IDENTIFIKASI MEME *SELF-HARM* MENGGUNAKAN *INTERMEDIATE FUSION* MODEL MULTIMODAL CLIP DAN ELECTRA

TUGAS AKHIR

Diajukan sebagai syarat menyelesaikan jenjang strata Satu (S-1) di Program Studi Teknik Informatika, Fakultas Teknologi Industri, Institut Teknologi Sumatera

Oleh:

Elsa Elisa Yohana Sianturi

122140135

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2025

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR RUMUS

BAB I
PENDAHULUAN

1.1 Latar Belakang

Transformasi dunia digital dalam satu dekade terakhir telah membawa dampak besar pada berbagai aspek kehidupan manusia, mulai dari pendidikan, kesehatan, industri kreatif, hingga kesejahteraan sosial. Perubahan ini ditandai oleh munculnya ruang interaksi baru di platform daring, yang memungkinkan masyarakat berbagi informasi, emosi, dan pengalaman hidup dalam berbagai bentuk konten. Perkembangan pesat teknologi informasi dan komunikasi telah mengubah cara manusia berinteraksi, berkomunikasi, dan memproduksi informasi dalam skala global yang menciptakan struktur sosial dan budaya baru serta cara baru dalam berbagi pengetahuan dan pengalaman kolektif. Dalam ranah komunikasi sosial dan interaksi interpersonal, media digital memungkinkan terbentuknya jaringan sosial luas tanpa batas geografis, serta mempercepat penyebaran informasi dan gagasan pada khalayak global [1].

Media sosial kini menjadi infrastruktur komunikasi utama dunia. Menurut laporan DataReportal, sebuah platform global yang menyajikan analisis digital berbasis data dari mitra riset internasional seperti Kepios, We Are Social, dan Meltwater mencatat bahwa terdapat sekitar 5,04 miliar pengguna media sosial aktif pada awal 2024, atau 62,3% populasi dunia, dengan pertumbuhan tahunan sebesar +5,6%, setara dengan 266 juta pengguna baru sepanjang 2023 [2], [3]. Rata-rata waktu penggunaan mencapai 2 jam 23 menit per hari [3], menunjukkan bahwa platform digital kini menjadi ruang publik yang sangat aktif sebagai tempat berbagai bentuk ekspresi diri, termasuk teks, gambar, dan format multimodal seperti meme. Besarnya intensitas penggunaan media sosial ini berkaitan langsung dengan meningkatnya fenomena kesehatan mental di ruang digital.

Platform seperti Instagram, Twitter, dan TikTok sering menjadi ruang bagi remaja dan dewasa muda untuk mengekspresikan emosi dan membagikan pengalaman personal, termasuk konten terkait *self-harm.* Studi terbaru menemukan bahwa paparan konten *self-harm* di media sosial berhubungan dengan meningkatnya risiko dorongan serta perilaku *nonsuicidal self-injury* (NSSI) pada remaja [4]. Sebuah survei berskala besar yang dilakukan oleh Samaritans bersama Swansea University menemukan bahwa sebanyak 83% pengguna media sosial pernah direkomendasikan konten bernuansa *self-harm* oleh algoritma *feed*, seperti halaman *Explore* di Instagram atau *For You* di TikTok, meskipun mereka tidak secara aktif mencarinya [5].

Menurut *World Health Organization*, secara global bunuh diri merupakan penyebab kematian tertinggi ketiga di kalangan remaja akhir dan dewasa muda usia 15–29 tahun, yang menunjukkan bahwa isu kesehatan mental pada kelompok usia ini merupakan tantangan kesehatan masyarakat yang serius [6]. Pada tahun 2024 sebuah meta-analisis terhadap remaja usia 10–19 tahun menemukan bahwa 17.7% remaja pernah melakukan NSSI, dengan 21.4% pada remaja perempuan dan 13.7% pada remaja laki-laki. Temuan ini berasal dari kumpulan data yang mencakup remaja di 17 negara yang tersebar di Amerika Utara, Australia, Eropa, dan Asia [7]. Sementara itu, berdasarkan data *Global Burden of Disease* (GBD) 2021, jumlah kematian akibat *self-harm* secara global mencapai sekitar 746,400 kasus pada tahun 2021 [8], hal ini menegaskan bahwa *self-harm* tetap menjadi isu kesehatan masyarakat yang signifikan di berbagai wilayah dunia. Paparan konten *self-harm* di media sosial bahkan terbukti dapat memprediksi perilaku menyakiti diri dalam satu bulan berikutnya pada populasi muda [9].

Pada konteks meme, pesan sering kali disampaikan secara implisit melalui perpaduan visual dan teks, sehingga pemahamannya bergantung pada hubungan antar-keduanya dan tidak dapat ditafsirkan hanya dari salah satu modalitas saja. Dalam penelitian yang dibahas dalam *"The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes"* oleh Kiela et al. , meme

seringkali menyembunyikan makna yang lebih mendalam yang hanya bisa dipahami dengan menggabungkan konteks visual dan tekstual. Gambar 1.1 di bawah ini menggambarkan kalimat *"Look how many people love you"* bisa terlihat tidak berbahaya jika dipisahkan, namun jika dipasangkan dengan gambar padang gurun, maknanya bisa berubah menjadi sesuatu yang lebih negatif. . Fenomena ini menggambarkan bagaimana model berbasis unimodal sering kali kesulitan untuk mendeteksi nuansa yang hanya bisa dipahami melalui pemahaman multimodal, yakni gabungan antara teks dan gambar [10].

Gambar 1.1 Contoh Meme dengan Makna Tersirat [10]

Shah et al. menegaskan bahwa gambar yang memuat teks membutuhkan pemahaman multimodal mendalam untuk menafsirkan makna tersirat tersebut [11]. Tantangan ini semakin diperjelas oleh Sharma et al. [12], yang menunjukkan bahwa sebagian besar penelitian masih berfokus pada deteksi *hate speech*, sementara jenis konten berbahaya lainnya seperti *self-harm* dan ekstremisme masih sangat kurang dieksplorasi akibat keterbatasan dataset publik. Untuk mengatasi keterbatasan dataset berlabel, pendekatan *pseudo-labeling* digunakan. *Pseudo-labeling* adalah teknik dimana model yang sudah dilatih digunakan untuk memberi label pada data yang tidak berlabel [13] sehingga memperluas dataset tanpa memerlukan anotasi manual. Teknik ini memungkinkan pelatihan model dengan data tambahan meskipun dengan label yang lemah, yang mengatasi keterbatasan annotator dan menggarisbawahi adanya kesenjangan riset yang signifikan terkait deteksi meme dengan nuansa *self-harm.*

Berdasarkan kompleksitas tersebut, pendekatan multimodal semakin relevan untuk analisis konten digital. Teknologi kecerdasan buatan kini memungkinkan pemrosesan informasi visual dan tekstual secara bersamaan, sehingga memberikan pemahaman konteks yang lebih komprehensif dibandingkan pendekatan unimodal, khususnya untuk konten bermakna implisit seperti meme.

1.2 Rumusan Masalah

1. Bagaimana membangun model klasifikasi biner untuk mengidentifikasi *self-harm* pada meme secara multimodal?
2. Bagaimana menyusun dataset meme multimodal melalui pengumpulan manual, pembuatan data tambahan, dan proses *pseudo-labeling* berbasis *Large Language Model* (LLM) untuk pelatihan model klasifikasi?
3. Bagaimana mengevaluasi kinerja model klasifikasi menggunakan metrik *accuracy, precision, recall, F1-score, confusion matrix,* pada data uji?

1.3 Tujuan Penelitian

1. Untuk membangun model klasifikasi biner yang dapat mengidentifikasi *self-harm* pada meme secara multimodal.
2. Untuk menyusun dataset meme multimodal melalui pengumpulan manual, pembuatan data tambahan, dan proses *pseudo-labeling* berbasis LLM untuk pelatihan model klasifikasi.
3. Untuk mengevaluasi kinerja model klasifikasi menggunakan metrik *accuracy, precision, recall, F1-score, confusion matrix,* pada data uji.

1.4 Batasan Masalah

Berikut adalah batasan masalah dalam penelitian ini:

1. Penelitian ini hanya menggunakan meme yang terdiri dari gambar dan teks, dengan fokus teks dalam bahasa Inggris. Pemilihan bahasa Inggris sebagai fokus teks didasarkan pada pada permasalahan yang diangkat

secara global, serta kemudahan akses terhadap model pra-latih yang telah dilatih dengan teks bahasa Inggris sebelumnya. Konten dalam bentuk video, audio, GIF, atau format multimodal lainnya tidak termasuk dalam ruang lingkup penelitian.

2. Penelitian ini dibatasi pada skema klasifikasi biner, yaitu kategori *self-harm* dan *non-self-harm*. Penelitian tidak mencakup klasifikasi multi-kelas, kategorisasi tingkat keparahan, atau analisis mendalam terkait karakteristik konten *self-harm*.
3. Dataset disusun melalui pengumpulan manual, pembuatan data tambahan yang dilabeli menggunakan metode *pseudo-labeling* berbasis *Large Language Model* (LLM) tanpa validasi oleh manusia, sehingga menghasilkan *weak labels* dan model yang bersifat *weakly supervised*. Selain itu, dataset juga diperoleh dari sumber terbuka di Kaggle sebagai data tambahan untuk memperluas variasi dan jumlah dataset yang digunakan dalam penelitian ini.

1.5 Manfaat Penelitian

Berikut adalah manfaat dari penelitian ini:

1. Menyediakan model klasifikasi khusus untuk meme dengan tema *self-harm* menggunakan multimodal untuk menganalisis gambar dan teks secara kontekstual.
2. Menjadi referensi untuk pengembangan model klasifikasi serupa yang dapat diterapkan pada jenis meme berisiko lainnya.
3. Mengisi kesenjangan riset terkait kurangnya dataset publik yang berfokus pada meme *self-harm,* serta memberikan kontribusi terhadap pembuatan dataset multimodal yang relevan untuk penelitian selanjutnya.
4. Menyediakan dataset meme *self-harm* dengan pelabelan *pseudo-labeling* yang dapat menjadi referensi bagi peneliti lain dalam mengembangkan penelitian lebih lanjut.

1.6 Sistematika Penulisan

Sistematika penulisan berisi pembahasan apa yang akan ditulis disetiap Bab. Sistematika pada umumnya berupa paragraf yang setiap paragraf mencerminkan bahasan setiap Bab.

1.6.1 Bab I: Pendahuluan

Bab ini memuat penjelasan mengenai latar belakang yang menjadi dasar penelitian ini. Latar belakang tersebut menjelaskan tentang konteks permasalahan yang dihadapi dan urgensi penelitian yang dilakukan. Selanjutnya, disajikan rumusan masalah yang merupakan identifikasi masalah yang ingin diselesaikan melalui penelitian ini. Peneliti juga menguraikan tujuan dari penelitian, batasan ruang lingkup penelitian agar fokus, serta manfaat yang diharapkan dari penelitian ini baik untuk pengembangan ilmu pengetahuan maupun untuk praktik di lapangan. Pada bagian ini juga dijelaskan sistematika penulisan yang memuat susunan dan isi tiap bab yang ada dalam tugas akhir ini.

