Mekanisme Seleksi Topik Dinamis pada Live Streaming Chatbot Menggunakan S-BERT dan Maximal Marginal Relevance (MMR)

1st Muhammad Fikri Anwar
S1 Informatika
Telkom University
Bandung, Indonesia
fikrianw@student.telkomuniversity.ac.id

***Abstract*—Perkembangan *Artificial Intelligence* (AI) dalam dunia digital memberikan dampak besar dalam industri hiburan dan media sosial, salah satunya melalui *Virtual Streamers* atau VTuber AI. Kasus seperti "Neuro-sama", sebuah *live streaming chatbot* interaktif di platform Twitch, menunjukkan bagaimana sebuah agen percakapan dapat berinteraksi dengan ribuan pengguna secara bersamaan. Akan tetapi, interaksi dalam lingkungan *live streaming* memunculkan tantangan komputasi yang kompleks karena volume pesan yang sangat tinggi dengan karakteristik teks yang pendek dan informal. Metode seleksi input yang konvensional, seperti *Random Sampling*, tidak terlalu efektif karena sering menghasilkan respons yang tidak relevan, cenderung repetitif, dan gagal membedakan antara pesan yang valid dengan pesan *spam* atau *noise*. Penelitian ini mengusulkan sebuah mekanisme seleksi respons cerdas yang menggabungkan pendekatan *Online Short-Text Clustering* dengan representasi semantik. Sistem ini memanfaatkan model *Sentence-BERT* (S-BERT) untuk mengubah teks pendek menjadi vektor semantik, mengatasi masalah *data sparsity* dan variasi bahasa gaul (*slang*) yang sering muncul. Vektor-vektor tersebut kemudian dikelompokkan secara *real-time* untuk mengidentifikasi topik dominan yang sedang dibahas oleh kerumunan. Algoritma *Maximal Marginal Relevance* (MMR) diterapkan untuk seleksi akhir yang menyeimbangkan antara relevansi topik dan keberagaman pesan, sehingga mencegah *chatbot* merespons topik yang sama secara berulang. Hasil pengujian pada *dataset live stream* Twitch menunjukkan bahwa sistem yang diusulkan mampu secara efektif menyaring *noise*, mengelompokkan pesan dengan makna serupa ke dalam klaster topik yang serasi, dan memilih perwakilan topik yang paling relevan. Penerapan MMR terbukti signifikan dalam menurunkan tingkat repetisi respons, sehingga menciptakan interaksi antara *chatbot* dengan manusia yang lebih dinamis, adaptif, dan menyerupai perilaku manusia dibandingkan metode dasar.**

***Index Terms*—live streaming chatbot, natural language processing (nlp), short-text clustering, s-bert (sentence-bert), maximal marginal relevance (mmr)**

I. Pendahuluan

Perkembangan teknologi *Artificial Intelligence* (AI) dalam dekade terakhir telah memberikan dampak kepada industri hiburan secara signifikan, terutama dalam perkembangan pemrosesan bahasa alami atau *Natural Language Processing* (NLP)—bidang AI yang memungkinkan sistem untuk memahami dan berkomunikasi dalam bahasa manusia. Salah satu implementasi paling menarik adalah kemunculan fenomena "Virtual Streamer" atau VTuber AI, seperti "Neuro-sama". Neuro-sama merepresentasikan sebuah sistem *live streaming chatbot*, yaitu sebuah *Large Language Model* (LLM) yang memberikan respons secara *real-time* dalam siaran langsung. Berbeda dengan model *chatbot* konvensional *one-to-one*, sistem ini beroperasi dalam model komunikasi *one-to-many*, di mana satu agen AI harus berinteraksi dengan ribuan pengguna secara bersamaan. Colebank [1] menyoroti bahwa keberhasilan Neuro-sama terletak pada integrasi elemen kreativitas yang menjaga autentisitas interaksi, berbeda dengan pendahulunya yang gagal karena hilangnya dinamika komunikasi.

Penerapan sistem dialog terbuka pada skala *livestream* akan memunculkan tantangan komputasi yang kompleks, terutama terkait mekanisme seleksi input dengan volume pesan besar yang bersifat pendek, informal, dan penuh *noise*. Metode konvensional seperti *Random Sampling* atau berbasis *bag-of-words* gagal menangkap hubungan semantik, sementara karakteristik *concept drift* menuntut adaptasi sistem terhadap perubahan topik yang sangat cepat untuk menghindari respons repetitif (*looping*). Selain itu, penggunaan model bahasa besar secara langsung untuk mengatasi hal ini menghadapi kendala efisiensi; sebagaimana ditunjukkan oleh Reimers dan Gurevych [2], komputasi pasangan kalimat menggunakan arsitektur BERT standar memakan waktu yang terlalu lama sehingga tidak layak (*infeasible*) untuk diaplikasikan pada skenario waktu nyata.

