JIKTEKS : Jurnal Ilmu Komputer dan Teknologi Informasi
Vol. 05, No. 01, Desember 2026, hlm. 01 - 12
ISSN : 2986-5417
DOI : doi.org/10.70404/jikteks.v5i01.739
Laman Jurnal: jurnal.faatuatua.com/index.php/JIKTEKS

Penerapan Algoritma SVM Untuk Sentimen Analisis Media Sosial Terhadap PP No. 17 Tahun 2025

Muhammad Rifqi Asy'ari[1*], Albert Riyandi[2]
[1*, 2]Informatika, Universitas Bina Sarana Informatika, Indonesia
[1*]*rifqi6886@gmail.com*
[2]*albert.abe@bsi.ac.id*

Keywords:

Sentiment Analysis;
Support Vector Machine;
Hybrid Labeling;
TF-IDF;
PP No. 17 of 2025.

ABSTRACT

The use of social media among Indonesian children continues to rise, bringing various digital risks ranging from cyberbullying to personal data exploitation. In response, the government issued Government Regulation Number 17 of 2025 concerning Electronic System Governance in Child Protection (PP TUNAS), which requires age restrictions on social media accounts for children under 16 years old. This policy triggered widespread public debate, reflected in more than 32 thousand comments on a news video posted by the KompasTV account on TikTok. This study aims to analyze public sentiment distribution, measure classification performance, and identify dominant word patterns related to the policy. The method used is Support Vector Machine with the LinearSVC algorithm, supported by a Hybrid Labeling approach and TF-IDF feature extraction. Data were collected through scraping using Apify, resulting in 274,087 rows of raw data, which were then processed down to 4,326 clean data points. The results show that the model achieved an accuracy of 97.23%, with negative sentiment dominating at 77.39%, indicating that the majority of the public tends to reject the policy during its early implementation phase.

Kata Kunci:

Analisis Sentimen;
Support Vector Machine;
Hybrid Labeling;
TF-IDF;
PP No. 17 Tahun 2025.

ABSTRAK

Penggunaan media sosial oleh anak-anak Indonesia terus meningkat dan membawa berbagai risiko digital, mulai dari perundungan siber hingga eksploitasi data pribadi. Menanggapi kondisi tersebut, pemerintah menerbitkan Peraturan Pemerintah Nomor 17 Tahun 2025 tentang Tata Kelola Penyelenggaraan Sistem Elektronik dalam Pelindungan Anak (PP TUNAS) yang mewajibkan pembatasan usia akun media sosial bagi anak di bawah 16 tahun. Kebijakan ini memicu perdebatan publik yang luas, salah satunya tercermin dari lebih dari 32 ribu komentar pada video berita akun KompasTV di *TikTok*. Penelitian ini bertujuan untuk menganalisis distribusi sentimen masyarakat, mengukur performa klasifikasi, serta mengidentifikasi pola kata dominan terhadap kebijakan tersebut. Metode yang digunakan adalah *Support Vector Machine* dengan algoritma *LinearSVC*, didukung pendekatan *Hybrid Labeling* dan ekstraksi fitur *TF-IDF*. Data dikumpulkan melalui *scraping* menggunakan *Apify* yang menghasilkan 274.087 baris data mentah, kemudian diproses hingga menyisakan 4.326 data bersih. Hasil penelitian menunjukkan model mencapai akurasi 97,23% dengan sentimen negatif mendominasi sebesar 77,39%, mengindikasikan bahwa mayoritas masyarakat cenderung menolak kebijakan ini pada fase awal implementasinya.

Korespondensi Penulis *):

Muhammad Rifqi Asy'ari
Universitas Bina Sarana Informatika
Kelurahan Kwitang, Kecamatan Senen, Kota Jakarta Pusat, 10450

Diajukan: 29-08-2026 | Direvisi: 03-09-2026 | Diterima: 19-10-2026 | Diterbitkan: 11-11-2026
This Journal is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License | CC BY-SA 4.0

1. PENDAHULUAN

Pertumbuhan pengguna internet di kalangan anak Indonesia berlangsung sangat cepat dalam beberapa tahun terakhir. Data Survei Sosial Ekonomi Nasional mencatat persentase penduduk usia 5 sampai 17 tahun yang pernah mengakses internet naik dari sekitar 49% pada 2020 menjadi hampir 74% pada 2024 [1]. Angka ini menggambarkan betapa cepatnya ruang digital menjadi bagian keseharian anak-anak, bukan lagi sekadar hiburan tambahan.

Kenaikan akses ini membawa manfaat nyata untuk pendidikan dan interaksi sosial, tapi di sisi lain risiko juga ikut membesar. Tinjauan terhadap fenomena perundungan siber pada anak menunjukkan bahwa persoalan ini makin kompleks dan menuntut perlindungan sistematis dari berbagai pihak, bukan cuma tanggung jawab orang tua semata [2]. Paparan konten berbahaya, gangguan kesehatan mental, sampai eksploitasi data pribadi jadi ancaman yang terus mengintai anak selama mereka berada di ruang digital.

Merespons kondisi tersebut, sejumlah negara mulai mengambil langkah regulatif. Australia menjadi negara pertama yang memberlakukan larangan nasional bagi anak di bawah usia 16 tahun untuk memiliki akun pada platform media sosial tertentu, efektif berlaku sejak 10 Desember 2025 [3]. Indonesia mengambil arah kebijakan yang serupa lewat Peraturan Pemerintah Nomor 17 Tahun 2025 tentang Tata Kelola Penyelenggaraan Sistem Elektronik dalam Pelindungan Anak, yang lebih dikenal dengan sebutan PP TUNAS [4].

Skala dampak kebijakan ini di Indonesia jauh lebih besar dibanding Australia. Indonesia memiliki populasi anak di bawah usia 18 tahun lebih dari 83 juta jiwa [5], sedangkan kebijakan Australia diperkirakan hanya memengaruhi sekitar 4,7 hingga 5,7 juta anak di bawah usia 16 tahun [6]. Perbandingan ini memperlihatkan betapa luas dan kompleksnya jangkauan kebijakan pembatasan usia media sosial di Indonesia dibanding negara yang menjadi acuan awal kebijakan tersebut.



Sumber. [5], [6]

Gambar 1. Perbandingan Jumlah Anak di Indonesia dan Australia

Namun, implementasi kebijakan semacam ini bukan tanpa tantangan. Kajian yuridis menyoroti masih adanya kesenjangan antara norma hukum dengan pelaksanaan di lapangan, terutama soal standar teknis verifikasi usia yang belum jelas, mekanisme pengawasan yang lemah, serta audit dan sanksi yang belum konkret [7]. Tanpa infrastruktur pendukung yang memadai, kebijakan berisiko berhenti jadi aturan di atas kertas saja.

Urgensi kebijakan ini semakin terlihat dari data lapangan. Studi *baseline UNICEF* mengenai pengetahuan dan kebiasaan daring anak Indonesia menemukan bahwa 42% anak pernah merasa tidak nyaman atau takut akibat pengalaman daring, lebih dari 48% pernah mengalami perundungan siber, dan sekitar 50,3% pernah melihat konten seksual di media sosial. Sementara itu, hanya 37,5% anak yang pernah mendapat informasi soal cara menjaga keselamatan saat berinternet [8]. Ditambah lagi, laporan KPAI mencatat kesenjangan antara kecepatan perkembangan teknologi dan rendahnya literasi digital masyarakat menjadi penyebab utama meningkatnya kasus kejahatan siber terhadap anak [9].