1.6.2 Bab II: Tinjauan Pustaka

Bab ini menyajikan kajian teori yang relevan dengan topik penelitian. Tinjauan pustaka mencakup penelitian terdahulu yang memiliki hubungan dengan penelitian ini, serta teori-teori yang mendasari pokok bahasan penelitian. Dalam bagian ini, peneliti akan mengkaji berbagai literatur yang dapat memberikan pemahaman lebih mendalam terkait topik yang diteliti.

1.6.3 Bab III: Metodologi Penelitian

Bab ini berisi penjelasan rinci mengenai metode yang digunakan dalam penelitian ini. Selanjutnya, dijelaskan pula rancangan pengujian yang akan dilakukan untuk mengukur dan menganalisis hasil penelitian.

1.6.4 Bab IV: Hasil dan Pembahasan

Bab ini memaparkan hasil yang diperoleh dari implementasi dan pengujian yang telah dilakukan dalam penelitian. Peneliti juga menganalisis dan mengevaluasi hasil yang diperoleh untuk memberikan gambaran yang jelas tentang pencapaian tujuan penelitian.

1.6.5 Bab V: Kesimpulan dan Saran

Bab ini memaparkan kesimpulan yang diambil berdasarkan hasil analisis dan pembahasan yang telah dilakukan. Peneliti merangkum temuan utama dari penelitian ini dan memberikan saran untuk penelitian lebih lanjut. Saran tersebut bisa mencakup perbaikan atau pengembangan lebih lanjut dari penelitian yang dilakukan, serta rekomendasi yang dapat diambil untuk implementasi di masa depan.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Tinjauan pustaka ini memaparkan penelitian terdahulu yang memiliki relevansi topik dan menjadi landasan utama penyusunan tugas akhir. Adapun penelitian yang dijadikan acuan diuraikan sebagai berikut.

Gilardi et al. pada tahun 2023 membandingkan kinerja *Large Language Models* (LLM) dengan pelabel manusia (*crowd-workers*) dalam tugas anotasi teks untuk penelitian *social computing*. Melalui percobaan pada empat dataset berisi berita dan tweet dengan berbagai skema pelabelan, penelitian tersebut menunjukkan bahwa ChatGPT mencapai akurasi sekitar 25% lebih tinggi dibandingkan manusia dalam skenario *zero-shot*. Selain itu, biaya pelabelan menggunakan ChatGPT hanya kurang dari $0.003 per label. Temuan ini mengindikasikan bahwa LLM mampu menghasilkan label yang konsisten, cepat, dan valid sebagai alternatif tenaga anotasi manusia untuk mengatasi tantangan biaya tinggi dan inkonsistensi dalam proses pelabelan manual [14]. Namun demikian, studi ini hanya berfokus pada data berbasis teks.

Lian et al. pada tahun 2023 menguji kemampuan GPT-4V dalam tugas pengenalan emosi yang disebut *Generalized Emotion Recognition* (GER), yang mencakup berbagai modalitas input seperti gambar tunggal, video, dan kombinasi teks-gambar. Studi ini menggunakan 21 dataset *benchmark* yang beragam untuk mengevaluasi apakah model mampu memahami sentimen visual, ekspresi wajah, hingga emosi dalam konteks multimodal. Hasil pengujian menunjukkan bahwa GPT-4V mampu menggabungkan informasi dari berbagai modalitas secara efektif untuk mengenali konteks emosi dengan akurat, bahkan tanpa memerlukan pelatihan tambahan atau *fine-tuning*. Penelitian ini memberikan bukti bahwa *Large Multimodal Models* (LMM) dapat digunakan sebagai *zero-*

shot automatic annotators untuk berbagai tugas multimodal, sehingga berpotensi mengurangi ketergantungan pada pelabelan manual yang mahal dan memakan waktu [15]. Namun, model ini masih kurang optimal dalam mendeteksi ekspresi mikro (*micro-expressions*) yang memerlukan detail visual yang sangat halus dan spesifik, serta memiliki keterbatasan dalam menangani gambar dengan kuat.

Pramanick et al. pada tahun 2021 memperkenalkan MOMENTA (*Multimodal Framework for Detecting Harmful Memes*), sebuah *framework* multimodal yang menggabungkan representasi global dan lokal untuk menganalisis meme berbahaya dengan konteks visual dan teks. Fitur visual diperoleh dari *Contrastive Language–Image Pretraining* (CLIP), deteksi objek dan *fine-grained feature* melalui Google Vision API dan VGG-19, sedangkan fitur teks menggunakan DistilBERT. Arsitektur MOMENTA menonjol karena penggabungan fitur pada tahap menengah (*intermediate fusion*) melalui *Cross-Modality Attention Fusion* (CMAF), yang memodelkan interaksi antar-modalitas secara mendalam sebelum tahap klasifikasi. Evaluasi pada dua dataset menunjukkan MOMENTA mengungguli sepuluh baseline dengan peningkatan akurasi absolut 1.3–2.6 poin untuk kedua tugas [16]. Namun, model ini memiliki kompleksitas arsitektur tinggi, serta ketergantungan pada layanan eksternal (*Application Programming Interface* pihak ketiga) yang dapat memengaruhi efisiensi pemrosesan dan biaya operasional.

Kumar dan Nandakumarpada tahun 2022 mengusulkan *Hate-CLIPper*, sebuah arsitektur yang dirancang untuk mengoptimalkan klasifikasi meme kebencian melalui interaksi fitur multimodal yang dihasilkan oleh model CLIP *pre-trained*. Inovasi utama penelitian ini terletak pada penerapan *Feature Interaction Matrix* (FIM) berbasis *outer-product*, yang berfungsi memetakan hubungan semantik lintas-modalitas melalui operasi perkalian *tensor* antara representasi visual dan tekstual. FIM menghasilkan matriks berukuran $d_v \times d_t$ (dimensi visual $\times$ dimensi tekstual) yang mampu menangkap pola interaksi halus antara elemen-elemen semantik dari kedua modalitas. Berdasarkan pengujian

pada beberapa dataset benchmark, model ini mencatat nilai AUROC sebesar 85.8. Penelitian ini memberikan kontribusi dengan menunjukkan bahwa penyelarasan fitur visual-tekstual melalui *interaction modeling* dapat meningkatkan akurasi klasifikasi secara signifikan, terutama untuk kasus meme yang mengandung ironi atau konten tersirat [17]. Namun, pendekatan ini memiliki keterbatasan yaitu tingginya komputasi akibat dimensi matriks FIM yang besar dan waktu inferensi yang relatif lambat (2.5 detik per meme).

Shah et al. pada tahun 2024 mengusulkan MemeCLIP, sebuah kerangka kerja yang dirancang untuk mengatasi tantangan kompleksitas gambar dengan teks tersemat melalui pemahaman berbagai aspek ekspresi di dalamnya. Berbeda dengan penelitian sebelumnya yang lebih berfokus pada aspek tunggal seperti ujaran kebencian dan subkelasnya, penelitian ini memperluas cakupan ke beberapa aspek linguistik, yaitu kebencian, target kebencian, sikap, dan humor. Dalam penelitian ini, mereka juga memperkenalkan dataset baru bernama PrideMM yang terdiri dari 5.063 gambar dengan teks tersemat terkait gerakan Pride LGBTQ+, sehingga mengisi kekurangan sumber daya yang ada. Hasil eksperimen menunjukkan bahwa MemeCLIP mencapai kinerja lebih baik dibandingkan kerangka kerja sebelumnya pada dua dataset dunia nyata. Penulis juga membandingkan kinerja MemeCLIP dengan GPT-4 dalam tugas klasifikasi kebencian serta membahas kekurangan model ini melalui analisis kualitatif terhadap sampel yang salah diklasifikasikan [11]. Namun, penelitian ini juga mencatat bahwa model masih memiliki keterbatasan dalam mengenali simbol visual berukuran kecil (*fine-grained*) seperti logo atau ikon beresolusi rendah, serta memahami konteks budaya spesifik yang tidak tercakup dalam data pelatihan CLIP, terutama untuk meme berbahasa non-Inggris atau yang mengandung referensi lokal.

Tabel 2.1 Literasi Penelitian Terdahulu

No.	Judul	Masalah	Metode	Hasil
1.	*Can ChatGPT Reproduce Human Labels?* (Gilardi et al., 2023)	Keterbatasan data berlabel dan biaya anotasi manual tinggi	ChatGPT *zero-shot text annotation*	Akurasi 25% lebih tinggi dari pelabelan manusia dan biaya $0.003 per label
2.	*GPT-4V with Emotion: A Zero-shot Benchmark for Generalized Emotion Recognition* (Lian et al., 2023)	Keterbatasan anotasi manual untuk data multimodal	GPT-4V sebagai *zero-shot automatic annotator*	Mampu mengenali emosi multimodal dengan akurat pada 21 dataset *benchmark* tanpa memerlukan pelatihan tambahan atau *fine-tuning*
3.	MOMENTA: *A Multimodal Framework for Detecting Harmful Memes and Their Targets* (Pramanick et al., 2021)	Analisis meme berbahaya memerlukan pemahaman konteks multimodal	CLIP + Google Vision API + VGG-19 + DistilBERT + CMAF	Peningkatan akurasi 1.3–2.6 poin vs UNITER dan ViLBERT

No.	Judul	Masalah	Metode	Hasil
4.	*Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features* (Kumar & Nandakumar, 2022)	Keterbatasan ekstraksi interaksi semantik lintas-modalitas pada meme kebencian	CLIP dengan *Feature Interaction Matrix* (FIM)	AUROC 85.8 pada Hateful Memes Challenge, melampaui manusia (82.65)
5.	*MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification* (Shah et al., 2024)	kompleksitas gambar teks-embedded, yang membutuhkan pemahaman multimodal	MemeCLIP dengan modifikasi model CLIP + *projection layers* + *lightweight feature adapters*	F1 77.21% (humor), 75.09% (hate) pada PrideMM dan HarMeme

2.2 Dasar Teori

2.2.1 *Deep Learning*

Deep learning merupakan pendekatan dalam *machine learning* yang menggunakan jaringan saraf tiruan berlapis banyak *(deep neural networks)* untuk mempelajari pola dan representasi data secara otomatis melalui proses hierarkis. Berbeda dari metode pembelajaran tradisional *(traditional machine learning)* yang mengandalkan rekayasa fitur manual, *deep learning* memungkinkan

ekstraksi fitur kompleks secara langsung dari data mentah menggunakan transformasi non-linear yang dioptimalkan melalui algoritma *backpropagation* [18]. Konsep representasi berlapis inilah yang membuat *deep learning* unggul dalam menangani data dengan skala besar dengan struktur yang beragam seperti teks dan gambar karena model dapat mempelajari fitur tingkat rendah hingga abstraksi tingkat tinggi secara bertahap [19]. Dengan fleksibilitas dan kemampuan generalisasi yang kuat, *deep learning* telah menjadi fondasi utama pengembangan berbagai arsitektur modern dalam bidang *computer vision* dan pemrosesan bahasa alami.