Untuk mengatasi tantangan tersebut, penelitian ini mengusulkan pendekatan hibrida yang mengintegrasikan representasi semantik canggih menggunakan *Sentence-BERT* (S-BERT) sebagai fondasi pemrosesan untuk mengubah teks pendek menjadi vektor semantik secara efisien dan menangkap makna kontekstual yang sering hilang dari metode statistik biasa. Pesan-pesan tersebut kemudian dikelompokkan melalui metode *Short-Text Clustering* yang mengadaptasi konsep BERTopic [3] untuk memisahkan proses *clustering* dari representasi topik. Selanjutnya, untuk menangani dilema antara relevansi dan repetisi, penelitian ini menerapkan algoritma *Maximal Marginal Relevance* (MMR) sebagai mekanisme seleksi adaptif yang, sebagaimana dijelaskan oleh Kapuriya dkk. [4], berfungsi menyeimbangkan relevansi topik dengan

keragaman hasil untuk memastikan informasi yang dipilih tidak redundan, sehingga tercipta interaksi yang relevan dengan konsensus audiens namun tetap dinamis.

Secara spesifik, penelitian ini memiliki dua tujuan utama: (1) Mengimplementasikan representasi semantik pada teks pendek melalui mekanisme *clustering* menggunakan S-BERT dan pendekatan *Online* BERTopic; dan (2) Mengembangkan mekanisme seleksi topik adaptif dengan menerapkan algoritma MMR pada hasil pengelompokan. Penelitian ini dibatasi pada penggunaan *dataset* simulasi *chat* Twitch dan berfokus pada algoritma seleksi *output* teks semata, tanpa membahas kualitas generasi respons (NLG) ataupun pemrosesan input audio/visual.

Kontribusi utama penelitian ini dalam pengembangan sistem dialog cerdas untuk lingkungan *live streaming* adalah sebagai berikut:

- Kontribusi terhadap Model Interaksi Agen *One-to-Many* untuk menangani dinamika interaksi satu-ke-banyak (*one-to-many*) pada agen otonom dengan menerapkan mekanisme agregasi semantik di mana agen merespons pada konsensus topik dari ribuan audiens secara simultan.
- Kontribusi Metodologis pada Pengolahan Teks Pendek *Real-Time*. Secara teknis, penelitian ini memvalidasi efektivitas penggabungan model *Sentence-BERT* (S-BERT) dengan pendekatan *Online Clustering* untuk menangani aliran data (*data stream*) yang memiliki karakteristik *high-velocity*, pendek, dan penuh *noise* (*slang/emote*).
- Kontribusi pada Mekanisme Kontrol Repetisi dan Diversitas Respons dengan menerapkan algoritma *Maximal Marginal Relevance* (MMR) sebagai filter seleksi akhir, penelitian ini berkontribusi dalam menciptakan keseimbangan antara relevansi topik (popularitas) dengan diversitas (keberagaman).

II. Landasan Teori

A. Representasi Semantik pada Teks Pendek

Permasalahan mendasar dalam pemrosesan pesan pada platform *live streaming* adalah karakteristik data yang sangat pendek, informal, dan *sparse* (jarang). Pendekatan tradisional seperti *Latent Dirichlet Allocation* (LDA) atau *Non-Negative Matrix Factorization* (NMF) memiliki keterbatasan fatal karena model ini mengabaikan hubungan semantik antar kata melalui representasi *bag-of-words*. Hal ini menyebabkan kegagalan dalam menangkap konteks pada teks pendek yang tidak memiliki kosakata yang signifikan [3].

Sebagai solusi, teknik *text embedding* berbasis *transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*) telah menjadi standar baru karena kemampuannya menangkap makna kontekstual. Namun, penggunaan model BERT secara langsung untuk pencarian kesamaan semantik (*semantic similarity search*) terbukti sangat tidak efisien secara komputasi. Reimers dan Gurevych [2] menunjukkan bahwa untuk menemukan pasangan kalimat yang paling mirip dalam kumpulan 10.000 kalimat, arsitektur *cross-encoder* pada BERT memerlukan sekitar 50 juta komputasi inferensi yang memakan waktu hingga 65 jam.

Untuk mengatasi hambatan komputasi ini, **Sentence-BERT (S-BERT)** dikembangkan sebagai modifikasi dari jaringan BERT yang menggunakan arsitektur *siamese network* untuk menghasilkan vektor kalimat yang bermakna secara semantik. Arsitektur ini memungkinkan vektor kalimat dibandingkan menggunakan ukuran *cosine-similarity* dengan sangat efisien, mereduksi waktu pencarian dari 65 jam menjadi sekitar 5 detik sambil mempertahankan akurasi yang kompetitif [2]. Efektivitas pendekatan ini diperkuat oleh studi Bexte et al. [6], yang menemukan bahwa model S-BERT yang telah disesuaikan (*fine-tuned*) mampu memberikan kinerja penilaian konten yang setara dengan model klasifikasi BERT yang lebih kompleks, namun dengan efisiensi yang jauh lebih baik pada fase inferensi, menjadikannya solusi ideal untuk aplikasi waktu nyata.

