INTISARI

NORMALISASI LEKSIKAL PESAN TWITTER BERBAHASA INDONESIA MENGGUNAKAN METODE PHONETIC MATCHING

Oleh

Amie Yulikawati

12/331440/PA/14694

Normalisasi leksikal pesan singkat berbahasa Indonesia pada penggunaan teknologi seperti Twitter yang memiliki ruang terbatas telah membawa dampak terhadap penggunaan kata serta kualitas teks yang dihasilkan. Keterbatasan panjang karakter yang dapat diinputkan pada suatu tweet menyebabkan kecenderungan menggunakan kata tidak baku yang sengaja dipersingkat atau menggunakan koreksi ejaan yang salah untuk mengikuti bahasa modern saat ini. Karena permasalahan ini, diperlukan suatu sistem normalisasi leksikal pencocokan kata yang sesuai definisinya adalah proses yang dapat mengubah atau mengoreksi kata tidak baku menjadi kata baku sesuai Kamus Besar Bahasa Indonesia.

Salah satu pendekatan normalisasi leksikal yang dilakukan adalah menggunakan metode *phonetic matching* dengan algoritma soundex, di mana algoritma ini untuk mengurangi kesalahan pengetikan suatu kata. Pada penelitian ini, pemrosesan menggunakan serangkaian algoritma pencocokan kata, yaitu algoritma Levenshtein Distance, Soundex, Skip-gram, dan Peter Norvig. Penelitian ini difokuskan pada tweet berbahasa Indonesia.

Berdasarkan pengujian yang telah dilakukan, kode soundex bahasa Inggris menghasilkan rata-rata tingkat akurasi sebesar 74,83%, sedangkan pengujian sistem dengan kode soundex bahasa Indonesia menghasilkan rata-rata tingkat akurasi sebesar 70,83%. Lama waktu pemrosesan masing-masing memiliki kesamaan waktu kira-kira 12 detik untuk setiap tweet.

Kata kunci: normalisasi leksikal, phonetic matching, soundex, levenshtein distance, skip-gram, peter norvig, natural language processing

ABSTRACT

LEXICAL NORMALIZATION OF INDONESIAN TWEETS USING PHONETIC MATCHING METHOD

By

Amie Yulikawati

12/331440/PA/14694

Lexical normalization on Indonesian text messages in technologies such as Twitter, which has limited space, has brought an impact on word usage and the resulting text quality. The limitation on character length for a tweet caused a tendency of using non-standard words which are intentionally shortened or using non-normative spelling to follow the current trends instead. Because of this issue, a lexical normalization system is required, which corresponds to its definition as a process to change or correct non-normative words to normative ones according to Indonesia Dictionary.

One of the lexical normalizations approaches done were using phonetic matching method with Soundex algorithm, where it is used to reduce the mistype of a word. In this research, processes were done using a series of word matching algorithms, such as Levenshtein Distance, Soundex, Skip-gram, and Peter Norvig. This research is focused on Indonesian Tweets.

Based on the testing, in English, Soundex code resulted in an average accuracy of 74,83%, while in Indonesian, its accuracy is averaging at 70,83%. Processing time for each is similar at approximately 12 seconds per tweet.

Keywords: lexical normalization, phonetic matching, soundex, levenshtein distance, skip-gram, peter norvig, natural language processing