negatif umumnya berisi kritik terhadap sistem pendidikan, sedangkan komentar positif lebih banyak mengandung empati dan solusi[21].

Dari hasil penelitian tersebut dapat disimpulkan bahwa pendekatan SVM dengan TF-IDF efektif dalam analisis sentimen teks khususnya dalam memahami opini masyarakat terhadap isu literasi [22]. Model ini dapat dikembangkan lebih lanjut untuk analisis skala lebih besar guna memberikan wawasan yang lebih mendalam. Berikut faktor – faktor penyebab sulit membaca.

Tabel 7. Faktor Penyebab

Faktor Penyebab	Deskripsi
Kurangnya Dukungan dari Guru	Siswa mengeluhkan kurangnya bimbingan dari guru dalam proses belajar.
Kesulitan dalam Memahami Materi	Materi yang disajikan dianggap terlalu sulit untuk dipahami oleh siswa.
Minimnya Sumber Belajar	Siswa merasa tidak memiliki cukup sumber belajar yang mendukung.

3. HASIL DAN PEMBAHASAN

Hasil penelitian membahas tentang proses pengolahan data, proses pengujian, dan implementasi algoritma *Support Vector Machine* (SVM).

3.1 Deskripsi *Dataset* Komentar *YouTube*

Dataset yang digunakan dalam penelitian ini berasal dari komentar pengguna pada video *YouTube* berjudul "Wah, Puluhan Siswa SMP di Pangandaran Masih Lancar Membaca" yang diunggah oleh kanal tvOneNews. Video tersebut menampilkan berita tentang siswa SMP Kelas 7 di Kabupaten Pangandaran yang kesulitan membaca dan menuai beragam reaksi publik dalam bentuk komentar. Pengumpulan data dilakukan menggunakan web *scraping* dengan *YouTube API* melalui platform *Google Colab*. Data yang berhasil dikumpulkan mencakup 1.055 komentar. Gambar 1 menunjukkan data yang digunakan

	author	published_at	comment
0	@YohanesDwi-m6k	2024-12-12T07:08:50Z	Covid jangan dijadikan alasan. Covid kan hanya...
1	@YohanesDwi-m6k	2024-11-28T13:25:43Z	Yg salah guru SD asal murid tersebut. Masalah ...
2	@dan93doang	2024-11-16T01:43:32Z	kurikulum buatan nadiem makariem terbukti gagal😂
3	@reg_mq3	2024-11-02T04:42:45Z	ibu guru nya aja , gagap gagap ngomong nya loh...

Gambar 1 Data yang digunakan

3.2 Tahapan Pra-Pemrosesan Data

Pembahasan Hasil Dalam penelitian ini, data dilakukan pada tahap modifikasi. Tahap ini berfungsi untuk menyiapkan data sebelum proses klasifikasi. Proses ini diawali dengan kapitalisasi. Semua huruf dalam teks ditulis dengan huruf kecil untuk memastikan kecocokan kata yang konsisten. Misalnya, kata "Baca" dan "baca" dianggap identik setelah kapitalisasi.

Tahap selanjutnya adalah pembersihan data. Pembersihan data dilakukan dengan membuang elemen yang tidak diperlukan seperti URL, angka, karakter khusus, tanda baca, dan simbol ASCII yang tidak bermakna dari data. Tahap ini diikuti dengan tokenisasi. Proses ini melibatkan penguraian teks menjadi kata-kata atau token individual untuk analisis yang lebih rinci pada tahap selanjutnya.

Proses selanjutnya yaitu normalisasi kemudian dilakukan untuk membakukan kata-kata yang tidak baku menjadi bentuk yang sesuai dengan bahasa Indonesia formal. Kata-kata seperti "gw" diubah menjadi "saya" dan "banget" menjadi "sekali." Tujuan dari proses ini adalah untuk menyamakan makna kata-kata dengan ejaan yang berbeda. Setelah dinormalisasi, kata-kata henti atau kata-kata umum yang tidak relevan dengan analisis sentimen, seperti "yang," "dan," "di," dan "itu," dihilangkan.

