Universitas Indonesia Library >> UI - Skripsi Membership

Analisis Performa Pendekatan Topic Modeling dan Similarity Measure untuk Text Summarization secara Ekstraktif pada Teks Berbahasa Indonesia = Performance Analysis of Topic Modeling and Similarity Measure Approach for Extractive Text Summarization in Indonesian Text

Gibran Brahmanta Patriajati, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920518396&lokasi=lokal

---

Abstrak

Text Summarization secara ekstraktif merupakan suatu isu yang dapat meningkatkan kualitas pengalaman pengguna ketika menggunakan suatu sistem perolehan informasi. Pada bahasa Inggris, terdapat beberapa penelitian terkait Text Summarization secara ekstraktif salah satunya adalah penelitian Belwal et al. (2021) yang memperkenalkan suatu metode Text Summarization secara ekstraktif yang berbasiskan proses Topic Modeling serta Semantic Measure menggunakan WordNet. Sementara pada bahasa Indonesia, juga terdapat beberapa penelitian terkait Text Summarization secara ekstraktif tetapi belum ada yang menggunakan metode yang sama seperti yang diperkenalkan oleh Belwal et al. (2021). Agar metode yang diperkenalkan Belwal et al. (2021) dapat digunakan pada bahasa Indonesia, proses Semantic Measure menggunakan WordNet harus diganti dengan Similarity Measure menggunakan Vector Space Model karena tidak adanya model WordNet bahasa Indonesia yang dapat digunakan oleh umum. Dalam menggunakan metode yang diperkenalkan oleh Belwal et al. (2021) pada bahasa Indonesia, terdapat beberapa metode yang dapat digunakan untuk melakukan Topic Modeling, Vector Space Model, serta Similarity Measure yang terdapat di dalamnya. Penelitian ini berfokus untuk mencari kombinasi metode ketiga hal yang telah disebutkan sebelumnya yang dapat memaksimalkan performa metode Text Summarization yang diperkenalkan oleh Belwal et al. (2021) pada bahasa Indonesia dengan menggunakan pendekatan hill-climbing. Proses evaluasi dilakukan dengan menggunakan metrik ROUGE-N dalam bentuk F-1 Score pada dua buah dataset yaitu Liputan6 serta IndoSUM. Hasil penelitian menemukan bahwa kombinasi metode yang dapat memaksimalkan performa metode Text Summarization secara ekstraktif yang diperkenalkan oleh Belwal et al. (2021) adalah Non-Negative Matrix Factorization untuk Topic Modeling, Word2Vec untuk Vector Space Model, serta Euclidean Distance untuk Similarity Measure. Kombinasi metode tersebut memiliki nilai ROUGE-1 sebesar 0.291, ROUGE-2 sebesar 0.140, dan ROUGE-3 sebesar 0.079 pada dataset Liputan6. Sementara pada dataset IndoSUM, kombinasi metode tersebut memiliki nilai ROUGE-1 sebesar 0.455, ROUGE-2 sebesar 0.337, dan ROUGE-3 sebesar 0.300. Performa yang dihasilkan oleh kombinasi metode tersebut bersifat cukup kompetitif dengan performa metode lainnya seperti TextRank serta metode berbasiskan model Deep Learning BERT apabila dokumen masukannya bersifat koheren.
.....Extractive text summarization is an issue that can improve the quality of user experience when using an information retrieval system. Research related to extractive text summarization is a language-specific research. In English, there are several studies related to extractive text summarization, one of them is the research of Belwal et al. (2021) They introduced an extractive Text Summarization method based on the Topic Modeling process and Semantic Measure using WordNet. While in Indonesian, there are also several studies related to extractive text summarization, but none have used the same method as introduced by Belwal et al. (2021). In order to use the method introduced by Belwal et al. (2021) in Indonesian, the

Semantic Measure process using WordNet must be replaced with Similarity Measure using the Vector Space Model because there is no Indonesian WordNet model that can be used by the public. When using the method introduced by Belwal et al. (2021) in Indonesian, there are several methods that can be used to perform Topic Modeling, Vector Space Model, and Similarity Measure that contained in there. This study focuses on finding a combination of the three methods previously mentioned that can maximize the performance of the Text Summarization method introduced by Belwal et al. (2021) in Indonesian using hill-climbing approach. The evaluation process is carried out using the ROUGE-N metric in the form of F-1 Score on two datasets, namely Liputan6 and IndoSUM. The results of the study found that the combination of methods that can maximize the performance of the extractive text summarization method introduced by Belwal et al. (2021) are Non-Negative Matrix Factorization for Topic Modeling, Word2Vec for Vector Space Model, and Euclidean Distance for Similarity Measure. The combination of those methods has a ROUGE-1 value of 0.291, ROUGE-2 value of 0.140, and ROUGE-3 value of 0.079 in the Liputan6 dataset. Meanwhile, in the IndoSUM dataset, the combination of those methods has a ROUGE-1 value of 0.455, ROUGE-2 value of 0.337, and ROUGE-3 value of 0.300. The performance generated by the combination of those methods is quite competitive with the performance of other methods such as TextRank and Deep Learning BERT model based method if the input document is coherent.