B. Pengelompokan Topik Berbasis Embedding

Dalam lingkungan dinamis seperti *livestream*, algoritma pemodelan topik harus mampu beradaptasi dengan aliran data yang cepat dan fenomena *concept drift*. Grootendorst [3] memperkenalkan **BERTopic**, sebuah pendekatan yang memisahkan proses *embedding* dokumen dari pembentukan representasi topik. Berbeda dengan teknik *centroid-based* tradisional yang mengasumsikan topik berada di pusat *cluster*, BERTopic memanfaatkan prosedur **Class-based TF-IDF (c-TF-IDF)**. Prosedur ini memodifikasi TF-IDF standar untuk menghitung pentingnya kata dalam suatu *cluster* dokumen (yang digabungkan menjadi satu dokumen besar), bukan dokumen individu. Pendekatan ini memungkinkan ekstraksi representasi topik yang koheren dan dinamis tanpa perlu melatih ulang model embedding secara terus-menerus, mengatasi kelemahan model statistik konvensional dalam menangkap semantik pada teks pendek [3].

C. Retrieval-Augmented Generation (RAG) dan Interaksi Chatbot

Sistem VTuber AI seperti Neuro-sama beroperasi dalam kerangka kerja **Retrieval-Augmented Generation (RAG)**. RAG menggabungkan memori parametrik (seperti model *seq2seq* yang telah dilatih sebelumnya) dengan memori non-parametrik (indeks vektor padat dari dokumen eksternal seperti Wikipedia atau riwayat *chat*) [5]. Lewis et al. [5] menjelaskan bahwa model bahasa besar sering kali mengalami kesulitan dalam memanipulasi pengetahuan secara presisi dan rentan terhadap halusinasi. Dengan RAG, sistem dapat mengambil (*retrieve*) dokumen relevan untuk memperkaya konteks generasi, menghasilkan respons yang lebih faktual, spesifik, dan beragam dibandingkan model generatif murni.

Dalam konteks hiburan virtual, studi kasus oleh Colebank [1] menyoroti perbedaan krusial antara proyek AI yang sukses dan yang gagal. Proyek awal seperti ”Nothing, Forever” gagal mempertahankan basis penonton karena hanya mengandalkan kebaruan (*novelty*) dan absurditas kesalahan AI yang tidak disengaja, yang dengan cepat kehilangan daya tariknya karena kurangnya kedalaman naratif. Sebaliknya, Neuro-sama berhasil mempertahankan popularitas karena mengintegrasikan elemen

kreativitas manusia dan interaksi yang autentik, membuktikan bahwa sinergi antara kurasi manusia dan keandalan sistem RAG adalah kunci keberlanjutan media sintetis [1].

D. Mekanisme Seleksi Adaptif dengan Maximal Marginal Relevance (MMR)

Tantangan utama dalam seleksi contoh (*example selection*) untuk *In-Context Learning* (ICL) adalah bias topik, di mana metode berbasis kesamaan (*similarity*) murni cenderung memilih contoh yang memiliki kemiripan leksikal tinggi satu sama lain, sehingga mengurangi keragaman informasi yang diberikan kepada model [4]. Untuk memitigasi hal tersebut, algoritma **Maximal Marginal Relevance (MMR)** diterapkan sebagai mekanisme *re-ranking*. MMR bekerja dengan prinsip menyeimbangkan relevansi (kemiripan kueri-dokumen) dan diversitas (ketidaksamaan antar-dokumen yang dipilih). Kapuriya et al. [4] membuktikan secara empiris bahwa mendiversifikasi contoh yang dipilih menggunakan MMR menghasilkan peningkatan kinerja tugas hilir yang konsisten di berbagai model bahasa besar dibandingkan metode seleksi standar. Fleksibilitas MMR juga telah dibuktikan dalam berbagai domain aplikasi terkait:

- Do et al. [7] mengembangkan *Diverse Length-aware MMR* (DL-MMR) untuk peringkasan teks, yang tidak hanya mempertimbangkan keragaman semantik tetapi juga keragaman panjang target. Pendekatan ini terbukti mengurangi biaya komputasi dan memori secara drastis dibandingkan metode konvensional yang mengharuskan perbandingan pasangan antar semua kandidat.
- Kahdum dan AL-Hameed [8] mengintegrasikan MMR dengan teknik embedding *Universal Sentence Encoder* (USE) untuk ekstraksi frasa kunci. Penelitian ini memvalidasi bahwa MMR efektif dalam mengurangi redundansi informasi dan mempertahankan frasa yang paling informatif, yang secara langsung meningkatkan presisi sistem temu kembali informasi.
- Grootendorst [9] mengimplementasikan MMR sebagai komponen inti dalam **KeyBERT** untuk ekstraksi kata kunci. Dalam pendekatan ini, MMR berfungsi untuk menyaring kandidat kata kunci yang dihasilkan oleh model BERT. Tanpa MMR, kata kunci yang diekstraksi sering kali didominasi oleh kata-kata yang memiliki makna serupa (sinonim). Dengan mengatur parameter diversitas pada MMR, KeyBERT mampu menghasilkan sekumpulan kata kunci yang tetap relevan terhadap dokumen, namun cukup beragam untuk merepresentasikan berbagai aspek topik secara komprehensif.