Semua langkah praproses ini dilakukan dengan menggunakan bahasa pemrograman *Python* dan pustaka seperti Sastrawi, *Pandas*, dan Re. Hasil akhir dari proses ini digunakan dalam fase pelabelan sentimen menggunakan metode berbasis leksikon dan kemudian berfungsi sebagai data input untuk pemodelan dengan algoritma *Support Vector Machine* (SVM) gambar 2 adalah Tahapan Pra-Pemrosesan Data.

```python
import re
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

stemmer = StemmerFactory().create_stemmer()
stopword = StopWordRemoverFactory().create_stop_word_remover()

def preprocess(text):
    text = text.lower()
    text = re.sub(r'http\S+|www\S+|[^a-zA-Z\s]', ' ', text)
    text = re.sub(r'\s+', ' ', text).strip()
    text = stopword.remove(text)
    text = stemmer.stem(text)
    return text

df['clean_comment'] = df['comment'].astype(str).apply(preprocess)
df.head()
```

Gambar 2 Tahapan Pra-Pemrosesan Data

3.3 Transformasi TF-IDF dan Pembentukan Model

Setelah data komentar berhasil melalui praproses, langkah selanjutnya adalah mengubahnya menjadi representasi numerik sehingga dapat diproses oleh algoritma *machine learning*. Dalam penelitian ini, transformasi dilakukan dengan menggunakan metode Term *Frequency-Inverse Document Frequency* (TF-IDF).

```
[13] from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.model_selection import train_test_split
     from sklearn.svm import LinearSVC

     tfidf = TfidfVectorizer()
     X = tfidf.fit_transform(df['clean_comment'])
     y = df['label']

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

     model = LinearSVC()
     model.fit(X_train, y_train)
```

```
LinearSVC
LinearSVC()
```

Gambar 3 Transformasi TF-IDF

TF-IDF merupakan metrik yang digunakan untuk menilai pentingnya sebuah kata terhadap sebuah dokumen dalam kumpulan dokumen (korpus). Semakin sering sebuah kata muncul dalam sebuah dokumen tetapi jarang muncul di dokumen lain, maka bobot TF-IDF kata tersebut akan semakin tinggi. Metode ini secara efektif menghilangkan kata-kata umum yang tidak berkontribusi signifikan terhadap makna keseluruhan dan lebih berfokus pada kata-kata yang unik dan bermakna bagi sentimen gambar 3 adalah Transformasi TF-IDF.

3.4 Evaluasi Model SVM

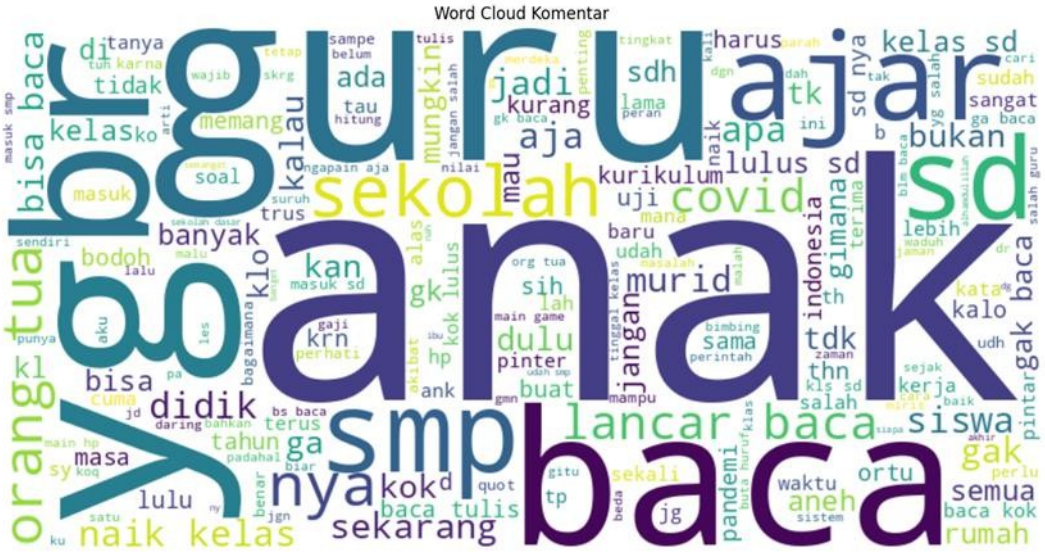
Setelah proses pelatihan model dengan algoritma *Support Vector Machine* (SVM) selesai, langkah selanjutnya adalah mengevaluasi performa model pada data uji. Evaluasi dilakukan untuk mengukur kemampuan model dalam mengklasifikasikan komentar *YouTube* ke dalam tiga kategori sentimen: positif, negatif, dan netral. Evaluasi performa model dalam studi ini menggunakan beberapa metrik klasifikasi umum. *Accuracy*: Persentase prediksi yang benar dibandingkan dengan semua data uji. *Precision*: Kemampuan model untuk mengidentifikasi komentar yang benar-benar termasuk dalam suatu kelas. *Recall*: Kemampuan model untuk menemukan semua komentar yang relevan dalam suatu kelas. *F1-score*: Rata-rata harmonis dari *precision* dan *recall*, yang memberikan gambaran umum tentang keseimbangan antara keduanya. Hasil evaluasi model pada data uji ditunjukkan pada gambar 4 Laporan Klasifikasi.

