Jurnal Ilmu Komputer dan Sistem Informasi (JIRSI)
Volume: 5, Nomor: 2, Mei 2026: 662-677
https://jurnal.unity-academy.sch.id/index.php/jirsi/index

e-ISSN 2830-3954
p-ISSN 2830-6031

Multilabel Aspect-Based Emotion Analysis Pada Ulasan Aplikasi IKD: Pengaruh Focal Loss dan Threshold Tuning Menggunakan Indobert

Multilabel Aspect-Based Emotion Analysis of IKD Application Reviews: The Impact of Focal Loss and Threshold Tuning Using Indobert

Viviana Purba*[1], Eka Dyar Wahyuni[2], Tri Luhur Indayanti Sugata[3]
[1,2,3]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur
E-mail: [1]22082010152@student.upnjatim.ac.id, [2]ekawahyuni.si@upnjatim.ac.id, [3]tri.luhur.fasilkom@upnjatim.ac.id

Abstrak

Ulasan pengguna aplikasi Identitas Kependudukan Digital (IKD) mengandung berbagai emosi terhadap berbagai aspek layanan. Ulasan ini tidak hanya mencerminkan tingkat kepuasan terhadap layanan, tetapi juga mencakup pengalaman, keluhan, harapan, dan persepsi masyarakat terhadap kualitas sistem yang digunakan. Penelitian ini bertujuan mengembangkan model Aspect-Based Emotion Analysis (ABEA) berbasis multilabel menggunakan end-to-end IndoBERT untuk mengidentifikasi emosi pengguna pada setiap aspek layanan aplikasi IKD, sekaligus menganalisis pengaruh penerapan Focal Loss dan threshold tuning terhadap performa klasifikasi pada kondisi distribusi label yang sangat tidak seimbang. Data dikumpulkan dari 13.197 ulasan pengguna di Google Play Store dalam rentang Juni 2024 hingga November 2025 menggunakan metode web scraping, kemudian dibersihkan dan difilter hingga menghasilkan 6.891 data. Aspek layanan diidentifikasi secara empiris menggunakan BERTopic. Pelabelan dilakukan oleh tiga anotator manusia dan dua anotator AI dengan label final ditentukan melalui majority voting. Model dikembangkan melalui 6 skenario eksperimen yang memvariasikan preprocessing, Focal Loss, threshold tuning dan rasio split data. Evaluasi menggunakan metrik F1 Score Macro, F1 Score Micro, Precision, Recall, dan Hamming Loss. BERTopic menghasilkan Coherence Score 0,6196 dan Topic Diversity 0,92 dengan 5 aspek representatif. Model paling optimal diperoleh dengan konfigurasi Focal Loss, threshold 0,4, dan split 60:20:20 mencapai F1 Score Macro 0,3916, meningkat 24,1% dari baseline, dengan F1 Score Micro 0,9134 dan Recall 0,9423. Model terpilih berhasil diintegrasikan ke sistem berbasis web menggunakan framework Flask untuk memvisualisasikan hasil klasifikasi. Emosi marah mendominasi ulasan pada aspek Login & Akses Akun dan Scan Barcode ke Dukcapil, sementara aspek Dokumen & Layanan Digital mencatat emosi gembira tertinggi. Kombinasi Focal Loss dan threshold tuning terbukti efektif menangani ketidakseimbangan distribusi label pada klasifikasi multilabel ABEA berbahasa Indonesia.

***Kata kunci**: Aspect-Based Emotion Analysis, IndoBERT, Focal Loss, Threshold Tuning, Klasifikasi Multilabel*

Abstract

User reviews of the Identitas Kependudukan Digital (IKD) application contain various emotions toward different service aspects. These reviews not only reflect the level of service satisfaction but also encompass user experiences, complaints, expectations, and public perceptions regarding the quality of the system. This study aims to develop a multi-label Aspect-Based Emotion Analysis (ABEA) model using an end-to-end IndoBERT architecture to identify user emotions across each service aspect of the IKD application. Additionally, it

analyzes the impact of implementing Focal Loss and threshold tuning on classification performance under highly imbalanced label distributions. Data were collected from 13,197 user reviews on the Google Play Store spanning from June 2024 to November 2025 using web scraping methods, which were subsequently cleaned and filtered to yield 6,891 data entries. Service aspects were empirically identified using BERTopics. Labeling was conducted by three human annotators and two AI annotators, with the final labels determined through majority voting. The model was developed across 6 experimental scenarios varying in preprocessing, Focal Loss, threshold tuning, and data split ratios. Evaluation was performed using F1 Score Macro, F1 Score Micro, Precision, Recall, and Hamming Loss metrics. BERTopic achieved a Coherence Score of 0.6196 and a Topic Diversity of 0.92 with 5 representative aspects. The most optimal model was obtained using a configuration of Focal Loss, a threshold of 0.4, and a 60:20:20 split ratio, achieving an F1 Score Macro of 0.3916, a 24.1% increase from the baseline, alongside an F1 Score Micro of 0.9134 and a Recall of 0.9423. The selected model was successfully integrated into a web-based system using the Flask framework to visualize the classification results. Anger dominated the reviews concerning the Login & Akses Akun and Scan Barcode ke Dukcapil aspects, whereas the Dokumen & Layanan Digital aspect recorded the highest joy emotion. The combination of Focal Loss and threshold tuning proved effective in handling imbalanced label distributions in Indonesian multi-label ABEA classification.

***Keywords**: Aspect-Based Emotion Analysis, IndoBERT, Focal Loss, Threshold Tuning, Multilabel Classification*

1. PENDAHULUAN

Perkembangan pesat teknologi informasi dan komunikasi saat ini telah mengubah berbagai aspek kehidupan masyarakat, termasuk pelayanan administrasi publik. Teknologi digital yang semakin canggih telah menciptakan peluang baru dalam pelayanan administrasi publik dan mendorong organisasi sektor publik untuk beralih dari layanan manual ke digital [1]. Salah satu bentuk digitalisasi layanan administrasi publik pemerintah Indonesia adalah peluncuran Aplikasi Identitas Kependudukan Digital (IKD) oleh Ditjen Dukcapil, Kementerian Dalam Negeri pada tahun 2022 [2]. Berdasarkan pasal 1 ayat 18 Permendagri Nomor 72 Tahun 2022, Identitas Kependudukan Digital adalah informasi elektronik yang digunakan untuk merepresentasikan dokumen kependudukan dan data terkait dalam aplikasi digital melalui gawai, yang menampilkan data diri pengguna sebagai identitas yang bersangkutan [3].

