

# Analisis Emosi Terhadap Komentar Video Youtube “Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia” Menggunakan Metode Random Forest

Nicodemus Alfriyanto<sup>\*1</sup>, Bintang Cahaya Purnama<sup>2</sup>, Fadhilah Khairunisa Hasanah<sup>3</sup>,  
Khairun Nisa Meiah Ngafidin<sup>4</sup>

<sup>1,2,3,4</sup>*Sistem Informasi, Telkom University Purwokerto  
Jl.D.I Panjaitan No.128, Kec. Purwokerto Selatan, Kabupaten Banyumas, Jawa Tengah 53147*

<sup>\*1</sup>21103077@ittelkom-pwt.ac.id

<sup>3</sup>21103100@ittelkom-pwt.ac.id

<sup>4</sup>nisa@ittelkom-pwt.ac.id

<sup>2</sup>21103048@itelkom-pwt.ac.id

Diterima pada 20-11-2024, direvisi pada 27-11-2024, diterima pada 04-12-2024

## Abstrak

Sistem pendidikan Finlandia telah menjadi topik perbandingan di berbagai negara, termasuk Indonesia, namun implementasinya di Indonesia menghadapi berbagai tantangan. Penelitian ini bertujuan untuk menganalisis emosi yang terkandung dalam komentar YouTube mengenai video berjudul “Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia” menggunakan metode *random forest*. Penelitian ini mengumpulkan 12.337 komentar dari video tersebut melalui YouTube Data API dan melakukan preprocessing data untuk memastikan keseragaman dan konsistensi. Selanjutnya, analisis emosi dilakukan dengan memanfaatkan *lexicon SenticNet* untuk mengklasifikasikan emosi dalam empat kategori: *introspeksi*, *temper*, *attitude*, dan *sensitivitas*. Hasil analisis menunjukkan bahwa pandangan publik terhadap sistem pendidikan Finlandia di Indonesia beragam, dengan emosi positif seperti *thoughtfulness* dan *enthusiasm* mendominasi, sementara emosi negatif seperti *fear* dan *disgust* lebih sedikit. Evaluasi model menggunakan *confusion matrix* menunjukkan bahwa *random forest* dengan teknik resampling menggunakan Random Over Sampling (ROS) memberikan hasil yang optimal. Penelitian ini menegaskan pentingnya pemilihan metode analisis yang tepat dalam menangani data tidak terstruktur dan memberikan wawasan lebih dalam mengenai tanggapan publik terhadap kebijakan pendidikan.

**Kata Kunci:** Analisis Emosi, Komentar, Pendidikan Finlandia, Random Forest, Youtube

*Ini adalah artikel akses terbuka di bawah lisensi [CC BY-SA](#).*



---

### Penulis Koresponden:

Nicodemus Alfriyanto Riskiwibowo  
Program Studi Desain Komunikasi Visual, Universitas Telkom Kampus Purwokerto, Jl. D.I Panjaitan No.128 Purwokerto, 53147  
Kabupaten Banyumas, Jawa Tengah, Indonesia Email: 21103077@ittelkom-pwt.ac.id

---

## I. PENDAHULUAN

Sistem pendidikan Finlandia telah menarik perhatian banyak negara di dunia karena dianggap berhasil menciptakan siswa yang berkualitas dan lingkungan belajar yang mendukung [1]. Ciri-ciri utama sistem pendidikan Finlandia, seperti pendekatan holistik, fleksibilitas dalam pembelajaran, dan fokus pada kesejahteraan siswa, telah membuatnya menjadi model yang banyak dipertimbangkan untuk diadopsi di berbagai negara, termasuk Indonesia. Namun, implementasi sistem ini di Indonesia tidaklah mudah, mengingat adanya tantangan terkait perbedaan budaya, infrastruktur pendidikan yang belum memadai, serta kesiapan guru dan siswa dalam mengadopsi pendekatan yang berbeda [2].

Video di YouTube yang membahas sistem pendidikan Finlandia dan alasan kegagalannya untuk diadopsi di Indonesia menimbulkan berbagai respons dari masyarakat. Platform digital seperti YouTube memberikan ruang bagi publik untuk menyuarakan pendapat mereka secara bebas [3]. Komentar-komentar yang ditulis oleh pengguna YouTube dapat memberikan wawasan penting tentang bagaimana masyarakat Indonesia memandang sistem pendidikan Finlandia dan potensi implementasinya di tanah air. Analisis terhadap komentar-komentar ini dapat membantu kita memahami emosi publik, apakah dominan positif, negatif, atau campuran, yang terkait dengan topik ini [4].

Dalam beberapa tahun terakhir, analisis emosi pada teks media sosial telah menjadi topik penelitian yang berkembang pesat. Melalui platform media sosial, persepsi masyarakat terhadap suatu isu dapat dipelajari dengan lebih mendalam, dan hal ini penting untuk memahami reaksi publik terhadap kebijakan atau topik tertentu [5]. Metode *Random Forest* dipilih dibandingkan *Naive Bayes* karena *Random Forest* mampu menangani data yang tidak seimbang dengan lebih baik dan memiliki kemampuan untuk menangkap kompleksitas hubungan antar fitur dalam dataset. Selain itu, *Random Forest* cenderung lebih tahan terhadap overfitting karena menggunakan teknik *ensemble*, sehingga hasil analisis lebih stabil dan akurat. D. P. Sinambela et al. [6] mengatakan Metode *Random Forest* telah menjadi salah satu pendekatan yang populer untuk analisis emosi karena kemampuannya dalam mengklasifikasikan data teks dengan tingkat akurasi yang baik. Metode ini juga dianggap efektif dalam mengolah data yang tidak terstruktur seperti komentar di media sosial.

Kajian sebelumnya menunjukkan bahwa sistem pendidikan Finlandia telah lama menjadi topik perbandingan untuk meningkatkan kualitas pendidikan di berbagai negara, termasuk Indonesia. Agustyaningrum dan Himmi [7] membahas *best practices* dari sistem pendidikan Finlandia yang bisa diadopsi di Indonesia, seperti prinsip kesetaraan, kualitas guru, dan budaya kepercayaan. Studi tersebut menekankan pentingnya adaptasi aspek-aspek positif dari pendidikan Finlandia guna meningkatkan kualitas pendidikan di Indonesia yang hingga kini masih menghadapi berbagai kendala. Karimah et al. [8] melakukan analisis sentimen terhadap komentar video terkait mobil listrik di YouTube menggunakan metode *Naive Bayes*. Studi ini menunjukkan bahwa metode *Naive Bayes* cukup efektif untuk mengklasifikasikan sentimen positif, negatif, dan netral dari komentar-komentar yang diunggah pengguna. Penggunaan metode ini menginspirasi penelitian saat ini untuk mengeksplorasi emosi dalam konteks pendidikan, khususnya mengenai sistem pendidikan Finlandia di Indonesia. Sanjaya dan Lhaksmana [9] meneliti respons publik terhadap isu politik di YouTube dengan menggunakan pendekatan berbasis leksikon. Hasil penelitian mereka menunjukkan bahwa analisis emosi pada teks media sosial dapat memberikan wawasan mendalam tentang persepsi publik terhadap isu tertentu. Penelitian ini relevan dengan penelitian saat ini yang bertujuan untuk menggali emosi publik terhadap sistem pendidikan asing. Selain itu, Afdhal et al. [10] menggunakan algoritma *Random Forest* untuk menganalisis sentimen terhadap komentar YouTube tentang *Islamofobia*. Studi ini menunjukkan bagaimana analisis sentimen dapat digunakan untuk mengidentifikasi persepsi publik terkait isu sosial yang sensitif, serta menekankan pentingnya penggunaan metode yang tepat dalam menangani data yang tidak terstruktur. Kajian ini mendukung penggunaan metode *Random Forest* dalam penelitian ini untuk mengklasifikasikan emosi dari komentar YouTube secara efektif.