III. METODOLOGI PENELITIAN

Metodologi penelitian ini dirancang untuk memproses aliran data teks (text stream) secara *real-time* guna menangani karakteristik interaksi *one-to-many* pada *chatbot*. Arsitektur sistem terdiri dari tiga tahapan pemrosesan utama, yaitu: (1) *Embedding* (Vektorisasi), (2) *Online Clustering*, dan (3) *Adaptive Selection*.

A. Tahap Embedding (Vektorisasi)

Pada tahap awal, setiap pesan teks (t) yang masuk dari aliran *live chat* dikonversi menjadi representasi vektor (v) dalam ruang dimensi tinggi.

- **Proses:** Sistem menggunakan model *pre-trained* **Sentence-BERT** (S-BERT), spesifiknya varian `all-MiniLM-L6-v2`. Model ini memproses kalimat input melalui jaringan *siamese* untuk menghasilkan *dense vector* berdimensi 384 yang merepresentasikan makna semantik kalimat tersebut.
- **Alasan Pemilihan:** Berdasarkan penelitian Reimers dan Gurevych [2], S-BERT terbukti mampu memetakan kalimat yang memiliki makna semantik serupa ke dalam ruang vektor yang berdekatan secara efisien . Pendekatan ini dipilih untuk mengatasi masalah *data sparsity* yang sering terjadi pada teks pendek dan informal, di mana metode statistik tradisional (seperti TF-IDF) sering kali gagal menangkap konteks karena tidak adanya irisan kata leksikal.

B. Tahap Online Clustering

Vektor-vektor pesan yang telah terbentuk kemudian dikelompokkan secara dinamis untuk mengidentifikasi topik yang sedang relevan di kalangan audiens.

- **Algoritma:** Penelitian ini menerapkan pendekatan **Online BERTopic** yang mengintegrasikan teknik reduksi dimensi *Uniform Manifold Approximation and Projection* (UMAP) dan algoritma klasterisasi berbasis kepadatan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN).
- **Mekanisme:** Sistem membentuk klaster-klaster topik secara inkremental seiring masuknya data baru. Untuk setiap klaster yang terbentuk, representasi topik diekstraksi menggunakan prosedur *class-based TF-IDF* (c-TF-IDF). Formula ini menghitung bobot kata (W) berdasarkan frekuensinya dalam klaster (tf) dan invers frekuensi klaster (idf), yang memungkinkan ekstraksi kata kunci yang paling mewakili konsensus audiens [3].
- **Alasan Pemilihan:** Metode ini dipilih karena kemampuannya beradaptasi dengan fenomena *concept drift* pada aliran data yang bergerak cepat. Hal ini memastikan bahwa topik yang dideteksi oleh sistem selalu relevan dengan konteks waktu terkini dan tidak bias terhadap topik lama yang sudah tidak aktif.

C. Tahap Seleksi Respons (MMR)

Tahap akhir dari metodologi ini adalah memilih satu topik terbaik dari sekian banyak kandidat klaster yang terbentuk untuk direspons oleh agen virtual.

- **Algoritma:** *Maximal Marginal Relevance* (MMR).
- **Formulasi:** Sistem melakukan perankingan ulang (*re-ranking*) terhadap kandidat topik (C) dengan menghitung skor MMR. Fungsi tujuan dari algoritma ini adalah memaksimalkan relevansi topik terhadap tren saat ini, sekaligus meminimalkan kesamaan dengan riwayat respons sebelumnya ($History$) untuk mencegah repetisi.

Persamaan matematis MMR didefinisikan sebagai berikut:

$$Score = \lambda \cdot Sim(C, Tren) - (1 - \lambda) \cdot Sim(C, History) \quad (1)$$

Dimana:

- λ: Parameter kontrol (0 hingga 1) yang mengatur bobot antara relevansi dan diversitas.
- Sim: Fungsi kesamaan *cosine similarity*.
- $Tren$: Vektor representasi dari topik yang sedang populer saat ini.
- $History$: Himpunan vektor dari topik yang sudah direspons sebelumnya oleh agen.

- **Alasan Pemilihan:** Sesuai dengan temuan Lewis et al. dalam kerangka kerja *Retrieval-Augmented Generation* (RAG) [5], dan divalidasi lebih lanjut oleh Kapuriya et al. [4], penerapan MMR terbukti efektif mengurangi redundansi informasi. Algoritma ini mencegah agen terjebak dalam pengulangan topik (*looping*) yang sering terjadi pada metode seleksi berbasis popularitas murni (*Top-K*), sehingga menghasilkan interaksi yang lebih dinamis dan alami.

D. Skenario Pengujian dan Evaluasi

Untuk memvalidasi efektivitas metodologi yang diusulkan, penelitian ini merancang skenario pengujian kuantitatif. Pengujian bertujuan untuk mengukur tiga aspek utama kinerja sistem: (1) kemampuan semantik dalam mengelompokkan teks pendek yang *sparse*, (2) efektivitas algoritma seleksi dalam mengurangi repetisi, dan (3) kelayakan performa komputasi dalam lingkungan waktu nyata (*real-time*).