Sejak diluncurkan, aplikasi IKD telah diunduh lebih dari 10 juta kali di Google Play Store dengan rating 3,2 dari sekitar 70,7 ribu ulasan pengguna per Desember 2025. Tingginya jumlah ulasan tersebut menunjukkan bahwa aplikasi IKD telah digunakan secara luas oleh masyarakat dan menjadi salah satu layanan publik digital yang mendapatkan perhatian besar dari pengguna. Ulasan pengguna tidak hanya mencerminkan tingkat kepuasan terhadap layanan, tetapi juga memuat berbagai pengalaman, keluhan, harapan, dan persepsi masyarakat terhadap kualitas sistem yang digunakan [4]. Selain menjadi sumber evaluasi bagi pengembang, ulasan juga memengaruhi keputusan calon pengguna lain dalam menggunakan layanan digital pemerintah [5]. Oleh karena itu, analisis terhadap ulasan pengguna menjadi penting untuk membantu pemerintah memahami pengalaman masyarakat secara lebih komprehensif serta mengidentifikasi prioritas perbaikan layanan secara lebih terarah.

Dalam satu ulasan, pengguna dapat membahas beberapa aspek layanan sekaligus dengan emosi yang berbeda-beda. Sebagai contoh, pengguna dapat merasa senang terhadap kemudahan akses dokumen digital, tetapi pada saat yang sama merasa marah terhadap proses verifikasi akun yang gagal. Kondisi tersebut menyebabkan analisis sentimen umum menjadi kurang memadai karena tidak dapat mengidentifikasi hubungan emosi pengguna dengan aspek yang dibahas. Oleh karena itu, penelitian ini menggunakan pendekatan Aspect-Based Emotion Analysis (ABEA), yaitu pendekatan yang tidak hanya mengidentifikasi aspek yang dibahas dalam ulasan, tetapi juga mengenali emosi yang diarahkan pada setiap aspek tersebut [6].

Berbagai penelitian sebelumnya telah menunjukkan perkembangan pendekatan analisis opini pada ulasan aplikasi digital, mulai dari analisis sentimen, emotion mining, hingga klasifikasi berbasis aspek. Namun, integrasi analisis emosi berbasis aspek dengan pendekatan multi label pada ulasan layanan publik digital berbahasa Indonesia masih belum banyak dikaji secara komprehensif. Penelitian yang dilakukan oleh Hakiki et al., membahas prediksi sentimen umum dan pemodelan topik menggunakan Latent Dirichlet Allocation pada ulasan aplikasi Identitas Kependudukan Digital, tetapi analisis yang dilakukan masih terbatas pada polaritas sentimen dan belum mampu mengidentifikasi keterkaitan antara aspek layanan dengan emosi pengguna secara spesifik [2]. Selanjutnya, penelitian (Sondakh et al., 2023) telah menerapkan emotion mining pada ulasan aplikasi BRImo, namun pendekatan klasifikasi yang digunakan masih bersifat single-label dan belum mempertimbangkan aspek-aspek layanan dalam ulasan. Sementara itu, penelitian Mei et al.. menunjukkan bahwa pendekatan multilabel berbasis IndoBERT efektif diterapkan pada Aspect-Based Sentiment Analysis (ABSA), tetapi penelitian tersebut hanya berfokus pada pasangan aspek–sentimen dan belum mengeksplorasi dimensi emosi yang lebih mendalam[7]. Selain itu, domain penelitian yang digunakan masih terbatas pada ulasan produk kosmetik sehingga memiliki karakteristik berbeda dengan layanan publik digital. Berdasarkan kondisi tersebut, penelitian mengenai Aspect-Based Emotion Analysis (ABEA) berbasis multilabel pada ulasan layanan administrasi publik digital berbahasa Indonesia masih memiliki ruang eksplorasi yang luas.

Penelitian ini menerapkan pendekatan klasifikasi multilabel menggunakan model end-to-end IndoBERT untuk mengidentifikasi pasangan aspek–emosi secara simultan dalam satu ulasan pengguna. Pemilihan IndoBERT dilakukan karena model tersebut dilatih secara khusus menggunakan korpus berbahasa Indonesia sehingga lebih mampu memahami konteks bahasa informal, variasi kosakata, dan struktur kalimat yang umum ditemukan pada ulasan pengguna aplikasi digital [8]. Selain itu, pendekatan multilabel dipilih karena satu ulasan memungkinkan mengandung lebih dari satu aspek dan lebih dari satu emosi secara bersamaan, sehingga pendekatan single-label tidak cukup representatif untuk menggambarkan pengalaman pengguna secara utuh.

Setelah proses pelabelan dilakukan, distribusi pasangan aspek–emosi pada dataset menunjukkan kondisi yang tidak seimbang, di mana beberapa label memiliki jumlah data yang jauh lebih dominan dibandingkan label lainnya. Kondisi tersebut menyebabkan model dasar berbasis IndoBERT berpotensi lebih fokus mempelajari label mayoritas dan kurang optimal dalam mengenali label minoritas. Oleh karena

itu, penelitian ini menerapkan strategi penanganan data imbalanced melalui eksperimen penggunaan fungsi loss dan threshold prediksi. Salah satu pendekatan yang digunakan adalah Focal Loss, yang bekerja dengan memfokuskan proses pembelajaran pada sampel yang sulit diklasifikasikan dan label minoritas. Penelitian Kesanam dkk. menunjukkan bahwa penerapan Focal Loss pada model berbasis transformer mampu meningkatkan sensitivitas model terhadap label minoritas sehingga berpotensi meningkatkan performa klasifikasi multilabel pada penelitian ini [9].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan mengembangkan model ABEA berbasis multilabel menggunakan IndoBERT untuk mengidentifikasi emosi pengguna pada setiap aspek layanan aplikasi IKD, sekaligus menganalisis pengaruh penerapan Focal Loss dan threshold tuning terhadap performa klasifikasi pada kondisi distribusi label yang sangat tidak seimbang. Sebelum proses klasifikasi dilakukan, aspek layanan yang menjadi target analisis diidentifikasi secara empiris menggunakan BERTopic, sebuah metode pemodelan topik berbasis transformer yang mampu menangkap makna semantik dari teks ulasan pendek secara lebih mendalam dibandingkan metode konvensional seperti LDA [10]. Selain itu, model terbaik akan diimplementasikan dalam sistem berbasis web untuk memvisualisasikan hasil analisis emosi berbasis aspek secara interaktif. Dengan demikian, penelitian ini diharapkan mampu menghasilkan insight yang lebih mendalam mengenai pengalaman emosional pengguna terhadap aspek aplikasi IKD.