2.2.2 *Computer Vision*

Computer vision merupakan bidang dalam kecerdasan buatan yang berfokus pada bagaimana mesin dapat memahami dan menafsirkan informasi visual dari gambar maupun video melalui representasi numerik yang terstruktur [20]. Proses komputasional ini mencakup tahapan seperti ekstraksi fitur, deteksi pola, dan pemahaman objek melalui algoritma yang mempelajari hubungan spasial maupun semantik di dalam citra. Pendekatan modern dalam *computer vision* didominasi oleh *deep learning*, khususnya *Convolutional Neural Networks* (CNN), yang mampu mempelajari fitur visual secara hierarkis mulai dari tepi (*low-level features*) hingga representasi abstrak (*high-level features*) melalui pembelajaran berbasis data [18]. Perkembangan lebih lanjut melahirkan *Vision Transformer* (ViT), arsitektur berbasis mekanisme *self-attention* yang memecah gambar dalam bentuk *patch* dan memprosesnya seperti token dalam NLP, sehingga menghasilkan performa kompetitif pada skala data besar [21]. Dengan kemampuan menghasilkan *embedding* visual yang kaya semantik, *computer vision* menjadi fondasi penting dalam berbagai aplikasi, termasuk klasifikasi gambar, deteksi objek, dan pemahaman multimodal seperti pada model *vision-language*.

2.2.3 *Natural Language Processing* (NLP)

NLP merupakan cabang kecerdasan buatan yang berfokus pada bagaimana mesin dapat memahami, memproses, dan menghasilkan bahasa manusia melalui representasi matematis yang terstruktur [22]. NLP modern bertumpu pada pembelajaran representasi *(representation learning)*, dimana teks diubah menjadi *embedding* yang mampu menangkap konteks semantik maupun sintaksis. Pergeseran besar dalam NLP terjadi dengan hadirnya pendekatan berbasis *deep learning*, yang dimulai dengan penggunaan *recurrent neural networks* (RNN) dan *Long Short-Term Memory* (LSTM), sebelum akhirnya banyak digantikan oleh arsitektur Transformer yang memungkinkan pemodelan konteks panjang melalui mekanisme *self-attention* [23]. Transformer kemudian menjadi fondasi model-model pra-latih (*pre-trained models*) seperti BERT, GPT, dan ELECTRA yang memanfaatkan *pre-training* berskala besar untuk mempelajari pola bahasa secara mendalam, sehingga meningkatkan performa berbagai tugas seperti klasifikasi teks, analisis sentimen, dan pemahaman multimodal [24]. Dengan kemampuan menghasilkan representasi linguistik yang kaya konteks, NLP menjadi komponen penting dalam pemrosesan modalitas teks pada sistem multimodal modern.

2.2.4 Transformer

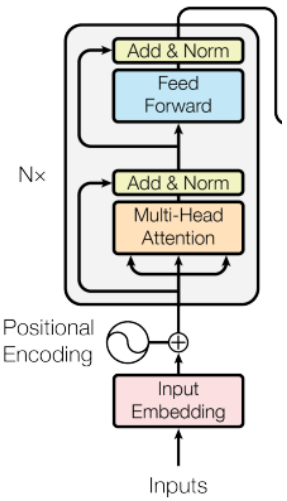
Transformer adalah arsitektur model deep learning yang diperkenalkan oleh Vaswani et al. pada tahun 2017 dan menjadi dasar bagi berbagai model bahasa dan multimodal modern. Keunggulan utama dari Transformer terletak pada mekanisme *self-attention*, yang memungkinkan model untuk menilai hubungan antar-token dalam sebuah urutan tanpa bergantung pada struktur sekuensial.

Salah satu komponen penting dalam Transformer adalah *multi-head attention*, yang memungkinkan model untuk memproses informasi dari berbagai subruang representasi secara paralel. Setiap token direpresentasikan sebagai

vektor, kemudian dihitung nilai *attention* berdasarkan interaksi antara *query (Q)*, *key (K)*, dan *value (V)*. Proses ini menggunakan rumus *Scaled Dot-Product Attention* yang dihitung sebagai berikut:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Rumus 2.1)}$$

Dimana d_k adalah dimensi dari *key* yang digunakan untuk menstabilkan nilai *dot-product*. Mekanisme ini memungkinkan model untuk menangkap dependensi jangka panjang secara efisien. Selain itu, *multi-head attention* memungkinkan model untuk memproses informasi dari berbagai subruang representasi, yang mendukung pengolahan informasi dalam skala besar dan memungkinkan penanganan konteks yang lebih kompleks.



Gambar 2.1 Arsitektur Transformer dengan Encoder [25]

Gambar 2.1 menunjukkan arsitektur Transformer dengan encoder yang fokus pada bagian *input embedding*. Proses dimulai dengan input yang berupa urutan token, yang kemudian diproses menjadi *input embedding*. *Input*

embedding mengubah token-token dalam urutan menjadi vektor representasi yang dapat diproses oleh model. Vektor-vektor ini kemudian diperkaya dengan *positional encoding* untuk mempertahankan informasi urutan input, yang sangat penting dalam model Transformer yang tidak bergantung pada urutan sekuensial. Gambar ini menggambarkan bagaimana *input embedding* berfungsi sebagai representasi awal bagi setiap token yang masuk, dan bagaimana *embedding* ini bekerja bersama dengan komponen lain seperti *multi-head attention* untuk menghasilkan representasi yang lebih kompleks dalam proses pemodelan [25].

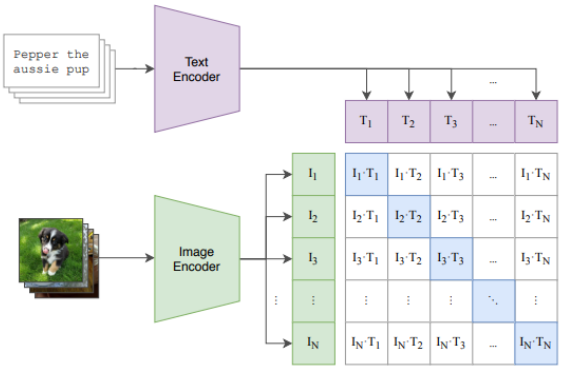
2.2.5 *Contrastive Language-Image Pre-training* (CLIP)

CLIP (*Contrastive Language–Image Pre-training*) adalah model vision–language yang dikembangkan oleh OpenAI untuk mempelajari keterkaitan semantik antara gambar dan teks melalui pembelajaran kontrasif berskala besar. CLIP terdiri atas dua *encoder* terpisah, yaitu *image encoder* berbasis Vision Transformer (ViT), serta *text encoder* berbasis Transformer yang keduanya memetakan gambar dan teks ke dalam ruang *embedding* yang sama. Selama pelatihan, CLIP memanfaatkan jutaan pasangan gambar–teks dari internet untuk mempelajari *alignment* antar-modalitas dengan menempatkan pasangan yang cocok semakin dekat dan pasangan tidak cocok semakin jauh dalam ruang vektor. Mekanisme pembelajaran ini diformalkan menggunakan *contrastive loss* (InfoNCE), yang dirumuskan sebagai:

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\cos(f_I(L_i), f_T(T_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(f_I(L_i), f_T(T_j))/\tau)} + \log\frac{\exp(\cos(f_T(T_i), f_I(L_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(f_T(T_i), f_I(L_j))/\tau)}\right] \quad \text{(Rumus 2.2)}$$

Di mana f_I dan f_T adalah encoder gambar dan teks, cos adalah fungsi kesamaan kosinus, dan τ adalah parameter temperatur. Melalui pendekatan ini, CLIP mampu menghasilkan representasi multimodal yang kaya konteks dan sangat efektif untuk berbagai tugas *downstream* tanpa perlu *fine-tuning*,

termasuk klasifikasi multimodal, analisis meme, serta deteksi konten berisiko.



Gambar 2.2 Arsitektur CLIP [26]

Gambar 2.2 di atas mengilustrasikan arsitektur model CLIP yang menerapkan mekanisme dual-encoder untuk mempelajari representasi visual dan tekstual secara bersamaan. Dalam proses ini, *text encoder* dan *image encoder* bekerja secara paralel mengubah sekumpulan input gambar dan teks menjadi vektor fitur *(embedding)* dalam ruang dimensi yang sama. Setiap pasangan gambar-teks yang sesuai dipetakan ke vektor yang berdekatan, sementara pasangan yang tidak sesuai dipisahkan lebih jauh, melalui optimasi fungsi InfoNCE yang memaksimalkan kesamaan kosinus antar *embedding* [26]

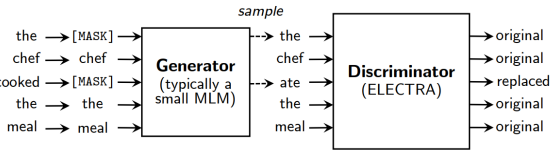
2.2.6 *Efficiently Learning an Encoder that Classifies Token Replacements Accurately* (ELECTRA)

ELECTRA adalah model pra-latih berbasis Transformer yang diperkenalkan sebagai alternatif yang lebih efisien daripada BERT melalui mekanisme *Replaced Token Detection* (RTD), yaitu tugas diskriminatif di mana model belajar membedakan token asli dan token yang telah diganti oleh model *generator* kecil. Tidak seperti pendekatan *Masked Language Modeling* (MLM) pada BERT yang hanya memperbarui representasi untuk token yang di *masking,*

ELECTRA memperbarui seluruh token dalam urutan, sehingga menghasilkan pembelajaran yang lebih stabil dan *sample-efficient*. Arsitektur ELECTRA terdiri dari dua komponen, yaitu *generator* yang memprediksi token pengganti menggunakan MLM kecil, dan *discriminator* yang menilai apakah setiap token adalah asli atau hasil substitusi. Secara formal, RTD meminimalkan fungsi *loss* biner berikut:

$$\mathcal{L}_{\text{RTD}} = -\sum_{i=1}^{n} \left[y_i \log D(x_i) + (1 - y_i) \log \left(1 - D(x_i)\right) \right] \quad \text{(Rumus 2.3)}$$

dengan $y_i = 1$ jika token asli dan $y_i = 0$ jika token hasil generator, serta $D(x_i)$ adalah probabilitas bahwa token tersebut asli. Pendekatan ini menghasilkan representasi linguistik yang lebih informatif dengan biaya komputasi lebih rendah dibanding model sejenis, menjadikan ELECTRA sangat efektif untuk tugas klasifikasi teks dan integrasi dalam sistem multimodal.



Gambar 2.3 Arsitektur ELECTRA [27]

Gambar 2.3 di atas mengilustrasikan arsitektur model ELECTRA yang terdiri dari *generator* dan *discriminator*. Proses pelatihan dengan *generator* dan *discriminator* ini diformalkan melalui fungsi *Replaced Token Detection* (RTD) *loss*, yang memaksimalkan kemampuan *discriminator* dalam mengidentifikasi token asli berdasarkan konteks sekitarnya [27]. Pendekatan ini memungkinkan ELECTRA untuk menghasilkan representasi linguistik yang lebih kaya dan efisien dibandingkan metode pra-latih tradisional seperti BERT.

