BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pertumbuhan infrastruktur internet dan platform penyiaran langsung (*live streaming*) seperti YouTube Live dan Twitch telah melahirkan budaya baru di dalam ekosistem kreator konten (*content creator*). Berbeda dengan format *Video on Demand* (VoD) yang sudah diedit, siaran langsung ditayangkan secara *real-time* dengan durasi yang sangat panjang, seringkali mencapai rentang empat hingga belasan jam tanpa henti. Durasi yang luar biasa panjang ini menciptakan tantangan kognitif bagi audiens modern yang memiliki rentang perhatian terbatas, memunculkan budaya baru berupa pendistribusian ulang potongan klip video secara luas [1].

Fenomena pembuatan klip (*clipping*) telah berkembang menjadi elemen paling krusial dalam strategi distribusi ulang konten (*content re-purposing*). Pembuat konten secara aktif mengiris siaran panjang mereka menjadi potongan-potongan video pendek (15 hingga 60 detik) yang menampilkan momen paling lucu, epik, atau dramatis. Klip-klip ini kemudian didistribusikan ke platform video pendek (*short-form video*) yang digerakkan oleh algoritma seperti TikTok, YouTube Shorts, dan Instagram Reels. Penyebaran klip terbukti sangat efektif sebagai ujung tombak pemasaran untuk menjangkau audiens baru dan menggiring penonton kembali ke siaran panjang mereka [1].

Namun, terdapat *bottleneck* atau masalah hambatan utama dalam ekosistem ini. Mengais momen ”emas” berdurasi 30 detik di dalam tumpukan siaran langsung yang berdurasi 6 jam merupakan proses pencarian jarum di tumpukan jerami. Proses ekstraksi klip secara manual sangat melelahkan, lambat, dan membebani penyunting video (*editor*) maupun kanal-kanal *clipper* independen. Kebutuhan akan alat pembuat klip otomatis (*auto-clipper*) atau sistem peringkas video menjadi kebutuhan yang sangat mendesak untuk meningkatkan produktivitas di industri kreatif digital [2].

Telah banyak riset kecerdasan buatan (*Artificial Intelligence*) yang mencoba membangun sistem peringkas video. Sayangnya, sistem peringkas konvensional umumnya mengandalkan pendekatan visual tunggal, seperti mendeteksi transisi adegan (*scene cut*) atau pergerakan objek secara drastis [3]. Pendekatan visual

murni terbukti memiliki kelemahan ketika diterapkan pada tayangan dengan kamera yang relatif statis [4]. Dalam konteks penyiaran modern di Indonesia, hal ini sangat relevan pada tayangan seperti siniar (*podcast*) atau siaran permainan video (*gaming*) di mana kamera seringkali statis, namun kejadian emosional sangat intens sedang berlangsung.

Pada siaran *live streaming*, sebuah momen klimaks atau *highlight* tidak selalu tercermin dari perubahan gambar, melainkan terdistribusi pada reaksi penonton dan subjek. Deteksi *highlight* yang akurat harus meniru cara manusia menonton: mendengarkan lonjakan emosi pada suara pembicara (*audio*) dan melihat seberapa heboh interaksi audiens di fitur obrolan langsung (*live chat*). Obrolan penonton adalah cerminan metrik paling otentik dari tingkat kehebohan (*hype*) pada suatu kejadian [4].

Oleh karena itu, untuk menciptakan *auto-clipper* yang benar-benar cerdas, sistem membutuhkan fusi multimodal (*multimodal fusion*) yang menggabungkan fitur visual, teks, dan audio sekaligus. Sistem terdahulu seperti CatchLive [4] telah mencoba fusi ini dengan aturan heuristik, namun kurang andal dalam memodelkan interaksi lintas modalitas yang kompleks.

Saat ini, salah satu arsitektur fusi mutakhir yang paling adaptif adalah TripleSumm [5]. Arsitektur ini menyeimbangkan kontribusi ketiga modalitas secara dinamis melalui lapisan atensi silang (*cross-attention layer*). Namun, arsitektur tersebut dirancang menggunakan tulang punggung pemrosesan bahasa alami (*Natural Language Processing* atau *NLP*) bawaan yang berbasis bahasa Inggris. Ketika dihadapkan pada ekosistem *streamer* di Indonesia, terdapat potensi kesenjangan domain (*domain mismatch*) karena komponen tekstual tersebut tidak dirancang untuk memproses bahasa Indonesia informal, bahasa gaul (*slang*), humor lokal, serta singkatan-singkatan khas yang mendominasi fitur obrolan langsung (*live chat*) penonton lokal.