Penelitian ini bertujuan untuk mengeksplorasi emosi yang terkandung dalam komentar YouTube mengenai video berjudul “Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia”. Dengan menggunakan metode *Random Forest*, penelitian ini diharapkan dapat menghasilkan analisis yang komprehensif tentang emosi publik, sehingga dapat memberikan wawasan yang lebih mendalam bagi pembuat kebijakan dan pihak terkait lainnya.

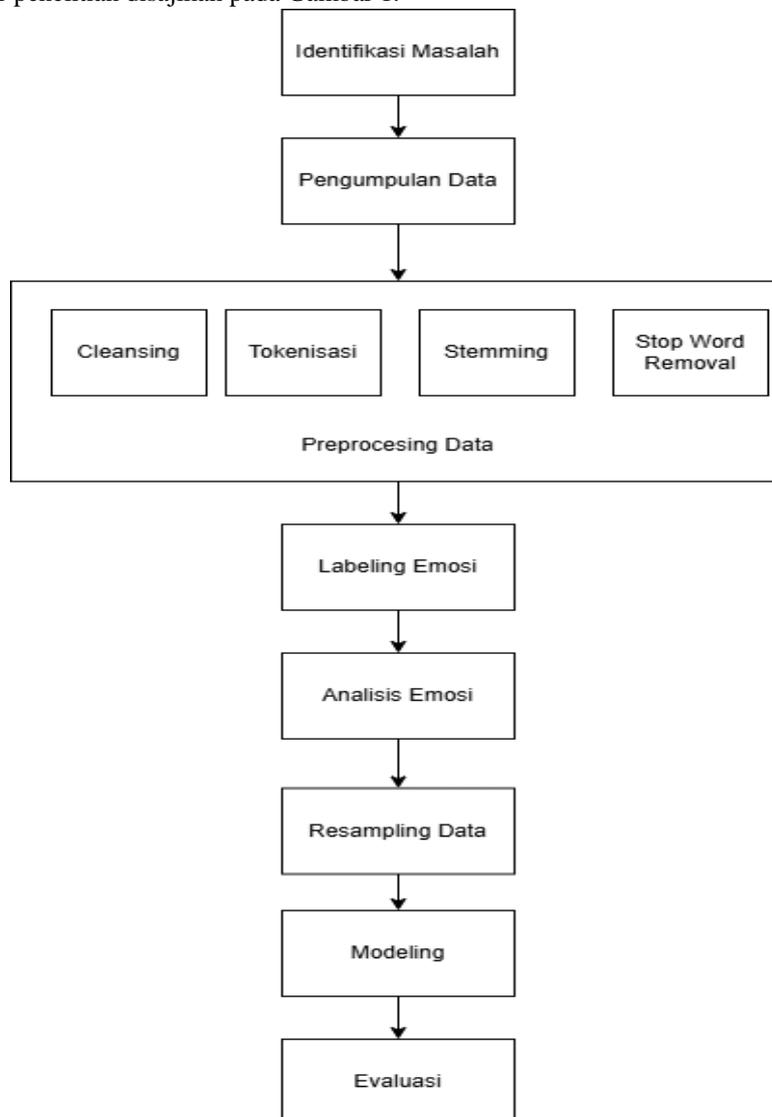
## II. METODE PENELITIAN

### A. Analisis Masalah dan Studi Literatur

Penelitian ini melakukan analisis reaksi emosi publik terhadap komentar YouTube mengenai video berjudul “Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia”. Beberapa isu utama yang ditemukan di sini adalah untuk menganalisis adanya beragam reaksi masyarakat dengan cara melihat komentar mereka terhadap video yang menunjukkan pandangan yang campur aduk, termasuk positif,

negatif, dan netral. Penggunaan metode *Random Forest*, untuk analisis emosi dalam studi ini melibatkan pengkajian literatur yang telah ada, sehingga menghasilkan wawasan tentang bagaimana masyarakat Indonesia memahami potensi penerapan sistem pendidikan Finlandia.

Penelitian ini dilaksanakan secara bertahap dan terstruktur, di mana setiap tahapannya menggunakan metode yang kritis dan mengarahkan pada proses yang tepat. Metode penelitian ini mencakup teknik-teknik yang digunakan dalam merumuskan, menganalisis, mengumpulkan data, hingga implementasi hasil penelitian. Alur penelitian disajikan pada Gambar 1.



Gambar 1. Alur Penelitian

Penelitian ini menggunakan dataset yang berasal dari komentar pada video YouTube yang relevan dengan topik "Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia". Komentar-komentar tersebut diambil menggunakan YouTube Data API dan kemudian diekspor dalam bentuk (*Comma Separated Values*) CSV untuk diproses lebih lanjut. Data tersebut berisi teks komentar yang nantinya akan digunakan untuk analisis sentimen dan klasifikasi emosi publik terkait topik yang dibahas.

### 1. Identifikasi Masalah

Langkah ini bertujuan untuk mengerti permasalahan yang ingin diatasi, seperti menjalankan analisis emosi dari teks untuk memperoleh wawasan tentang perasaan orang terkait suatu isu. Pada tahap ini, penting untuk menjelaskan dengan jelas tujuan, ruang lingkup, dan manfaat dari analisis emosi agar proyek memiliki arah yang lebih jelas.

## 2. Pengumpulan Data

Setelah masalah diketahui, langkah berikutnya adalah mengumpulkan data yang relevan. Data ini dapat berasal dari berbagai sumber, seperti postingan di media sosial, ulasan produk, survei, atau data teks lainnya. Kualitas data yang dikumpulkan sangat penting karena akan mempengaruhi hasil akhir analisis. Selain itu, penting juga untuk memastikan bahwa data yang digunakan bersifat representatif dan cukup bervariasi.

### a. *Preprocessing data*

*Preprocessing* data adalah langkah awal yang sangat penting dalam analisis teks, terutama untuk mempersiapkan data mentah agar dapat diproses lebih lanjut dengan metode statistik atau algoritma pembelajaran mesin. *Cleansing* data dibersihkan dari elemen-elemen yang tidak perlu. Ini dapat mencakup penghapusan tanda baca, emotikon, *URL*, simbol, atau karakter khusus lainnya yang tidak berkontribusi pada analisis. Tujuan utama dari langkah ini adalah untuk memastikan bahwa teks siap untuk diproses lebih lanjut [11]. Tokenisasi merupakan proses memecah teks menjadi bagian-bagian yang lebih kecil yang disebut "token"[12]. Dalam konteks teks, token dapat berupa kata atau kalimat. Proses ini penting karena algoritma analisis teks memerlukan data yang terstruktur dalam bentuk kata-kata yang bisa dianalisis. Stemming merupakan teknik untuk mengubah kata menjadi bentuk dasarnya (*root word*) [13]. Misalnya, kata-kata seperti "berlari", "lari", dan "lari-lari" diubah menjadi "lari". Stemming membantu mengurangi variasi kata sehingga analisis menjadi lebih konsisten. *Stop words* adalah kata-kata yang sering muncul tetapi memiliki makna yang sedikit dalam konteks analisis teks, seperti "dan", "yang", "adalah", "atau"[13]. Menghapus *stop words* membantu dalam menyederhanakan data dan fokus pada kata-kata yang lebih bermakna.