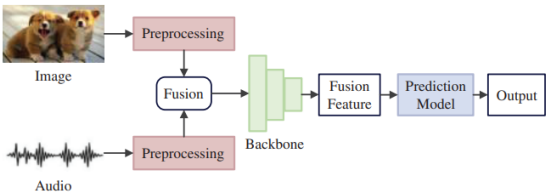
2.2.7 *Multimodal Deep learning*

Multimodal *deep learning* adalah pendekatan pembelajaran mesin yang memanfaatkan *deep neural networks* untuk mengintegrasikan berbagai modalitas data, seperti teks, gambar, audio, atau video, sehingga model mampu memahami informasi secara lebih kontekstual [28]. Modalitas merujuk pada jenis data yang berbeda, gambar membawa informasi visual spasial, teks membawa informasi semantik dan linguistik, sedangkan audio atau video menangkap informasi temporal [29]. Dalam multimodal *deep learning*, setiap modalitas diproses oleh *encoder* neural network tersendiri untuk menghasilkan representasi (*embedding*) yang memodelkan karakteristik modalitas tersebut secara mendalam. *Embedding* dari berbagai modalitas kemudian digabung melalui mekanisme *fusion* agar model dapat menangkap hubungan lintas-modal (*cross-modal relationships*) yang kompleks, sehingga memungkinkan penggabungan keunggulan tiap modalitas dan mengatasi keterbatasan ketika hanya menggunakan satu modalitas. Pendekatan ini pertama kali distandardisasi oleh Ngiam et al. pada tahun, yang menunjukkan bahwa jaringan neural yang dilatih dengan data multimodal mampu mempelajari representasi bersama yang lebih kuat dibandingkan representasi unimodal [28].

2.2.8 Klasifikasi *Fusion* dalam *Multimodal Deep learning*

Salah satu aspek kunci dari *multimodal deep learning* adalah bagaimana modalitas berbeda digabung, yaitu melalui proses *fusion*. Berdasarkan kapan dan bagaimana penggabungan dilakukan, teknik *fusion* secara umum dapat diklasifikasikan menjadi empat kategori utama, yaitu *early fusion*, *intermediate fusion*, *late fusion*, dan *hybrid fusion*. Berikut penjelasannya:

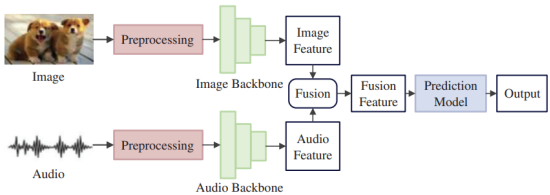
2.2.8.1 *Early Fusion*



Gambar 2.4 Contoh Early Fusion [30]

Pada pendekatan ini, modalitas digabungkan pada tingkat input sebelum proses *encoding*. Contohnya pada Gambar 2.4, data gambar dan audio digabung menjadi satu representasi gabungan yang kemudian diberikan ke neural network. Early fusion memungkinkan model mempelajari interaksi antar-modalitas dari tahap awal, tetapi menghadapi tantangan signifikan karena heterogenitas struktur data (misalnya perbedaan dimensi, skala, dan format) yang dapat menyulitkan integrasi langsung [31].

2.2.8.2 *Intermediate Fusion*

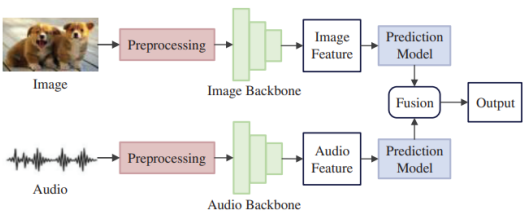


Gambar 2.5 Contoh Intermediate Fusion [30]

Pada Gambar 2.5, setiap modalitas terlebih dahulu diproses oleh *encoder* masing-masing untuk menghasilkan *embedding* atau fitur spesifik modalitas. Setelah itu, embedding dari berbagai modalitas digabung misalnya melalui *concatenation*, *projection*, *pooling*, atau mekanisme *fusion* lanjutan sebelum diteruskan ke lapisan prediksi. Pendekatan ini menawarkan keseimbangan

antara kemampuan merepresentasikan keunikan tiap modalitas dan mempelajari hubungan lintas-modalitas, sehingga banyak digunakan pada model modern, termasuk sistem vision language [30].

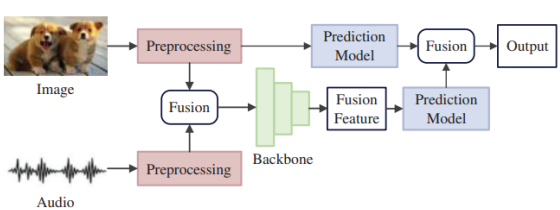
2.2.8.3 *Late Fusion*



Gambar 2.6 Contoh Late Fusion [30]

Pada Gambar 2.6 setiap modalitas diproses secara independen hingga menghasilkan prediksi masing-masing, seperti probabilitas atau skor klasifikasi. Prediksi tersebut kemudian digabung menggunakan metode seperti *weighted sum*, *voting*, atau *averaging* untuk memperoleh prediksi akhir. Late fusion efektif ketika modalitas relatif independen dan prediksi masing-masing sudah cukup kuat, namun kurang mampu menangkap interaksi semantik yang mendalam antar-modalitas [30].

2.2.8.4 *Hybrid Fusion*



Gambar 2.7 Contoh Hybrid Fusion [30]

Hybrid fusion menggabungkan dua atau lebih strategi fusion, misalnya mengombinasikan *feature-level fusion* dan *decision-level fusion*, atau menggabungkan embedding dengan mekanisme *attention* sekaligus memanfaatkan penggabungan skor prediksi. Pada Gambar 2.7, hybrid fusion banyak digunakan pada aplikasi multimodal yang kompleks karena mampu memanfaatkan kelebihan beragam strategi sekaligus mengatasi keterbatasan masing-masing [30].

Dalam praktiknya, pemilihan teknik *fusion* sangat bergantung pada karakteristik data, tujuan aplikasi, serta kompleksitas model yang diinginkan. Pendekatan *intermediate fusion* sering kali menjadi pilihan populer karena fleksibilitas dan kemampuannya dalam menangkap hubungan lintas-modalitas secara efektif.

2.2.9 Meme

Meme adalah bentuk pesan digital yang menyebar secara cepat melalui internet dan media sosial, dan biasanya terdiri atas kombinasi elemen visual (gambar) serta teks yang membentuk makna tertentu dalam konteks budaya. Secara etimologis, istilah *meme* pertama kali diperkenalkan oleh Richard Dawkins (1976) sebagai "unit budaya yang menyebar melalui imitasi," namun dalam konteks modern, meme berkembang menjadi fenomena visual linguistik yang merepresentasikan humor, kritik sosial, opini, atau emosi melalui struktur multimodal yang ringkas. Penelitian dalam bidang komunikasi digital menunjukkan bahwa makna meme tidak hanya berasal dari gambar atau teks secara terpisah, tetapi dari hubungan interaktif antara keduanya, termasuk elemen seperti ironi, metafora visual, sarkasme, dan referensi budaya (*cultural references*) yang bersifat kontekstual [32]. Kompleksitas ini menjadikan meme sebagai bentuk komunikasi multimodal yang sulit dipahami oleh model berbasis unimodal, karena penafsirannya sering bergantung pada kemampuan untuk menangkap keterkaitan antara modalitas visual dan linguistik. Oleh sebab itu,

dalam deteksi meme berbahaya seperti *self-harm* atau *hate speech* diperlukan pendekatan *multimodal deep learning* yang mampu memahami hubungan lintas-modal (*cross-modal semantics*) antara teks dan gambar.

2.2.10 *Self-Harm*

Self-harm atau perilaku menyakiti diri sendiri merupakan tindakan yang dilakukan seseorang untuk melukai tubuhnya secara sengaja, baik dengan maupun tanpa niat untuk mengakhiri hidup. Organisasi Kesehatan Dunia (WHO) mendefinisikan *self-harm* sebagai *"intentional self-inflicted injury, with or without suicidal intent"* [33]. Dalam ranah psikiatri, *Diagnostic and Statistical Manual of Mental Disorders* edisi kelima (DSM-5) membedakan antara *non-suicidal self-injury* (NSSI), yaitu perilaku menyakiti diri tanpa niat bunuh diri yang setidaknya terjadi selama lima hari dalam setahun, dengan *suicidal behavior* yang berorientasi pada keinginan mengakhiri hidup [34]. Self-harm banyak dikaitkan dengan kondisi psikologis seperti depresi, kecemasan, dan kesulitan regulasi emosi.

Perkembangan media sosial memperluas cara individu mengekspresikan pikiran atau perasaan terkait self-harm. Penelitian menunjukkan bahwa konten *self-harm* di platform digital seperti Instagram, TikTok, dan Reddit sering muncul dalam bentuk teks, gambar, atau kombinasi keduanya, dan dapat menjadi indikator risiko ideasi bunuh diri maupun NSSI pada remaja dan dewasa muda [9]. Paparan konten tersebut dapat meningkatkan risiko perilaku, normalisasi self-harm, serta perburukan kondisi mental pengguna yang rentan. Bentuk ekspresi self-harm di media sosial sering kali bersifat implisit melalui humor gelap, metafora visual, atau ungkapan ironi, sehingga sulit diidentifikasi oleh sistem deteksi sederhana [12]. Meme self-harm biasanya memadukan elemen visual dan teks secara multimodal sehingga maknanya muncul dari hubungan antara kedua modalitas tersebut, bukan dari salah satu modalitas secara terpisah. Penelitian menunjukkan bahwa meme bertema self-harm sering mengandung

humor gelap, sarkasme, atau representasi simbolik dari rasa sakit emosional, yang membuatnya menantang untuk ditangkap oleh model berbasis unimodal [10]. Oleh karena itu, pendekatan *multimodal deep learning* diperlukan untuk memahami interaksi visual-linguistik pada meme dan mengidentifikasi konten *self-harm* secara lebih akurat.

2.2.11 Prapemrosesan Data

Prapemrosesan data adalah tahap penting dalam pipeline analisis data yang bertujuan untuk menyiapkan data mentah agar siap digunakan dalam model. Prapemrosesan data terdiri dari berbagai teknik yang masing-masing dirancang untuk mempersiapkan data dalam bentuk yang sesuai untuk analisis lebih lanjut. Berikut adalah teknik-teknik utama dalam prapemrosesan data.

2.2.11.1 *Resize* Gambar

Resize gambar adalah proses mengubah ukuran gambar untuk memenuhi kebutuhan model yang digunakan. Secara matematis, proses ini melibatkan perhitungan rasio antara dimensi gambar awal dan ukuran target. Jika ukuran gambar asli adalah (W, H) dan gambar yang diinginkan berukuran (W', H'), maka rasio skala untuk lebar dan tinggi adalah sebagai berikut:

$$\text{Scale Factor} = \left(\frac{W'}{W}\right)\left(\frac{H'}{H}\right) \tag{Rumus 2.4}$$

Resize gambar akan mengubah ukuran citra asli berdasarkan skala ini, yang dapat memperkecil atau memperbesar citra sesuai dengan ukuran yang diinginkan. Dengan memperkecil ukuran gambar, kita dapat mengurangi jumlah parameter yang perlu diproses, meningkatkan efisiensi, dan mengurangi waktu komputasi [35].

