

**ANALISIS EMOSI KOMENTAR YOUTUBE TERKAIT
KEKERASAN SEKSUAL MEI 1998: DALAM MODEL BERT,
GRU, DAN HDBSCAN**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh
AINUR RAHMI SELIAN
22.11.5125

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**ANALISIS EMOSI KOMENTAR YOUTUBE TERKAIT
KEKERASAN SEKSUAL MEI 1998: DALAM MODEL BERT,
GRU, DAN HDBSCAN**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh
AINUR RAHMI SELIAN
22.11.5125

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

HALAMAN PERSETUJUAN

SKRIPSI

**ANALISIS EMOSI KOMENTAR YOUTUBE TERKAIT KEKERASAN
SEKSUAL MEI 1998: DALAM MODEL BERT, GRU, DAN HDBSCAN**

yang disusun dan diajukan oleh

Ainur Rahmi Selian

22.11.5125

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 20 Februari 2026

Dosen Pembimbing,



Windha Mega Pradnya Dhuhiha, S.Kom., M.Kom.
NIK. 190302185

HALAMAN PENGESAHAN

SKRIPSI

**ANALISIS EMOSI KOMENTAR YOUTUBE TERKAIT KEKERASAN
SEKSUAL MEI 1998: DALAM MODEL BERT, GRU, DAN HDBSCAN**

yang disusun dan diajukan oleh

Ainur Rahmi Selian

22.11.5125

Telah dipertahankan di depan Dewan Penguji
pada tanggal 20 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Rizqi Sukma Kharisma, S.Kom., M.Kom.
NIK. 190302215

Bambang Pitu Hartato, S.Kom., M.Eng.
NIK. 190302707

Windha Mega Pradnya Dhuhiha, S.Kom., M.Kom.
NIK. 190302185

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal, 20 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Ainur Rahmi Selian
NIM : 22.11.5125

Menyatakan bahwa Skripsi dengan judul berikut:

**Analisis Emosi Komentar Youtube Terkait Kekerasan Seksual Mei 1998:
Dalam Model BERT, GRU, dan HDBSCAN**

Dosen Pembimbing : Windha Mega Pradnya Dhuhiha, S.Kom., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 20 Februari 2026

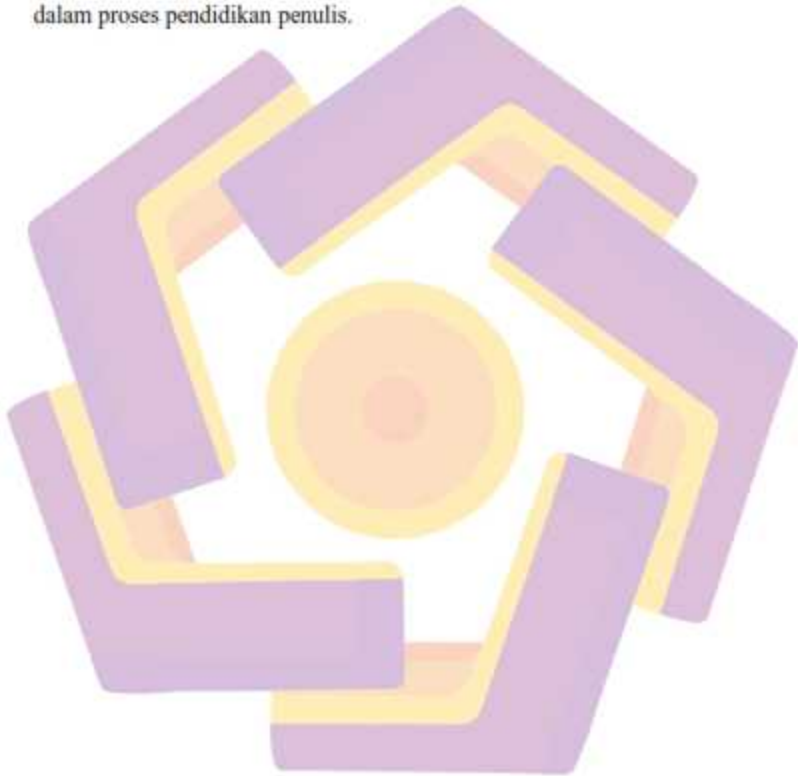
Yang Menyatakan,



Ainur Rahmi Selian

HALAMAN PERSEMBAHAN

Laporan ini penulis persembahkan kepada orang tua yang selalu memberikan doa dan dukungan, serta kepada semua pihak yang telah membantu dalam proses pendidikan penulis.



KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul “Analisis Emosi pada Komentar YouTube Menggunakan Model IndoBERT dengan Pendekatan UMAP dan HDBSCAN” dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana pada Program Studi Informatika di Universitas Amikom Yogyakarta.

Dalam proses penyusunan skripsi ini, penulis menyadari bahwa banyak pihak yang telah memberikan bantuan, dukungan, serta bimbingan. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

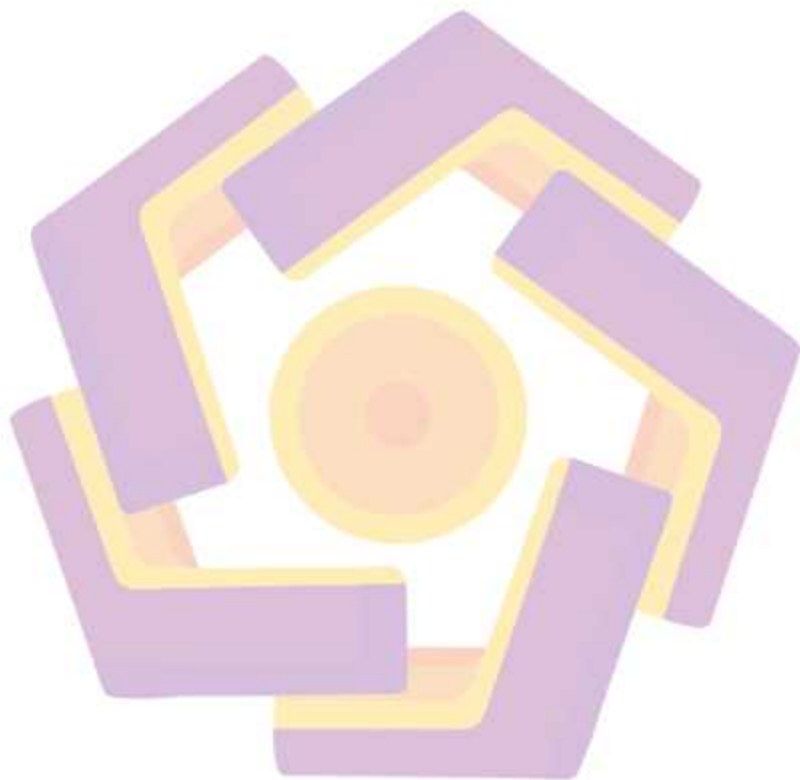
1. Orang tua dan keluarga penulis ibu Wahyuni S.Pd.I., ayah Muhammad Yaman Brueh, Adik Ammar Dzaky, dan Bapak Ahmad Thahir S.Pd.I., yang selalu memberikan doa, dukungan, serta semangat kepada penulis dalam menyelesaikan studi.
2. Ibu dosen pembimbing Windha Mega Pradnya Dhuhiita, S.Kom., M.Kom. yang telah memberikan bimbingan, arahan, serta perbaikan untuk penulis bisa mengerti dan paham dengan penjelasan yang diberikan selama proses penyusunan skripsi. Sehingga memberi penulis ruang untuk bisa berpendapat, ide, gagasan, dengan berbagai keterbatasan penulis.
3. Tim dosen penguji bapak Rizqi Sukma Kharisma, S.Kom., M.Kom. dan bapak Bambang Pilu Hartato, S.Kom., M.Eng. yang telah memberikan masukan dan saran yang sangat berharga demi penyempurnaan skripsi ini.
4. Teman-teman penulis Hanifah Putri Aksiomatika, Putra Ajril Fatah, Paundra Asfa Ravikansya, dan Akhmad Rizki Ramadhani, yang telah memberikan semangat dan selamat, serta menemani penulis dalam proses skripsi hingga ujian pendadaran di Universitas Amikom Yogyakarta.