2. METODOLOGI PENELITIAN

2.1. Pengumpulan Data

Data penelitian diperoleh dari ulasan pengguna aplikasi Identitas Kependudukan Digital pada platform Google Play Store. Proses pengambilan data dilakukan menggunakan google-play-scraper dan parameter bahasa diatur menggunakan kode id untuk membatasi ulasan pada bahasa Indonesia. Data dikumpulkan dalam rentang waktu Juni 2024 hingga November 2025 dan menghasilkan sebanyak 13.197 ulasan. Kolom content yang bersis teks ulasan pengguna menjadi adalah sumber utama dalam proses analisis emosi berbasis aspek.

2.2. Data Cleaning

Tahap pembersihan data dilakukan untuk menghilangkan noise serta memastikan bahwa dataset siap digunakan untuk proses analisis lebih lanjut. Pada tahap ini, kolom-kolom seperti reviewId, userName, userImage, thumbsUpCount, reviewCreatedVersion, at, replyContent, repliedAt, appVersion, dan score dihapus karena tidak digunakan lebih lanjut pada penelitian ini. Selanjutnya dilakukan penghapusan ulasan duplikat menggunakan drop_duplicates() guna memastikan setiap ulasan bersifat unik dan tidak menimbulkan bias pada proses klasifikasi. Terakhir, elemen seperti mention, URL, angka, serta karakter non-alfanumerik dihapus karena tidak relevan dalam analisis teks.

2.3. Data Filtering

Tahap filtering data diterapkan untuk memastikan bahwa ulasan yang digunakan benar-benar relevan untuk analisis emosi berbasis aspek dan representatif dengan

penggunaan aplikasi IKD. Tahap ini dilakukan dengan menghapus ulasan yang memiliki kurang dari dua kata dan menghapus ulasan yang tidak relevan dengan konteks aplikasi IKD. Penyaringan ini dilakukan guna mengeliminasi ulasan yang tidak mengandung informasi bermakna, seperti "bagus", "ok", atau kata tunggal lainnya serta menyeleksi ulasan yang tidak relevan dengan kebutuhan analisis. Contoh ulasan yang tidak relevan ditunjukkan pada Tabel 1.

Tabel 1. Contoh Ulasan yang Tidak Relevan

Ulasan	Jumlah Kata	Kriteria
tidak puas	2	Ulasan yang terlalu umum
Semoga parA pekerja pemerintahan Adil slalu dalam membantu sesama	9	Ulasan yang di luar konteks aplikasi IKD
Sangat sangat buruk dan tidak sama sekali membantu	8	Ulasan yang terlalu umum
Aplikasi gak jelas buang buang uang negara gak becus yang buat kebanyakan makan uang rakyat	15	Ulasan yang terlalu umum
APLIKASI GBLOK INI LAH BRPO KALI DI CUBO DAK MASOK EMANG BENGAK NIAN YG BUAT APLIKASI INI	17	Ulasan yang menggunakan bahasa daerah

2.4. Preprocessing untuk BERTopic

Tahap preprocessing dilakukan untuk mengubah data teks menjadi bentuk yang lebih terstandar dan siap digunakan dalam proses pemodelan topik. Untuk BERTopic, preprocessing terbatas pada lowercase, normalization, dan stopword removal tanpa stemming, guna mempertahankan struktur kalimat asli yang dibutuhkan oleh model embedding berbasis transformer [10]. Tabel 2 adalah contoh ulasan sebelum dan sesudah preprocessing.

Tabel 2. Contoh Hasil Preprocessing

Ulasan	Preprocessed
Kenapa tiba tiba aplikasinya logout dan jadi aneh Gak bisa login lagi	['aplikasinya', 'logout', 'aneh', 'masuk']
aku hilangin bintang aplikasinya error buat foto aja susah tidak ada tekan tombolnya	['hilangin', 'bintang', 'aplikasinya', 'eror', 'foto', 'tekan', 'tombolnya']
servernya ampas menu pelayanan udh seminggu ga bisa dibuka bukan memudahkan malah merepotkan	['servernya', 'ampas', 'menu', 'pelayanan', 'seminggu', 'dibuka', 'memudahkan', 'merepotkan']
pendaftaran hrs dihadapan pegawai dutcapil utk meminta barkot sama saja TDK memudahkan app ini	['pendaftaran', 'hrs', 'dihadapan', 'pegawai', 'dutcapil', 'utk', 'barkot', 'memudahkan', 'app']
Selama ini tidak ada masalah login lancar update terbaru malah suruh daftar lagi hadeeehh	['masuk', 'lancar', 'update', 'terbaru', 'suruh', 'daftar', 'hadeeehh']

2.5. Identifikasi Aspek

Tahap identifikasi aspek dilakukan menggunakan metode BERTopic untuk memperoleh aspek layanan yang dominan pada ulasan aplikasi IKD. Implementasi BERTopic pada penelitian ini menggunakan model embedding SentenceTransformer all-MiniLM-L6-v2 untuk menghasilkan representasi vektor setiap ulasan. Reduksi dimensi menggunakan UMAP dengan parameter n_neighbors=15 dan n_components=10. Proses clustering dilakukan menggunakan algoritma HDBSCAN dengan parameter min_cluster_size=60. Representasi topik kemudian dibentuk menggunakan pendekatan class-based TF-IDF (c-TF-IDF) untuk memperoleh kata-kata representatif pada setiap topik [11].

2.6. Pelabelan Data

Pelabelan data dilakukan secara manual dengan melibatkan lima anotator yang terdiri atas tiga anotator manusia berlatar belakang Sistem Informasi yang memiliki pemahaman dan pengalaman terhadap aplikasi Identitas Kependudukan Digital dan dua anotator berbasis kecerdasan buatan. Penggunaan model AI sebagai anatator tambahan mengacu pada praktik penelitian sebelumnya [12]. Setiap anotator diberikan panduan pelabelan yang seragam mencakup definisi setiap aspek dan kategori emosi. Skema pelabelan menggunakan 5 aspek yang diperoleh dari BERTopic dan 5 kategori emosi, yaitu Gembira, Marah, Sedih, Takut, dan Lainnya. Sedangkan label “Tidak Ada” diberikan apabila ulasan tidak mengandung aspek-emosi yang relevan.