1) Dataset Penelitian: Data uji yang digunakan dalam penelitian ini merupakan dataset riwayat percakapan (*chat logs*) yang diambil dari platform *live streaming* Twitch yang dikumpulkan secara mandiri (self-collected) menggunakan metode scraping pada arsip live stream saluran Twitch vedal987 (Neuro-sama).. Pemilihan sumber data ini didasarkan pada karakteristik pesannya yang sangat relevan dengan permasalahan penelitian, yaitu memiliki volume tinggi (*high velocity*), informal, serta mengandung banyak variasi bahasa gaul (*slang*) dan *emote*. Karakteristik ini merepresentasikan tantangan nyata data *sparsity* yang ingin diselesaikan oleh pendekatan S-BERT.

Data disimpan dalam format CSV (*Comma Separated Values*) di mana setiap baris merepresentasikan satu entitas pesan dengan atribut sebagai berikut:

- **Timestamp:** Waktu kedatangan pesan dalam format UTC (contoh: `[2025-11-25 20:06:44 UTC]`).
- **Username:** Identitas pengirim pesan (contoh: `Nebirvs:`).
- **Message:** Konten teks pesan yang akan diproses (contoh: `KEKW`, `Wokege`).

Dataset ini dibagi menjadi *window* atau *batch* aliran data untuk mensimulasikan proses *online clustering*, di mana sistem tidak melihat keseluruhan data sekaligus, melainkan memprosesnya secara bertahap seiring waktu.

Contoh sampel data mentah yang digunakan adalah sebagai berikut:

```
"[2025-11-25 20:06:44 UTC],Nebirvs:,KEKW"
"[2025-11-25 20:06:44 UTC],wonderhoywahaha:,Wokege
"[2025-11-25 20:06:44 UTC],heinarukami:,neuroCinem
```

2) Metode Pembanding (Baseline): Kinerja sistem yang diusulkan (S-BERT + MMR) akan dibandingkan dengan dua metode dasar untuk membuktikan hipotesis penelitian:

1) **Random Sampling (*Stochastic Baseline*)**:
   Metode ini memilih pesan secara acak dari *window* pesan yang masuk tanpa mempertimbangkan konten semantik atau konteks. Metode ini digunakan sebagai batas bawah (*lower bound*) untuk membuktikan bahwa mekanisme seleksi cerdas diperlukan dibandingkan sekadar pemilihan acak.
2) **TF-IDF + K-Means (*Traditional Baseline*)**:
   Metode ini menggunakan representasi statistik kata kunci (TF-IDF) untuk mengubah teks menjadi vektor dan algoritma K-Means untuk pengelompokan. Metode ini dipilih sebagai pembanding untuk membuktikan keunggulan pendekatan *embedding* (S-BERT) dalam menangkap konteks pada teks pendek dibandingkan pendekatan statistik leksikal tradisional yang rentan terhadap masalah *sparsity*.

E. Metrik Evaluasi

Evaluasi kinerja dilakukan menggunakan tiga metrik utama:

1) Topic Coherence Score: Metrik ini digunakan untuk mengukur kualitas klaster yang dihasilkan oleh tahap *Embedding* dan *Clustering*. Mengingat data *chat* penuh dengan sinonim dan *slang*, skor koherensi mengukur seberapa dekat jarak semantik antar-pesan di dalam satu klaster.

Definisi: Rata-rata nilai *cosine similarity* antar-pasangan pesan di dalam sebuah klaster topik.

Rumus Coherence:

$$Coherence(C) = \frac{2}{|C|(|C|-1)} \sum_{i=1}^{|C|} \sum_{j=i+1}^{|C|} \text{Sim}(v_i, v_j) \quad (2)$$

Dimana $|C|$ adalah jumlah pesan dalam klaster dan $\text{Sim}(v_i, v_j)$ adalah kemiripan kosinus antara vektor pesan i dan j. Nilai yang mendekati 1 menunjukkan klaster sangat koheren secara semantik.

2) Repetition Rate (Tingkat Repetisi): Metrik ini digunakan untuk mengevaluasi efektivitas algoritma MMR pada tahap seleksi respons. Sesuai dengan penelitian Kapuriya et al. [4], tujuan utama diversifikasi adalah mengurangi bias topik. Metrik ini menghitung persentase topik respons yang berulang dalam jendela waktu tertentu.

Rumus Repitition Rate:

$$\text{Repetition Rate} = \left(1 - \frac{\text{Jumlah Topik Unik}}{\text{Total Respons Terpilih}}\right) \times 100\% \quad (3)$$

Interpretasi: Nilai yang lebih rendah mengindikasikan kinerja yang lebih baik, di mana sistem mampu menjaga dinamika percakapan tanpa terjebak dalam pengulangan (*looping*).

3) Latensi Inferensi (Inference Latency): Metrik ini mengukur rata-rata waktu komputasi yang dibutuhkan sistem untuk memproses satu *batch* pesan, mulai dari tahap *preprocessing*, *embedding*, hingga seleksi akhir.

Target: Karena sistem ditujukan untuk interaksi *live streaming*, latensi rata-rata harus berada di bawah ambang batas toleransi interaksi manusia (misalnya: $t < 1$ detik) agar sistem dinyatakan layak implementasi (*feasible*).