Dari sisi hukum, PP TUNAS secara tegas mewajibkan anak di bawah usia 16 tahun untuk tidak memiliki akun media sosial dan mengharuskan platform menonaktifkan akun yang tidak memenuhi syarat usia [10]. Sementara dari sisi kesehatan, paparan berulang terhadap konten negatif turut berkontribusi pada meningkatnya kecemasan dan depresi remaja. Survei Kesehatan Mental Remaja Nasional mencatat sekitar 34,9% atau 15,5 juta remaja Indonesia mengalami masalah kesehatan mental, dengan media sosial menjadi salah satu pemicu utamanya [11]. Temuan ini diperkuat tinjauan sistematis yang menunjukkan anak perempuan cenderung lebih rentan terhadap dampak tersebut [12], sementara studi lain mencatat waktu yang tersita untuk media sosial ikut menurunkan konsentrasi belajar anak di sekolah [13].

Aturan pelaksana PP TUNAS kemudian diperkuat lewat Peraturan Menteri Komunikasi dan Digital Nomor 9 Tahun 2026 yang efektif berlaku sejak 28 Maret 2026, dengan delapan platform berisiko tinggi menjadi sasaran utama yaitu *YouTube*, *TikTok*, *Facebook*, *Threads*, *Instagram*, *X*, *Bigo Live*, dan *Roblox* [14]. Kebijakan sebesar ini otomatis memicu perdebatan luas di ruang publik. Sebagian mendukung karena dianggap penting untuk melindungi anak dari

kecanduan dan paparan konten berbahaya, sementara sebagian lain mengkritik potensi pembatasan kebebasan berekspresi dan tantangan teknis verifikasi usia yang berkaitan dengan privasi.

Perdebatan ini terlihat nyata di berbagai platform media sosial, salah satunya *TikTok*. Video berita yang diunggah akun KompasTV pada 28 Maret 2026, bertepatan dengan hari pertama efektivitas kebijakan, berhasil meraih 5,5 juta tayangan dan lebih dari 32 ribu komentar publik. Volume komentar sebesar itu menunjukkan tingginya partisipasi masyarakat dalam menyuarakan pendapat, tapi skala data teks yang begitu besar membuat analisis manual menjadi tidak realistis dilakukan secara menyeluruh dalam waktu singkat.

Kondisi inilah yang mendorong kebutuhan akan pendekatan komputasional untuk memahami distribusi sentimen masyarakat secara objektif. Studi terdahulu memperlihatkan bahwa algoritma *Support Vector Machine (SVM)* yang dikombinasikan dengan ekstraksi fitur *TF-IDF* cukup konsisten menunjukkan performa baik dalam analisis sentimen teks berbahasa Indonesia. Sebagai contoh, penerapan *SVM* untuk menganalisis sentimen masyarakat terhadap kebijakan pemerintah di platform X berhasil mencapai tingkat akurasi yang tinggi [15]. Pendekatan berbasis hak anak juga ditekankan sejumlah kajian sebagai landasan penting dalam merumuskan strategi perlindungan anak di ruang digital [16].

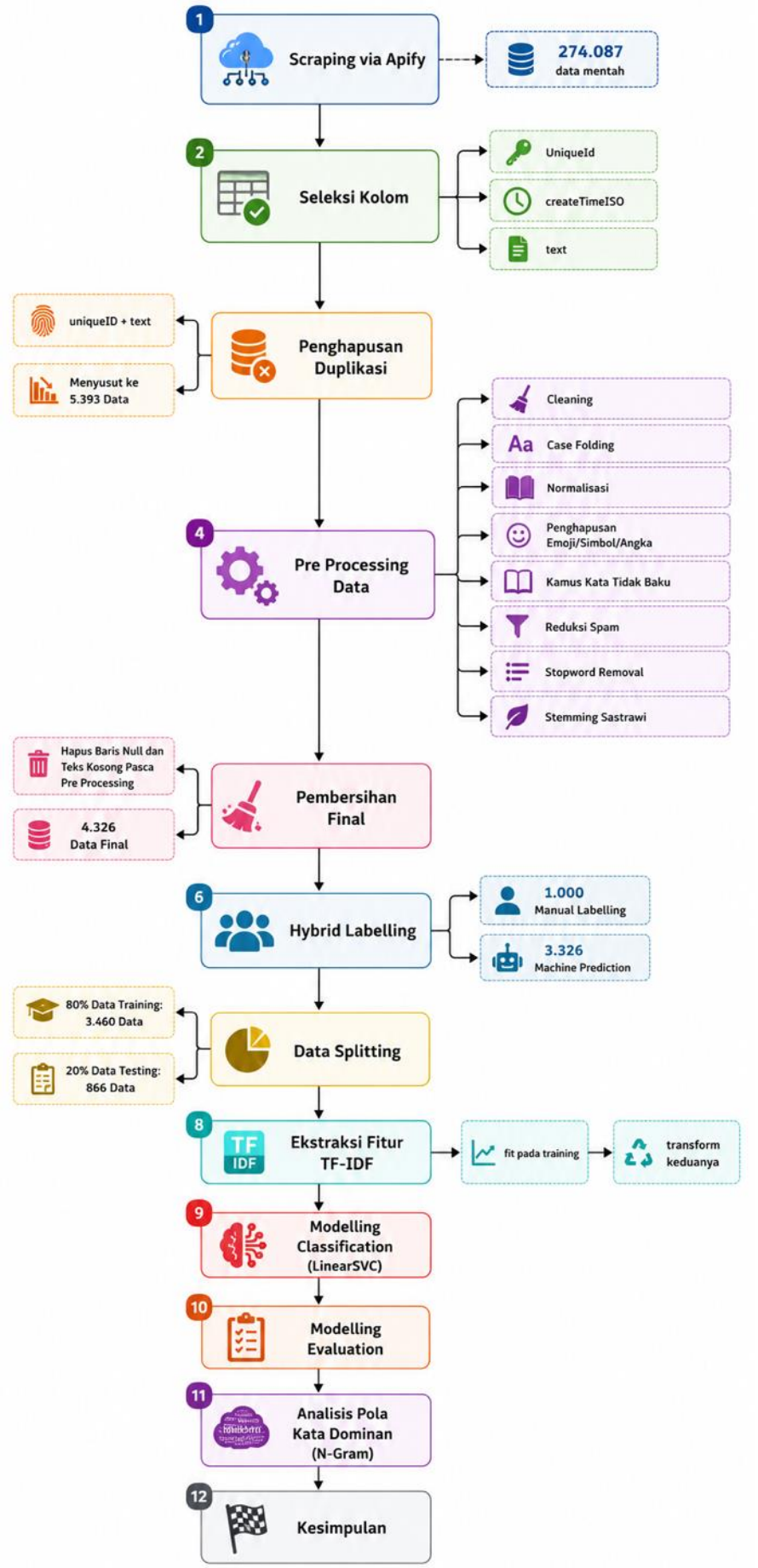
Beberapa penelitian lain turut memperkaya gambaran penerapan *SVM* pada data media sosial berbahasa Indonesia. Penerapan *Naive Bayes* pada komentar *Twitter* terkait *SEA Games* 2023 mencapai akurasi 70,15% [17], sementara kombinasi *SVM* dengan *Recursive Feature Elimination* pada isu kenaikan pajak hiburan mampu menembus akurasi 95% [18]. Pada konteks yang lebih dekat dengan platform *TikTok*, kombinasi *TF-IDF* dan *SVM* pada ulasan aplikasi *TikTok* di *Google Play Store* mencatat akurasi 84% [19], sedangkan penelitian pada komentar *TikTok* terkait Program Makan Bergizi Gratis yang juga memakai data hasil *scraping Apify* hanya mencapai akurasi 77,75% [20]. Studi lain pada layanan *Gojek* di *Instagram* [21], ulasan *Tokopedia* [22], adopsi *QRIS* lewat pendekatan *IndoBERT* [23], sistem balik nama ponsel bekas [24], kasus pelecehan di *TikTok* [25], sampai kebijakan PPKM di *Twitter* [26], turut memperkuat bukti bahwa metode berbasis *machine learning* cukup andal dipakai untuk membaca opini publik di ruang digital.

