INTISARI

PERINGKASAN TEKS OTOMATIS EKSTRAKTIF MENGGUNAKAN PENGUKURAN KEMIRIPAN KALIMAT DENGAN SENTENCE-BERT

Rokhana Diyah Rusdiati
21/486134/PPA/06236

Banyaknya berita *online* dari berbagai sumber membuat manusia tidak bisa membaca keseluruhan berita yang ada karena memiliki waktu yang terbatas. Oleh sebab itu diperlukan adanya informasi dalam bentuk yang lebih ringkas agar pengguna bisa mendapatkan informasi secara singkat, yaitu dengan menggunakan peringkas teks otomatis.

Belum ada penelitian yang membandingkan Doc2Vec dan Sentence-BERT pada peringkasan teks otomatis berbahasa Indonesia. Dalam penelitian ini dilakukan perbandingan ringkasan yang dihasilkan dari peringkas teks otomatis menggunakan Doc2Vec dan Sentence-BERT pada dataset IndoSum. Pada proses pembuatan ringkasan otomatis dilakukan beberapa tahap pada prapemrosesan, yaitu pemisahan kalimat, penghilangan tanda baca, *case folding*, dan pemisahan tiap kata. Tahap berikutnya adalah ekstraksi fitur teks yang digunakan, yaitu *title score*, *position score*, kemiripan kalimat dengan judul (*similarity sentence to title* atau SRST), kemiripan kalimat dengan klaster kalimat (*similarity sentence to sentence cluster* atau SRSC), dan *sentence score*. Kemiripan kalimat dihitung menggunakan teknik kemiripan berbasis *embeddings*, yaitu Doc2Vec dan Sentence-BERT. Proses pembuatan ringkasan dilakukan menggunakan metode regresi pada LightGBM.

Hasil penelitian menunjukkan bahwa Sentence-BERT cukup baik diterapkan pada peringkasan teks otomatis untuk bahasa Indonesia. Ringkasan terbaik dihasilkan menggunakan Sentence-BERT + *stemming* dengan nilai evaluasi ROUGE yang lebih tinggi daripada ringkasan yang dihasilkan menggunakan Doc2Vec + *stemming* , yaitu rata-rata nilai *precision* = 64.36%, *recall* = 86.04%, dan *f-measure* = 72.82%.

Kata kunci: peringkasan teks otomatis, ekstraksi fitur, kemiripan kalimat, Doc2Vec, Sentence-BERT, LightGBM

ABSTRACT

EXTRACTIVE AUTOMATIC TEXT SUMMARIZATION USING SENTENCE SIMILARITY MEASUREMENT WITH SENTENCE-BERT

Rokhana Diyah Rusdiati
21/486134/PPA/06236

The large amount of online information from many different sources means that people cannot read all the information online because of limited time. Therefore, it is necessary to provide information in a more concise form so that users can retrieve the information briefly, especially by using automatic text summarizers.

There has been no research comparing Doc2Vec and Sentence-BERT for automatic text summarization in Indonesian. Therefore, in this study, a comparative analysis of summaries extracted from automatic text summarizers using Do2Vec and Sentence-BERT was performed using IndoSum dataset. During automatic summary generation, several steps are performed in preprocessing, which are sentence separation, punctuation removal, capitalization, and word separation. The next step is to extract the used text features, which are title score, position score, sentence-to-title similarity (SRST), sentence-to-sentence cluster similarity (SRSC), and sentence score. Sentence similarity is calculated using embedding-based similarity techniques, specifically Doc2Vec and Sentence-BERT. The summary generation process is performed using a regression method on LightGBM.

The research results show that Sentence-BERT is a particularly good option for automatic text summarization in Indonesian. The best summaries created with Sentence-BERT + stemming have a higher ROUGE value than summaries created with Doc2Vec + stemming, specifically an average value of precision = 64.36%, recall = 86.04%, and f-measure = 72.82%.

Keywords: automatic text summarization, feature extraction, sentence similarity, Doc2Vec, Sentence-BERT, LightGBM