### b. *Labeling Emosi*

Labeling emosi merupakan proses memberikan label pada data berdasarkan emosi yang terkandung di dalamnya, seperti senang, marah, sedih, atau netral. Ini bisa dilakukan secara manual atau otomatis[14]. Tujuannya adalah agar data siap digunakan dalam pelatihan model machine learning. Label-label ini nantinya akan digunakan untuk mengajarkan model dalam mengenali emosi dalam teks. Setelah proses labeling dilakukan, digunakan Principal Component Analysis (PCA) untuk memvisualisasikan persebaran data berlabel tersebut. Proses visualisasi ini memberikan gambaran yang jelas mengenai distribusi masing-masing emosi dalam dataset, sehingga dapat membantu dalam memahami pola data serta memastikan bahwa semua kelas emosi memiliki distribusi yang dapat dideteksi dengan baik oleh model. Visualisasi ini juga bermanfaat untuk melihat apakah data sudah merata atau masih ada ketidakseimbangan yang signifikan antar kelas emosi.

### c. *Analisis Emosi*

Setelah data diberi label, dilakukan analisis emosi untuk mengetahui pola emosional yang terdapat dalam data. Analisis ini dapat memberikan wawasan tentang sentimen atau perasaan umum yang ada dalam data, misalnya menemukan apakah reaksi terhadap suatu produk lebih banyak negatif atau positif. Pada tahap ini, teknik statistik dan visualisasi data sering digunakan untuk mengungkap informasi dari data yang telah diproses.

### d. *Resampling Data*

*Resampling* data dilakukan ketika ada ketidakseimbangan dalam distribusi kelas emosi pada dataset. Contohnya, jika jumlah data dengan label "senang" jauh lebih banyak dibandingkan dengan "marah" atau "sedih". Ketidakseimbangan ini dapat menghambat model agar tidak dapat belajar seluruh kelas dengan merata. Teknik resampling dapat dilakukan dengan menambah jumlah sampel pada kelas yang kurang (*oversampling*) atau mengurangi jumlah sampel pada kelas yang berlebihan (*undersampling*) agar model dapat belajar dengan lebih efisien dan adil. Setelah proses labeling dilakukan, Principal Component Analysis (PCA) digunakan untuk memvisualisasikan persebaran data yang telah diberi label, guna melihat bagaimana distribusi data pada setiap kelas emosi. Hal ini memberikan gambaran mengenai persebaran dan apakah setiap kelas memiliki representasi yang cukup dalam dataset. Setelah itu, proses resampling

dilakukan untuk mengatasi ketidakseimbangan kelas. Setelah resampling, PCA kembali digunakan untuk mereduksi dimensi data berbasis TF-IDF (Term Frequency-Inverse Document Frequency) menjadi dua komponen utama, sehingga memungkinkan visualisasi persebaran data dalam ruang dua dimensi. Proses visualisasi ini membantu membandingkan distribusi sebelum dan setelah resampling, memungkinkan analisis yang lebih baik terkait keberhasilan teknik resampling dalam menciptakan distribusi yang lebih seimbang di antara kelas emosi. Dengan demikian, PCA pada tahap labeling dan setelah resampling memberikan wawasan penting tentang efektivitas langkah-langkah yang diambil dalam menyeimbangkan dataset, yang pada akhirnya dapat mendukung kinerja model dalam mempelajari seluruh kelas dengan lebih adil.

#### e. Modeling

*Modeling* merupakan proses di mana algoritma *machine learning* digunakan untuk membangun model yang mampu memprediksi emosi dari teks. Algoritma yang sering digunakan untuk analisis emosi mencakup *Naive Bayes*, Support Vector Machine (SVM), *Random Forest*, atau algoritma berbasis jaringan saraf seperti LSTM. Data yang telah diproses dan dilabeli sebelumnya dimanfaatkan untuk melatih model agar dapat mengenali pola emosi dalam teks yang baru.

#### f. Evaluasi

Tahap evaluasi dilakukan untuk memastikan bahwa model yang telah dibuat berfungsi dengan baik. Evaluasi model biasanya melibatkan pengukuran dengan metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai kinerja model. Jika model tidak memberikan hasil yang diinginkan, akan dilakukan iterasi tambahan untuk meningkatkan hasil, misalnya dengan mengubah parameter atau mencoba teknik preprocessing yang baru. Tujuan dari evaluasi adalah untuk memastikan bahwa model memiliki performa yang memadai untuk digunakan di dunia nyata. Selain itu, digunakan juga teknik cross-validation untuk mendapatkan gambaran yang lebih akurat mengenai performa model. Cross-validation adalah teknik evaluasi yang membagi dataset menjadi beberapa subset atau fold untuk melatih dan menguji model secara bergantian pada subset yang berbeda. Ini membantu mengurangi risiko overfitting dan memastikan bahwa model memiliki performa yang stabil pada data yang beragam. Hasil cross-validation memberikan mean accuracy yang lebih representatif dari keseluruhan dataset dan membantu dalam menilai apakah model cukup generalis untuk digunakan di dunia nyata.

### B. Algoritma Decision Tree

*Decision Tree* adalah salah satu algoritma dalam pembelajaran mesin yang digunakan untuk mengubah data yang sangat besar menjadi pohon keputusan yang menggambarkan aturan [15]. Metode ini berfungsi dengan mengkonversi keputusan ke dalam bentuk pohon yang memiliki struktur bercabang. Setiap simpul dalam pohon melambangkan fitur atau atribut, setiap cabang menunjukkan keputusan yang diambil berdasarkan fitur tersebut, dan setiap daun adalah hasil dari prediksi atau klasifikasi. Kelebihan *Decision Tree* tidak membutuhkan pengolahan data sebelumnya seperti normalisasi atau skala, yang biasanya diperlukan dalam algoritma lain (contohnya, SVM atau KNN). Pohon keputusan dapat berfungsi dengan baik untuk kedua jenis data, dan mampu mengatasi hubungan non-linear antara variabel. Dengan menerapkan algoritma pohon keputusan, kinerja yang baik dapat dicapai untuk data kategori (categorical data) serta data numerik (continuous data). *Decision Tree* menggunakan struktur hierarki dalam bentuk diagram pohon, yang sangat mudah dipahami, bahkan oleh individu tanpa latar belakang teknis. Setiap cabang atau node mencerminkan keputusan logis yang dapat dilacak hingga mencapai hasil akhir.

Kekurangan *Decision Tree* rentan terhadap *Overfitting* dan sering kali menjadi terlalu rumit jika tidak ada pengendalian (contohnya, jika pohon dibiarkan tumbuh hingga kedalaman maksimum tanpa batas). Ini menyebabkan model sangat cocok dengan data latih, sehingga performanya buruk saat menghadapi data uji. *Decision Tree* Kurang Stabil terhadap Perubahan Data dan sangat peka terhadap perubahan kecil dalam data. Perubahan kecil dalam dataset (misalnya, menambah atau menghapus sampel) dapat menyebabkan struktur pohon berubah dengan signifikan. *Decision Tree* cenderung menghasilkan pohon yang berusaha belajar dari semua data, termasuk *noise*. Ini menyebabkan *overfitting* dan menjadikannya Tidak Efektif pada Data dengan Banyak *Noise*.

$$Gini(S) = 1 - \sum_{i=1}^n p_i^2$$

S = Dataset yang akan dihitung Gini Index-nya

Pi = Proporsi dari kelas ke-i dalam dataset.

n = Jumlah kelas dalam dataset.