Meski begitu, dari seluruh penelitian terdahulu tersebut, penulis belum menemukan kajian yang secara khusus menyasar sentimen publik terhadap kebijakan pembatasan usia media sosial berdasarkan PP No. 17 Tahun 2025. Minimnya pemanfaatan data komentar *TikTok*, padahal platform ini paling populer di kalangan anak muda yang justru menjadi subjek utama kebijakan, juga masih menjadi celah tersendiri. Ditambah lagi, belum banyak penelitian yang menggabungkan *scraping* otomatis, pra-pemrosesan lengkap, dan pendekatan *Hybrid Labeling* sekaligus dalam satu *pipeline* untuk konteks kebijakan regulasi yang baru diterapkan.

Berangkat dari celah tersebut, penelitian ini menerapkan algoritma *SVM* dengan pendekatan *Hybrid Labeling* dan ekstraksi fitur *TF-IDF* untuk mengklasifikasikan sentimen masyarakat terhadap PP TUNAS berdasarkan komentar pada video KompasTV di *TikTok*. Penelitian ini disusun untuk menjawab dua hal utama, yaitu bagaimana distribusi sentimen masyarakat terhadap kebijakan tersebut, dan bagaimana tingkat performa algoritma *SVM* dalam mengklasifikasikan sentimen komentar berbahasa Indonesia dengan pendekatan *Hybrid Labeling*. Sebagai pendalaman tambahan, penelitian ini juga menelusuri pola kata dan topik yang paling dominan muncul pada masing-masing kelas sentimen lewat analisis *N-Gram*.

2. METODE PENELITIAN

Penelitian ini memakai pendekatan kuantitatif dengan desain deskriptif eksploratif, difokuskan pada klasifikasi sentimen berbasis *machine learning* terhadap data teks komentar publik. Alur penelitian disusun dalam beberapa tahap berurutan mulai dari pengumpulan data mentah sampai evaluasi model, seperti dijabarkan pada bagian berikut.



Gambar 2. Kerangka Alur Penelitian

2.1 Pengumpulan Data

Sumber data penelitian ini adalah komentar publik pada video berita akun *TikTok* KompasTV yang diunggah tanggal 28 Maret 2026, bertepatan dengan hari pertama efektivitas Peraturan Menteri Komunikasi dan Digital Nomor 9 Tahun 2026. Video tersebut memiliki 5,5 juta tayangan dan lebih dari 32 ribu komentar asli.

Pengambilan data dilakukan lewat teknik *scraping* memakai platform pihak ketiga *Apify*, mengingat *TikTok* tidak menyediakan akses *API* publik yang stabil untuk pengambilan komentar dalam jumlah besar. Proses ini berhasil mengumpulkan 274.087 baris data mentah, jauh melebihi jumlah komentar asli akibat mekanisme *crawling Apify* yang mengambil data secara berulang dalam beberapa sesi berbeda.

2.2 Pra-Pemrosesan Data

Seluruh tahap pra-pemrosesan dikerjakan di lingkungan *Visual Studio Code* memakai ekstensi format berkas .ipynb dengan bahasa pemrograman *Python*. Dari 274.087 baris data mentah, dilakukan seleksi kolom terlebih dulu untuk mempertahankan tiga atribut yang relevan yaitu identitas pengguna (*uniqueId*), waktu komentar (*createTimeISO*), dan isi komentar (*text*). Tahap ini menyisakan 268.393 baris data.

Langkah berikutnya adalah penghapusan duplikasi. Peneliti sengaja memilih deduplikasi berdasarkan kombinasi *uniqueId* dan *text*, bukan *text* saja, karena pendekatan berbasis *text* saja terbukti hanya menghasilkan 3.565 data dan berisiko menghapus opini unik dari pengguna berbeda yang kebetulan menuliskan kalimat serupa. Deduplikasi berbasis kombinasi *uniqueId* dan *text* menyisakan 5.393 komentar unik per pengguna.

Data hasil deduplikasi kemudian melewati tahap pembersihan linguistik secara berurutan, yaitu pembersihan dasar berupa penghapusan *URL* dan tag *HTML* disertai *case folding* dan normalisasi awal memakai kamus kata tidak baku eksternal, pembersihan karakter berupa penghapusan emoji lewat *regex Unicode* serta penghapusan simbol dan angka, normalisasi lanjutan memakai kamus kata tidak baku tambahan yang disusun manual untuk menangani singkatan dan slang khas komentar *TikTok* seperti "gak" menjadi "tidak" dan "hp" menjadi "handphone", reduksi spam repetitif lewat fungsi regex untuk mendeteksi pola pengulangan frasa, *stopword removal* memakai daftar *stopword* Bahasa Indonesia dari *library NLTK* ditambah daftar kata hapus kustom, dan terakhir *stemming* memakai *library Sastrawi* yang memang dirancang khusus untuk morfologi Bahasa Indonesia dan sudah banyak dipakai dalam penelitian *text mining* berbahasa Indonesia [27].

Perlu dicatat bahwa tokenisasi tidak dijalankan sebagai proses eksplisit tersendiri dalam *pipeline* ini. Fungsi *split()* dipakai langsung pada tahap *stopword removal* untuk memisahkan token, dan *TF-IDF vectorizer* pada tahap ekstraksi fitur juga menjalankan tokenisasi internal secara otomatis, sehingga pendekatan ini tetap valid tanpa mengurangi keabsahan *pipeline* secara keseluruhan.

Tahap terakhir adalah pembersihan final dengan menghapus baris yang mengandung nilai kosong pada kolom identitas pengguna, isi komentar, atau hasil *cleaning*, termasuk baris yang teksnya menjadi kosong setelah diproses. Tahap ini menghasilkan 4.326 data bersih yang siap dilabeli.