Penulis menyadari bahwa skripsi ini masih memiliki kekurangan. Oleh karena itu, penulis mengharapkan saran yang membangun demi perbaikan di masa

yang akan datang. Semoga skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang Informatika.

Yogyakarta, 20 Februari 2026

Penulis



DAFTAR ISI

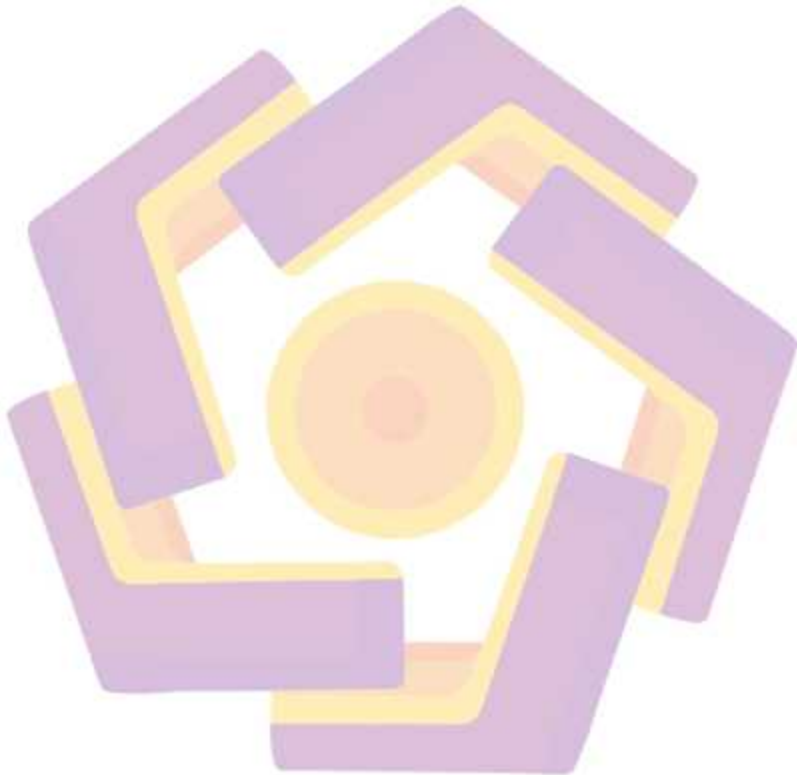
HALAMAN JUDUL.....	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN.....	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMPIRAN.....	xii
DAFTAR LAMBANG DAN SINGKATAN.....	xiii
DAFTAR ISTILAH.....	xiv
INTISARI.....	xvi
ABSTRACT.....	xvii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	4
1.5 Manfaat Penelitian.....	4
1.6 Sistematika Penulisan.....	5
BAB II TINJAUAN PUSTAKA.....	6

2.1	Studi Literatur	6
2.2	Dasar Teori.....	13
BAB III METODE PENELITIAN		22
3.1	Objek Penelitian.....	22
3.2	Alur Penelitian	23
3.4	Alat dan Bahan.....	35
BAB IV HASIL DAN PEMBAHASAN		37
BAB V PENUTUP.....		61
5.1	Kesimpulan	61
REFERENSI		63
LAMPIRAN.....		66



DAFTAR TABEL

Tabel 2.1. Keaslian Penelitian	9
Tabel 4.1. Komentar	50
Tabel 4.2. Tabel akurasi IndoBERT	52
Tabel 4.3. Tabel akurasi GRU	54
Tabel 4.4. Perbandingan	56