2.2.11.2 Normalisasi Gambar

CLIP (*Contrastive Language–Image Pretraining*) menerapkan normalisasi pada dua tahapan krusial dalam arsitekturnya. Pertama, normalisasi input

gambar dilakukan dengan menstandarisasi nilai piksel menggunakan *mean* (μ) dan standar deviasi (σ) spesifik agar distribusi data selaras dengan *backbone encoder* visual (seperti ResNet atau ViT). Secara matematis, transformasi ini dapat dituliskan sebagai:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

(Rumus 2.5)

dimana x adalah nilai piksel asli dan x_{norm} adalah input yang diteruskan ke jaringan.

Kedua, setelah citra dan teks diproses menjadi *embedding*, CLIP menerapkan *L2-normalization* pada vektor tersebut agar diproyeksikan ke dalam *unit hypersphere*. Jika v adalah vektor fitur keluaran dari *encoder*, maka normalisasi didefinisikan sebagai:

$$v_{norm} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{\sum_{i=1}^{d} v_i^2}} + h$$

(Rumus 2.6)

Normalisasi vektor pada persamaan Rumus 2.6 ini krusial karena menjadikan operasi *dot product* setara dengan *cosine similarity* dalam perhitungan *loss* kontrastif, sehingga model dapat mempelajari penyelarasan semantik antara teks dan gambar dengan lebih stabil dan konsisten [26].

2.2.11.3 *Flip*

Flip adalah teknik augmentasi gambar yang digunakan untuk memperbesar variasi data dengan membalik gambar secara horizontal atau vertikal. Ini membantu model menjadi lebih robust terhadap variasi dalam data, seperti orientasi objek [35].

2.2.11.4 *Rotation*

Rotasi merupakan teknik augmentasi yang dilakukan dengan memutar gambar ke kanan atau kiri pada suatu sumbu dengan sudut tertentu, biasanya berada pada rentang 1° hingga 359°. Tingkat keamanan *(safety)* dari augmentasi

ini sangat ditentukan oleh besar sudut rotasi, karena rotasi yang terlalu ekstrem dapat menyebabkan hilangnya informasi penting pada citra, seperti tepi objek atau struktur yang relevan [36].

2.2.11.5 *Color Jitter*

Color Jitter merupakan teknik augmentasi data fotometrik yang berfungsi memodifikasi atribut visual citra meliputi kecerahan, kontras, saturasi, dan rona secara stokastik tanpa mengubah struktur geometris objek di dalamnya. Penerapan teknik ini bertujuan untuk mensimulasikan variabilitas pencahayaan yang alami terjadi di dunia nyata, sehingga memaksa model untuk mempelajari fitur struktural yang invarian alih-alih bergantung pada bias warna atau intensitas piksel absolut. Mekanisme distorsi warna ini terbukti efektif dalam meningkatkan kemampuan generalisasi model serta memitigasi risiko *overfitting* akibat keterbatasan variasi visual pada data latih [36].

2.2.12 Evaluasi

Evaluasi merupakan tahap penting dalam menilai performa model klasifikasi, terutama pada tugas sensitif seperti deteksi konten self-harm dalam meme. Evaluasi umumnya dilakukan menggunakan *confusion matrix* yang menggambarkan hubungan antara prediksi model dan kondisi sebenarnya. Matriks ini terdiri dari empat komponen, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [37]. Struktur *confusion matrix* dapat dilihat pada Tabel 2.2.

Tabel 2.2 *Confusion Matrix*

Actual \ Predicted	***Positive***	***Negative***
Positive	TP	FN
Negative	FP	TN

Berdasarkan elemen-elemen tersebut, beberapa metrik evaluasi dapat

dihitung. *Accuracy* mengukur proporsi prediksi benar terhadap seluruh sampel, namun sering tidak cukup representatif ketika data tidak seimbang (*imbalanced*) seperti pada deteksi self-harm meme [38]. *Precision* mengukur tingkat ketepatan prediksi kelas positif, sedangkan *Recall* mengukur kemampuan model dalam menangkap seluruh sampel positif. Keduanya penting karena kesalahan *False Negative* (FN) pada deteksi konten self-harm berpotensi membahayakan pengguna. Untuk menyeimbangkan precision dan recall digunakan *F1-score*, yaitu rata-rata harmonik keduanya, yang efektif untuk dataset tidak seimbang [39].

Secara matematis, metrik-metrik evaluasi dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(Rumus 2.7)}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{(Rumus 2.8)}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{(Rumus 2.9)}$$

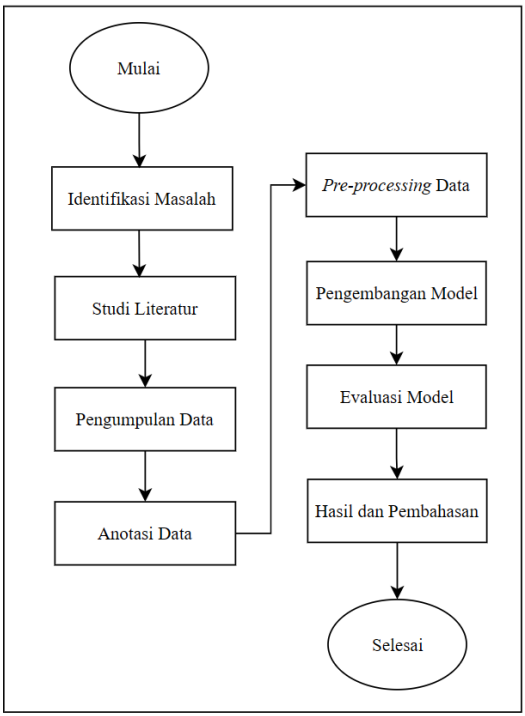
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{(Rumus 2.10)}$$

Kombinasi dari seluruh metrik tersebut memberikan gambaran evaluasi yang lebih komprehensif dibandingkan hanya menggunakan Accuracy saja, terutama karena model untuk deteksi self-harm harus sensitif terhadap kesalahan pada kelas positif. Evaluasi yang menyeluruh ini penting untuk memastikan model tidak hanya akurat secara keseluruhan, tetapi juga efektif dalam mengidentifikasi konten berisiko tinggi.

BAB III
METODE PENELITIAN

3.1 Alur Penelitian

Tahapan penelitian ini diawali dengan proses identifikasi masalah dan studi literatur untuk membangun landasan teori yang kuat. Selanjutnya, dilakukan pengumpulan data, anotasi data dan *pre-processing*. Setelah data siap, penelitian berfokus pada pengembangan model arsitektur multimodal, yang kemudian diuji melalui tahap evaluasi model. Rangkaian penelitian ini ditutup dengan analisis hasil dan pembahasan untuk menarik kesimpulan. Adapun ilustrasi lebih rinci mengenai alur penelitian ini dapat dilihat pada Gambar 3.1.



Gambar 3.1 Alur Penelitian

3.2 Penjabaran Langkah Penelitian

Pada alur penelitian yang telah digambarkan dalam *flowchart* pada subbab 3.1, penjelasan berikut menggambarkan secara rinci langkah-langkah penelitian yang akan dilakukan.

3.2.1 Identifikasi Masalah

Tahap pertama dalam penelitian ini adalah identifikasi masalah mellaui studi literatur, yang berfokus pada urgensi memahami meme dengan tema *self-harm* sebagai konten digital dengan makna yang tersirat. Meme jenis ini sering menggunakan ironi, sarkasme, humor gelap, atau metafora visual, sehingga indikasi *self-harm* tidak selalu tampak secara eksplisit dan hanya dapat dipahami jika teks dan gambar dianalisis secara kontekstual. Hingga saat ini, belum ditemukan penelitian yang secara spesifik mengembangkan model klasifikasi untuk meme *self-harm* , sehingga belum ada acuan metodologis yang benar-benar fokus pada konteks tersebut. Selain itu, belum tersedia dataset publik yang secara khusus menyajikan kumpulan meme *self-harm* , karena sebagian besar dataset multimodal yang ada lebih banyak berfokus pada kategori lain seperti ujaran kebencian atau propaganda [12]. Ketiadaan model klasifikasi yang khusus, terbatasnya dataset yang relevan, dan sifat makna meme yang implisit, membentuk kesenjangan riset yang jelas, yang menegaskan perlunya pengembangan model multimodal dalam penelitian ini.

3.2.2 Studi Literatur

Tahap kedua dalam penelitian ini adalah studi literatur, yang mencakup kajian mendalam terhadap berbagai sumber ilmiah yang relevan dengan topik penelitian. Studi literatur ini bertujuan untuk memahami konsep, metode, dan temuan terkini terkait pemanfaatan *Large Language Models* (LLM) untuk anotasi data, kemampuan model multimodal modern dalam mengenali emosi dan konten berisiko, serta perkembangan arsitektur yang digunakan untuk menganalisis

meme berbahaya. Pada tahap ini, ditinjau penelitian yang membahas penggunaan LLM sebagai *automatic annotator* untuk mengatasi keterbatasan anotasi manual, model multimodal untuk pengenalan emosi dan pemahaman konteks visual dan tekstual, serta berbagai kerangka kerja untuk deteksi *harmful, toxic,* dan *hateful memes* yang banyak memanfaatkan representasi CLIP. Hasil studi literatur tersebut menjadi landasan teoritis untuk merumuskan celah riset pada domain meme *self-harm* dan menyusun rancangan model multimodal yang akan dikembangkan dalam penelitian ini.

3.2.3 Pengumpulan Data

Tahap berikutnya adalah pengumpulan data, data diperoleh melalui dua pendekatan. Pertama, *scraping* manual dari berbagai platform seperti Reddit, X, Threads, dan Facebook dengan kata kunci dan tagar terkait *self-harm*, misalnya #killmyself, #depressed, #suicidal, dan #selfharm. Kedua, pembuatan *constructed dataset* berupa 500 meme yang disusun secara mandiri dengan menggabungkan teks dan gambar. Komponen gambar berasal dari dataset publik berlisensi terbuka seperti *Alcohol Bottle Images* [40] dan *Weapon and Knife Detection* [41] dari Kaggle, sedangkan elemen teks diambil dari *Suicidal Ideation Detection Reddit Dataset* [42] yang berisi lebih dari 15.000 unggahan berlabel *suicidal* dan *non-suicidal*. Tahap ini memastikan ketersediaan dataset yang beragam untuk mendukung model multimodal. Ketiga, penambahan dataset yang diambil dari Kaggle *6992 Meme Images Dataset with Labels* [43] sebagai data tambahan. Dataset ini digunakan secara spesifik untuk memperkaya variasi visual sekaligus menyeimbangkan jumlah sampel pada kelas *non-self-harm*, guna mencegah ketimpangan distribusi data terhadap kelas *self-harm*. Tahapan ini memastikan ketersediaan dataset yang beragam dan proporsional untuk mendukung pelatihan model multimodal.