Before

```
--- Info Data Sebelum Pembersihan Baris Kosong ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5393 entries, 0 to 5392
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   User ID            5369 non-null   object
 1   Date               5393 non-null   object
 2   Time               5393 non-null   object
 3   Content            5362 non-null   object
 4   Cleaning Progress  5362 non-null   object
 5   Case Folding       5362 non-null   object
 6   Normalisasi        5362 non-null   object
 7   Cleaning           5393 non-null   object
 8   Final              5393 non-null   object
 9   Stopwords Removal  5393 non-null   object
 10  Stemming           5393 non-null   object
dtypes: object(11)
memory usage: 463.6+ KB
```

After

```
--- PROSES PEMBERSIHAN FINAL SELESAI ---
Jumlah data awal : 5393 baris
Jumlah data bersih: 4326 baris
File "3OutputPreProcessing.csv" berhasil diperbarui.
<class 'pandas.core.frame.DataFrame'>
Index: 4326 entries, 0 to 5383
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   User ID            4326 non-null   object
 1   Date               4326 non-null   object
 2   Time               4326 non-null   object
 3   Content            4326 non-null   object
 4   Cleaning Progress  4326 non-null   object
 5   Case Folding       4326 non-null   object
 6   Normalisasi        4326 non-null   object
 7   Cleaning           4326 non-null   object
 8   Final              4326 non-null   object
 9   Stopwords Removal  4326 non-null   object
 10  Stemming           4326 non-null   object
dtypes: object(11)
memory usage: 405.6+ KB
```

Gambar 3. Proses Transformasi Sebelum dan Sesudah Pembersihan Data

2.3 Pelabelan Data dengan *Hybrid Labeling*

Pelabelan sentimen memakai pendekatan *Hybrid Labeling* yang menggabungkan anotasi manual dan prediksi otomatis. Pada tahap pertama, 1.000 baris data pertama dilabeli manual oleh peneliti dengan dua kelas, yaitu label 1 untuk sentimen positif atau komentar yang mendukung kebijakan, dan label 0 untuk sentimen negatif atau komentar yang menolak, khawatir, atau mengkritik kebijakan. Data berlabel manual ini berfungsi sebagai *seed* data untuk melatih model awal.

Pada tahap kedua, *seed* data tersebut dipakai untuk melatih model *LinearSVC* dengan *TF-IDF* sebagai ekstraksi fitur awal. Model ini kemudian dipakai untuk memprediksi label pada 3.326 data sisanya yang belum berlabel. Penelitian ini tidak memakai kelas netral, sehingga seluruh komentar dipetakan ke salah satu dari dua kelas sentimen agar hasil analisis lebih fokus mencerminkan arah dukungan atau penolakan publik.

2.4 Pembagian Data

Dataset berlabel sebanyak 4.326 data dibagi memakai fungsi *train_test_split* dari *library scikit-learn* dengan rasio 80:20 dan parameter *random_state*=42, untuk menjaga konsistensi pembagian apabila proses penelitian ini direplikasi peneliti lain.

2.5 Ekstraksi Fitur *TF-IDF*

Fitur teks diekstraksi memakai *TfidfVectorizer* dari *scikit-learn*. *TF-IDF* menggabungkan dua komponen, yaitu *Term Frequency* yang mengukur seberapa sering sebuah kata muncul dalam satu dokumen, dan *Inverse Document Frequency* yang mengukur seberapa jarang kata tersebut muncul di seluruh dokumen dalam korpus [28]. Rumus dasarnya dapat dituliskan sebagai berikut.

$$TF\text{-}IDF(t,d) = \mathrm{TF}(t,d) \times \log\left(\frac{N}{\mathrm{DF}(t)}\right) \quad (1)$$

dengan N adalah jumlah total dokumen dan DF(t) adalah jumlah dokumen yang memuat term t.

Vectorizer ini di-fit secara eksklusif pada *data training* memakai *method fit_transform()*, lalu hasilnya dipakai untuk mentransformasi *data testing* memakai *method transform()* saja. Pendekatan dua tahap ini sengaja diterapkan untuk mencegah *data leakage*, yaitu kondisi di mana informasi statistik kosakata dari *data testing* ikut bocor ke proses pelatihan dan membuat hasil evaluasi menjadi tidak valid.

2.6 Pemodelan *LinearSVC*

Model klasifikasi yang dipakai adalah *LinearSVC* dari *library scikit-learn* dengan parameter *random_state*=42. *LinearSVC* merupakan implementasi *SVM* yang dioptimalkan untuk data berdimensi tinggi dan bersifat *sparse* seperti representasi *TF-IDF*, sehingga cocok dipakai untuk tugas klasifikasi teks. Model dilatih memakai *data training* yang sudah ditransformasi *TF-IDF*, lalu dipakai untuk memprediksi label pada *data testing* yang belum pernah dilihat model sebelumnya.

2.7 Evaluasi Model

Evaluasi performa model dilakukan memakai *confusion matrix* yang terdiri dari *True Positive*, *True Negative*, *False Positive*, dan *False Negative*, kemudian diturunkan menjadi empat metrik utama, yaitu akurasi, presisi, *recall*, dan *F1-Score*. Visualisasi *confusion matrix* disajikan dalam bentuk *heatmap* memakai *library seaborn* untuk mempermudah interpretasi distribusi prediksi benar dan salah pada masing-masing kelas.

2.8 Analisis Pola Kata Dominan

Sebagai pendalaman tambahan, dilakukan analisis pola kata dominan memakai *CountVectorizer* dengan variasi *N-Gram*, yaitu *unigram*, *bigram*, dan *trigram*. *Word cloud* dihasilkan secara terpisah untuk kelas sentimen positif dan negatif berdasarkan kolom hasil *stemming* dari dataset berlabel, dengan tujuan memperdalam interpretasi terhadap distribusi sentimen yang sudah diperoleh.

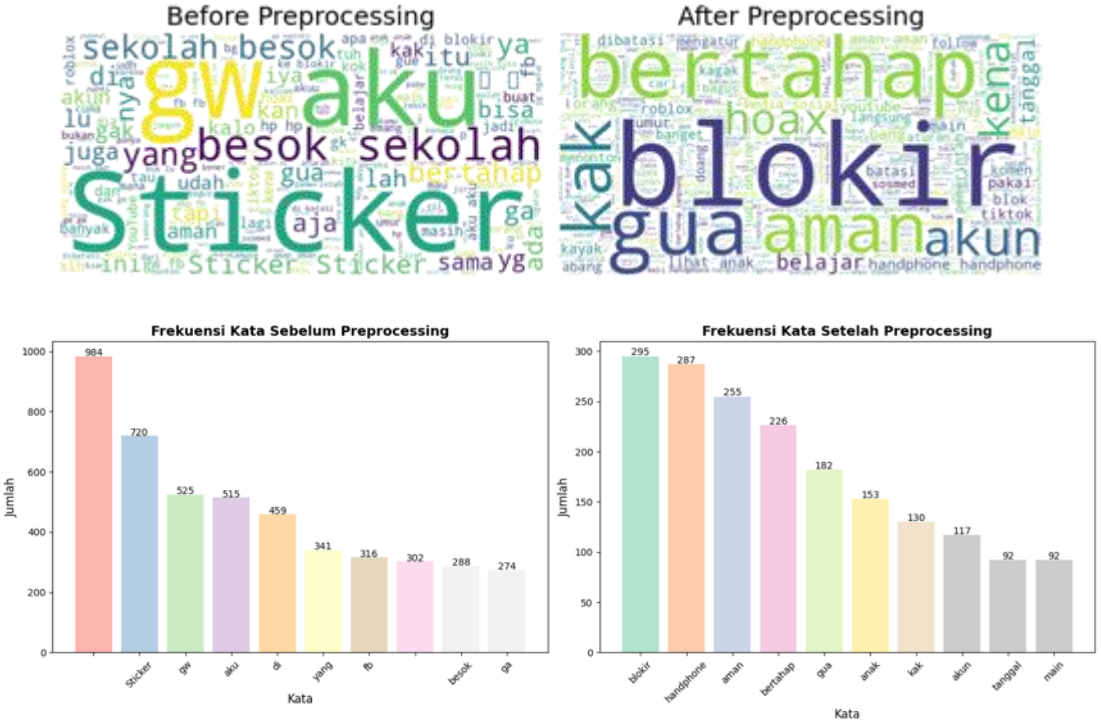
3. HASIL DAN ANALISIS

3.1 Hasil Pengumpulan dan Pemrosesan Data

Proses *scraping* berhasil mengumpulkan 274.087 baris data mentah dari kolom komentar video KompasTV di *TikTok*. Tahap pra-pemrosesan data menyusutkan jumlah data secara bertahap sebagaimana dirangkum pada tabel 1.

