

ABSTRACT

The phenomenon of information overload in Indonesian online news portals often triggers a high level of information redundancy across articles, making it difficult for readers to efficiently grasp the core of an event. This study proposed a multi-document extractive text summarization method that integrated Agglomerative Clustering for topic isolation, Sentence-BERT (SBERT) semantic representation utilizing the *paraphrase-multilingual-MiniLM-L12-v2* variant, and the Maximal Marginal Relevance (MMR) algorithm as an anti-redundant sentence selection mechanism. This dense embedding-based computational approach was implemented to overcome the limitations of conventional statistical representations in detecting paraphrased meanings. The architectural testing was conducted on the Liputan6 benchmark dataset distributed into six main news categories. The quality of the summaries was evaluated using ROUGE metrics against human ground truth, which was validated for reliability with a Percentage Agreement of 63.21%. Computational experimental results across 20%, 30%, 40%, and 50% compression scenarios demonstrated that the 30% compression rate was the most optimal sweet spot for balancing main information density and phrase diversity. At this compression threshold, the proposed method achieved a competitive overall average performance with an F1-Score of 0.4750 for ROUGE-1, F1-Score of 0.3034 for ROUGE-2, and F1-Score of 0.3391 for ROUGE-L. This research empirically proves that the integration of deep semantic understanding and computational diversity metrics effectively reduces lexical redundancy while maintaining the factual accuracy of news texts for digital readers.

Keywords : Multi-Document Extractive Summarization, Sentence-BERT, Maximal Marginal Relevance, ROUGE.