### C. Random Forest

*Random Forest* merupakan algoritma *machine learning* berbasis *ensemble* yang digunakan dalam klasifikasi, regresi, dan tugas lainnya [16]. Algoritma ini bekerja dengan menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil. Penggunaan *Random Forest* pada penelitian ini ditujukan untuk mengklasifikasikan emosi dalam komentar YouTube. Berikut adalah langkah dan penjelasan lebih detail mengenai rumus dan penggunaan *Random Forest*:

#### a. Pembentukan Dataset Pelatihan dengan Bootstrap

*Random Forest* menggunakan teknik bootstrap aggregating atau bagging, di mana beberapa subset acak dari dataset pelatihan dipilih dengan pengembalian. Setiap subset kemudian digunakan untuk melatih satu pohon keputusan. Teknik ini membantu dalam mengurangi variansi model dan meningkatkan stabilitas.

#### b. Pembentukan Decision Tree

Setiap pohon keputusan dalam *Random Forest* dilatih dengan subset yang berbeda dari dataset pelatihan. Pada setiap node di dalam pohon, pemisahan dilakukan berdasarkan fitur yang dipilih secara acak. Proses ini dinyatakan dengan rumus:

$$Gini(t) = 1 - \sum_{i=1}^m (P_i)^2$$

di mana  $Gini(t)$  adalah indeks Gini pada node  $t$ ,  $m$  adalah jumlah kelas, dan  $P_i$  adalah proporsi sampel dari kelas ke-  $i$ . Indeks Gini digunakan untuk mengukur ketidakmurnian pada node, dan pemisahan terbaik adalah yang meminimalkan nilai indeks Gini.

#### c. Pemilihan Fitur Secara Acak

Pada setiap node, hanya sebagian dari semua fitur yang tersedia dipertimbangkan untuk pemisahan. Ini dinyatakan dalam parameter  $\sqrt{M}$ , di mana  $M$  adalah jumlah total fitur. Tujuannya adalah untuk menciptakan keragaman di antara pohon-pohon dalam hutan (*forest*), sehingga mengurangi risiko *overfitting*.

#### d. Voting untuk Prediksi Akhir

Setelah semua pohon dilatih, prediksi dibuat berdasarkan voting dari setiap pohon. Untuk masalah klasifikasi, setiap pohon memberikan prediksi, dan kelas dengan jumlah suara terbanyak dipilih sebagai output akhir. Rumus untuk voting adalah:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_B(x)\}$$

di mana  $h_1(x)$ , adalah prediksi dari pohon keputusan ke-  $i$ , dan  $B$  adalah jumlah pohon dalam *Random Forest*. Kelas yang paling banyak dipilih oleh seluruh pohon dalam hutan adalah hasil akhir dari model.

#### e. Keuntungan Random Forest

Penggunaan banyak pohon keputusan membantu mengatasi masalah *overfitting* yang sering terjadi pada pohon keputusan tunggal. Selain itu, metode ini memberikan estimasi pentingnya fitur, yang dapat digunakan untuk menganalisis seberapa besar pengaruh setiap fitur terhadap prediksi.

### III. HASIL DAN PEMBAHASAN

#### A. Identifikasi Masalah

Penelitian ini bertujuan untuk mengeksplorasi emosi yang terkandung dalam komentar YouTube mengenai video berjudul "Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia". Dataset yang digunakan berupa komentar-komentar yang dikumpulkan melalui API YouTube. Komentar ini kemudian dianalisis untuk mengidentifikasi sentimen dan emosi yang dominan.

#### B. Data Collecting

Data komentar diambil dari video YouTube "Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia" menggunakan YouTube Data API. Dataset ini berisi teks komentar yang kemudian disimpan dalam format CSV. Jumlah komentar yang dikumpulkan mencapai 12337 komentar, yang menjadi dasar dari analisis sentimen dalam penelitian ini.

#### C. Preprocessing Data

Dalam penelitian ini, preprocessing dilakukan untuk membersihkan dan menormalkan data teks yang diambil dari komentar YouTube. Tujuannya adalah untuk memastikan bahwa data lebih terstruktur, relevan, dan konsisten, sehingga mendukung analisis sentimen dan emosi secara efektif. Berikut proses *preprocessing* data:

##### a. Cleansing

Langkah-langkah cleansing dalam penelitian ini mencakup menghapus elemen yang mengandung HTML seperti tag `

` atau `

`, serta simbol seperti emoticon dan hashtag. Selain itu, karakter non-ASCII yang tidak kompatibel juga dihapus untuk memastikan keutuhan data. Kata-kata bermakna rendah (stopwords) yang tidak signifikan untuk analisis, seperti "yang" atau "dan", turut dihilangkan. Singkatan atau kata tidak baku seperti "gak" diubah menjadi "tidak", dan "tp" menjadi "tetapi" menggunakan kamus khusus. Kata yang berulang seperti "bagus bagus bagus" disederhanakan menjadi satu kata "bagus" untuk mencegah bias. Seluruh teks diubah menjadi huruf kecil (lowercase) agar konsisten, sementara angka dan spasi berlebih juga dihapus untuk menjaga struktur teks tetap rapi. Proses cleansing ini bertujuan untuk memastikan data teks bersih dan siap untuk langkah analisis berikutnya. Hasil dari pemrosesan cleansing disajikan pada Tabel I.

Table I. *Cleansing*

Data Input	Data Output
'Itu cuma 0,0...% kali dan anaknya ga pernah belajar dgn sungguh2'	'itu cuma anaknya tidak pernah belajar dengan sungguh'

##### b. Tokenisasi

Tokenisasi adalah proses memecah teks menjadi unit-unit terkecil yang disebut token. Dalam konteks analisis teks, token biasanya berupa kata-kata, meskipun dapat mencakup karakter atau frasa tergantung pada kebutuhan. Tokenisasi merupakan langkah penting dalam preprocessing data karena membantu mempersiapkan teks untuk analisis lebih lanjut, seperti penghapusan stopwords, stemming, atau ekstraksi fitur. Berikut merupakan hasil dari proses tokenisasi disajikan pada Tabel II.

Tabel II. Tokenisasi

Data Input	Data Output
'itu cuma anaknya tidak pernah belajar dengan sungguh'	[itu, cuma, anaknya, tidak, pernah, belajar, dengan, sungguh]

### c. Stemming

Dalam penelitian ini, stemming diterapkan untuk teks berbahasa Indonesia menggunakan library Sastrawi, yang dirancang khusus untuk menangani kompleksitas morfologi Bahasa Indonesia. Hasil dari stemming disajikan pada Tabel III.

Tabel III. *Stemming*

Data Input	Data Output
'itu cuma anaknya tidak pernah belajar dengan sungguh'	'itu cuma anak tidak pernah ajar dengan sungguh'

### d. Stopwords

Dalam penelitian ini, penghapusan *stopwords* dilakukan sebagai bagian dari proses *preprocessing* teks. Tujuannya adalah untuk mengurangi kompleksitas data dan memastikan bahwa analisis hanya mempertimbangkan kata-kata yang memberikan kontribusi signifikan terhadap pola atau tema dalam teks. Namun, penelitian ini juga mengakomodasi kata-kata tertentu yang ingin dikecualikan dari penghapusan *stopwords*. Kata-kata seperti "pns", "sd", "smp", "sma", "smk", "tk", dan "kkm" dimasukkan ke dalam daftar `exclude_words` karena dianggap relevan dengan konteks analisis. Selain itu, *stopwords* tambahan ditambahkan ke dalam daftar menggunakan variabel `additional_stopwords`, yang berisi kata-kata informal dan slang seperti "sdh", "trus", "kek", "nya", "deh", dan lainnya yang sering muncul dalam komentar media sosial tetapi tidak memiliki nilai signifikan dalam analisis.

Proses ini menggabungkan *stopwords* bawaan Bahasa Indonesia dari library NLTK (`stopwords.words('indonesian')`) dengan daftar *stopwords* tambahan. Gabungan ini dilakukan menggunakan fungsi `union`, sehingga menghasilkan daftar lengkap dalam variabel `stopwords_combined`. Pendekatan ini memberikan fleksibilitas lebih dalam *preprocessing* teks, memungkinkan penghapusan kata-kata umum sambil mempertahankan kata-kata tertentu yang relevan dengan penelitian.