	precision	recall	f1-score	support
negatif	1.00	0.37	0.54	19
netral	0.93	1.00	0.96	186
positif	1.00	0.50	0.67	6
accuracy			0.93	211
macro avg	0.98	0.62	0.72	211
weighted avg	0.93	0.93	0.91	211

Gambar 4 Laporan Klasifikasi

3.5 Visualisasi Word Cloud

Setelah menyelesaikan praproses dan pelabelan sentimen, langkah selanjutnya dalam penelitian ini adalah memvisualisasikan data teks menggunakan teknik word cloud. Visualisasi ini bertujuan untuk mengidentifikasi kata- kata yang paling sering muncul dalam komentar pengguna YouTube yang telah dibersihkan dan memberikan gambaran umum tentang tema atau opini yang dominan dalam komunitas mengenai kesulitan membaca di antara siswa kelas 7. Word cloud adalah representasi grafis dari frekuensi kata. Ukuran setiap kata dalam visualisasi mencerminkan frekuensi kemunculannya dalam kumpulan teks. Semakin besar kata, semakin tinggi frekuensinya dalam korpus komentar yang dianalisis. Untuk melakukan ini, semua komentar yang telah dibersihkan digabungkan menjadi string teks dan kemudian diproses menggunakan pustaka WordCloud dalam Python gambar 4.6 adalah Visualisasi Word Cloud.



Gambar Nilai Akurasi Algoritma SVM Nilai Akurasi Algoritma SVM 5 Visualisasi Word Cloud

Setelah menyelesaikan praproses dan pelabelan sentimen, langkah selanjutnya dalam penelitian ini adalah memvisualisasikan data teks menggunakan teknik word cloud. Visualisasi ini bertujuan untuk mengidentifikasi kata- kata yang paling sering muncul dalam komentar pengguna *YouTube* yang telah dibersihkan dan memberikan gambaran umum tentang tema atau opini yang dominan dalam komunitas mengenai kesulitan membaca di antara siswa kelas 7. *Word cloud* adalah representasi grafis dari frekuensi kata. Ukuran setiap kata dalam visualisasi mencerminkan frekuensi kemunculannya dalam kumpulan teks. Semakin besar kata, semakin tinggi frekuensinya dalam korpus komentar yang dianalisis [25]. Untuk melakukan ini, semua komentar yang telah dibersihkan digabungkan menjadi string teks dan kemudian diproses menggunakan pustaka WordCloud dalam Python gambar 4.4 adalah Visualisasi *Word Cloud.*

3.6 Pembahasan

Pembahasan pada subbab ini menguraikan hasil pembahasan data dan sejauh mana kinerja algoritma *Support Vector Machine* (SVM) dalam melakukan analisis sentimen opini publik terkait siswa Sekolah Menengah Pertama (SMP) yang mengalami kesulitan membaca.