Tingkat kesepakatan antar seluruh anotator kemudian diukur menggunakan Krippendorff's Alpha sebagai metrik reliabilitas yang mampu menangani lebih dari dua anotator dengan tipe data kategorikal [13]. Proses anotasi dilakukan secara independen untuk meminimalkan bias, dan label final ditentukan melalui metode majority voting. Hasil pelabelan kemudian ditransformasi menggunakan teknik multi-hot encoding menjadi vektor biner. Setiap kombinasi aspek dan emosi direpresentasikan sebagai sebuah kolom biner, dengan format penamaan seperti A1_E1, A1_E2, hingga seterusnya, yang menunjukkan pasangan antara aspek ke-n dan emosi ke-m.

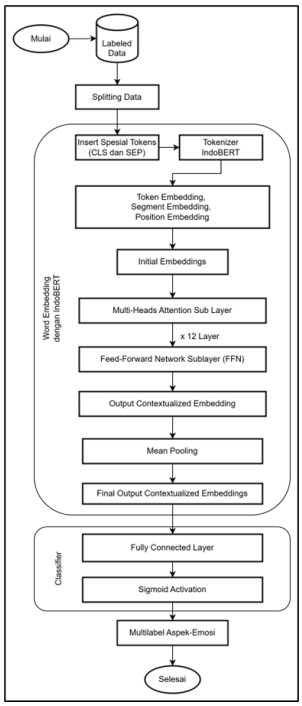
2.7. Pembagian Data

Dataset yang telah ditransformasi sebanyak 6.891 data dibagi menjadi dua bagian: 90% (6.201 data) sebagai data pemodelan dan 10% (690 data) sebagai data validasi sistem yang disimpan terpisah dan tidak digunakan dalam proses pelatihan maupun pengujian model.

2.8. Pengembangan Model Klasifikasi

Model yang dikembangkan memanfaatkan IndoBERT secara end-to-end, di mana seluruh parameter model dilatih bersama lapisan klasifikasi multilabel dalam satu arsitektur terintegrasi. Tahapan pemrosesan dimulai dari tokenisasi teks menggunakan algoritma WordPiece IndoBERT dengan penambahan token khusus [CLS] di awal dan [SEP] di akhir kalimat. Setiap token kemudian direpresentasikan melalui penjumlahan tiga jenis embedding: Token Embeddings, Position Embeddings, dan Segment Embeddings.

Representasi embedding diproses oleh Encoder IndoBERT yang terdiri dari lapisan-lapisan transformer dengan mekanisme multi-head self-attention bidireksional. Teknik mean pooling diterapkan untuk merata-ratakan contextualized embedding seluruh token menjadi representasi kalimat tunggal. Vektor hasil pooling kemudian dimasukkan ke dalam classification head berupa linear layer dengan aktivasi Sigmoid yang menghasilkan probabilitas independen untuk setiap label. Label dianggap aktif jika probabilitas melampaui nilai threshold yang telah ditentukan. Hyperparameter awal yang digunakan meliputi learning rate 2e-5, batch size 8, epoch 5, dan optimizer Adam.



Gambar 1. End-to-end IndoBERT Multilabel Untuk ABEA [7]

2.9. Skenario Eksperimen

Pada penelitian ini dilakukan 6 skenario pengujian yang memvariasikan preprocessing, focal loss, split data, dan threshold tuning untuk menemukan konfigurasi model yang optimal dalam menangani ketidakseimbangan distribusi label. Focal Loss merupakan pengembangan dari cross-entropy yang secara adaptif menekan kontribusi sampel yang mudah diprediksi dan memfokuskan pembelajaran pada sampel yang sulit. Threshold tuning dilakukan dengan mencari nilai threshold optimal pada validation set untuk menyeimbangkan precision dan recall. Konfigurasi model untuk setiap skenario dapat dilihat pada Tabel 3.

Tabel 3. Contoh Hasil Preprocessing

Skenario	Deskripsi
1	BCE, Threshold 0.5, Split 75:15:15
2	Preprocessing, BCE, Threshold 0.5, Split 75:15:15
3	Preprocessing, BCE with Focal Loss, Threshold 0.5, Split 75:15:15
4	Preprocessing, BCE with Focal Loss, Threshold 0.4 (Tuning), Split 75:15:15
5	Preprocessing, BCE with Focal Loss, Threshold 0.5, Split 60:20:20
6	Preprocessing, BCE with Focal Loss, Threshold 0.4 (Tuning), Split

60:20:20

2.10. Evaluasi Model
Evaluasi model menggunakan metrik F1 Score Micro, F1 Score Macro, Precision, Recall, dan Hamming Loss (proporsi kesalahan per label). Analisis confusion matrix per label juga dilakukan untuk mengidentifikasi pola kesalahan prediksi secara spesifik.

2.11. Deployment
Model IndoBERT terpilih yang telah dilatih dan dievaluasi diintegrasikan ke dalam antarmuka web menggunakan framework Flask berbasis Python. Sistem akan menerima input berupa file CSV yang berisi teks ulasan pengguna, kemudian memproses setiap ulasan secara otomatis dan memvisualisasikan hasil klasifikasinya.

2.12. Validasi Model
Validasi model dilakukan menggunakan 10% data (690 ulasan). Tujuan utama tahap ini adalah memastikan bahwa performa model pada data pelatihan dapat dipertahankan secara konsisten ketika dihadapkan pada data baru yang belum pernah dilihat sebelumnya, sehingga tidak terjadi overfitting terhadap data pengembangan.

3. HASIL DAN PEMBAHASAN
3.1. Hasil Identifikasi Aspek dengan BERTopic
Evaluasi kualitas topik dilakukan menggunakan metrik Coherence Score (C_v) dan Topic Diversity. Hasil evaluasi menunjukkan bahwa model BERTopic memperoleh nilai Coherence Score sebesar 0,6196 dan Topic Diversity sebesar 0,92. Setiap topik yang dihasilkan kemudian diinterpretasikan berdasarkan keyowrd utama. Hasil interpretasi dapat dilihat pada Tabel 4.