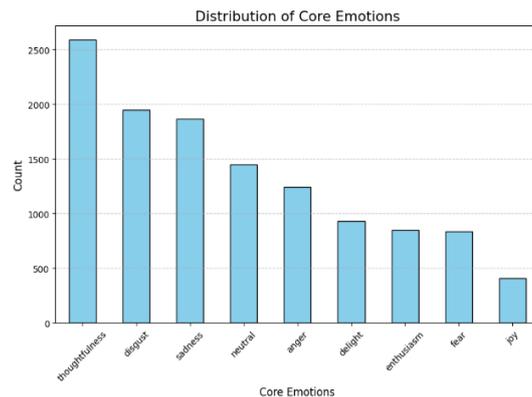
Dengan langkah ini, proses penghapusan *stopwords* dapat disesuaikan untuk meningkatkan kualitas analisis, memastikan bahwa hanya kata-kata yang benar-benar tidak relevan yang dihapus, sementara kata-kata penting tetap dipertahankan untuk mendukung analisis tema dan pola dalam data teks.

Tabel VI. *Stopwords*

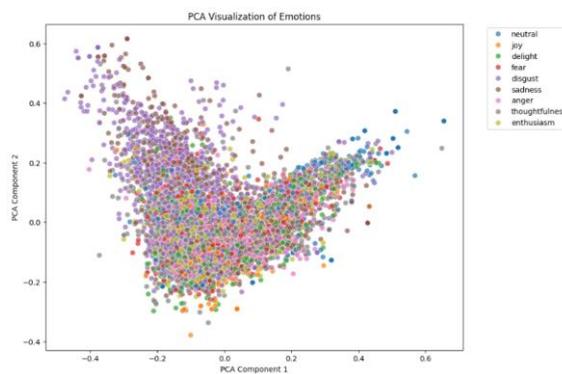
Data Input	Data Output
'itu cuma anaknya tidak pernah belajar dengan sungguh'	'anak ajar sungguh'

### D. Labeling Emosi

Labeling emosi dilakukan berdasarkan empat kategori utama dari kamus *lexicon* SenticNet, yaitu *INTROSPECTION*, *TEMPER*, *ATTITUDE*, dan *SENSITIVITY*. Nilai maksimal dari keempat kategori digunakan untuk memetakan emosi positif seperti *joy*, *delight*, *enthusiasm*, dan *thoughtfulness*, sedangkan nilai minimal digunakan untuk menentukan emosi negatif seperti *fear*, *anger*, *disgust*, dan *sadness*. Pendekatan ini memungkinkan prediksi emosi dengan mengacu pada nilai tertinggi dan terendah dari empat dimensi tersebut. Distribusi lengkap dari emosi utama ini ditampilkan pada Gambar 2.



Gambar 2. Distribusi Emosi Utama dalam Komentar YouTube



Gambar 3. Visualisasi PCA 2 dimensi

Pada visualisasi yang disajikan pada Gambar 3 hasil Principal Component Analysis (PCA) yang mereduksi dimensi data berbasis TF-IDF menjadi dua komponen utama, yaitu PCA Component 1 dan PCA Component 2. Titik-titik dalam gambar mewakili sampel komentar yang telah diproses, dengan setiap warna menunjukkan kategori emosi, seperti *neutral*, *joy*, *delight*, *fear*, *disgust*, *sadness*, *anger*, *thoughtfulness*, dan *enthusiasm*. Visualisasi ini memberikan gambaran tentang distribusi data dalam ruang dua dimensi, sehingga memudahkan analisis pola dan hubungan antar kategori emosi.

Dari hasil visualisasi pada Gambar 3, terlihat bahwa beberapa kategori seperti *neutral* dan *thoughtfulness* memiliki pola pengelompokan yang jelas, sedangkan kategori seperti *sadness*, *disgust*, dan *fear* menunjukkan tumpang tindih yang signifikan. Hal ini mengindikasikan bahwa beberapa kategori emosi memiliki kemiripan fitur yang mungkin menyulitkan model untuk membedakannya. Visualisasi ini memberikan wawasan awal yang penting untuk mengidentifikasi potensi tantangan dalam proses klasifikasi, seperti peningkatan fitur atau pengelolaan data kategori tertentu, guna memastikan hasil analisis yang lebih akurat.

#### E. Analisis Emosi

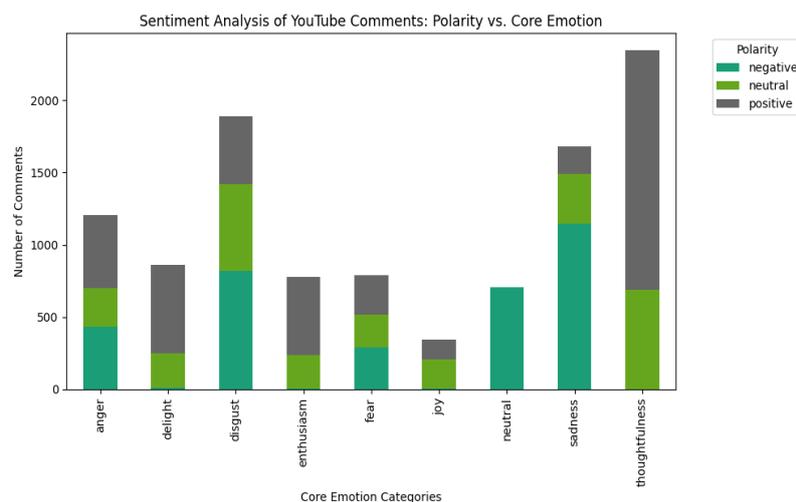
Dalam penelitian ini, metode lexicon SenticNet digunakan untuk memetakan emosi dari teks komentar YouTube berdasarkan empat dimensi utama, yaitu INTROSPECTION, TEMPER, ATTITUDE, dan SENSITIVITY. Proses pemetaan dimulai dengan memeriksa validitas teks input, memastikan bahwa teks yang diberikan adalah string dan tidak kosong. Jika teks valid, setiap kata dalam teks kemudian dipecah (tokenisasi) dan dicocokkan dengan kamus SenticNet untuk mendapatkan nilai skor emosi pada keempat dimensi tersebut. Kamus SenticNet menyediakan nilai yang mencerminkan bagaimana kata tersebut terkait dengan yaitu INTROSPECTION, TEMPER, ATTITUDE, dan SENSITIVITY. Jika kata ditemukan dalam kamus, skor untuk masing-masing dimensi akan ditambahkan, dan total rata-rata dihitung dengan membagi nilai total setiap dimensi dengan jumlah kata yang valid. Hal ini bertujuan untuk memastikan bahwa setiap kata memberikan kontribusi yang seimbang terhadap nilai dimensi secara keseluruhan.

Setelah semua kata dianalisis dan nilai dimensi dihitung, langkah berikutnya adalah menentukan prediksi emosi berdasarkan nilai dari keempat dimensi ini. Prediksi emosi bisa berupa emosi negatif, emosi

positif, atau netral. Untuk emosi negatif, pemetaan dilakukan dengan melihat nilai minimum dari keempat dimensi. Jika nilai minimum adalah negatif, maka emosi yang dipetakan disesuaikan dengan dimensi tersebut, seperti *anger* untuk TEMPER atau *fear* untuk SENSITIVITY. Sebaliknya, untuk emosi positif, penentuan dilakukan dengan mencari nilai tertinggi dari keempat dimensi. Jika nilai tertinggi menunjukkan emosi positif, maka emosi tersebut akan dipetakan menjadi, misalnya, *delight* untuk ATTITUDE atau *enthusiasm* untuk SENSITIVITY. Jika seluruh nilai dimensi adalah nol, emosi dianggap netral, yang berarti komentar tersebut tidak menunjukkan emosi yang signifikan.