3.7 Analisis Pola Sentimen dan Temuan Utama

Hasil analisis sentimen terhadap 1.055 komentar pengguna *YouTube* pada video tentang kesulitan membaca siswa sekolah menengah menunjukkan bahwa mayoritas komentar, dengan total 927, bersifat netral. 89 komentar bersifat negatif, dan 39 bersifat positif. Distribusi ini menunjukkan bahwa sebagian besar komentar bersifat informatif atau tidak mengandung pendapat yang eksplisit baik dalam bentuk dukungan maupun kritik.

Komentar negatif biasanya mengungkapkan kekecewaan, kekhawatiran, dan kritik terhadap kondisi pendidikan saat ini. Beberapa komentar mengkritik sistem pendidikan karena gagal mengajarkan keterampilan membaca dasar kepada siswa, serta kurangnya peran yang dimainkan oleh guru dan orang tua dalam mendukung pembelajaran anak-anak. Kata-kata seperti "sedih," "sulit," "gagal," dan "buruk" sering muncul dalam komentar negatif dan mencerminkan sentimen publik.

Di sisi lain, komentar positif biasanya mengandung dukungan moral, empati, dan dorongan untuk meningkatkan pencapaian pendidikan. Meskipun jumlahnya relatif kecil, komentar-komentar ini menunjukkan bahwa ada beberapa orang yang optimis dan berharap bahwa pemerintah dan sekolah akan segera mengambil tindakan untuk memperbaiki situasi. Tabel 8 adalah data yang telah dianalis.

Tabel 8. Data yang Telah di Analisis

Author	Published_at	Clean_comment	Label
@YohanesDwi-m6k	2024-12-12T07:08:50Z	covid jangan jadi alas covid kan sampai tahun lancar baca kan kelas sd sudah lancar	Netral
@YohanesDwi-m6k	2024-11-28T13:25:43Z	yg salah guru sd asal murid sebut masalah covid jangan jadi alas	Netral
@dan93doang	2024-11-16T01:43:32Z	kurikulum buat nadiem makariem bukti gagal	Negatif
@sadilahsadilah9666	2023-08-07T11:43:54Z	hasil kurikulum merdeka parah smp kok gak baca	Negatif

@vitriaanindita8629	2023-08-15T07:34:21Z	semangaat de ade jangan patah semangat jangan nyalahin guru hasil kalau mau naek terus padahal dulu kalo ngak naek kelas memang mampu naek kan kelas sekarang kalo ngak di naek kan kelas nanto orang tua marah ngak terima trus anak ngadu takut ketepel yg nberak	Positif

3.8 Pembahasan Temuan Terhadap Tujuan Penelitian

Hasil penelitian ini menjawab tiga tujuan utama yang telah disebutkan di atas. Pertama, model klasifikasi sentimen yang dikembangkan dengan algoritma *Support Vector Machine* (SVM) memiliki kinerja yang sangat baik, yaitu mencapai akurasi sebesar 92,89%. Nilai ini menunjukkan bahwa model dapat mengidentifikasi dan mengklasifikasikan komentar berdasarkan sentimennya secara akurat, bahkan pada kumpulan data yang sebagian besar berisi komentar netral.

Kedua, akurasi yang tinggi menunjukkan bahwa algoritma SVM dengan pendekatan fitur TF-IDF sangat efektif dalam menangani data teks dari komentar media sosial, khususnya dalam bahasa Indonesia. Model mampu membedakan komentar dengan sentimen positif, negatif, dan netral secara jelas, meskipun terdapat ketidakseimbangan data antarkelas sentimen.

Ketiga, jumlah data latih yang digunakan dalam model (80% dari keseluruhan kumpulan data) juga berkontribusi terhadap kinerja model yang optimal. Semakin besar data latih yang digunakan, semakin baik model mengenali pola kata yang terkait dengan setiap jenis sentimen. Hal ini membuktikan bahwa ketersediaan data yang cukup merupakan faktor penting bagi efektivitas pelatihan model machine learning.