Tabel 4. Aspek Hasil BERTopic

Aspek	Daftar Kata	Deskripsi
Login & Akses Akun	masuk, data,update, pin, daftar, buka, ulang, tolong, akun, aplikasinya	Keluhan pengguna terkait kesulitan login, gagal masuk setelah update, masalah PIN, masalah koneksi saat buka aplikasi, serta kebutuhan daftar ulang akun
Scan Barcode ke Dukcapil	online, barcode, scan, scan barcode, dukcapil, kantor, petugas, digital, daftar, qr	Keluhan pengguna karena aktivasi akun masih harus scan barcode ke petugas Dukcapil sehingga proses online dianggap tidak sepenuhnya digital.
Verifikasi Foto Wajah	foto, ambil, ambil foto, wajah, frame, masuk, ktp, selfie, update, tombol	Masalah pada proses verifikasi identitas melalui foto/selfie seperti gagal ambil foto, frame tidak pas, atau tombol tidak berfungsi
Dokumen &	pelayanan, pengajuan,	Keluhan terkait layanan

Aspek	Daftar Kata	Deskripsi
Layanan Digital	kk, menu, cetak, menu pelayanan, pengguna, cetak kk, respon, buka	administrasi seperti pengajuan dokumen, cetak KK, menu pelayanan error, dan lambatnya pelayanan
Kompatibilitas Perangkat Android	android, hp, versi, android versi, hp android, versi android, support, terbaru, samsung, android terbaru	Keluhan terkait aplikasi tidak kompatibel dengan versi Android tertentu

3.2. Distribusi Label dan Karakteristik Data

Hasil pelabelan yang diuji konsistensinya menunjukkan bahwa nilai Krippendorff's Alpha berada pada rentang 0,71–0,73 yang tergolong moderat dan cukup reliabel. Selanjutnya, aspek final dari ulasan yang tidak mencapai kesepakatan ditentukan oleh majority voting, Distribusi label hasil majority voting yang disajikan pada Tabel 5.

Tabel 5. Distribusi Label per Aspek-Emosi

Aspek	Gembira	Marah	Sedih	Takut	Lainnya	Tidak ada
Login & Akses Akun	47	1510	374	28	65	4867
Scan Barcode ke Dukcapil	16	1122	139	3	19	5592
Verifikasi Foto Wajah	18	795	165	1	13	5899
Dokumen & Layanan Digital	227	656	124	25	65	5794
Kompatibilitas Perangkat Android	4	46	33	0	6	6802

Dari hasil transformasi, didapatkan total label aktif sebanyak 29. Hal ini karena Label A5_E4 (Kompatibilitas Android – Takut) tidak memiliki sampel positif sama sekali, sehingga selanjutnya model tidak dapat mengidentifikasi pasangan aspek emosi ini.

Berdasarkan hasil pelabelan yang ditunjukkan pada Tabel 3, terdapat ketimpangan distribusi label yang sangat ekstrem pada seluruh aspek. Label "Tidak Ada" mendominasi di semua aspek, dengan proporsi tertinggi pada aspek Kompatibilitas Perangkat Android (6.802 dari 6.891 data atau 98,7%). Emosi "Marah" menjadi kategori emosi paling dominan hampir di seluruh aspek, dengan jumlah tertinggi pada aspek Login & Akses Akun (1.510 data) dan Scan Barcode ke Dukcapil (1.122 data). Kondisi ini mengkonfirmasi bahwa keluhan dan ekspresi negatif mendominasi ulasan aplikasi IKD. Satu-satunya pengecualian terdapat pada aspek Dokumen & Layanan Digital, di mana emosi "Gembira" mencapai 227 data tertinggi

dibandingkan aspek lainnya yang mengindikasikan adanya pengalaman positif pengguna terhadap fitur layanan administrasi digital.

3.3. Hasil Evaluasi Model

Pada penelitian ini, evaluasi dilakukan menggunakan beberapa metrik yang saling melengkapi, yaitu F1 Score Micro, F1 Score Macro, Precision, Recall, serta Hamming Loss, dan diperkuat dengan analisis confusion matrix per label guna memberikan gambaran yang lebih komprehensif terhadap kemampuan model dalam menangani masing-masing kelas. Tabel 6 menunjukkan hasil pengujian performa model pada seluruh skenario.

Tabel 6. Performa Model pada Seluruh Skenario

Skenario	F1 Micro	F1 Macro	Precision	Recall	Hamming Loss
1	0,9174	0,3156	0,9404	0,8954	0,0278
2	0,9255	0,3185	0,9455	0,9063	0,0251
3	0,9234	0,3581	0,9297	0,9173	0,0262
4	0,9151	0,3768	0,8849	0,9476	0,0303
5	0,9195	0,3355	0,9390	0,9007	0,0272
6	0,9134	0,3916	0,8862	0,9423	0,0308

3.4. Performa Baseline dan Pengaruh Preprocessing

Skenario 1 sebagai baseline menghasilkan F1 Score Micro sebesar 0,9174 dan F1 Score Macro sebesar 0,3156. Nilai F1 Score Micro yang tinggi mencerminkan performa model yang baik pada label-label mayoritas, terutama label "Tidak Ada" yang mendominasi dataset. Namun, kesenjangan yang sangat besar antara F1 Score Micro dan F1 Score Macro mengindikasikan bahwa model gagal mengenali label-label minoritas. Analisis confusion matrix menunjukkan bahwa sejumlah label seperti A1_E1, A1_E4, A2_E1, dan A3_E1 sama sekali tidak berhasil diprediksi dengan TP = 0. Penambahan preprocessing pada Skenario 2 memberikan peningkatan yang konsisten pada seluruh metrik: F1 Score Micro naik menjadi 0,9255, F1 Score Macro naik menjadi 0,3185, dan Hamming Loss turun menjadi 0,0251. Peningkatan ini mengkonfirmasi bahwa normalisasi teks, terutama penanganan kata slang dan variasi penulisan informal, berkontribusi positif terhadap kualitas representasi input. Beberapa label yang sebelumnya tidak terdeteksi mulai menunjukkan prediksi yang benar, seperti A3_E3 yang mencapai TP = 7 dari sebelumnya TP = 0.

3.5. Pengaruh Focal Loss

Focal Loss digunakan untuk mengurangi dominasi kontribusi label mayoritas selama proses pelatihan model. Fungsi loss ini merupakan pengembangan dari Binary Cross-Entropy pada fungsi loss standar, sehingga secara adaptif menekan

kontribusi sampel yang mudah diprediksi dan memfokuskan pembelajaran pada sampel yang sulit diklasifikasikan maupun label minoritas.

