ABSTRAK

Komentar pengguna pada video YouTube merupakan sumber informasi yang mencerminkan opini dan diskusi terhadap isi video. Berbagai penelitian telah memanfaatkan *topic modelling* untuk mengidentifikasi topik yang dibahas, namun belum mampu menghubungkannya dengan bagian video yang memicu munculnya pembahasan tersebut sehingga hubungan antara diskusi pengguna dan isi video belum dapat dipahami secara temporal. Selain itu, kualitas pemetaan topik terhadap transkrip dipengaruhi oleh representasi transkrip yang dibentuk melalui proses *chunking*. Oleh karena itu, penelitian ini bertujuan mengimplementasikan *Topic-Temporal Distribution Mapping* (TTDM) untuk memetakan topik komentar terhadap bagian transkrip video yang relevan serta mengevaluasi pengaruh *semantic chunking* terhadap kualitas pemetaan temporal.

Penelitian menggunakan data komentar dan transkrip dari lima video pada kanal YouTube GadgetIn. Komentar diproses melalui tahap *preprocessing* dan dimodelkan menggunakan BERTopic dengan representasi embedding dari model Sentence Transformer *paraphrase-multilingual-MiniLM-L12-v2*. Transkrip diproses menggunakan empat skenario *chunking*, yaitu *fixed-window chunking*, *segment-aware chunking*, *segment-aware chunking* dengan *topic-shift detection*, serta penambahan *overlap strategy*. Selanjutnya, dilakukan pemetaan temporal menggunakan *cosine similarity* antara *topic centroid* dan embedding *chunk* transkrip. Evaluasi dilakukan menggunakan metrik CV Coherence, Semantic Consistency, Coverage Ratio, dan Window Fragmentation melalui pendekatan *ablation study*.

Hasil penelitian menunjukkan bahwa BERTopic menghasilkan kualitas topik yang konsisten pada seluruh skenario dengan nilai CV Coherence sebesar 0,4766. Hasil *ablation study* menunjukkan bahwa *segment-aware chunking* meningkatkan Coverage Ratio dari 0,8388 menjadi 0,9659, sedangkan kombinasi seluruh mekanisme yang diusulkan menghasilkan pemetaan temporal yang lebih baik dibandingkan *fixed-window chunking* dengan meningkatkan Coverage Ratio menjadi 0,8875 serta menurunkan Window Fragmentation dari 4,22 menjadi 3,65. Temuan tersebut menunjukkan bahwa strategi pembentukan *chunk* berpengaruh terhadap kualitas TTDM. Penelitian ini memberikan evaluasi mengenai pengaruh berbagai mekanisme *semantic chunking* terhadap kualitas *Topic-Temporal Distribution Mapping* pada komentar video YouTube sebagai dasar pengembangan metode pemetaan temporal pada penelitian selanjutnya.

Kata kunci: BERTopic, *semantic chunking*, *Topic-Temporal Distribution Mapping*, YouTube, *topic modelling*.

ABSTRACT

User comments on YouTube videos provide valuable information that reflects viewers' opinions and discussions regarding video content. Previous studies have applied topic modelling to identify discussion topics from user comments; however, these approaches are generally unable to associate the identified topics with the specific parts of a video that trigger the discussions. Consequently, the temporal relationship between user discussions and video content cannot be adequately captured. Furthermore, the quality of mapping topics to video transcripts is influenced by how the transcripts are represented, particularly through the chunking process. Therefore, this study aims to implement Topic-Temporal Distribution Mapping (TTDM) to map discussion topics from user comments to relevant segments of video transcripts and to evaluate the effect of semantic chunking on the quality of temporal topic mapping.

This study uses user comments and transcripts collected from five YouTube videos published on the GadgetIn channel. The comments were preprocessed and modelled using BERTopic with sentence embeddings generated by the Sentence Transformer paraphrase-multilingual-MiniLM-L12-v2 model. The transcripts were processed using four chunking scenarios: fixed-window chunking, segment-aware chunking, segment-aware chunking with topic-shift detection, and segment-aware chunking with an overlap strategy. Temporal mapping was then performed by measuring the cosine similarity between topic centroids and transcript chunk embeddings. The proposed approach was evaluated using CV Coherence, Semantic Consistency, Coverage Ratio, and Window Fragmentation through an ablation study.

The results show that BERTopic produced consistent topic quality across all scenarios, achieving a CV Coherence score of 0.4766. The ablation study demonstrates that segment-aware chunking increased the Coverage Ratio from 0.8388 to 0.9659, while the combination of all proposed mechanisms improved the overall temporal mapping performance compared with fixed-window chunking by increasing the Coverage Ratio to 0.8875 and reducing the Window Fragmentation from 4.22 to 3.65. These findings indicate that transcript chunking strategies significantly influence the quality of TTDM. This study provides an evaluation of the impact of different semantic chunking mechanisms on Topic-Temporal Distribution Mapping for YouTube user comments and serves as a foundation for future research on temporal topic mapping methods.

Keywords: *BERTopic, semantic chunking, Topic-Temporal Distribution Mapping, YouTube, topic modeling.*