Dari segi konten, hasil ini menggambarkan gambaran konkret opini publik tentang masalah membaca di kalangan siswa sekolah menengah. Dominasi komentar netral menunjukkan bahwa publik lebih cenderung menyampaikan informasi dan pandangan tanpa ekspresi emosional yang kuat, tetapi tetap menyatakan keprihatinan tentang masalah tersebut. Komentar negatif jelas menggambarkan keprihatinan publik tentang kualitas pendidikan dasar, sementara komentar positif mencerminkan harapan akan solusi dari pihak berwenang gambar 4.4 adalah Nilai Akurasi Algoritma SVM.

```
from sklearn.metrics import accuracy_score

# Prediksi data uji
y_pred = model.predict(X_test)

# Hitung akurasi
accuracy = accuracy_score(y_test, y_pred)
print(f'Akurasi model SVM: {accuracy:.4f}')
```

```
Akurasi model SVM: 0.9289
```

Gambar 6 Nilai Akurasi Algoritma SVM

4. KESIMPULAN

Berdasarkan hasil analisis sentimen komentar pengguna YouTube pada video tentang kesulitan membaca siswa kelas 7 SMPN 1 Mangunjaya, dapat disimpulkan bahwa algoritma Support Vector Machine (SVM) menunjukkan kinerja yang sangat baik dalam klasifikasi sentimen. Proses analisis meliputi beberapa langkah: akuisisi data, praproses, pelabelan sentimen berbasis leksikon, dan pembangunan model menggunakan TF-IDF dan SVM. Model klasifikasi SVM yang dikembangkan dalam penelitian ini mencapai akurasi sebesar 92,89%. Hal ini menunjukkan bahwa model ini dapat secara efektif mengklasifikasikan komentar ke dalam tiga kategori sentimen positif, negatif, dan netral. Dari total 1.055 komentar, 927 tergolong netral, 89 tergolong negatif, dan 39 tergolong positif. Dominasi sentimen netral menunjukkan bahwa sebagian besar komentar bersifat deskriptif atau informatif tanpa secara eksplisit mengungkapkan pendapat. Komentar negatif biasanya berisi kritik terhadap sistem pendidikan, peran guru, dan kebijakan pemerintah dalam mengatasi masalah literasi. Komentar positif menyatakan dukungan moral dan harapan untuk perbaikan sistem pendidikan. Visualisasi word cloud menunjukkan kata-kata dominan seperti "siswa," "membaca," "guru," dan "pendidikan," yang mencerminkan opini publik. Analisis sentimen menggunakan metode SVM terbukti cocok untuk menangkap persepsi publik secara akurat dan merupakan alternatif yang efektif untuk memahami opini publik tentang isu pendidikan.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan dukungan, bimbingan, dan doa dalam proses penyusunan penelitian ini, khususnya kepada kedua orang tua tercinta yang selalu memberikan doa, dukungan moral, semangat, serta kasih sayang yang tiada henti.

Tanpa mereka, penulis tidak akan mampu menyelesaikan penelitian ini dengan baik dan kepada dosen pembimbing yang telah dengan sabar membimbing, memberikan arahan, kritik, dan saran yang sangat berarti selama proses penyusunan penelitian ini. Bimbingan beliau sangat membantu penulis dalam menyusun dan menyempurnakan karya ini, serta kepada Politeknik TEDC Bandung atas fasilitas dan dukungan yang diberikan sehingga penelitian ini dapat terselesaikan dengan baik.

REFERENSI

[1] S. Krisnasari, D. Suhermah, and I. Latifah, "Pemanfaatan Aplikasi Quizizz dalam Pembelajaran Literasi dan Numerasi di PAUD," JIIP- J. Ilm. Ilmu Pendidik., vol. 5, no. 6, pp. 1730–1734, 2022, doi: 10.54371/jiip.v5i6.635.