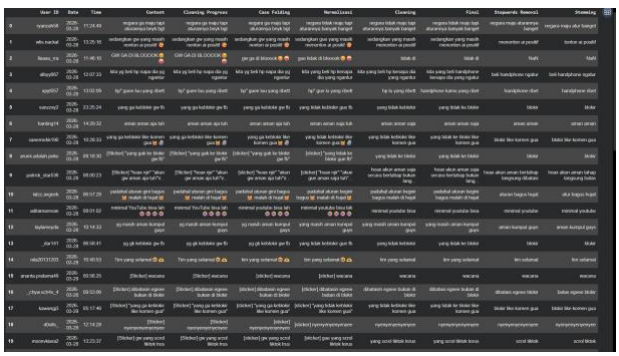
Tabel 1. Riwayat Penyusutan Data

Tahapan	Jumlah Data
Data mentah hasil *scraping Apify*	274.087 baris
Setelah seleksi kolom (uniqueId, date, text)	268.393 baris
Setelah penghapusan duplikasi (uniqueId + text)	5.393 baris
Setelah pre processing linguistik (cleaning, normalisasi, *stopword*, *stemming*)	5.393 baris
Setelah pembersihan final (null & teks kosong)	4.326 baris



Gambar 4. *Word cloud* Setelah *Preprocessing*

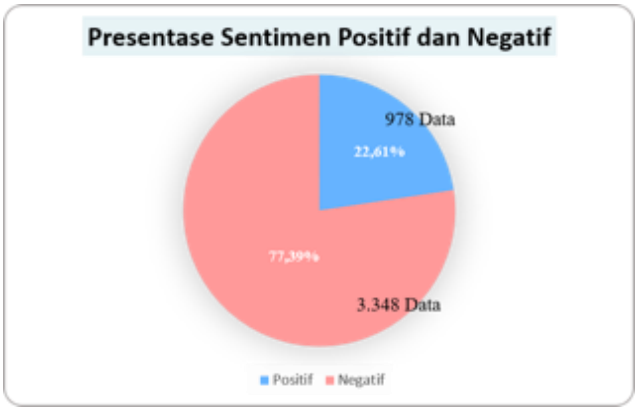
Penyusutan terbesar terjadi pada tahap penghapusan duplikasi, yaitu dari 268.393 menjadi 5.393 baris. Tahap pembersihan final kemudian mengurangi jumlah data dari 5.393 menjadi 4.326 baris akibat penghapusan baris dengan nilai kosong pada kolom identitas pengguna, isi komentar, atau hasil *cleaning*.



Gambar 5. *Cuplikan Dataset Mentah hingga Bersih*

3.2 Hasil *Hybrid Labeling*

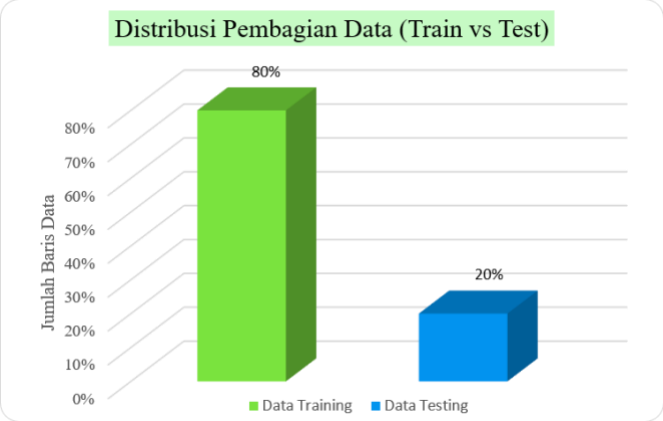
Pelabelan manual terhadap 1.000 data pertama menghasilkan 302 data sentimen positif (30,2%) dan 698 data sentimen negatif (69,8%). Pelabelan otomatis oleh model terhadap 3.326 data sisanya menghasilkan 676 data sentimen positif (20,3%) dan 2.650 data sentimen negatif (79,7%). Secara akumulatif, dataset final memuat 978 data sentimen positif (22,61%) dan 3.348 data sentimen negatif (77,39%) dari total 4.326 data.



Gambar 6. Diagram Visualisasi Presentase Kelas Positif dan Negatif

3.3 Hasil Pembagian Data

Dataset berlabel dibagi dengan rasio 80:20, menghasilkan 3.460 *data training* dan 866 data testing.

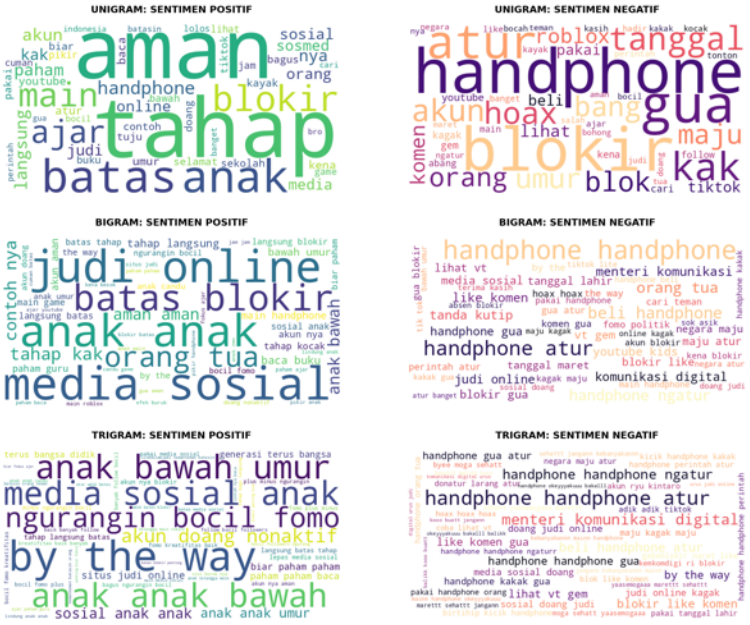


Gambar 7. Distribusi Pembagian Data

3.4 Hasil Analisis Pola Kata Dominan

Analisis *N-gram* pada kelas sentimen positif menunjukkan tiga pola kata yang dominan, yaitu ekspresi ketenangan dan keyakinan bahwa kebijakan aman, pemahaman bahwa penerapan kebijakan dilakukan secara bertahap, dan dukungan terhadap tujuan perlindungan anak dalam berinteraksi di media sosial.