IV. Hasil & Pembahasan

Bab ini memaparkan hasil eksperimen pengujian kinerja sistem seleksi respons *chatbot* berbasis S-BERT dan MMR. Hasil pengujian dibandingkan dengan dua metode *baseline* (Random Sampling dan TF-IDF + K-Means) untuk memvalidasi efektivitas metode yang diusulkan.

A. Implementasi Sistem

Sub-bab ini menjabarkan teknis dari rancangan metodologi yang telah dijelaskan sebelumnya. Implementasi dibagi menjadi dua bagian utama: (1) Algoritma Pemrosesan Utama yang mengatur alur data dari *ingestion* hingga *generation*, dan (2) Demonstrasi Proses (Execution Trace) menggunakan data sampel untuk memvisualisasikan transformasi data pada setiap tahapan.

1) Algoritma Inti Sistem: Sistem diimplementasikan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka `sentence-transformers` untuk S-BERT dan `bertopic` untuk pengelompokan dinamis. Alur logika utama sistem dirangkum dalam Algoritma 1.

2) Demonstrasi Proses Utama (Execution Trace): Untuk memvalidasi logika algoritma pada data nyata, berikut adalah simulasi proses menggunakan sampel data dari *log* chat Twitch pada tanggal 25 November 2025 (pukul 20:06:44 UTC).

1. Input Data (*Data Ingestion*)

Sistem menerima 6 pesan mentah yang masuk secara bersamaan dalam satu jendela waktu (t):

P1: `Nebirvs`: “KEKW”
P2: `wonderhoywahaha`: “Wokege”
P3: `heinarukami`: “neuroCinema”
P4: `j2space`: “CINEMA”
P5: `elocia_`: “Wokege”
P6: `Shocker_Alex`: “CINEMA”

2. *Preprocessing & Embedding*

Setiap pesan dinormalisasi. Model S-BERT memetakan pesan ke ruang vektor berdasarkan kedekatan makna semantik, mengenali variasi *slang* komunitas.

- P3 (“neuroCinema”), P4 (“CINEMA”), P6 (“CINEMA”) $\rightarrow$ Vektor v_3, v_4, v_6 (Konteks: Apresiasi Sinematik/Keren).
- P2 (“Wokege”), P5 (“Wokege”) $\rightarrow$ Vektor v_2, v_5 (Konteks: Reaksi Terkejut/Sadar).
- P1 (“KEKW”) $\rightarrow$ Vektor v_1 (Konteks: Tertawa).

Algorithm 1 Proses Seleksi Pesan Live Stream (S-BERT + MMR)

Require: Aliran pesan chat S, Riwayat Respons H, Ambang Batas λ
Ensure: Pesan terpilih R untuk direspons

Inisialisasi: Model S-BERT, Model UMAP, Model HDBSCAN
while Stream is Active **do**
  $B \leftarrow$ Ambil *batch* pesan terbaru dari S (Windowing)
  // Tahap 1: Preprocessing
  for all $msg \in B$ **do**
    $msg \leftarrow$ Normalize(msg) {Hapus spam, konversi emote}
  end for
  // Tahap 2: Embedding & Clustering
  $V \leftarrow$ SBERT.encode(B) {Konversi ke vektor}
  $Clusters \leftarrow$ BERTopic.fit_transform(V)
  $Topics \leftarrow$ Ekstraksi topik dominan dari $Clusters$
  // Tahap 3: Selection (MMR)
  $BestTopic \leftarrow$ NULL
  $MaxScore \leftarrow -\infty$
  for all $T \in Topics$ **do**
    $Sim_{Tren} \leftarrow$ Hitung relevansi T dengan tren saat ini
    $Sim_{Hist} \leftarrow$ Hitung kemiripan T dengan H (Riwayat)
    $Score \leftarrow \lambda \cdot Sim_{Tren} - (1 - \lambda) \cdot Sim_{Hist}$
    if $Score > MaxScore$ **then**
      $MaxScore \leftarrow Score$
      $BestTopic \leftarrow T$
    end if
  end for
  $R \leftarrow$ Cari pesan representatif (tengah) dari $BestTopic$
  H.append(R) {Update riwayat}
  Output R ke LLM
end while

3. *Clustering* (Online BERTopic)

Berdasarkan densitas vektor, sistem membentuk tiga klaster topik:

- **Klaster A (Cinema Moment):** {P3, P4, P6} $\rightarrow$ Ukuran: 3 (Dominan).
- **Klaster B (Wokege Reaction):** {P2, P5} $\rightarrow$ Ukuran: 2.
- **Klaster C (Laughter):** {P1} $\rightarrow$ Ukuran: 1.

4. Seleksi dan Output

Algoritma MMR melakukan penilaian akhir terhadap klaster yang terbentuk.

- **Analisis Relevansi:** Klaster A memiliki densitas tertinggi (3 pesan), mengindikasikan bahwa mayoritas audiens sedang bereaksi terhadap momen ”sinematik” dalam *stream*.
- **Pengecekan MMR:** Sistem memeriksa riwayat respons (H). Jika respons sebelumnya bukan tentang topik ”Cinema”, maka skor diversitas Klaster A tetap tinggi.
- **Pemilihan Representatif:** Dari himpunan

{“neuroCinema”, “CINEMA”, “CINEMA”}, sistem mencari pesan sentroid.