Penggantian fungsi loss BCE standar dengan Focal Loss pada Skenario 3 menghasilkan peningkatan F1 Score Macro yang signifikan dari 0,3185 menjadi 0,3581, sekaligus mempertahankan F1 Score Micro yang tinggi pada 0,9234 dan Hamming Loss yang terjaga di 0,0262. Peningkatan F1 Score Macro sebesar 12,5% ini merupakan capaian terbesar yang dicapai dari satu perubahan tunggal tanpa melibatkan threshold tuning maupun perubahan rasio split data. Focal Loss terbukti lebih efektif dibandingkan pendekatan loss konvensional karena mampu secara adaptif memfokuskan pembelajaran pada sampel yang sulit diklasifikasikan tanpa menyebabkan ledakan false positive yang berlebihan. Beberapa label yang sebelumnya tidak terdeteksi sama sekali kini mulai memiliki nilai True Positive yang positif, seperti A1_E3 (TP: 23) dan A5_E2 (TP: 3). Temuan ini sejalan dengan penelitian [9] yang menunjukkan bahwa Focal Loss secara konsisten meningkatkan sensitivitas model berbasis transformer terhadap kelas minoritas pada dataset emosi yang tidak seimbang, tanpa mengorbankan performa keseluruhan secara signifikan.

Keunggulan Focal Loss dibandingkan BCE standar pada penelitian ini juga terlihat dari pola distribusi prediksi per label. Pada Skenario 2 (BCE dengan preprocessing), label-label minoritas seperti A1_E1 (Login–Gembira), A2_E1 (Scan Barcode–Gembira), dan A3_E1 (Verifikasi–Gembira) masih menghasilkan TP = 0 karena model didominasi oleh sinyal dari label mayoritas "Tidak Ada". Setelah Focal Loss diterapkan pada Skenario 3, model mulai menunjukkan kapasitas pengenalan yang lebih merata. Hal ini mengindikasikan bahwa Focal Loss berhasil mendorong model untuk tidak hanya mengoptimalkan performa pada label-label yang mudah dan sering muncul, tetapi juga memperhitungkan pasangan aspek–emosi yang jarang namun tetap memiliki nilai informatif tinggi bagi pengembang aplikasi IKD.

3.6. Pengaruh Threshold Tuning

Penelitian Hinojosa Lee et al. yang membandingkan berbagai varian F1-score pada klasifikasi emosi multilabel dengan kondisi class imbalance menunjukkan bahwa pemilihan threshold yang tepat memiliki dampak signifikan terhadap nilai metrik evaluasi yang dihasilkan [14]. Pada penelitian ini, threshold tuning dilakukan dengan mencari nilai threshold optimal melalui evaluasi sistematis pada validation set menggunakan rentang kandidat nilai dari 0,4 hingga 0,7 dengan interval 0,05. Rentang pencarian ini ditetapkan berdasarkan pertimbangan bahwa nilai di bawah 0,4 berpotensi menghasilkan terlalu banyak false positive pada label mayoritas, sementara nilai di atas 0,7 terlalu konservatif dan berisiko menekan recall pada label minoritas secara berlebihan.

Threshold tuning dilakukan menggunakan validation set dengan mengeksplorasi beberapa nilai threshold pada rentang 0,4 hingga 0,7 untuk memperoleh keseimbangan yang optimal antara precision dan recall pada klasifikasi multilabel. Proses ini dilakukan karena penggunaan threshold default sebesar 0,5 belum tentu menghasilkan performa terbaik pada dataset multilabel yang tidak seimbang.

Penerapan threshold 0,4 pada Skenario 4 (Focal Loss + threshold tuning + split 75:15:15) berhasil meningkatkan F1 Score Macro dari 0,3581 menjadi 0,3768 sekaligus mendorong Recall hingga 0,9476 yang merupakan nilai tertinggi di antara

semua skenario yang diuji. Penurunan threshold dari nilai default 0,5 ke nilai optimal 0,4 berarti model memerlukan keyakinan yang lebih rendah untuk mengaktifkan prediksi label positif, sehingga lebih banyak kasus label minoritas yang berhasil terdeteksi. Analisis per label memperlihatkan peningkatan yang merata, di mana label-label seperti A1_E3 (TP: 32), A3_E2 (TP: 113), dan A4_E2 (TP: 65) mengalami kenaikan yang berarti dibandingkan Skenario 3. Meskipun terjadi sedikit penurunan F1 Score Micro menjadi 0,9151 dan kenaikan Hamming Loss menjadi 0,0303, pertukaran ini menguntungkan dalam konteks analisis emosi berbasis aspek yang memprioritaskan kemampuan mendeteksi seluruh emosi yang muncul pada berbagai aspek termasuk pasangan aspek emosi yang minoritas..
Kombinasi Focal Loss dan threshold tuning yang diterapkan bersamaan pada Skenario 4 dan Skenario 6 terbukti menghasilkan kolaborasi yang efektif. Focal Loss mempersiapkan model untuk menghasilkan distribusi probabilitas yang lebih informatif pada label-label minoritas selama pelatihan, sementara threshold tuning kemudian memanfaatkan distribusi probabilitas tersebut secara optimal.

3.7. Pengaruh Split Data

Perubahan rasio split data dari 75:15:15 menjadi 60:20:20 pada Skenario 5 (Focal Loss, threshold 0,5) menghasilkan sedikit penurunan pada F1 Score Macro dari 0,3581 menjadi 0,3355 dibandingkan Skenario 3 dengan konfigurasi yang setara. Pengurangan proporsi data latih menyebabkan model kehilangan sedikit kapasitas generalisasi, terutama pada label-label dengan sedikit sampel. Namun demikian, perbedaan yang tidak terlalu drastis menunjukkan bahwa model cukup robust terhadap perubahan rasio split dalam batas tertentu. Ketika threshold tuning diterapkan pada split 60:20:20 di Skenario 6, model justru mencapai performa terbaik secara keseluruhan. Ukuran data uji yang lebih besar (20%) memberikan estimasi performa yang lebih representatif terhadap kondisi data baru.