Berdasarkan kesenjangan tersebut, penelitian ini merancang dan mengembangkan sistem peringkas video (*auto-clipper*) terpadu berbasis *Multimodal Deep Learning* yang disesuaikan khusus untuk pasar penonton Indonesia. Mesin deteksi *highlight* dalam sistem ini diadaptasi dengan mengganti komponen teks menggunakan model IndoBERT [6] agar mampu menafsirkan semantik dan sentimen *live chat* lokal.

Model kebahasaan tersebut kemudian diintegrasikan secara fusi atensi silang dengan model *Contrastive Language-Image Pre-Training* (CLIP) [7] untuk ekstraksi visual, serta *Audio Spectrogram Transformer* (AST) [8] untuk mendeteksi

intensitas suara. Dengan otomatisasi ini, diharapkan tercipta sebuah *pipeline* inferensi yang mampu mengambil tautan siaran langsung dan menghasilkan potongan klip *highlight* secara cepat, akurat, dan holistik.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana merancang dan mengintegrasikan arsitektur *Multimodal Deep Learning* berbasis atensi silang (*cross-attention*) yang memadukan fitur teks obrolan (*live chat*) IndoBERT, visual CLIP, dan akustik AST untuk mendeteksi momen penting (*highlight*) pada siaran langsung (*live streaming*)?
2. Berapa tingkat performa kuantitatif (*accuracy*, *precision*, *recall*, dan *F1-score*) dari sistem peringkas video *multimodal* dalam mengenali momen penting tersebut, serta bagaimana efektivitas fusi *multimodal* dibandingkan dengan model berbasis modalitas tunggal (teks-saja)?

1.3 Batasan Permasalahan

Agar penelitian ini lebih terarah, spesifik, dan dapat dievaluasi secara objektif, ditetapkan batasan permasalahan sebagai berikut:

1. Data pengujian difokuskan pada rekaman tayangan *live streaming* dari platform YouTube (total 70 video) beserta log data *live chat*-nya yang didominasi oleh bahasa Indonesia, bahasa daerah, atau bahasa gaul (*slang/hype culture*).
2. Pelabelan momen penting dilakukan secara otomatis (*weakly-supervised*) untuk menghasilkan label acuan otomatis (*pseudo-label*) berbasis pendekatan hibrida *Local Rolling Z-Score* volume obrolan dan intensitas *YouTube Heatmap* (*Most Replayed*).
3. Model fusi *multimodal* dibangun dengan menerapkan metode *transfer learning* [9] memanfaatkan bobot *pre-trained* dari model Mosu (arsitektur TripleSumm). Untuk efisiensi sumber daya komputasi, parameter ekstraktor fitur visual (CLIP ViT-L/14) dan audio (*Audio Spectrogram Transformer* atau

AST) dibekukan (*frozen*) tanpa melatih ulang model dari awal, sementara proses penyesuaian bobot (*fine-tuning*) difokuskan sepenuhnya pada model bahasa (IndoBERT-base-p2) untuk memahami konteks teks.

4. Pemrosesan AI berjalan secara penuh (*full-inference*) pada seluruh rangkaian video tanpa penyaringan manual di awal (*keyword heuristic filtering*) guna menguji batas maksimal ekstraksi informasi kontekstual dari arsitektur fusi.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Merancang dan mengintegrasikan arsitektur *Multimodal Deep Learning* dengan menggabungkan fitur teks obrolan langsung (*live chat*) IndoBERT, fitur visual CLIP, dan fitur audio AST menggunakan mekanisme fusi atensi silang (*cross-attention fusion*) untuk mendeteksi momen penting (*highlight*).
2. Mengevaluasi performa kuantitatif model fusi *multimodal* pada *dataset* uji menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, serta memvalidasi efektivitas fusi *multimodal* dalam menekan tingkat alarm palsu dibandingkan model modalitas tunggal.
3. Membangun *pipeline* inferensi otomatis berbasis Python (*auto-clipper*) yang memproses tautan video siaran langsung secara *end-to-end* hingga menghasilkan potongan klip video ringkasan.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini meliputi:

1. Bagi penonton, Memberikan pengalaman konsumsi konten yang jauh lebih efisien. Penonton dapat menghemat waktu berjam-jam dengan langsung menonton rangkuman momen terbaik dari siaran yang mereka lewatkan.
2. Bagi pembuat konten (*Content Creator*), memangkas waktu, biaya, dan beban kognitif pada fase pascaproduksi. Sistem AI ini bertindak layaknya asisten penyunting video yang menyeleksi kandidat *highlight* mentah untuk disebarkan secara instan ke platform video pendek.

3. Bagi perkembangan akademis (Ilmu Komputer), memberikan kontribusi pada kajian integrasi pemrosesan bahasa alami berbasis linguistik lokal (Bahasa Indonesia dan *slang*) di dalam arsitektur fusi skala besar.