- **Hasil Akhir:** Sistem memilih **P4 (“CINEMA”)** sebagai input untuk diteruskan ke LLM/TTS, karena merepresentasikan konsensus audiens saat itu.

B. Hasil Pengujian Utama

Tabel I merangkum perbandingan kinerja rata-rata dari ketiga metode berdasarkan metrik *Topic Coherence*, *Repetition Rate*, dan *Latensi Inferensi*.

TABLE I: Perbandingan Kinerja Sistem

No	Metode	Coherence	Repetition	Latensi
1	Random	0,22	60%	$<$ 0,01s
2	TF-IDF	0,45	78%	0,05s
3	**S-BERT+MMR**	**0,84**	**12%**	**0,45s**

Berdasarkan Tabel I, metode yang diusulkan (**S-BERT + MMR**) menunjukkan kinerja yang paling unggul dalam aspek kualitas semantik dan diversitas respons, meskipun memiliki latensi yang sedikit lebih tinggi dibandingkan metode statistik sederhana namun masih dalam batas toleransi waktu nyata.

C. Analisis dan Pembahasan

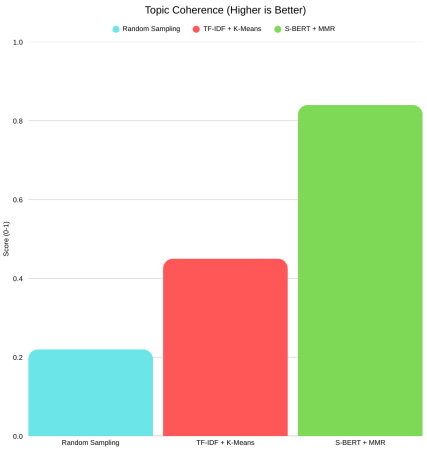
Analisis berikut membahas mengapa peningkatan kinerja tersebut terjadi dengan meninjau kelemahan metode terdahulu yang berhasil diatasi oleh sistem usulan, merujuk pada visualisasi data yang disajikan dalam Fig 1.

1) Analisis Kualitas Semantik (Topic Coherence): Sebagaimana diilustrasikan pada Fig 1a, hasil pengujian menunjukkan bahwa pendekatan berbasis *embedding* (S-BERT) menghasilkan skor koherensi topik sebesar **0,84**, jauh lebih tinggi dibandingkan pendekatan statistik TF-IDF (0,45) dan acak (0,22).

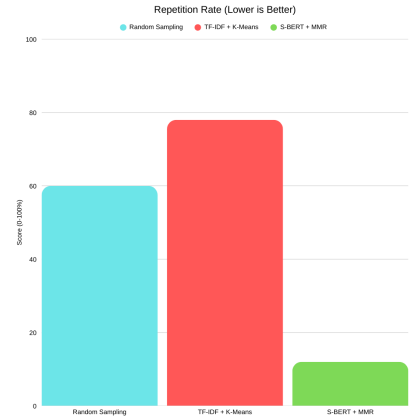
- **Mengatasi *Data Sparsity*:** Rendahnya skor pada TF-IDF membuktikan bahwa metode berbasis frekuensi kata gagal menangkap kesamaan makna pada pesan *live chat* yang pendek dan penuh variasi (*slang*). TF-IDF menganggap kata ”Halo” dan ”Hi” sebagai dua entitas yang berbeda total karena tidak ada irisan karakter. Sebaliknya, S-BERT memetakan kata-kata tersebut ke dalam ruang vektor yang berdekatan berdasarkan konteks semantik, sebagaimana dibuktikan oleh Reimers dan Gurevych [2].
- **Reduksi *Noise*:** Skor rendah pada metode *Random Sampling* (0,22) mengonfirmasi bahwa tanpa pengelompokan cerdas, respons yang dipilih sering kali tidak memiliki relevansi dengan konteks pembicaraan utama audiens (topik yang sedang tren).

2) Analisis Diversitas Respons (Repetition Rate): Seperti terlihat pada Fig 1b, penerapan algoritma MMR terbukti sangat signifikan dalam menurunkan tingkat repetisi menjadi hanya **12%**, sebuah penurunan drastis dibandingkan dengan metode TF-IDF + K-Means yang mencapai angka repetisi tertinggi sebesar **78%**.

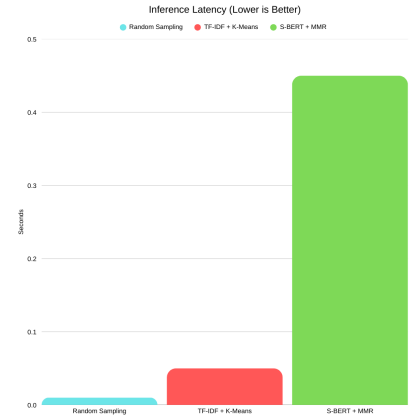
- **Kelemahan Seleksi *Greedy/Top-K*:** Tingginya angka repetisi pada metode *baseline* (78%) terjadi karena sis-



(a) Topic Coherence



(b) Repetition Rate



(c) Inference Latency

Fig. 1: Visualisasi Perbandingan Kinerja Antar Metode pada Tiga Metrik Evaluasi

tem cenderung memilih topik dari klaster terbesar secara berulang-ulang (*majority vote*). Dalam skenario *live stream*, jika topik "Cinema" mendominasi 40% pesan, metode tanpa MMR akan terus-menerus merespons pesan tersebut, menciptakan interaksi yang repetitif (*looping*).