Pada kelas sentimen negatif, pola yang muncul lebih beragam, mencakup komentar yang menyimpang dari konteks berita, ekspresi penolakan personal terhadap intervensi kebijakan pada penggunaan ponsel pribadi, serta keraguan terhadap keseriusan implementasi kebijakan yang oleh sebagian pengguna dianggap sebagai gertakan semata.



Gambar 8. *Word cloud* Setelah *Preprocessing*

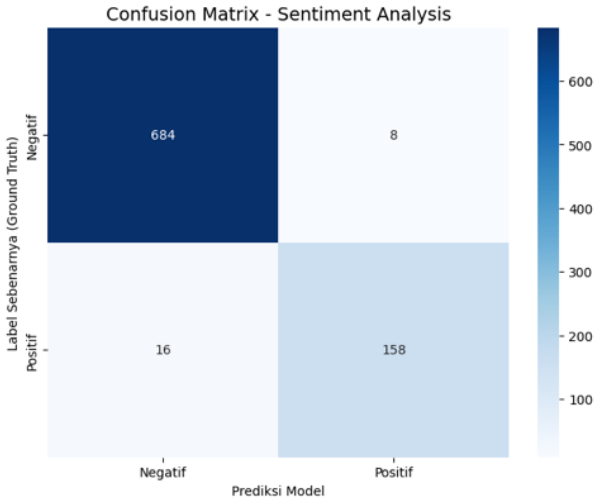
3.5 Hasil Analisis Pola Kata Dominan

Pengujian model dilakukan pada 866 *data testing* yang belum pernah dilihat model selama proses pelatihan. Tabel 2 merangkum hasil evaluasi performa model LinearSVC.

Tabel 2. Hasil Evaluasi Model

Metrik	Negatif	Positif	Macro Avg	Weight Avg
Precision	0,98	0,95	0,96	0,97
Recall	0,99	0,91	0,95	0,97
F1-Score	0,98	0,93	0,96	0,97
Support	692	174	866	866
Accuracy				97,23%

Dari 692 data aktual negatif pada *data testing*, model memprediksi benar 684 data dan salah mengklasifikasikan 8 data sebagai positif. Dari 174 data aktual positif, model memprediksi benar 158 data dan salah mengklasifikasikan 16 data sebagai negatif.



Gambar 9. *Confusion Matrix*

3.6 Pembahasan

Distribusi sentimen masyarakat terhadap kebijakan PP TUNAS yang diperoleh dari penelitian ini menjawab rumusan masalah pertama secara jelas. Dominasi sentimen negatif sebesar 77,39% berbanding sentimen positif 22,61% menunjukkan bahwa mayoritas warganet yang berkomentar pada video KompasTV cenderung menyatakan penolakan atau kekhawatiran terhadap kebijakan pembatasan usia media sosial ini. Pola serupa konsisten muncul baik pada data yang dilabeli manual (69,8% negatif) maupun data yang diprediksi mesin (79,7% negatif), yang mengindikasikan bahwa model berhasil mempelajari dan meneruskan kecenderungan sentimen dari *seed* data dengan cukup baik.

Hasil ini secara langsung menolak hipotesis nol yang menduga tidak ada kecenderungan sentimen yang mendominasi untuk menolak kebijakan, dan sebaliknya menerima hipotesis alternatif bahwa sentimen masyarakat justru dominan menolak kebijakan PP TUNAS. Temuan ini selaras dengan pola yang lebih dulu terlihat pada penelitian dengan topik kebijakan pemerintah lainnya di media sosial, yang juga kerap menunjukkan sentimen kritis publik terhadap kebijakan baru yang menyentuh ruang privat masyarakat [15].

Ditelisik lebih jauh lewat analisis pola kata dominan, dominasi sentimen negatif ini bukan sekadar penolakan terhadap substansi kebijakan perlindungan anak itu sendiri. Sebagian besar komentar negatif justru menyasar isu di luar konteks berita, ekspresi penolakan personal terhadap intervensi pada penggunaan ponsel pribadi, dan keraguan terhadap keseriusan implementasi kebijakan.

Pola ini menunjukkan bahwa resistensi publik yang tertangkap dari komentar *TikTok* tidak selalu mencerminkan penolakan murni terhadap tujuan perlindungan anak, melainkan lebih banyak dipicu kekhawatiran soal batasan kebebasan personal dan ketidakpercayaan terhadap konsistensi penerapan kebijakan pemerintah. Sebaliknya, sentimen positif justru lebih fokus pada apresiasi terhadap tujuan kebijakan dan pemahaman bahwa penerapannya dilakukan bertahap, bukan mendadak.

Rumusan masalah kedua, hasil evaluasi menunjukkan bahwa algoritma *SVM* dengan pendekatan *Hybrid Labeling* mampu mengklasifikasikan sentimen komentar berbahasa Indonesia dengan performa yang sangat baik, tercermin dari akurasi 97,23% dan *F1-Score macro average* sebesar 0,96. Perbandingan dengan penelitian terdahulu yang juga memakai *SVM* pada data media sosial berbahasa Indonesia memperlihatkan performa penelitian ini berada di kisaran atas, mengungguli penerapan *SVM* pada isu Program Makan Bergizi Gratis yang hanya mencapai akurasi 77,75% [20] serta layanan *Gojek* di *Instagram* dengan akurasi 70,82% [21], meski tetap berada dalam rentang yang sebanding dengan penelitian lain yang mencatat akurasi di atas 95% pada topik kebijakan kenaikan pajak hiburan [18].

Ketimpangan performa antara kelas negatif dan positif turut terlihat jelas dari *confusion matrix*, di mana *recall* kelas negatif mencapai 0,99 sementara *recall* kelas positif hanya 0,91. Pola ini merupakan konsekuensi logis dari ketidakseimbangan kelas dalam dataset, dengan rasio negatif berbanding positif mendekati 3,4 banding 1. Kelas negatif yang punya representasi data jauh lebih besar membuat model lebih terlatih mengenali pola bahasa pada kelas tersebut dibanding kelas positif yang jumlah datanya lebih sedikit.

Tingginya performa model ini juga perlu dibaca secara kritis mengingat karakteristik pendekatan *Hybrid Labeling* yang dipakai. Sekitar 77% dari total data berlabel dalam penelitian ini sebenarnya berasal dari prediksi model awal, bukan anotasi manusia secara langsung. Kondisi ini membuka kemungkinan bahwa sebagian dari tingginya angka akurasi yang diperoleh turut dipengaruhi oleh kesamaan pola antara model evaluasi akhir dengan model *seed* yang dipakai untuk melabeli data secara otomatis. Meski begitu, distribusi sentimen pada *seed* data manual dan data hasil prediksi otomatis yang konsisten searah memberi indikasi awal bahwa propagasi label tidak menyimpang jauh dari pola aslinya, walau validasi lebih mendalam terhadap kualitas *seed* model tetap diperlukan pada penelitian lanjutan.

Dari sisi metodologi pengumpulan data, tahap deduplikasi yang menyusutkan 268.393 baris data menjadi hanya 5.393 baris menegaskan bahwa *scraping* lewat *Apify* memang rawan menghasilkan duplikasi masif akibat mekanisme *crawling* yang mengambil data secara berulang. Pemilihan strategi deduplikasi berbasis kombinasi identitas pengguna dan isi teks, bukan teks saja, terbukti penting untuk menjaga keberagaman opini pengguna berbeda yang kebetulan menuliskan komentar serupa, sekaligus menjadi catatan metodologis yang relevan bagi peneliti lain yang berencana memakai teknik *scraping* serupa pada platform *TikTok*.