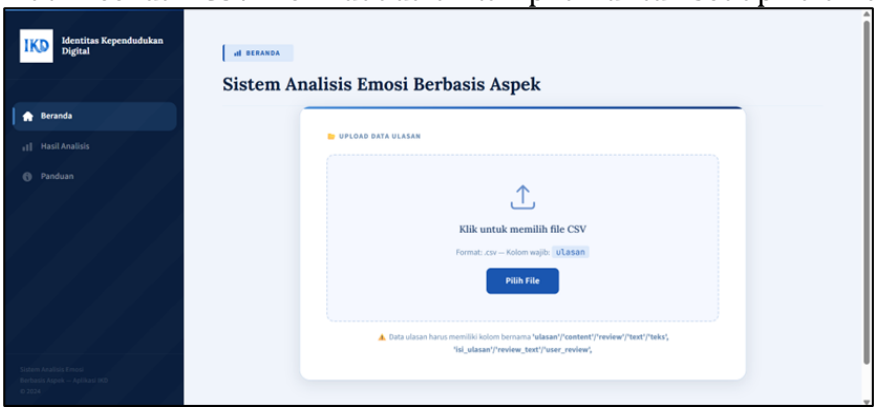
3.8. Evaluasi Model Terbaik (Skenario 6)

Skenario 6 ditetapkan sebagai konfigurasi terbaik dalam penelitian ini berdasarkan kemampuannya melakukan optimasi model di tengah kondisi distribusi data yang tidak seimbang. Pemilihan ini dilakukan dengan mempertimbangkan tujuan penelitian, yaitu memperoleh model Analisis Emosi Berbasis Aspek yang mampu mengidentifikasi emosi pengguna terhadap setiap aspek layanan secara lebih mendalam. Dalam kasus klasifikasi multilabel dengan distribusi label yang tidak seimbang, nilai F1 Score Micro cenderung didominasi oleh performa model pada label-label mayoritas sehingga belum sepenuhnya mencerminkan kemampuan model dalam mengenali seluruh kombinasi aspek dan emosi [14]. Oleh karena itu, penelitian ini lebih memprioritaskan peningkatan F1 Score Macro karena metrik tersebut memberikan evaluasi yang lebih seimbang pada setiap label, termasuk label minoritas yang tetap penting dalam analisis emosi pengguna. Meskipun nilai F1 Score Micro pada Skenario 6 sedikit lebih rendah dibandingkan beberapa skenario lainnya, penurunan tersebut relatif kecil dan masih menunjukkan performa keseluruhan yang baik. Di sisi lain, peningkatan F1 Score Macro menjadi 0,3916 menunjukkan bahwa model memiliki kemampuan yang lebih baik dalam mengenali variasi emosi pada berbagai aspek layanan secara lebih merata. Dengan demikian, Skenario 6 dipilih sebagai yang paling optimal dalam menghasilkan

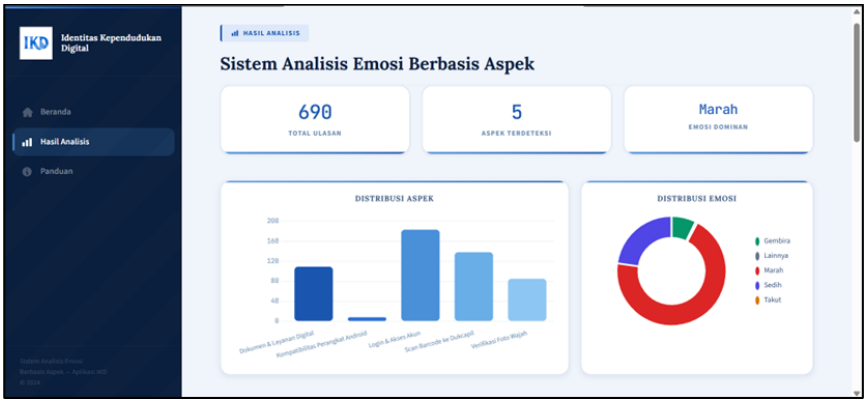
representasi emosi pengguna yang lebih komprehensif pada setiap aspek layanan aplikasi IKD.

3.9. Hasil Deployment

Sistem berbasis web berhasil dikembangkan menggunakan framework Flask. Sistem terdiri dari tiga halaman utama: Halaman Beranda untuk mengunggah file CSV, Halaman Hasil Analisis yang menyajikan ringkasan statistik, grafik distribusi aspek dan emosi, tabel matriks aspek-emosi, dan Halaman Panduan yang memuat panduan penggunaan sistem. Selain itu, sistem juga disediakan fitur ekspor hasil dalam format Excel dan CSV. Berikut adalah tampilan untuk setiap halaman.

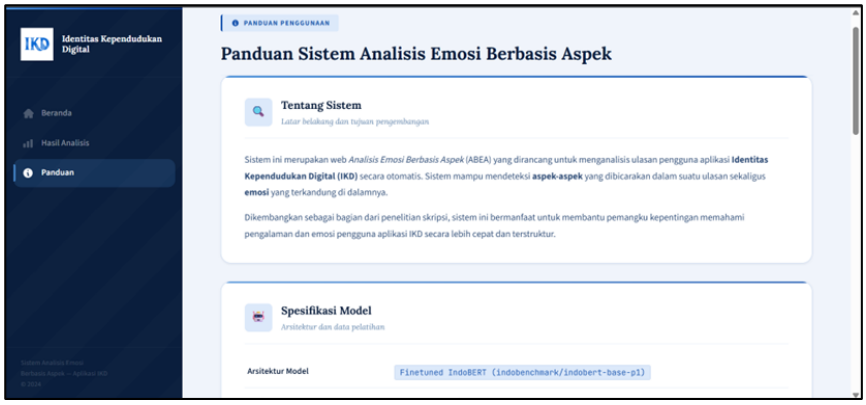


Gambar 2. Halaman Beranda



Gambar 3. Halaman Hasil Analisis (Digaram)

Gambar 4. Halaman Hasil Analisis (Tabel)



Gambar 5. Halaman Panduan

3.10. Hasil Validasi Model

Validasi model pada sistem menggunakan 10% data yang tidak digunakan dalam pelatihan menunjukkan bahwa model mampu mempertahankan performanya secara konsisten. F1 Score Micro turun hanya 0,36% dari 0,9134 menjadi 0,9098, dan Recall turun 0,43% dari 0,9423 menjadi 0,9380. Penurunan F1 Score Macro sebesar 8,82% dari 0,3916 menjadi 0,3571 dapat dijelaskan sebagai konsekuensi statistik dari distribusi label minoritas yang semakin langka pada data validasi yang lebih kecil. Secara keseluruhan, hasil validasi mengkonfirmasi bahwa model tidak mengalami overfitting