[2] Y. Findawati and A. Rosid, Muhammad, BUKU AJAR TEXT MINING, 1st ed. Sidoarjo: UMSIDA Press, 2020. ISBN: 978-623-6833- 19-3.

[3] F. F. Mailoa and L. Lazuardi, "Analisis sentimen data twitter menggunakan metode text mining tentang masalah obesitas di indonesia," J. Inf. Syst. Public Heal., vol. 6, no. 1, p. 44, 2021, doi: 10.22146/jisph.44455.

[4] R. Yusuf, K. Bahumatra, N. Komaria, E. A. Aqma, and L. Cahyani, "Analisis Sentimen Terhadap Aplikasi Google Meet Berdasarkan Komentar Pengguna Menggunakan Metode Logistic Regresion," J. Ilm. Edutic Pendidik. dan Inform., vol. 11, no. 1, pp. 53–64, 2024, doi: 10.21107/edutic.v11i1.28113.

[5] I. R. Ainunnisa and S. Sulastri, "Analisis Sentimen Aplikasi Tiktok dengan Metode Support Vector Machine (SVM), Logistic Regression dan Naïve Bayes," J. Teknol. Sist. Inf. dan Apl., vol. 6, no. 3, pp. 423–430, 2023, doi: 10.32493/jtsi.v6i3.31076.

[6] D. Nugraha, Metodologi penelitian: pendekatan kuantitatif, kualitatif, dan campuran, 1st ed., no. June. Agam, Indonesia: CV LAUK PUYU PRESS, 2024.

[7] M. R. Fikri, R. T. Handayanto, and D. Irwan, "Web Scraping Situs Berita Menggunakan Bahasa Pemograman Python," J. Students Res. Comput. Sci., vol. 3, no. 1, pp. 123–136, 2022, doi: 10.31599/jsrcs.v3i1.1514.

[8] Rahmawati, B. A. Habsy, and M. Nursalim, "Jenis-Jenis Metode Pengumpulan Data (Qualitative Research)," J. Pendidik. Tambusai, vol. 9, no. 1, p. 9935, 2025, doi: 10.47827/jer.v5i4.281.

[9] M. D. Alizah, A. Nugroho, U. Radiyah, and W. Gata, "Sentimen Analisis Terkait Lockdown pada Sosial Media Twitter," Indones. J. Softw. Eng., vol. 6, no. 2, pp. 223–229, 2020, doi: 10.31294/ijse.v6i2.8991.

[10] Omari Firas, "A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach," World J. Adv. Eng. Technol. Sci., vol. 8, no. 1, pp. 009–014, 2023, doi: 10.30574/wjaets.2023.8.1.0147.

[11] A. Salsabila Juwita, A. Rizky Kurniawan, A. Aryaputra Ashari, D. Tyan Putro, V. Nurcahyawati, and H. Artikel, "Implementasi Data Mining untuk Memprediksi Kesehatan Mental Mahasiswa menggunakan Algoritma Naïve Bayes," KOMPUTEK J. Tek. Univ. Muhammadiyah Ponorogo, vol. 8, no. No 1, p. Hal 61-70, 2024, [Online]. Available: http://studentjournal.umpo.ac.id/index.php/komputek

[12] Y. A. Suwitono and F. J. Kaunang, "Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras," J. Komtika (Komputasi dan Inform., vol. 6, no. 2, pp. 109–121, 2022, doi: 10.31603/komtika.v6i2.8054.

[13] M. Z. Siregar et al., "Analisis Sentimen Terhadap Ulasan Aplikasi Media Sosial Di Google Play Menggunakan Algoritma Naive Bayes," JATI (Jurnal Mhs. Tek. Inform., vol. 9, no. 2, pp. 3373–3381, 2025, doi: 10.36040/jati.v9i2.12841.

[14] Pande sindu, Agus Aan Jiwa Permana, and I Nyoman Saputra Wahyu Wijaya, "Identifikasi Dan Normalisasi Teks Slang Dengan Fasttext Pada Twitter Dalam Bahasa Indonesia," J. Pendidik. Teknol. dan Kejuru., vol. 21, no. 1, pp. 33–44, 2024, doi: 10.23887/jptkundiksha.v21i1.66381.