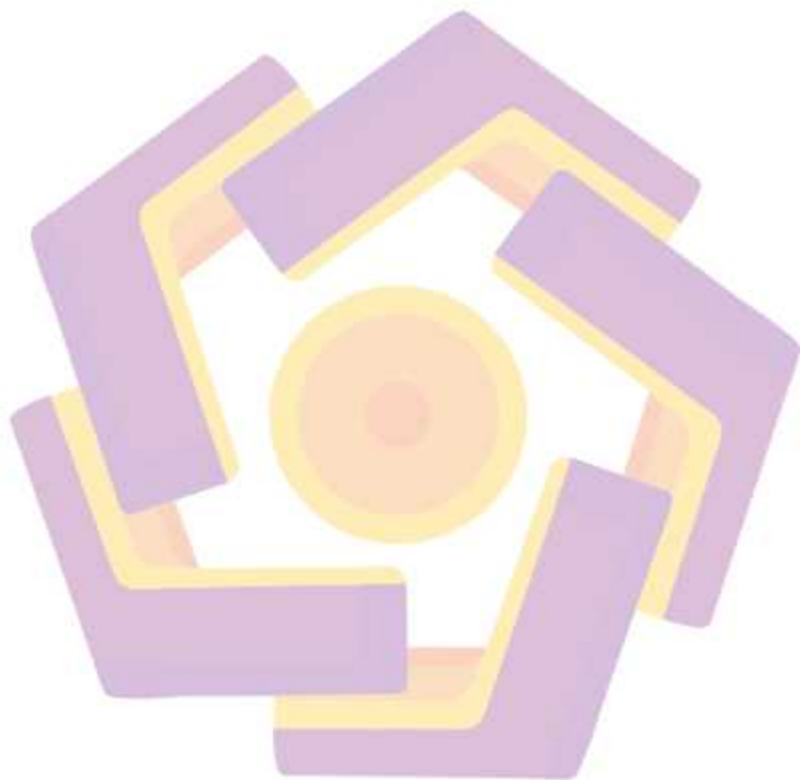


DAFTAR GAMBAR

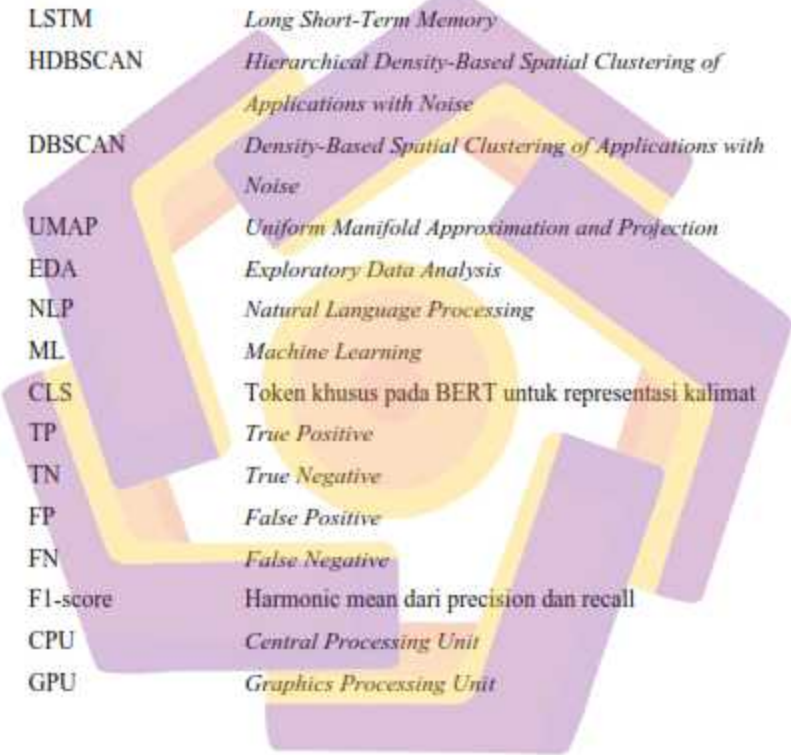
Gambar 2.1 Token Model BERT	16
Gambar 3.1 Alur Penelitian	22
Gambar 3.2 Alur <i>Deployment</i>	34
Gambar 4.1 Kode <i>Scrapping attribut</i>	37
Gambar 4.2 Melihat isi data	38
Gambar 4.3 Melihat tipe data	38
Gambar 4.4 Statistik panjang komentar	39
Gambar 4.5 Histogram distribusi panjang komentar	39
Gambar 4.6 Jumlah <i>missing values</i>	40
Gambar 4.7 Jumlah duplikat	40
Gambar 4.8 Hasil <i>cleaning</i>	41
Gambar 4.9 Hasil <i>lower casing</i>	41
Gambar 4.10 Hasil text normalisasi	42
Gambar 4.11 Tokenisasi <i>wordpiece</i>	43
Gambar 4.12 <i>Encoder</i> IndoBERT	44
Gambar 4.13 <i>Mean Pooling</i>	44
Gambar 4.14 <i>Sentence embedding</i>	45
Gambar 4.15 Vektor kalimat	46
Gambar 4.16 UMAP	47
Gambar 4.17 HDBSCAN	48
Gambar 4.18 Visualisasi <i>cluster</i> HDBSCAN	48
Gambar 4.19 Validation <i>cluster</i> HDBSCAN	49
Gambar 4.20 Persentase emosi	50
Gambar 4.21 Parameter IndoBERT	52
Gambar 4.22 <i>Confusion Matrix</i> IndoBERT	53
Gambar 4.23 Parameter GRU	54
Gambar 4.24 <i>Confusion Matrix</i> GRU	55
Gambar 4.25 Tampilan <i>VS code</i> file main.py	58
Gambar 4.26 Tampilan <i>streamlit</i> analisis emosi	59
Gambar 2.27 Tampilan <i>streamlit</i> hasil analisis emosi	60

