Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: *Tweet* Produk *Gadget* Pada Twitter

Riri Riyaddulloh[1], Ade Romadhony[2]

1,2 Fakultas Informatika, Universitas Telkom, Bandung
riririyaddulloh@student.telkomuniversity.ac.id, aderomadhony@telkomuniversity.ac.id

Abstrak

Sosial media adalah alat bantu untuk memperkaya informasi tentang *gadget*, informasi yang diperoleh dapat berupa atribut produk *gadget*, hingga harga dari suatu *gadget*. Twitter merupakan salah satu dari sosial media yang berperan sebagai alat bantu untuk memperkaya berbagai informasi, mulai dari informasi tentang *gadget* hingga menjadi sumber berita keluhan seseorang. Normalisasi teks adalah istilah yang digunakan untuk menyampaikan gagasan dengan mengubah format teks untuk memenuhi tujuan tertentu. Terkadang dalam sebuah *tweets* terdapat unggahan kata yang berisi kata-kata non baku atau dapat disebut kata *slang*, kata *slang* adalah ragam bahasa tidak resmi dan tidak baku yang sifatnya musiman, dipakai oleh kaum remaja atau kelompok sosial tertentu untuk komunikasi intern. Kata *slang* tersebut perlu dilakukan normalisasi yang mana langkah awalnya dengan cara me-reduksi setiap kata yang memiliki imbuhan menjadi kata yang seragam, yang bertujuan agar dapat diproses pada pemrosesan selanjutnya. Pada Tugas Akhir ini, penulis membangun sistem untuk menormalisasi kata *slang* dari tweets produk *gadget*. Proses normalisasi teks menggunakan model *word2vec* untuk mencari kata formal dengan *similarity* tertinggi terhadap sebuah kata *slang*. Hasil normalisasi dievaluasi pada sebuah task klasifikasi yang akan mengelompokkan sentiment *tweets* ke dalam 3 kelas, yaitu: Positif, Negatif, dan Netral. Hasil pengujian menunjukkan bahwa terdapat peningkatan akurasi klasifikasi pada data yang sudah dinormalisasi, dengan nilai akurasi sebesar 91%, dibandingkan dengan dataset tanpa normalisasi, dengan nilai akurasi sebesar 88%.

Kata kunci: *gadget*, kata *Slang*, *Slang List*, normalisasi teks, korpus, *word2vec*

Abstract

Social media is a tool to enrich information about gadgets, the information obtained can be in the form of gadget product attributes, to the price of a gadget. Twitter is one of the social media that acts as a tool to enrich various information, ranging from information about gadgets to being a source of news for someone's complaints Text normalization is a term used to convey ideas by changing the format of the text to fulfill a specific purpose. Sometimes in a tweet there are uploads of words that contain non-standard words or can be called slang words, slang words are a variety of informal and non-standard languages that are seasonal in nature, used by teenagers or certain social groups for internal communication. in which the *slang* word needs to be normalized, the initial step is to reduce every word that has an affix to a assorted word, which aims to be processed in the next processing. In this final project, the author builds a system to normalize slang words from tweets of gadget products. The text normalization process uses the word2vec model to find the formal word with the highest similarity to a slang word. The results of normalization are evaluated in a classification task that will classify sentiment tweets into 3 classes, namely: Positive, Negative, and Neutral. The test results show that there is an increase in classification accuracy in normalized data, with an accuracy value of 91%, compared to datasets without normalization, with an accuracy value of 88%.

Keywords: *gadget*, *slang words*, *Slang List*, text normalization, corpus, *word2vec*