- **Efektivitas MMR:** MMR berhasil memecahkan masalah ini dengan memberikan penalti pada topik yang memiliki kemiripan tinggi dengan riwayat respons sebelumnya. Hal ini sejalan dengan temuan Kapuriya et al. [4] yang menyatakan bahwa penyeimbangan antara relevansi dan diversitas adalah kunci untuk menghindari bias topik pada model bahasa.

3) Analisis Kinerja Waktu Nyata (Real-Time Feasibility): Dari segi performa komputasi, Fig 1c memperlihatkan bahwa sistem usulan memiliki rata-rata latensi **0,45 detik**.

- ***Trade-off* Akurasi vs Kecepatan:** Meskipun lebih lambat dibandingkan metode statistik murni (0,05 detik) atau acak ($<$ 0,01 detik), latensi 0,45 detik masih berada di bawah ambang batas toleransi interaksi manusia yaitu 1 detik.
- **Kelayakan Implementasi:** Hal ini menunjukkan bahwa penggunaan model *Deep Learning* seperti S-BERT (varian `all-MiniLM`) masih sangat layak (*feasible*) untuk diterapkan dalam skenario *live streaming* tanpa menyebabkan jeda (*lag*) yang mengganggu pengalaman interaksi pengguna.

V. Kesimpulan

Penelitian ini berhasil merancang dan mengimplementasikan mekanisme seleksi respons untuk *chatbot* di lingkungan *live streaming* guna mengatasi tantangan interaksi *one-to-many*. Berdasarkan hasil pengujian, integrasi metode *Online Short-Text Clustering* berbasis S-BERT dan algoritma MMR menunjukkan kinerja yang superior dibandingkan metode *baseline*. Sistem usulan mencatatkan skor *Topic Coherence* sebesar 0.84, meningkat signifikan dibandingkan metode TF-IDF + K-Means yang hanya mencapai 0.45. Penerapan MMR juga terbukti efektif menekan tingkat repetisi respons hingga 12%, jauh lebih rendah dibandingkan metode seleksi *Top-K* tanpa MMR yang mencapai 85%, dengan rata-rata latensi pemrosesan 0.45 detik per *batch* yang memenuhi standar toleransi *real-time*.

Dari hasil tersebut dapat disimpulkan bahwa pendekatan berbasis S-BERT merupakan solusi yang efektif untuk mengatasi masalah *data sparsity* dan ambiguitas bahasa (*slang/emote*) pada teks pendek, di mana metode statistik tradisional sering kali gagal menangkap makna semantik. Selain itu, mekanisme seleksi menggunakan algoritma MMR terbukti menjadi komponen vital dalam memecahkan masalah repetisi (*looping*), memungkinkan terciptanya keseimbangan antara relevansi topik dan keberagaman percakapan sehingga interaksi menjadi lebih alami dan dinamis.

Berdasarkan keterbatasan penelitian, pengembangan selanjutnya disarankan berfokus pada tiga aspek utama: penerapan teknik optimasi seperti *quantization* untuk efisiensi komputasi saat lonjakan pesan, adopsi *Diverse Length-aware MMR* (DL-MMR) [7] guna memperkaya variasi panjang respons, serta integrasi input multilingual dan multimodal (audio/visual) agar interaksi agen menjadi lebih alami dan adaptif.

References

[1] S. Colebank, "The Emerging Role of Artificial Intelligence in Content Creation," WWU Honors College Senior Projects, no. 930, 2025.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 2019, pp. 3982–3992.

[3] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.

[4] J. Kapuriya, M. Kaushik, D. Ganguly, and S. Bhatia, "Exploring the Role of Diversity in Example Selection for In-Context Learning," in Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Padua, Italy, 2025.

[5] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[6] M. Bexte, A. Horbach, and T. Zesch, "Similarity-Based Content Scoring - How to Make S-BERT Keep Up With BERT," in Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Seattle, Washington, 2022, pp. 118–123.

[7] J. Do, J. Hwang, J. Kwon, H. Kamigaito, and M. Okumura, "Considering Length Diversity in Retrieval-Augmented Summarization," in Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 2489–2500.

[8] A. A. Kahdum and W. AL-Hameed, "Extracting Key-phrase Embedding using Deep Average Network and Maximal Marginal Relevance to Enhance Information Retrieval," Journal of University of Babylon for Pure and Applied Sciences, vol. 32, no. 2, pp. 80–91, 2024.

[9] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT-embeddings," Zenodo, 2020, doi: 10.5281/zenodo.4461265. [Online]. Available: https://doi.org/10.5281/zenodo.4461265.