Tabel VII. Proses Labeling

Reprocessing Text	Intro- spection	Tem- per	Attitude	Sensiti- vity	Prediction Emotion
hallo warga sipil selamat tonton kembali sistem naik kelas kelulus biar naikin biarpun nilai anjlok lulusin karna bikin anak malas ajar	0	0	0	0	neutral
mantan mendikbud kira pintar goblok kurikulum zonasi terap riset kemarin muka asli ngerusak didik negara gapapa opini bilang sekolah tentu sukses kaya lawak baca hitung matematika dasar	0.2876	0.249	-0.166	0.09	disgust
nilai uji pakai perankingan sd sistem ajar kebut malam ganti kurikulum rapor salah kurikulum tingkat satu didik nama sentak alim hari biar merah liat uji nasional adakan semangat ajar hapus sistem zonasi anak pintar desa pelosok masuk smp negeri kota	0.29783	-0.103	0.4216	0.12983	anger
jauh ceo milyader triliuner kursi perintah menteri muhajir cetus zonasi evaluasi biar jalan	-0.3504	0	-0.01239	0.363	sadness
buat inovasi bareng ikut siswa inovasi motivasi lingkung sekolah ramah anak keras leceh seksual	0.23349	0	0.08325	-0.0696	fear
	0.1726	0.2033	0.2604	0.1998	delight
	0	0	0.23	0.4525	enthusias m
	0	0.1516	0.0823	0.082	joy
	0.3528	0.1124	0.3092	0.1586	thoughtfu lness



Gambar 4. Sentimen Analisis antara Polarity dan Emosi

Pada Gambar 4 tersebut menggambarkan analisis hubungan antara kategori emosi inti dan polaritas sentimen. Grafik menunjukkan distribusi komentar berdasarkan kategori emosi, yang dipecah menjadi sentimen positif, netral, dan negatif. Kategori *thoughtfulness* memiliki jumlah komentar tertinggi, dengan mayoritas polaritasnya positif, sedangkan kategori seperti *sadness* dan *disgust* didominasi oleh sentimen negatif. Sebaliknya, emosi seperti *delight* dan *joy* sebagian besar memiliki polaritas positif. Gambar ini

mengilustrasikan bagaimana emosi inti mencerminkan pola polaritas sentimen, membantu mengidentifikasi hubungan antara ekspresi emosional dalam teks dan persepsi keseluruhan dari komentar tersebut. Analisis ini memberikan pemahaman yang lebih mendalam tentang distribusi sentimen berdasarkan emosi dalam data.

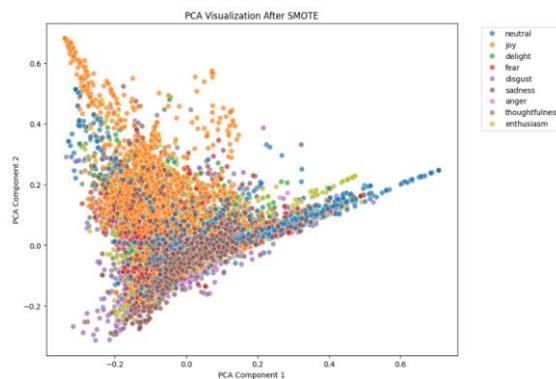
#### F. Resampling Data

Dalam penelitian ini, resampling data dilakukan untuk mengatasi ketidakseimbangan jumlah data antar kategori emosi, yang dapat mempengaruhi akurasi model pembelajaran mesin. Ketidakseimbangan ini menyebabkan model cenderung memberikan bobot lebih pada kategori dengan jumlah data yang dominan, sehingga hasil analisis menjadi kurang representatif.

Untuk menangani masalah ini, dua metode resampling diterapkan, yaitu *SMOTE* (Synthetic Minority Oversampling Technique) dan *Random Over Sampling*. *SMOTE* bekerja dengan mensintesis data baru untuk kategori minoritas dengan menciptakan sampel-sampel buatan berdasarkan interpolasi antara data yang ada, sedangkan *Random Over Sampling* menambahkan salinan data yang ada untuk kategori minoritas secara acak. Kedua metode ini diterapkan untuk membandingkan performa model dan menentukan pendekatan yang lebih optimal terhadap data yang digunakan dalam penelitian ini.

Hasil perbandingan menunjukkan bahwa metode *Random Oversampling* memberikan hasil yang lebih optimal dibandingkan dengan *SMOTE*. *Random Over Sampling* menghasilkan distribusi data yang lebih sederhana namun tetap efektif untuk meningkatkan performa model. Sementara itu, meskipun *SMOTE* memberikan distribusi data yang lebih bervariasi, metode ini tidak memberikan peningkatan performa yang signifikan pada dataset yang digunakan.

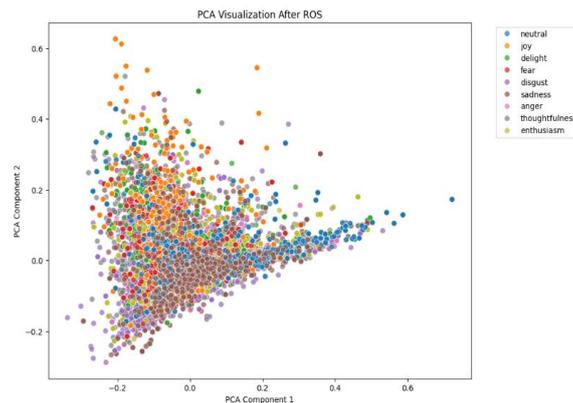
Sebagai hasil dari proses resampling dengan *Random Over Sampling*, jumlah data pada masing-masing kategori emosi menjadi lebih proporsional, yang meningkatkan kualitas pelatihan model. Dengan demikian, analisis yang dihasilkan menjadi lebih akurat dan representatif terhadap distribusi emosi dalam komentar. Pendekatan ini memastikan bahwa model dapat belajar secara lebih adil dari semua kategori.



Gambar 5. Visualisasi PCA setelah *SMOTE EN*

Pada visualisasi hasil *Principal Component Analysis (PCA)* yang disajikan pada Gambar 5 setelah diterapkan metode resampling *SMOTE (Synthetic Minority Oversampling Technique)*. *PCA* digunakan untuk mereduksi dimensi data menjadi dua komponen utama, yaitu *PCA Component 1* dan *PCA Component 2*, yang bertujuan untuk memvisualisasikan distribusi data dengan lebih sederhana. Setelah *SMOTE* diterapkan, data kategori minoritas yang sebelumnya memiliki jumlah terbatas diperbanyak secara sintesis, sehingga distribusi kategori emosi menjadi lebih seimbang.

Dalam visualisasi ini, setiap titik mewakili satu sampel data komentar, dengan warna yang menunjukkan kategori emosi, seperti *neutral*, *joy*, *delight*, *fear*, *disgust*, *sadness*, *anger*, *thoughtfulness*, dan *enthusiasm*. Setelah proses resampling, distribusi data menunjukkan peningkatan proporsi untuk kategori minoritas seperti *fear* dan *enthusiasm*, yang sebelumnya tumpang tindih atau jarang terlihat. Visualisasi ini juga menunjukkan bahwa data memiliki pola distribusi yang lebih merata, meskipun beberapa kategori seperti *disgust* dan *sadness* masih memiliki sedikit tumpang tindih. Hasil ini memberikan bukti bahwa *SMOTE* berhasil meningkatkan jumlah data kategori minoritas, membantu memastikan bahwa model pembelajaran mesin dapat belajar secara lebih adil dari semua kategori.