Secara keseluruhan, hasil penelitian ini memperkuat bukti empiris bahwa kombinasi *SVM*, *TF-IDF*, dan *Hybrid Labeling* cukup andal dipakai untuk membaca opini publik terhadap kebijakan pemerintah lewat data komentar *TikTok*, sekaligus mengisi celah penelitian yang selama ini belum banyak menyasar isu kebijakan pembatasan usia media sosial berdasarkan PP No. 17 Tahun 2025 secara spesifik.

4. KESIMPULAN

Penerapan algoritma *Support Vector Machine* dengan pendekatan *Hybrid Labeling* dan ekstraksi fitur *TF-IDF* untuk menganalisis sentimen masyarakat terhadap PP No. 17 Tahun 2025 pada komentar video KompasTV di platform *TikTok*, dapat disimpulkan beberapa hal berikut. Pertama, distribusi sentimen masyarakat terhadap kebijakan PP TUNAS didominasi oleh sentimen negatif. Dari total 4.326 data komentar yang dianalisis, sebanyak 77,39% atau 3.348 komentar diklasifikasikan sebagai sentimen negatif, sementara 22,61% atau 978 komentar diklasifikasikan sebagai sentimen positif. Hasil ini menolak hipotesis nol dan menerima hipotesis alternatif bahwa sentimen masyarakat dominan menolak kebijakan pembatasan usia media sosial tersebut. Kedua, algoritma *SVM* dengan pendekatan *Hybrid Labeling* terbukti efektif mengklasifikasikan sentimen komentar berbahasa Indonesia pada platform *TikTok*. Model *LinearSVC* yang dibangun mencapai akurasi 97,23% pada *data testing*, dengan *F1-Score macro average* sebesar 0,96, *precision* 0,96, dan *recall* 0,95. Ketiga, analisis pola kata dominan lewat pendekatan *N-gram* menunjukkan pola yang berbeda antara kedua kelas sentimen. Sentimen positif lebih banyak mengarah pada apresiasi terhadap tujuan perlindungan anak dan pemahaman bahwa kebijakan diterapkan bertahap, sementara sentimen negatif lebih beragam dan mengarah pada kekhawatiran terhadap intervensi kebebasan pribadi serta keraguan terhadap konsistensi implementasi kebijakan.

Temuan ini punya implikasi praktis yang cukup luas. Bagi pembuat kebijakan, khususnya Kementerian Komunikasi dan Digital, tingginya angka sentimen negatif bisa menjadi masukan penting untuk mengevaluasi strategi komunikasi publik dan memperjelas mekanisme teknis penerapan kebijakan, misalnya lewat sosialisasi yang lebih transparan mengenai proses verifikasi usia dan jaminan perlindungan privasi data anak. Bagi platform digital, kerja sama lebih aktif dengan pemerintah dalam menyediakan mekanisme verifikasi usia yang ramah privasi juga menjadi langkah yang relevan untuk diambil. Pendekatan komputasional yang dipakai dalam penelitian ini turut membuka peluang aplikasi yang lebih luas, mulai dari pemantauan sentimen publik terhadap kebijakan pemerintah lain secara berkelanjutan, sampai menjadi rujukan metodologis bagi kajian *text mining* pada domain kebijakan berbeda yang memakai data media sosial berbahasa Indonesia.

Penelitian selanjutnya, disarankan memperluas sumber data tidak hanya dari satu video pada satu akun saja, melainkan dari berbagai video dan akun yang membahas kebijakan PP TUNAS agar hasil analisis lebih representatif terhadap opini publik secara keseluruhan. Penanganan ketidakseimbangan kelas lewat teknik seperti *SMOTE*, pembobotan kelas, atau *undersampling* juga layak dicoba, begitu juga perbandingan performa *SVM* dengan model berbasis transformer seperti *IndoBERT* atau *RoBERTa* untuk melihat potensi peningkatan performa pada data komentar *TikTok* yang sangat informal. Penambahan analisis berbasis aspek turut disarankan agar penelitian selanjutnya bisa mengidentifikasi lebih spesifik aspek kebijakan mana saja yang paling banyak menuai kritik atau dukungan dari masyarakat.

REFERENSI

[1] Susenas, “Persentase Penduduk Usia 5 Tahun ke Atas yang Pernah Mengakses Internet dalam 3 Bulan Terakhir Menurut Provinsi dan Kelompok Umur - Tabel Statistik - Badan Pusat Statistik Indonesia,” Indonesia, Oct. 2025. Accessed: Jun. 18, 2026. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NDIwIzI=/persentase-penduduk-usia-5-tahun-ke-atas-yang-pernah-mengakses-internet-dalam-3-bulan-terakhir-menurut-provinsi-dan-kelompok-umur.html

[2] R. Al Adawiah and F. N. Eleanora, “Perundungan Dunia Maya pada Anak: Tinjauan Fenomena dan Tren dalam Rentang 2016–2020,” *Aspirasi: Jurnal Masalah-masalah Sosial*, vol. 14, no. 1, pp. 99–117, Jun. 2023, doi: 10.46807/aspirasi.v14i1.3065.

[3] Australian Legislation, “Online Safety Amendment (Social Media Minimum Age) Act 2024 - Federal Register of Legislation,” pp. 1–16, Dec. 2024, Accessed: Jun. 16, 2026. [Online]. Available: https://www.legislation.gov.au/C2024A00127/asmade/text

[4] M. E. Aminanto and D. Pitaloka, “Indonesia’s age-restriction policy for social media access needs more than just a decree,” Govinsider. Accessed: Jun. 16, 2026. [Online]. Available: https://govinsider.asia/intl-en/article/indonesias-age-restriction-policy-for-social-media-access-needs-more-than-just-a-decree

[5] Unicef, “How many children are there in Indonesia?,” UNICEF. Accessed: Jun. 16, 2026. [Online]. Available: https://data.unicef.org/how-many/how-many-children-under-18-are-there-in-indonesia/

[6] Australian Government, “Social media age restrictions | eSafety Commissioner,” eSafety Commissioner. Accessed: Jun. 16, 2026. [Online]. Available: https://www.esafety.gov.au/about-us/industry-regulation/social-media-age-restrictions

[7] Z. N. Adam, “Regulasi Usia Minimal Akses Media Sosial Anak Di Indonesia: Analisis Yuridis Verifikasi Usia,” *PERAHU (Penerangan Hukum) : Jurnal Ilmu Hukum*, vol. 14, no. 1, pp. 28–40, Mar. 2026, Accessed: Jun. 16, 2026. [Online]. Available: https://www.jurnal.unka.ac.id/index.php/Perahu/article/view/1803/1214

[8] Dr. K. Muller, A. G. Dionisio, and S. Park, “Online Knowledge And Practice Of Parents And Children In Indonesia,” Jakarta: UNICEF, 2023, pp. 3–232. Accessed: Jun. 16, 2026. [Online]. Available: https://www.unicef.org/indonesia/media/23586/file/online-knowledge-practice-parents-and-children-Indonesia-baseline-study-2023.pdf