DAFTAR LAMPIRAN

Lampiran 1. Profil obyek Penelitian	10
Lampiran 2. Dokumentasi Penelitian	11



DAFTAR LAMBANG DAN SINGKATAN



BERT	<i>Bidirectional Encoder Representations from Transformers</i>
IndoBERT	Model BERT yang dilatih khusus pada korpus bahasa Indonesia
GRU	<i>Gated Recurrent Unit</i>
RNN	<i>Recurrent Neural Network</i>
LSTM	<i>Long Short-Term Memory</i>
HDBSCAN	<i>Hierarchical Density-Based Spatial Clustering of Applications with Noise</i>
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
UMAP	<i>Uniform Manifold Approximation and Projection</i>
EDA	<i>Exploratory Data Analysis</i>
NLP	<i>Natural Language Processing</i>
ML	<i>Machine Learning</i>
CLS	Token khusus pada BERT untuk representasi kalimat
TP	<i>True Positive</i>
TN	<i>True Negative</i>
FP	<i>False Positive</i>
FN	<i>False Negative</i>
F1-score	Harmonic mean dari precision dan recall
CPU	<i>Central Processing Unit</i>
GPU	<i>Graphics Processing Unit</i>

DAFTAR ISTILAH



Embedding	Representasi numerik dari teks dalam bentuk vektor yang menangkap makna semantik sehingga dapat diproses oleh model pembelajaran mesin.
Clustering	Teknik pengelompokan data tanpa label (unsupervised) berdasarkan kemiripan tertentu, dalam penelitian ini digunakan untuk pelabelan awal emosi komentar YouTube.
Pretrained Model	Model yang telah dilatih pada dataset besar sebelumnya dan kemudian disesuaikan (<i>fine-tuning</i>) untuk tugas tertentu, seperti klasifikasi emosi.
Fine-Tuning	Proses melatih ulang sebagian atau seluruh parameter model <i>pretrained</i> pada dataset spesifik agar sesuai dengan konteks tugas penelitian.
Self-Attention	Mekanisme dalam arsitektur <i>transformer</i> yang memungkinkan model memperhatikan hubungan antar kata dalam satu kalimat secara kontekstual.
Mean Pooling	Teknik agregasi embedding dengan menghitung rata-rata vektor token untuk memperoleh satu representasi vektor kalimat.
Token [CLS]	Token khusus pada BERT yang merepresentasikan keseluruhan kalimat dan digunakan sebagai masukan untuk tugas klasifikasi.
Noise (Outlier)	Data yang tidak termasuk ke dalam kluster manapun karena memiliki kepadatan rendah atau karakteristik yang menyimpang dari pola umum data.

Klasifikasi Emosi Proses pengelompokan teks ke dalam kategori emosi tertentu (marah, sedih, takut, bahagia, cinta) berdasarkan pola linguistik dan semantik.