Gambar 6. Visualisasi PCA setelah ROS

Pada Gambar 6 disajikan *Random Over Sampling* diterapkan, data kategori minoritas yang sebelumnya memiliki jumlah terbatas diperbanyak secara sintesis, sehingga distribusi kategori emosi menjadi lebih seimbang. Melalui visualisasi ini, kita dapat mengamati pola penyebaran emosi dari komentar secara keseluruhan. Sebagai contoh, emosi seperti *neutral* atau *joy* mungkin terlihat lebih dominan dengan penyebaran yang rapat di area tertentu, sedangkan emosi seperti *anger* atau *fear* lebih jarang muncul dan tersebar di wilayah yang lebih luas. Pola distribusi ini memberikan wawasan tentang sentimen umum dalam komentar.

#### G. Modeling

Proses pemodelan dilakukan dengan memanfaatkan *decision tree* dan *random forest*, namun penelitian ini lebih mengutamakan penggunaan *random forest* karena terbukti lebih andal dalam mencapai akurasi yang lebih tinggi. Dalam implementasinya, model *random forest* dirancang dengan parameter pembatas seperti minimal sample *split* sebesar 5, minimal sample *leaf* sebesar 2, *random state* dengan nilai 42, sementara kedalaman pohon (*tree depth*) tidak dibatasi untuk memberikan fleksibilitas maksimal.

Hasil pemodelan menunjukkan bahwa *decision tree* mencapai kedalaman 382 untuk menghasilkan akurasi sebesar 0.76, sedangkan *random forest* memiliki kedalaman rata-rata pohon sebesar 305.62 dan mampu mencapai akurasi sebesar 0.80. Pembatasan parameter pada pemodelan ini dirancang untuk mengurangi risiko *overfitting*, yang sering terjadi pada pohon keputusan tunggal, sekaligus meningkatkan stabilitas prediksi pada dataset yang kompleks.

Tabel VIII. *Parameter Modelling*

Parameter	Decision Tree	Random Forest
random_state	42	42
min_samples_split	5	5
min_samples_leaf	2	2
max_features	none	none
avg max_depth	382	305.62

#### H. Evaluasi

Setelah melakukan modeling menggunakan *random forest* dengan metode resampling *Random Over Sampling* (ROS), evaluasi model dilakukan menggunakan metrik seperti *confusion matrix*, *precision*, *recall*, dan *F1-score*. Hasil evaluasi menunjukkan bahwa performa model bervariasi di setiap kategori emosi. Secara rata-rata, kategori emosi mayoritas seperti *thoughtfulness* dan *joy* memiliki nilai *precision* dan *recall* yang tinggi, menunjukkan bahwa model mampu memprediksi kategori tersebut dengan baik. Sebaliknya, kategori minoritas seperti *fear* dan *disgust* memiliki *precision* dan *recall* yang rendah,

mengindikasikan bahwa model kesulitan dalam memprediksi emosi dengan jumlah data yang terbatas meskipun telah dilakukan *resampling*.

Berdasarkan tabel evaluasi, rata-rata akurasi model mencapai 0.80, yang dianggap cukup baik untuk dataset ini, mengingat kompleksitas dan variasi emosi dalam data. Namun, performa yang kurang optimal pada kategori dengan jumlah sampel kecil menunjukkan bahwa ketidakseimbangan data masih menjadi tantangan, meskipun sudah diminimalkan dengan ROS. Kesimpulan dari evaluasi ini adalah bahwa model secara keseluruhan cukup efektif untuk kategori mayoritas, tetapi memerlukan pendekatan tambahan, seperti kombinasi dengan metode lain atau peningkatan fitur, untuk meningkatkan akurasi pada kategori minoritas. Dengan akurasi rata-rata yang cukup tinggi, model ini dapat dianggap andal, tetapi masih memiliki ruang untuk perbaikan terutama pada kategori emosi yang jarang muncul.

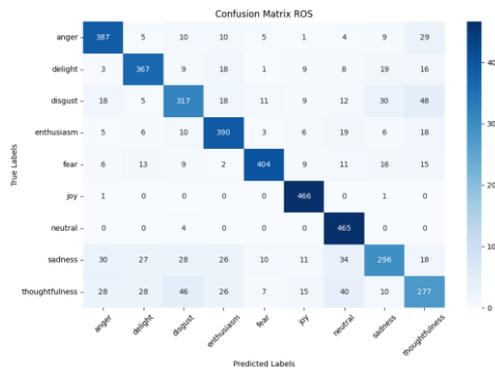
Tabel IX. Evaluasi Decision Tree

Kategori	Precision	Recal	F1 score	support
anger	0.75	0.80	0.78	460
delight	0.75	0.86	0.80	450
enthusiasm	0.74	0.82	0.78	463
fear	0.76	0.80	0.78	485
disgust	0.62	0.60	0.61	468
joy	0.90	1.00	0.95	468
sadness	0.72	0.61	0.66	480
thoughtfulness	0.61	0.40	0.48	477
nautral	0.88	0.94	0.91	469
accuracy			0.76	4220
macro avg	0.75	0.76	0.75	4220
weighted avg	0.75	0.76	0.75	4220

Tabel X. Evaluasi Random Forest

Kategori	Precision	Recal	F1 score	support
anger	0.81	0.84	0.83	460
delight	0.81	0.82	0.81	450
enthusiasm	0.80	0.84	0.82	463
fear	0.92	0.83	0.87	485
disgust	0.73	0.68	0.81	450
joy	0.89	1.00	0.94	468
sadness	0.76	0.62	0.68	480
thoughtfulness	0.66	0.58	0.62	477
nautral	0.78	0.99	0.88	469
accuracy			0.80	4220
macro avg	0.80	0.80	0.79	4220
weighted avg	0.80	0.80	0.79	4220

Untuk memastikan bahwa model tidak mengalami overfitting dan performanya stabil di berbagai subset data, dilakukan cross-validation. Hasil dari Cross-Validation Accuracy Scores adalah sebagai berikut: [0.7841, 0.7860, 0.8118, 0.8251, 0.8488]. Dengan nilai Mean CV Accuracy sebesar 0.8111, hal ini menunjukkan bahwa akurasi model relatif konsisten pada berbagai iterasi validasi silang, yang memperkuat keandalan model meskipun terdapat variasi dalam distribusi data. Evaluasi menggunakan cross-validation ini memberikan keyakinan lebih pada kinerja model secara keseluruhan, terutama dalam menghadapi variasi data yang berbeda.



Gambar 7. Heatmap dan Confusion Matrix

Gambar 7 menunjukkan confusion matrix hasil evaluasi model *Random Forest* setelah dilakukan resampling menggunakan Random Over Sampling (ROS). Matriks ini menggambarkan performa model dalam memprediksi kategori emosi berdasarkan data uji, dengan nilai diagonal menunjukkan prediksi yang benar untuk setiap kategori emosi. Model menunjukkan performa terbaik pada kategori *joy* dan *neutral*, dengan masing-masing mencapai 466 dan 465 prediksi yang benar. Kategori *fear* dan *enthusiasm* juga memiliki tingkat akurasi yang cukup baik, meskipun terdapat beberapa kesalahan prediksi yang tersebar di kategori lain.