3.2.4 Anotasi Data

Setelah data terkumpul, proses anotasi data dilakukan menggunakan pendekatan *pseudo-labeling*. Pada tahap ini, label awal diberikan secara otomatis oleh *ChatGPT-4o mini* dengan menggunakan teknik *zero-shot prompting*. Studi sebelumnya oleh Gilardi et al. menunjukkan bahwa teknik *zero-shot prompting* menggunakan *LLM* seperti *ChatGPT* dapat menghasilkan akurasi yang lebih tinggi dibandingkan dengan pelabel manusia dalam tugas anotasi teks, dengan biaya yang jauh lebih murah [14].Dengan metode ini, model diberi instruksi dalam bentuk prompt yang dirancang untuk mengarahkan model dalam mengidentifikasi indikasi *self-harm* yang terkandung dalam kombinasi teks dan gambar, tanpa memerlukan pelatihan atau *fine-tuning* khusus untuk dataset tersebut. Teknik *pseudo-labeling* memungkinkan pelabelan dalam skala besar tanpa memerlukan anotator manusia, yang menghasilkan *weak labels* yang tetap cukup representatif untuk pembelajaran awal model.

Proses pelabelan ini menggunakan *API ChatGPT*. Model ini diakses melalui API untuk memberikan label secara otomatis berdasarkan data input yang diberikan. Untuk memastikan kualitas label yang dihasilkan, dilakukan *random checking* untuk memverifikasi keakuratan hasil anotasi. Dalam hal ini, harga API ChatGPT ditentukan berdasarkan jumlah *token* yang digunakan. Rincian harga untuk model *GPT-4o mini* yang digunakan dalam penelitian ini $0.15 untuk input, dan $0.60 untuk output.

GPT-4o mini (“o” untuk “omni”) adalah model kecil yang cepat dan terjangkau, dirancang untuk tugas-tugas yang lebih fokus. Model ini menerima input teks dan gambar, serta menghasilkan output teks [44]. Model ini dipilih karena biaya yang lebih rendah dibandingkan dengan *GPT-4*, sehingga lebih efisien dalam hal pengeluaran operasional.

3.2.5 *Pre-processing Data*

Setelah anotasi selesai, seluruh data memasuki tahap *pre-processing* untuk memastikan kualitas input yang optimal. Pada modalitas gambar, *pre-processing* meliputi penyesuaian resolusi, normalisasi piksel, dan konversi format untuk menyesuaikan standar masukan CLIP. Pada modalitas teks, *pre-processing* mencakup pembersihan karakter khusus, normalisasi huruf, tokenisasi, serta penghilangan duplikasi. Tahap ini penting agar model dapat mempelajari pola secara konsisten.

3.2.5.1 *Resize* Gambar

Pada modalitas gambar, ukurannya diubah menjadi 224 x 224 piksel sesuai dengan input model CLIP menggunakan CLIP *processor*, yang secara otomatis menyesuaikan dimensi gambar. Proses ini memastikan bahwa ukuran gambar sesuai dengan yang dibutuhkan oleh model CLIP, yaitu 224 x 224 piksel. CLIP *processor* menggunakan metode untuk mengubah ukuran gambar tanpa merusak proporsinya, sehingga gambar tetap terlihat jelas meskipun ukurannya berubah.

3.2.5.2 Normalisasi Gambar

Normalisasi gambar adalah proses untuk menstandarkan nilai piksel agar berada dalam rentang yang konsisten sesuai dengan kebutuhan model. Normalisasi dilakukan melalui *CLIP processor* yang secara otomatis menangani tahap *preprocessing*. Proses ini bekerja dengan menyesuaikan nilai piksel menggunakan pengurangan rata-rata (*mean*) dan pembagian dengan deviasi standar (*std*) yang telah ditentukan sesuai dengan kondisi pelatihan model CLIP. Dengan demikian, setiap kanal warna *(Red, Green, Blue)* memiliki distribusi nilai yang konsisten sehingga model dapat memproses informasi visual secara lebih efektif.

3.2.5.3 *Flip*

Flipping horizontal diterapkan pada gambar untuk meningkatkan kemampuan model dalam mengenali objek dari berbagai orientasi. Pada tahap ini, gambar dibalik secara acak secara horizontal dengan probabilitas 50%. Teknik ini memungkinkan model untuk tidak bergantung pada orientasi tertentu, sehingga model tetap dapat mendeteksi objek meskipun posisinya berubah, seperti ketika objek menghadap ke kiri atau kanan. Contoh dari augmentasi ini dapat dilihat pada Gambar 3.2, yang menunjukkan perbandingan antara gambar asli dan hasil augmentasi.

(a) Gambar asli (b) Gambar setelah *flip horizontal*

Gambar 3.2 Perbandingan gambar asli dan hasil augmentasi *flip horizontal*

3.2.5.4 *Rotation*

Selain *flipping horizontal*, *rotasi acak* juga diterapkan pada gambar. Pada tahap ini, gambar diputar secara acak dengan rotasi sebesar 15 derajat. Dengan menambahkan rotasi acak, model dapat meningkatkan ketahanannya terhadap perubahan orientasi gambar. Contoh dari augmentasi ini dapat dilihat pada Gambar 3.3, yang menunjukkan perbandingan antara gambar asli dan hasil augmentasi rotasi.

(a) Gambar asli

(b) Gambar setelah *rotasi*

Gambar 3.3 Perbandingan gambar asli dan hasil augmentasi *rotasi acak*

3.2.5.5 *Color Jitter*

*Color jitter*juga diterapkan pada gambar untuk meningkatkan ketahanan model terhadap variasi pencahayaan dan warna. Pada tahap ini, kecerahan kontras, dan saturasi diubah secara acak dengan nilai 0.2 sedangkan rona diubah secara acak 0.1. Teknik ini bertujuan untuk memperkaya variasi gambar dalam dataset, sehingga model dapat belajar mengenali objek dengan lebih baik meskipun dalam kondisi pencahayaan yang berbeda atau variasi warna yang bervariasi. Contoh dari augmentasi ini dapat dilihat pada Gambar 3.4, yang menunjukkan perbandingan antara gambar asli dan hasil augmentasi *color jitter*.

(a) Gambar asli

(b) Gambar setelah *color jitter*

Gambar 3.4 Perbandingan gambar asli dan hasil augmentasi *color jitter*

3.2.5.6 *Lowercasing*

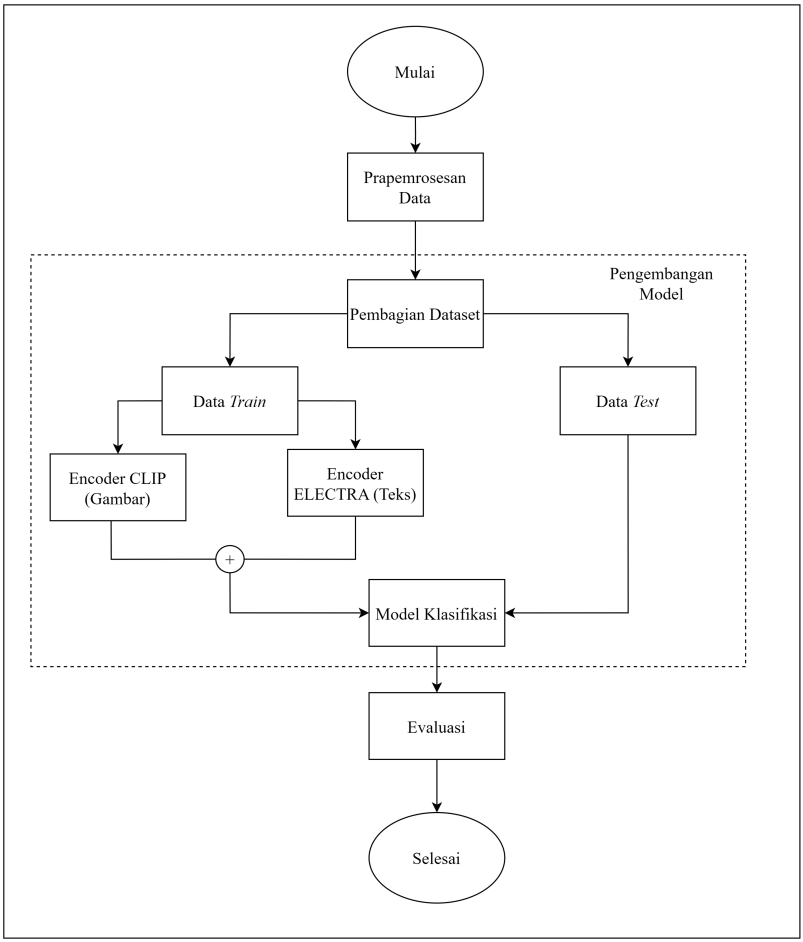
Pada modalitas teks, *lowercasing* dilakukan menggunakan ELECTRA *tokenizer* yang mengubah semua huruf kapital menjadi huruf kecil. Proses ini bertujuan untuk menyederhanakan representasi teks sehingga model tidak perlu membedakan antara huruf besar dan kecil, yang dapat mengurangi kompleksitas pemrosesan.

3.2.5.7 Tokenisasi

Pada modalitas teks, tokenisasi dilakukan menggunakan ELECTRA *tokenizer* yang memecah teks menjadi unit-unit yang lebih kecil, yaitu token. Token berupa sub kata. Proses tokenisasi ini penting agar teks dapat diproses oleh model ELECTRA, yang memerlukan input dalam bentuk token untuk memahami konteks dan makna dari teks tersebut.

