INTISARI

Analisis Perbandingan Kinerja Algoritma *Similarity Measure* sebagai Tahapan *Data Preprocessing: Text Normalization* Bahasa Indonesia untuk Analisa Sentimen

Oleh

Achmad Yohni Wahyu Finansyah
16/398491/PA/17452

Pada masa pandemi COVID-19, kegiatan belajar mengajar dilakukan secara daring. Untuk meningkatkan standar belajar mengajar secara daring, perlu dilakukan analisa sentimen terhadap perangkat lunak belajar daring. Penggalian data pada analisa sentimen sendiri, memiliki banyak permasalahan seperti data *noisy*. Sehingga untuk meningkatkan performa *data preprocessing* diperlukan algoritma normalisasi kata yang paling baik untuk *dataset* Bahasa Indonesia. Dalam penelitian ini dilakukan perbandingan normalisasi kata untuk *spell correction* dengan algoritma *Levenshtein Distance, Jaro-Winkler Distance,* Bigram, dan Smith-Waterman dengan variasi *threshold* 60%, 70%, 80%, 90%. Penelitian ini menggunakan dua *dataset* yang berbeda yaitu data kuisioner dan data twitter. Dan pada penelitian ini dilakukan pengujian validasi untuk melihat dampak normalisasi terhadap klasifikasi sentimen. Hasil yang didapatkan algoritma *Levenshtein Distance* pada threshold 60% secara konsisten memberi nilai akurasi normalisasi paling baik. Namun dari waktu normalisasi, pengurangan jumlah kata unik hasil terbaik dihasilkan *Jaro-Winkler Distance.* Sedangkan normalisasi dengan performa terbaik tidak selalu memberi nilai klasifikasi terbaik.

Kata - kata kunci : *Text Mining, Text Preprocessing,* Normalisasi Kata, *Spell Correction,* Klasifikasi, Analisis Sentimen

ABSTRACT

Comparation of Similarity Measure Algorithm Performance as Data Preprocessing Stage: Text Normalization in Bahasa for Sentiment Analysis

By

Achmad Yohni Wahyu Finansyah
16/398491/PA/17452

During the COVID-19 pandemic, the learning activities were carried out online. To improve the online learning standards, sentiment analysis needs to be done on online learning software. Data mining on sentiment analysis itself has many problems such as noisy data. So, to improve the performance of preprocessing data, it is needed the best word normalization algorithm in spell correction for Bahasa dataset. In this research, word normalization using Levenshtein Distance, Jaro-Winkler Distance, Bigram, and Smith-Waterman algorithm with threshold variations of 60%, 70%, 80%, 90%. This study uses two different datasets namely questionnaire data and twitter data. And in this research, validation testing is done to see the affects of the normalization on sentiment classification. The results obtained by the Levenshtein Distance algorithm at the fully approved 60% threshold give the best normalization accuration. But from the time of normalization, the number of unique word results the best results are normalization with Jaro-Winkler Distance. While normalization with the best performance does not always provide the best classification score.

Keywords : *Text Mining, Text Preprocessing, Text Normalization*, *Spell Correction, Classifcation*, *Sentiment Analysis*