Analisis Emosi Pendekatan analisis teks untuk mengidentifikasi dan memetakan emosi yang diekspresikan pengguna dalam komentar media sosial.

Unsupervised Learning Metode pembelajaran mesin tanpa label data, digunakan dalam penelitian ini pada tahap awal pelabelan otomatis menggunakan HDBSCAN.

Supervised Learning Metode pembelajaran mesin dengan data berlabel, digunakan pada tahap klasifikasi emosi menggunakan IndoBERT dan GRU.

INTISARI

Penelitian ini membahas analisis emosi pada komentar *YouTube* terkait kekerasan seksual pada Tragedi Mei 1998. Komentar *YouTube* dominannya menggunakan bahasa singkatan, sarkas, dan satu komentar bisa mengandung emosi campur, serta berpotensi bias dan *noise* menyebabkan proses pelabelan manual menjadi sulit dan beresiko tidak konsisten. Dampaknya, informasi sentimen publik yang seharusnya dapat digunakan untuk memahami respons masyarakat terhadap isu sensitif ini menjadi kurang terukur dan sulit dianalisis secara sistematis. Metode yang digunakan adalah kombinasi pendekatan *unsupervised* dan *supervised learning*. Data dikumpulkan dari 18 video *YouTube* sehingga diperoleh 13.818 komentar. Setelah proses pembersihan data menggunakan *EDA* dan *preprocessing*, data menjadi 11.130 komentar. Representasi kalimat dibentuk menggunakan *embedding IndoBERT*, direduksi dengan *UMAP*, lalu dikelompokkan menggunakan *HDBSCAN* untuk menghasilkan pelabelan awal serta mengidentifikasi *noise* 9.70% dari total data. Hasil *cluster* diberi label manual dari komentar yang maknanya dominan mengarah ke salah satu emosi yaitu marah, sedih, takut, bahagia, dan cinta. kemudian hasil label tersebut digunakan untuk membandingkan kinerja klasifikasi *IndoBERT* dan *GRU*. Hasil evaluasi menunjukkan *IndoBERT* unggul dengan akurasi 0.8095 dan *F1-score* 0.7107, sedangkan *GRU* memperoleh akurasi 0.6532 dan *F1-score* 0.5321. kontribusi penelitian ini adalah menyediakan alur analisis emosi yang lebih efisien dan konsisten serta implementasi aplikasi *Streamlit* untuk *EDA*, *preprocessing*, dan evaluasi model. Hasil penelitian dapat digunakan untuk pemetaan emosi publik pada isu sosial melalui media sosial.

Kata kunci: Emosi, Komentar, *YouTube*, *IndoBERT*, *GRU*, *UMAP*, *HDBSCAN*

ABSTRACT

This study explores emotion analysis in YouTube comments related to sexual violence issues in the May 1998 tragedy. YouTube comments often use abbreviations, informal language, and sarcasm, and a single comment may contain mixed emotions. These characteristics can introduce bias and noise, making manual labeling difficult, time-consuming, and potentially inconsistent. As a result, public emotional responses toward this sensitive topic become harder to measure and analyze systematically. To address this, the study combines unsupervised and supervised learning approaches. Data were collected from eight YouTube videos, resulting in 13,818 comments. After data cleaning through EDA and preprocessing, 11,130 comments remained. Sentence representations were built using IndoBERT embeddings, reduced with UMAP, and clustered using HDBSCAN to generate initial labels and identify noise, which accounted for 9.70% of the dataset. The clusters were then manually assigned to dominant emotions: anger, sadness, fear, happiness, and love. These labels were used to compare the classification performance of IndoBERT (fine-tuning) and GRU. The evaluation results show that IndoBERT performed better, achieving 0.8095 accuracy and a macro F1-score of 0.7107, while GRU achieved 0.6532 accuracy and a macro F1-score of 0.5321. This research contributes an efficient and more consistent emotion analysis workflow and provides a Streamlit application for EDA, preprocessing, and model evaluation. The results can support mapping public emotions on social issues using social media data.

Keyword: emotion, comments, YouTube, IndoBERT, GRU, UMAP, HDBSCAN