[9] H. KPAI, "Kasus Kekerasan di Sekolah Meningkat, KPAI Desak Reformasi Menyeluruh Sistem Pendidikan Aman Anak." Accessed: Jun. 23, 2026. [Online]. Available: https://www.kpai.go.id/publikasi/kasus-kekerasan-di-sekolah-meningkat-kpai-desak-reformasi-menyeluruh-sistem-pendidikan-aman-anak

[10] PP RI, "Tata Kelola Penyelenggaraan Sistem Elektronik Dalam Perlindungan Anak," 2025. Accessed: May 07, 2026. [Online]. Available: https://peraturan.bpk.go.id/Details/316698/pp-no-17-tahun-2025

[11] I-NAMHS, "National Survey Report," Yogyakarta: Center for Reproductive Health, 2022. Accessed: Jun. 23, 2026. [Online]. Available: https://publichealth.jhu.edu/sites/default/files/2025-05/I-NAMHS-Report-English.pdf

[12] R. Daka *et al.*, "Adolescent Mental Health In The Digital Age: A Systematic Review Of Social Media Impact And Psychosocial Determinants," *Jurnal Promkes*, vol. 13, no. SI2, pp. 201–208, Aug. 2025, doi: 10.20473/jpk.V13.ISI2.2025.201-208.

[13] UNICEF, "PERUNDUNGAN DI INDONESIA: Fakta-fakta Kunci, Solusi, dan Rekomendasi," 2023. Accessed: Jun. 23, 2026. [Online]. Available: https://indonesia.ureport.in/v2/opinion/3454/

[14] KomDigi, "Peraturan Menteri Komdigi Terbit, Pemerintah Mulai Implementasi Perlindungan Anak di Platform Digital." Accessed: Jun. 16, 2026. [Online]. Available: https://www.komdigi.go.id/berita/siaran-pers/detail/permen-komdigi-terbit-pemerintah-mulai-implementasi-perlindungan-anak-di-platform-digital

[15] I. Najiyah and M. Rizal, "Analisis Sentimen Media Sosial X Terhadap Kebijakan Presiden Republik Indonesia Prabowo Subianto," *JURNAL FASILKOM*, vol. 15, no. 3, pp. 443–455, Dec. 2025, doi: https://doi.org/10.37859/jf.v15i3.10385.

[16] C. Nursyahbani, R. W. Pusari, and Purwadi, "Perlindungan Anak Berbasis Hak di Ruang Digital: Systematic Literature Review terhadap Risiko dan Strategi Perlindungan," *PAUDIA (Jurnal Penelitian dalam Bidang Pendidikan Anak Usia Dini)*, vol. 15, no. 1, pp. 69–82, Feb. 2026, doi: https://doi.org/10.26877/paudia.v15i1.3192.

[17] W. Gumilang and A. Riyandi, "Sentimen Analisis Pengguna Twitter Terhadap Sea Games 2023 Dengan Metode Naive Bayes," *JURNAL AKADEMIKA*, vol. 16, no. 1, pp. 95–100, Nov. 2023, doi: https://doi.org/10.53564/akademika.v16i1.1125.

[18] W. Romadhona and A. R. Isnain, "Analisis Sentimen Pengguna Media Sosial Terhadap Kebijakan Kenaikan Pajak Hiburan Menggunakan Metode SVM (Support Vector Machine)," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 4, pp. 2185–2195, Nov. 2024, doi: https://doi.org/10.29100/jipi.v9i4.5603.

[19] Sukirman, Sajiah, N. Putri Husain, A. Febriana Syam, and R. Mustikosari, "Analisis Sentimen Ulasan Pengguna Tiktok pada Google Play Store Berbasis TF-IDF dan Support Vector Machine," *Journal of System and Computer Engineering (JSCE) ISSN*, vol. 5, no. 1, pp. 91–102, Jan. 2024, Accessed: May 12, 2026. [Online]. Available: https://doi.org/10.61628/jsce.v5i1.1105

[20] M. Azril and G. Crisnawati, "Klasifikasi Sentimen Komentar Pengguna Tiktok Mengenai Program Makan Bergizi Gratis (MBG) Dengan Support Vector Machine (SVM)," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 5, no. 1, pp. 8859–8866, Mar. 2026, doi: https://doi.org/10.31004/riggs.v5i1.7237.

[21] M. J. Savero, A. Ibrahim, Y. Utama, and E. Lestari, "Harnessing SVM for Sentiment Analysis: Insights from Gojek's Instagram Engagement," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 663–680, Mar. 2025, doi: https://doi.org/10.51519/journalisi.v7i1.1041 .

[22] A. L. D. Ulhaq and Suprayogi, "Comparing Machine Learning Models for Sentiment Analysis of Tokopedia Reviews," *JAIC (Journal of Applied Informatics and Computing)*, vol. 9, no. 6, pp. 3642–3647, Dec. 2025, [Online]. Available: https://doi.org/10.30871/jaic.v9i6.11239

[23] E. Supriyadi and P. N. Makatita, "Sentiment Analysis of TikTok User Comments on QRIS Adoption in Indonesia Using IndoBERT," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 121–130. doi: 10.1016/j.procs.2025.08.265.

[24] A. L. Muji, A. Musthofa, and D. Muriyatmoko, "Analisis Sentimen Publik pada TikTok terhadap Rencana Penerapan Sistem Balik Nama Ponsel Bekas menggunakan Naive Bayes dan Support Vector Machine," *PROSEMNASPROIT (Prosiding Seminar Nasional Ilmu Teknik)*, vol. 2, Dec. 2025, Accessed: Jun. 17, 2026. [Online]. Available: https://prosiding.aritekin.or.id/index.php/PROSEMNASPROIT/article/view/198

[25] A. Oktaviyanti, Firmansyah, and A. R. Kadafi, "Analisis Sentimen Ulasan Pelecehan Di Media Sosial Tiktok Menggunakan Svm Dan Pengelompokan Pola Kejadian Dengan K-Means Clustering," *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, vol. 13, no. 3S1, pp. 1938–1945, Oct. 2025, [Online]. Available: https://journal.eng.unila.ac.id/index.php/jitet/article/view/8149

[26] I. P. A. P. Widiarta, R. Dwiyansaputra, and A. Aranta, "Analisis Sentimen Masyarakat Terhadap Kebijakan Penerapan Ppkm Di Media Sosial Twitter Dengan Menggunakan Metode Xgboost," *JTIKA (Jurnal Teknologi Informasi, Komputer dan Aplikasinya)*, vol. 5, pp. 154–163, Sep. 2023, [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA

[27] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving The Accuracy Of Text Classification Using Stemming Method, A Case Of Non-Formal Indonesian Conversation," *J. Big Data*, vol. 8, no. 1, pp. 1–16, Jan. 2021, doi: 10.1186/S40537-021-00413-1.

[28] S. F. Kadir and A. Fairuzabadi, "Analisis Sentimen Ulasan Shopee di Google Play dengan TF-IDF dan Logistic Regression," *Journal of Artificial Intelligence and Digital Business (RIGGS)*, vol. 4, no. 2, pp. 7940–7945, Jul. 2025, doi: 10.31004/riggs.v4i2.2850. Available: https://journal.ilmudata.co.id/index.php/RIGGS/article/view/2850