Sebaliknya, kategori seperti *sadness* dan *disgust* menunjukkan tingkat kesalahan prediksi yang lebih tinggi, dimana *sadness* sering salah diklasifikasikan sebagai *thoughtfulness* atau *anger*. Hal ini mengindikasikan bahwa model masih kesulitan dalam membedakan kategori dengan pola fitur yang saling tumpang tindih. Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa model cukup andal dalam memprediksi kategori mayoritas, tetapi memerlukan peningkatan untuk kategori minoritas melalui optimalisasi lebih lanjut, seperti penyesuaian parameter atau pengayaan fitur. *Confusion matrix* ini memberikan wawasan penting tentang kekuatan dan kelemahan model untuk perbaikan di masa depan.

Tabel XI. Top 5 contoh kesalahan

Text	True Label	Predicted
'adik angkat sistem zonasi bodoh'	<i>disgust</i>	<i>thoughtfulness</i>
'uji nasional smua negatif positif tentu kualitas didik ukur standar sekolah daerah siswa bantu mnjawab berbagai trik menjadi rahasia ajar ketika membantu siswa pindah sampai diskriminasi sedia bantu mengangap nilai nilai poin bagaimana terapan jujur sehingga tekad ajar curang'	<i>thoughtfulness</i>	<i>neutral</i>
'alhamdulillah jogja uji nasional zonasi anak nilai bagus masuk sekolah favorit'	<i>thoughtfulness</i>	<i>enthusiasm</i>
'udah bener nadiem merdeka ajar maksud anak didik bidang perintah bikin materi upload tanding hadiah tiap minggu adalah lomba matematik lain menang ambil jadiin negara biar masuk riset teliti naah paket didik minat sekolah dasar tengah bawah bertelele contoh kasih matematika lulus kuliah teknik ambil emang bakat umur doyan itb gitu matpel bidang suka kelar situ yaudah ajarin budipekerti moralitas tolongmenolong sopan bersih gaul masyarakat menteri elon lahir indonesia bibit demen paketin sd smp ajar fisik vokasi juru dalam kontak siapin materi prabowo semangat'	<i>thoughtfulness</i>	<i>disgust</i>
'hadeh erti anak sekolah jmn cemas banget generasi depan'	<i>disgust</i>	<i>joy</i>

#### IV. KESIMPULAN

Penelitian ini menunjukkan bahwa teknik *Random Forest* efektif dalam menilai emosi publik melalui komentar YouTube terkait “kegagalan adopsi sistem pendidikan Finlandia di Indonesia”. Dengan akurasi mencapai **80%**, temuan menunjukkan adanya emosi positif yang dominan, seperti *thoughtfulness* dan *enthusiasm*, sementara emosi negatif seperti *fear* dan *disgust* terlihat lebih jarang. Meskipun metode *Random Forest*, yang menggunakan Random Over Sampling (ROS) untuk mengatasi ketidakseimbangan data, memberikan tingkat akurasi yang lebih baik, masih ada kesulitan dalam memprediksi emosi minoritas. Penelitian ini menekankan pentingnya memilih metode yang tepat untuk menganalisis data yang tidak terstruktur dan membuka kemungkinan untuk penelitian lebih lanjut dengan menggunakan algoritma *deep*

*learning*, seperti LSTM atau BERT, untuk memahami konteks yang lebih rumit dan meningkatkan ketepatan analisis.

#### UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan karya ini dengan baik. Terima Kasih kepada Dosen pembimbing yang telah memberikan bimbingannya kepada penulis, sehingga penulis dapat menyelesaikan penulisan ini sesuai dengan arahan dan bimbingannya. Terima Kasih kepada rekan-rekan yang telah ikut bekerjasama dalam penyusunan penulisan ini sehingga karya ini dapat diselesaikan dengan baik. Penulis berharap karya ini dapat memberikan manfaat dan kontribusi bagi ilmu pengetahuan. Diharapkan karya ini dapat menjadi referensi untuk penelitian selanjutnya menggunakan metode yang lebih *advance* seperti menggunakan *deep learning* untuk menganalisis data sehingga mendapatkan hasil yang lebih akurasi.

#### DAFTAR PUSTAKA

- [1] L. N. Cahyani, “Sistem Pendidikan Finlandia: Membangun Kemandirian Dan Semangat Belajar Siswa,” *Journal Of Contemporary Issues In Primary Education (Jcipe)*, Vol. 1, No. 2, Pp. 55–61, 2023.
- [2] S. Fadia And N. Fitri, “Problematika Kualitas Pendidikan Di Indonesia”.
- [3] A. Utomo, M. A. Andadinata, And D. Widhiandono, “Analisis Kualitas Konten Youtube Berdasarkan Kolom Komentar Dari Channel Youtube Baim Paula,” 2023.
- [4] H. Al Rasyid Harpizon *Et Al.*, “Analisis Sentimen Komentar Di Youtube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes,” *Jurnal Nasional Komputasi Dan Teknologi Informasi*, Vol. 5, No. 1, 2022.
- [5] F. Sakinah Alnaz And W. Maharani, “Analisis Emosi Melalui Media Sosial Twitter Dengan Menggunakan Metode Naïve Bayes Dan Perbandingan Fitur N-Gram Dan Tf-Idf.” *Skripsi*, 2021
- [6] D. P. Sinambela, H. Naparin, M. Zulfadhilah, And N. Hidayah, “Implementasi Algoritma Decision Tree Dan Random Forest Dalam Prediksi Perdarahan Pascasalin,” *Jurnal Informasi Dan Teknologi*, Vol. 5, No. 3, Pp. 58–64, Sep. 2023.
- [7] N. Agustyaningrum And N. Himmi, “Best Practices Sistem Pendidikan Di Finlandia Sebagai Refleksi Sistem Pendidikan Di Indonesia,” *Edukatif: Jurnal Ilmu Pendidikan*, Vol. 4, No. 2, Pp. 2100–2109, Feb. 2022..
- [8] A. Karimah And G. Dwilestari, “Analisis Sentimen Komentar Video Mobil Listrik Di Platform Youtube Dengan Metode Naive Bayes,” 2024.
- [9] G. Sanjaya And K. Muslim Lhaksana, “Analisis Sentimen Komentar Youtube Tentang Terpilihnya Menteri Kabinet Indonesia Maju Menggunakan Lexicon Based”. *Jurnal Teknologi Pintar*, Vol. 2, No. 2, 2022
- [10] I. Afdhal *Et Al.*, “Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di Youtube Tentang Islamofobia,” *Skripsi*, 2020.
- [11] A. H. Nasrullah, “Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris,” Vol. 7, No. 2, 2021.
- [12] L. Andraini And T. Komputer, “Analisis Part Of Tagging Bahasa Indonesia Tentang Swamedikasi Pada Dialog Interactive Qestion Dengan Brill Tagger.” *Jurnal Teknologi Pintar*, Vol. 2, No. 2, 2022.
- [13] F. Al-Hawari And H. Barham, “A Machine Learning Based Help Desk System For It Service Management,” *Journal Of King Saud University - Computer And Information Sciences*, Vol. 33, No. 6, Pp. 702–718, Jul. 2021, Doi: 10.1016/J.Jksuci.2019.04.001.

- [14] F. Amrullah and A. Solichin, "Analisis Emosi Pada Live Chat Youtube 'Mata Najwa: 3 Bacapres Bicara Gagasan' Menggunakan Pendekatan Lexicon Dan Algoritma Naive Bayes," *Jurnal Ticom: Technology Of Information And Communication*, Vol. 12, No. 3, P. 2024.
- [15] Y. P. Tanjung, S. Sentinuwo, And A. Jacobus, "Penentuan Daya Listrik Rumah Tangga Menggunakan Metode Decision Tree."
- [16] Sugianto, C. (2019). Penerapan Teknik Data Mining Untuk Menentukan Hasil Seleksi Masuk Sman 1 Gibeber Untuk Siswa Baru Menggunakan Decision Tree. *Jurnal TEDC*, 9(1), 39-43."