[15] N. Nofiyani and W. Wulandari, "Implementasi Electronic Data Processing Untuk meningkatkan Efektifitas dan Efisiensi Pada Text Mining," J. Media Inform. Budidarma, vol. 6, no. 3, p. 1621, 2022, doi: 10.30865/mib.v6i3.4332.

[16] D. Yulianto and A. Nugraheni, "DECODE : Jurnal Pendidikan Teknologi Informasi," Decod. J. Pendidik. Teknol. Inf., vol. 1, no. 1, pp. 33–42, 2021, doi: 10.51454/decode.v1i1.5.

[17] N. S. Fathullah, Y. A. Sari, and P. P. Adikara, "Analisis Sentimen Terhadap Rating dan Ulasan Film dengan menggunakan Metode Klasifikasi Naïve Bayes dengan Fitur Lexicon-Based," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 2, pp. 590–593, 2020, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6987

[18] A. N. Ihsan and S. Tresnawati, "Analisis Sentimen Pengguna X Terhadap Layanan Provider Iconnet Menggunakan Naïve Bayes Dan Support Vector Machine," JITET (Jurnal Inform. dan Tek. Elektro Ter., vol. 12, no. 3S1, pp. 4105–4113, 2024, doi: 10.23960/jitet.v12i3S1.5264.

[19] D. S. Putri, N. Sulistiyowati, and A. Voutama, "Analisis Sentimen dan Pemodelan Ulasan Aplikasi AdaKami Menggunakan Algoritma SVM dan KNN," J. Sensi, vol. 9, no. 2, pp. 209–225, 2023, doi: 10.33050/sensi.v9i2.2914.

[20] A. Wyawhare, "Comparative Analysis of Multilingual Text Classification & Identification through Deep Learning and Embedding Visualization," arXiv Prepr., vol. 2312.03789, pp. 1–9, 2023, [Online]. Available: https://arxiv.org/abs/2312.03789

[21] J. A. Wibowo, V. C. Mawardi, and T. Sutrisno, "Visualisasi Word Cloud Hasil Analisis Sentimen Berbasis Fitur Layanan Aplikasi Gojek Dengan Support Vector Machine," J. Serina Sains, Tek. dan Kedokt., vol. 2, no. 1, pp. 61–70, 2024, doi: 10.24912/jsstk.v2i1.32058.

[22] S. Andayani and A. Ryansyah, "Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen," JuSiTik J. Sist. dan Teknol. Inf. Komun., vol. 1, no. 1, p. 53, 2017, doi: 10.32524/jusitik.v1i1.218.

[23] M. Alfyando, F. T. Anggraeny, and A. N. Sihananto, "Perbandingan Algoritma Random Forest dan Logistic Regression Untuk Analisis Sentimen Ulasan Aplikasi Tumbuh Kembang Anak Di Play Store," J. Sist. Inf. dan Ilmu Komput., vol. 2, no. 1, pp. 77–86, 2024, [Online]. Available: https://doi.org/10.59581/jusiik-widyakarya.v2i1.2262

[24] A. Kartika Sari, Akhmad Irsyad, Dinda Nur Aini, Islamiyah, and Stephanie Elfriede Ginting, "Analisis Sentimen Twitter Menggunakan Machine Learning untuk Identifikasi Konten Negatif," Adopsi Teknol. dan Sist. Inf., vol. 3, no. 1, pp. 64–73, 2024, doi: 10.30872/atasi.v3i1.1373.

[25] A. R. S. Darwanto, Taza Luzia Viarindita, and Yekti Widyaningsih, "Analisis Regresi Logistik Binomial dan AlgoritmaRandom Forest pada Proses Pengklasifikasian Penyakit Ginjal Kronis," J. Stat. dan Apl., vol. 5, no. 1, pp. 1–14, 2021, doi: 10.21009/jsa.05101.