KESIMPULAN

Penelitian ini berhasil mengembangkan model Multilabel Aspect-Based Emotion Analysis pada ulasan aplikasi IKD menggunakan end-to-end IndoBERT dengan tiga temuan utama. Pertama, BERTopic berhasil mengidentifikasi 5 aspek layanan yang representatif dengan Coherence Score 0,6196 dan Topic Diversity 0,92, yaitu Login & Akses Akun, Scan Barcode ke Dukcapil, Verifikasi Foto Wajah, Dokumen & Layanan Digital, dan Kompatibilitas Perangkat Android. Kedua, kombinasi Focal Loss dan threshold tuning (0,4) pada split 60:20:20 menghasilkan konfigurasi optimal dengan F1 Score Macro 0,3916 (meningkat 24,1% dari baseline 0,3156), F1 Score Micro 0,9134, dan Recall 0,9423. Focal Loss terbukti efektif meningkatkan sensitivitas model terhadap label minoritas tanpa mengorbankan performa secara signifikan, sementara threshold tuning pada rentang 0,4–0,7 menghasilkan nilai optimal 0,4 yang mendorong recall tertinggi sebesar 0,9476. Ketiga, Emosi marah mendominasi hampir seluruh aspek layanan IKD, terutama pada Login & Akses Akun dan Scan Barcode ke Dukcapil, sementara Dokumen & Layanan Digital menjadi satu-satunya aspek dengan emosi gembira yang paling banyak. Ke depannya, kerangka ABEA yang dibangun dalam penelitian ini berpotensi dikembangkan lebih lanjut melalui teknik penanganan data tidak seimbang lainnya, seperti data augmentation berbasis model bahasa, oversampling label minoritas, atau penyesuaian threshold per label untuk meningkatkan performa model. Selain itu, penerapan pendekatan *hierarchical classification* yang memisahkan proses deteksi aspek dan klasifikasi emosi secara bertahap juga berpotensi menghasilkan pemahaman hubungan antar tugas yang lebih terstruktur dan akurat.

DAFTAR PUSTAKA

[1] Irma Nurdiana and Khithoh Ayumi, "Implementasi Aplikasi Identitas Kependudukan Digital (IKD) Di Disdukcapil Kota Tanjungpinang," *Harmoni Sos. J. Pengabdi. Dan Solidar. Masy.*, vol. 1, no. 2, pp. 50–58, Apr. 2024, doi: 10.62383/harmoni.v1i2.141.

[2] P. Hakiki, D. Satria, and A. A. Arifiyanti, "Prediksi Sentimen dan Pemodelan Topik dari Ulasan Aplikasi Identitas Kependudukan Digital," *Jutisi J. Ilm. Tek. Inform. Dan Sist. Inf.*, vol. 14, no. 1, p. 760, Jul. 2025, doi: 10.35889/jutisi.v14i1.2777.

[3] Kementerian Dalam Negeri, "Peraturan Menteri Dalam Negeri Nomor 72 Tahun 2022 tentang Standar dan Spesifikasi Perangkat Keras, Perangkat Lunak, dan Blangko Kartu Tanda Penduduk Elektronik serta Penyelenggaraan Identitas Kependudukan Digital." 2022. [Online]. Available: https://peraturan.bpk.go.id/Details/247759/permendagri-no-72-tahun-2022

[4] D. S. Akbar Rizki, M. S. Khabib, N. Rahmayuna, and V. G. Utomo, "Klasifikasi Sentimen Ulasan Pengguna Aplikasi Layanan Publik Google Play Store Menggunakan NLP dan ML," *J. Tekno Kompak*, vol. 20, no. 1, pp. 51–64, Oct. 2025, doi: https://doi.org/10.33365/jtk.v20i1.586.

[5] D. E. Sondakh, R. C. Maringka, F. P. Ayorbaba, J. S. C. B. T. Mangi, and S. R. Pungus, "Emotion Mining User Review of the BRImo Mobile Banking Application Using the Decision Tree Algorithm," *J. Sisfokom Sist. Inf. Dan Komput.*, vol. 12, no. 3, pp. 350–355, Nov. 2023, doi: 10.32736/sisfokom.v12i3.1721.

[6] L. De Bruyne, A. Karimi, O. De Clercq, A. Prati, and V. Hoste, "Aspect-Based Emotion Analysis and Multimodal Coreference: A Case Study of Customer Comments on Adidas Instagram Posts," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Jun. 2022, pp. 574–580. [Online]. Available: https://aclanthology.org/2022.lrec-1.61/

[7] N. C. Mei, S. Tiun, and G. Sastria, "Multi-Label Aspect-Sentiment Classification on Indonesian Cosmetic Product Reviews with IndoBERT Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, 2024, doi: 10.14569/IJACSA.2024.0151168.

[8] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Electr. Comput. Eng. IJECE*, vol. 13, no. 5, p. 5641, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[9] A. Kesanam, G. V. R. Ram, C. S. Banoth, and G. R. M. Reddy, "NITK-VITAL at SemEval-2025 Task 11: Focal-RoBERTa: Addressing Class Imbalance in Multi-Label Emotion Classification," in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1077–1081. [Online]. Available: https://aclanthology.org/2025.semeval-1.142/

[10] M. D. Pratiwi and K. D. Tania, "Knowledge Discovery Through Topic Modeling on GoPartner User Reviews Using BERTopic, LDA, and NMF," *J. Appl. Inform. Comput.*, vol. 9, no. 1, pp. 1–7, Jan. 2025, doi: 10.30871/jaic.v9i1.8782.

[11] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv*. doi: 10.48550/ARXIV.2203.05794.

[12] H. Aka Uymaz and S. Kumova Metin, "Collaborative Emotion Annotation: Assessing the Intersection of Human and AI Performance with GPT Models:," in *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2023, pp. 298–305. doi: 10.5220/0012183200003598.

[13] G. Marzi, M. Balzano, and D. Marchiori, "K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient," *MethodsX*, vol. 12, p. 102545, Jun. 2024, doi: 10.1016/j.mex.2023.102545.

[14] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Appl. Sci.*, vol. 14, no. 21, p. 9863, Oct. 2024, doi: 10.3390/app14219863.