3.2.6 Pengembangan Model Arsitektur Multimodal

Berikut adalah tahapan dalam pengembangan model arsitektur multimodal dengan menggunakan penggabungan fitur dalam tingkat menengah *intermediate fusion*



Gambar 3.5 Alur Pengembangan Model

3.2.6.1 Pembagian Data

Data yang telah melewati tahap *pre-processing* selanjutnya dibagi menjadi dua *subset* utama dengan rasio 80:20 secara random. Sebanyak 80% dari total data sebagai data latih (*training set*) untuk proses pembelajaran model, sedangkan 20% sisanya digunakan sebagai data uji (*testing set*). Pembagian ini bertujuan untuk memastikan bahwa model memiliki porsi data yang cukup besar untuk

mempelajari pola fitur multimodal secara optimal, sekaligus menyisakan data yang representatif untuk evaluasi kinerja yang objektif tanpa terjadi kebocoran data (*data leakage*).par

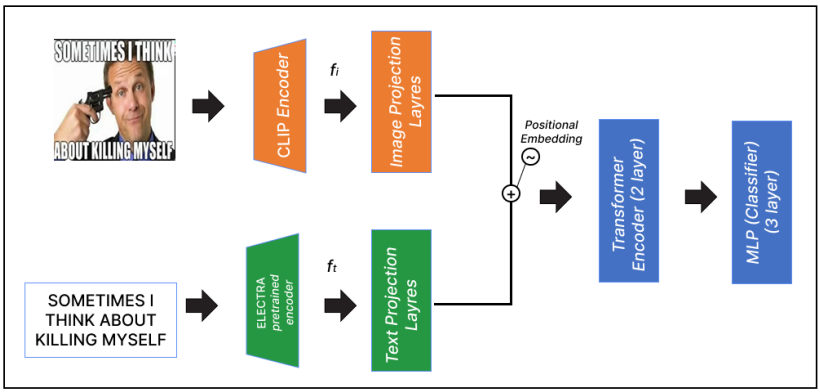
3.2.6.2 Pelatihan model

Tahap pelatihan model dilakukan dengan arsitektur *dual-encoder* yang menggabungkan dua modalitas input. Untuk ekstraksi fitur visual, penelitian ini menggunakan model CLIP (*Contrastive Language-Image Pre-training*) varian CLIP ViT *base patch-size* 32 dengan parameter 151 juta. Pemilihan CLIP didasarkan pada efektivitasnya yang telah terbukti dalam studi literatur terdahulu, seperti pada framework MOMENTA [16] dan Hate-CLIPper [17], dimana representasi visual CLIP terbukti krusial dalam menangkap hubungan semantik lintas-modalitas dan mengenali konten implisit tanpa memerlukan arsitektur yang terlalu kompleks. Pada penelitian ini, model CLIP digunakan murni sebagai ekstraktor fitur *(feature extractor)* tanpa dilakukan proses *fine-tuning,* sehingga seluruh parameter visual encoder tetap dibekukan *(frozen)* selama pelatihan.

Model *ELECTRA* varian *Suicidal-Electra*, adalah model yang telah dilatih untuk mendeteksi kecenderungan bunuh diri dalam teks. Model ini berbasis pada arsitektur transformer, yang unggul dalam memahami konteks kata-kata dalam suatu urutan kalimat. Model ini dilatih menggunakan dataset besar yang berisi postingan *suicidal* (bunuh diri) dan *non-suicidal* (tidak bunuh diri), yang memungkinkan model untuk mengenali pola bahasa, metafora, dan ekspresi yang terkait dengan pemikiran bunuh diri. Pada penelitian ini, model digunakan murni sebagai *feature extractor* tanpa dilakukan proses *fine-tuning*, sehingga seluruh parameter visual encoder tetap dibekukan (frozen) selama pelatihan. *Suicidal-Electra* memiliki *parameter* 110 juta [45].

3.2.6.3 *Fusion* model

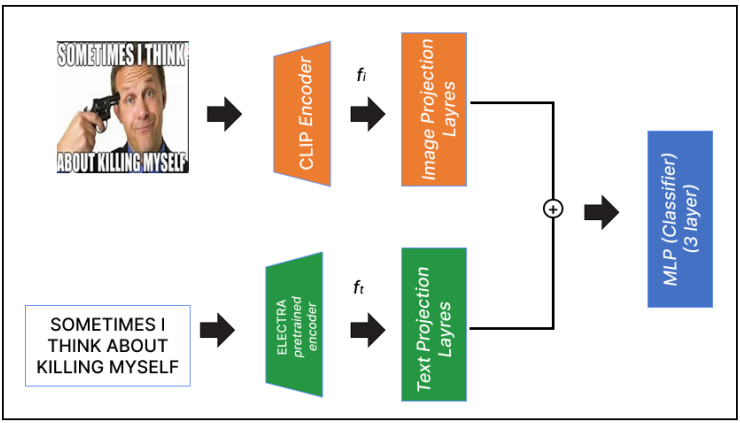
Dalam penelitian ini, digunakan dua arsitektur *model fusion* yang berbeda untuk menggabungkan informasi dari teks dan gambar. Kedua arsitektur ini dirancang untuk mendeteksi kecenderungan bunuh diri melalui analisis multimodal, yaitu teks dan gambar. Arsitektur yang dirancang sebagai berikut:



Gambar 3.6 Rancangan Fusion Model Varian 1

Gambar 3.6 menunjukkan rancangan arsitektur fusion model varian pertama. Model ini menggunakan Transformer Encoder untuk interaksi antara modalitas teks dan gambar. Proses pertama dimulai dengan pemrosesan CLIP *encoder* untuk mengekstrak fitur dari gambar, yang kemudian diproyeksikan ke dalam *Image Projection Layers*. Di sisi lain, teks diproses menggunakan *ELECTRA Pretrained Encoder* yang juga diproyeksikan ke dalam *Text Projection Layers*. Fitur dari teks dan gambar kemudian digabungkan dengan menggunakan *Positional Embedding*. Setelah itu, gabungan fitur dari kedua modalitas ini diproses melalui Transformer Encoder yang memungkinkan model untuk memahami hubungan dan konteks antara teks dan gambar. Hasil dari proses ini selanjutnya diteruskan ke *Multi Layer Perceptron* (MLP) untuk klasifikasi.

Gambar 3.7 menunjukkan rancangan arsitektur fusion model varian kedua. Berbeda dengan varian pertama, model ini tidak menggunakan Transformer



Gambar 3.7 Rancangan Fusion Model Varian 2

Encoder untuk interaksi antar-modalitas. Melainkan langsung menggabungkan fitur teks dan gambar untuk masuk ke MLP. Model ini lebih sederhana dibandingkan varian pertama.

3.2.7 Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan metrik standar yang diturunkan dari *Confusion Matrix*, yang meliputi *accuracy*, *precision*, *recall*, dan *F1-Score*. Mengingat urgensi dan sensitivitas identifikasi konten meme *self-harm*, evaluasi tidak hanya berfokus pada akurasi global, tetapi lebih evaluasi pada *recall* untuk memastikan sistem mampu meminimalkan risiko lolosnya konten berbahaya (*false negative*), serta *precision* untuk menjaga validitas prediksi positif. Sebagai metrik penentu, *F1-Score* digunakan untuk menyeimbangkan antara *precision* dan *recall*, sehingga memberikan gambaran objektif mengenai keandalan model dalam mengklasifikasikan fitur multimodal secara efektif pada dataset yang kompleks.

3.2.8 Analisis Hasil dan Pembahasan

Tahap akhir penelitian ini adalah analisis hasil dan pembahasan. Pada tahap ini, hasil evaluasi model dianalisis untuk melihat seberapa baik model dalam mengenali meme *self-harm*. Analisis meliputi penjelasan metrik evaluasi, kesalahan yang terjadi, serta kelebihan dan kekurangan model. Pembahasan juga mencakup manfaat hasil penelitian dan saran untuk pengembangan selanjutnya.

3.3 Alat dan Bahan Tugas Akhir

Berisi alat-alat dan bahan-bahan yang digunakan dalam penelitian.

3.3.1 *Software dan Library*

Software dan library yang digunakan untuk mendukung pelaksanaan penelitian ini dengan rincian sebagai berikut:

1. *Code Editor* Visual Studio Code untuk penulisan kode program.
2. *Library Deep Learning* `PyTorch` digunakan untuk implementasi dan pelatihan model.
3. Platform *Cloud Computing* `Runpod` untuk eksekusi kode Python dan pelatihan model memanfaatkan akselerasi GPU berbasis *cloud*, dengan perangkat keras GPU `NVIDIA RTX 4090`.
4. GitHub sebagai media penyimpanan repositori (*repository*) dan manajemen kode.
5. Platform desain `Canva` yang digunakan dalam proses pembuatan dataset meme secara manual (*constructed dataset*).

3.3.2 Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari tiga sumber dataset. Pertama, dataset meme yang dikumpulkan secara manual dari berbagai platform internet, yang mencakup kombinasi gambar dan teks. Kedua, dataset yang dibuat sendiri *(constructed dataset)* sebanyak 500 meme. Ketiga, dataset publik dari Kaggle berupa *"6992 Meme Images Dataset with Labels"* yang diintegrasikan

untuk memperkaya dan meningkatkan keragaman dataset penelitian [46].

3.4 Ilustrasi Perhitungan Metode

Pada bagian ini, akan dijelaskan ilustrasi perhitungan metode yang digunakan dalam penelitian ini.

3.4.1 Ilustrasi Perhitungan *contrastive loss* (InfoNCE)

Pada bagian ini, akan dihitung *loss* menggunakan Rumus 2.2. InfoNCE digunakan untuk menghitung kesamaan antara representasi gambar dan teks. Misalkan terdapat dua pasangan gambar-teks sebagai berikut:

- Pasangan 1: L_1 dengan T_1
- Pasangan 2: L_2 dengan T_2

Nilai kesamaan kosinus (*cosine similarity*) untuk setiap pasangan adalah:

$$\cos(f_L(L_1), f_T(T_1)) = 0.85$$

$$\cos(f_L(L_1), f_T(T_2)) = 0.45$$

$$\cos(f_L(L_2), f_T(T_1)) = 0.40$$

$$\cos(f_L(L_2), f_T(T_2)) = 0.75$$

Dengan parameter suhu $\tau = 0.07$, nilai *loss* untuk pasangan pertama L_1, T_1 dapat dihitung menggunakan rumus InfoNCE berikut:

$$\mathcal{L}_1 = -\frac{1}{2}\left(\log\left(\frac{e^{\cos(f_L(L_1), f_T(T_1))/\tau}}{\sum_{i=1}^{2} e^{\cos(f_L(L_1), f_T(T_i))/\tau}} \right) + \log\left(\frac{e^{\cos(f_T(T_1), f_L(L_1))/\tau}}{\sum_{i=1}^{2} e^{\cos(f_T(T_1), f_L(L_i))/\tau}} \right) \right)$$

Substitusi nilai *cosine similarity*:

$$\cos(f_L(L_1), f_T(T_1)) = 0.85, \quad \cos(f_L(L_1), f_T(T_2)) = 0.45$$

$$\cos(f_T(T_1), f_L(L_1)) = 0.85, \quad \cos(f_T(T_1), f_L(L_2)) = 0.40$$

Term pertama:

$$\frac{\exp(0.85/0.07)}{\exp(0.85/0.07) + \exp(0.45/0.07)} \approx \frac{1.751 \times 10^5}{1.751 \times 10^5 + 624.8}$$
$$\approx 0.9996$$

Term kedua:

$$\frac{\exp(0.85/0.07)}{\exp(0.85/0.07) + \exp(0.40/0.07)} \approx \frac{1.751 \times 10^5}{1.751 \times 10^5 + 302.5}$$
$$\approx 0.9998$$

Substitusi ke dalam rumus:

$$\mathcal{L}_1 = -\frac{1}{2}\Big[\log(0.9996) + \log(0.9998)\Big]$$

Logaritma:

$$\log(0.9996) \approx -0.0004, \quad \log(0.9998) \approx -0.0002$$

Hasil akhir:

$$\mathcal{L}_1 = -\frac{1}{2}\left[-0.0004 + (-0.0002)\right] = 0.0003$$

3.4.2 Ilustrasi Perhitungan *Token Detection Loss* pada ELECTRA

Pada bagian ini, akan dihitung *loss* menggunakan Rumus 2.3. Misalkan terdapat sebuah kalimat yang terdiri dari 3 token, yaitu:

$$\text{Kalimat: } \{\text{"}Saya\text{"}, \text{"}pergi\text{"}, \text{"}ke\text{"}\}$$

Kemudian, hasil generator model untuk token-token tersebut adalah:

$$\text{Hasil Generator: } \{\text{"}Saya\text{"}, \text{"}berlari\text{"}, \text{"}ke\text{"}\}$$

Dari kalimat di atas, diketahui bahwa:

- Token ”Saya” adalah token asli ($y_1 = 1$).
- Token ”pergi” digantikan dengan token ”berlari” (hasil generator), sehingga $y_2 = 0$.
- Token ”ke” adalah token asli ($y_3 = 1$).

