INTISARI

NORMALISASI TEKS TWITTER BAHASA INDONESIA BERBASISKAN *NOISY CHANNEL MODEL*

Oleh:

Johanes Irawan
12/339890/PPA/04017

Banyaknya media sosial dan konten web yang dibuat oleh user membuat banyaknya jumlah teks Bahasa Indonesia dalam bentuk elektronik. Salah satu contoh media tersebut adalah Twiter. Namun ternyata, struktur kata Bahasa Indonesia yang digunakan di Twitter berbeda dengan bahasa yang digunakan pada teks formal. Akibat dari perbedaan tersebut, tools yang sudah ada untuk Bahasa Indonesia mengalami penurunan akurasi bila diaplikasikan ke teks Twitter.

Salah satu cara untuk menyelesaikan masalah tersebut adalah dengan normalisasi. Normalisasi berusaha untuk mengubah kata tidak baku pada Twitter menjadi kata baku. Penelitian ini akan memilih untuk melakukan normalisasi pada Teks Twitter dengan menggunakan *Noisy Channel Model*.

Noisy Channel Model menjelaskan bahwa bentuk yang terlihat pada suatu kata yang terlihat dapatlah menjadi sebuah bentuk yang terdistorsi dari bentuk aslinya. Dengan menggunakan probabilitas, model ini akan mencari kata dari model yang ada dan memilih kata yang memiliki probabilitas tertinggi dengan kata 'noisy' yang ada.

Pada penelitian ini, sistem mampu membedakan beberapa variasi dari sebuah kata, namun tidak dapat membedakan kata yang harusnya tidak dinormalisasi dan tidak dapat memperbaiki singkatan. Akurasi yang didapatkan sistem adalah 70,12% dan F1 Score sebesar 38,36%.

keyword = noisy channel model, normalisasi, twitter, bahasa indonesia

ABSTRACT

NORMALIZATION OF INDONESIAN TWITTER TEXT BASED ON *NOISY CHANNEL MODEL*

By

Johanes Irawan
12/339890/PPA/04017

The rise of social media and web contents that are created by users have made an abundant amount of Indonesian Text in electronic form. One of the example of the social media is Twitter. But then, the structure of the Indonesian Text in Twitter is different than the one in formal form. This is because users might want to abbreviate the word or make a variation of the word. Because of this, tools that are used for formal Indonesian text are going to have their accuracy reduced when applied to Twitter text.

One way to solve this problem is by normalization. Normalization task tries to correct the informal words in Twitter into formal words. This research is going to do normalization on Twitter Text using Noisy Channel Model.

Noisy Channel Model explains that the form that we see could be a form that is distorted from the real form. By using probability, this model will find a word from a model that has the highest probability.

In this research, the model is able to find some variant of a noisy word, but fails to differentiate which word that shouldn't be normalized and cannot normalize abbreviation. The accuracy that is achieved by the system is 70,12% and the F1 Score is 38,36%.

keyword = noisy channel model, normalization, twitter, Indonesian Language