Kemudian, probabilitas $D(x_i)$ yang diberikan oleh diskriminator untuk setiap token adalah sebagai berikut:

$$D(\text{"}Saya\text{"}) = 0.95, \quad D(\text{"}berlari\text{"}) = 0.05, \quad D(\text{"}ke\text{"}) = 0.90$$

Sekarang, dapat dihitung $\mathcal{L}_{RTD}$ untuk kalimat ini dengan menggunakan rumus yang telah diberikan.

$$\begin{aligned}
\mathcal{L}_{RTD} = &- [y_1 \log D(\text{"}Saya\text{"}) + (1 - y_1) \log(1 - D(\text{"}Saya\text{"})) \\
&+ y_2 \log D(\text{"}berlari\text{"}) + (1 - y_2) \log(1 - D(\text{"}berlari\text{"})) \\
&+ y_3 \log D(\text{"}ke\text{"}) + (1 - y_3) \log(1 - D(\text{"}ke\text{"}))] \\
= &- [1 \cdot \log(0.95) + 0 \cdot \log(1 - 0.95) \\
&+ 0 \cdot \log(0.05) + 1 \cdot \log(1 - 0.05) \\
&+ 1 \cdot \log(0.90) + 0 \cdot \log(1 - 0.90)]
\end{aligned}$$

Setelah mengganti nilai logaritma:

$$\begin{aligned}
\mathcal{L}_{RTD} &= - [\log(0.95) + \log(0.95) + \log(0.90)] \\
&= - [-0.0223 + -0.0223 + -0.1054] \\
&= 0.15
\end{aligned}$$

Dengan demikian, nilai dari $\mathcal{L}_{RTD}$ untuk kalimat ini adalah 0.15.

3.4.3 Ilustrasi Perhitungan Metrik Evaluasi

Berikut ilustrasi perhitungan metrik evaluasi pada Rumus 2.7, Rumus 2.8, Rumus 2.9, dan **??**.

Dimana:

TP *(True Positive)*: Jumlah prediksi positif yang benar.

TN *(True Negative)*: Jumlah prediksi negatif yang benar.

FP *(False Positive)*: Jumlah prediksi positif yang salah.

FN *(False Negative)*: Jumlah prediksi negatif yang salah.

Misalkan terdapat hasil pengujian model dengan jumlah sebagai berikut:

$TP = 50$

$TN = 40$

$FP = 10$

$FN = 5$

Dengan nilai-nilai tersebut, metrik evaluasi dapat dihitung sebagai berikut:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{50 + 40}{50 + 40 + 10 + 5} \\ &= \frac{90}{105} \\ &= 0.8571 \end{aligned}$$

Jadi, *Accuracy* = 0.8571 atau 85.71%.

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ &= \frac{50}{50 + 10} \\ &= \frac{50}{60} \\ &= 0.8333 \end{aligned}$$

Jadi, *Precision* = 0.8333 atau 83.33%.

$$\begin{aligned} Recall &= \frac{TP}{TP + FN} \\ &= \frac{50}{50 + 5} \\ &= \frac{50}{55} \\ &= 0.9091 \end{aligned}$$

Jadi, *Recall* = 0.9091 atau 90.91%.

$$
\begin{aligned}
F1 &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \\
&= 2 \times \frac{0.8333 \times 0.9091}{0.8333 + 0.9091} \\
&= 2 \times \frac{0.7562}{1.7424} \\
&= 2 \times 0.4330 \\
&= 0.8660
\end{aligned}
$$

Jadi, *F1 Score* = 0.8660 atau 86.60%.

Dari perhitungan di atas, dapat disimpulkan bahwa model memiliki: - *Accuracy* = 85.71% - *Precision* = 83.33% - *Recall* = 90.91% - *F1 Score* = 86.60%

Metrik-metrik ini memberikan gambaran yang lebih lengkap tentang performa model dalam mengklasifikasikan data dengan benar, baik dalam hal ketepatan (*precision*) maupun dalam mengidentifikasi seluruh data positif yang relevan (*recall*).

DAFTAR PUSTAKA

[1] Abdul Wahab Syakhrani and Engelbertus Kukuh Widijatmoko. "Perkembangan Komunikasi Digital: Dampak Media Sosial pada Interaksi Sosial di Era Modern". *Jurnal Komunikasi* 2.12 (2024).

[2] DataReportal. *5 Billion Social Media Users – Global Digital Insights*. https://datareportal.com/reports/digital-2024-deep-dive-5-billion-social-media-users. [Online; accessed 2024]. 2024.

[3] DataReportal. *Digital 2024: Global Overview Report*. https://datareportal.com/reports/digital-2024-global-overview-report. [Online; accessed 2024]. 2024.

[4] Jessica L. Hamilton et al. "Self-Harm Content on Social Media and Proximal Risk for Self-Injurious Thoughts and Behaviors Among Adolescents". *JAACAP Open* 3.3 (2025). eCollection 2025 Sep, pp. 431–438.

[5] Samaritans. *Samaritans report reveals dangers of social media's self-harm content*. https://www.samaritans.org/news/samaritans-report-reveals-dangers-of-social-medias-self-harm-content/. Press Release. [Online; accessed November 2025]. Nov. 2022.

[6] World Health Organization. *Adolescent mental health*. https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health. Diakses: 29 November 2025. 2025.

[7] Ellen-ge Denton and Kiara Álvarez. "The Global Prevalence of Nonsuicidal Self-Injury Among Adolescents". *JAMA Network Open* 7.6 (2024), e2415406.

[8] T. An et al. "Global burden and trends of self-harm from 1990 to 2021, with predictions to 2050". *Frontiers in Public Health* 13 (2025).

[9] Florian Arendt, Sebastian Scherr, and Daniel Romer. “Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults”. *New Media & Society* 21.11-12 (2019).

[10] Douwe Kiela et al. *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*. 2020. arXiv: 2005.04790 [cs.AI].

[11] Pranav Shah et al. “MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification”. *arXiv preprint arXiv:2403.11245* (2024).

[12] Shivam Sharma et al. *Detecting and Understanding Harmful Memes: A Survey*. 2022. arXiv: 2205.04274 [cs.CL].

[13] Dong-Hyun Lee. “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)* (July 2013).

[14] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks”. *arXiv preprint arXiv:2303.15056* (2023).

[15] Zheng Lian et al. “GPT-4V with Emotion: A Zero-shot Benchmark for Generalized Emotion Recognition”. *arXiv preprint arXiv:2312.04293* (2023).

[16] Shubham Pramanick et al. “MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets”. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 10374–10388.

[17] Anup Kumar and Karthik Nandakumar. “Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2022.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. *Nature* 521.7553 (2015), pp. 436–444.

[19] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. *Neural Networks* 61 (2015), pp. 85–117.

[20] *Computer Vision – an overview*. Diakses pada 9 Desember 2025. ScienceDirect Topics. n.d.

[21] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. *arXiv preprint arXiv:2010.11929* (2020).

[22] IBM. *What is Natural Language Processing (NLP)?* Accessed 9 December 2025. n.d.

[23] Tom Young et al. “Recent Trends in Deep Learning Based Natural Language Processing”. *arXiv preprint arXiv:1708.02709* (2017).

[24] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *arXiv preprint arXiv:1810.04805* (2018).

[25] Ashish Vaswani et al. “Attention Is All You Need”. *arXiv preprint arXiv:1706.03762* (2017).

[26] Xuran Pan et al. “Contrastive Language-Image Pre-Training with Knowledge Graphs”. *arXiv preprint arXiv:2210.08901* (2022). Preprint.

[27] Kevin Clark et al. “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”. *arXiv preprint arXiv:2003.10555* (2020).

[28] Jiquan Ngiam et al. “Multimodal Deep Learning”. *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*. 2011, pp. 689–696.

[29] IBM. *What is AI multimodal?* https://www.ibm.com/think/topics/multimodal-ai. Accessed 9 December 2025. 2024.

[30] Tianzhe Jiao et al. “A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications”. *Computers, Materials & Continua* 80.1 (2024), pp. 1–35.

[31] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443.

[32] Limor Shifman. “Memes in a Digital World: Reconciling with a Conceptual Troublemaker”. *Journal of Computer-Mediated Communication* 18 (Apr. 2013), n/a–n/a.

[33] World Health Organization. *Preventing suicide: A global imperative*. Geneva, Switzerland: World Health Organization, 2014. ISBN: 978-92-4-156477-9.

[34] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (5th ed., text rev.)* American Psychiatric Association, 2022.

[35] Mingle Xu et al. “A Comprehensive Survey of Image Augmentation Techniques for Deep Learning”. *Pattern Recognition* 137 (2023), p. 109347.

[36] Connor Shorten and Taghi M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. *Journal of Big Data* 6.1 (2019), p. 60.

[37] Tom Fawcett. “Introduction to ROC Analysis”. *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874.

[38] Marina Sokolova and Guy Lapalme. “A Systematic Analysis of Performance Measures for Classification Tasks”. *Information Processing & Management* 45.4 (July 2009), pp. 427–437.

[39] David M. W. Powers. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 2020. arXiv: 2010.16061 [cs.LG].

[40] Zain Ali. *Weapon and Knife Detection (137K Images)*. https://www.kaggle.com/datasets/zinkzsa/weapon-and-knife-detection-137k-images. Accessed: Dec. 7, 2025. 2025.

[41] *Alcohol Bottle Images (Glass Bottles)*. https://www.kaggle.com/datasets/dataclusterlabs/alcohol-bottle-images-glass-bottles. Accessed: Dec. 7, 2025. 2025.

[42] Md Mafiul Hasan Matin Mafi and Md. Sabbir Alam. *Suicidal Ideation Detection Reddit Dataset*. Version 2. 2023.

[43] Hammad Javaid. *6992 Meme Images Dataset with Labels*. https://www.kaggle.com/datasets/hammadjavaid/6992-meme-images-dataset-with-labels. Accessed 9 December 2025. 2024.

[44] OpenAI. *GPT-4o mini*. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed 9 December 2025.

[45] Yasin Dus and Georgiy Nefedov. "An Automated Tool to Detect Suicidal Susceptibility from Social Media Posts". *Suicidal Text Detection Report* (2022). Accessed: 2025-12-10.

[46] Hammad Javaid. *6992 Labeled Meme Images Dataset*. https://www.kaggle.com/datasets/hammadjavaid/6992-labeled-meme-images-dataset. Accessed: Dec. 7, 2025. 2023.