INTISARI

KLASIFIKASI KEPRIBADIAN PENGGUNA TWITTER BERDASARKAN DATA *TWEET* MENGGUNAKAN *DEEP LEARNING*

Oleh

Rian Apriansyah

17/418662/PPA/05446

Total pengguna aktif media sosial mencapai hampir separuh dari total populasi dunia. Banyak hal yang telah tercipta dari penggunaan media sosial tersebut, salah satunya adalah data linguistik. Peneliti mengungkap bahwa terdapat keterkaitan antara aspek linguistik terhadap kepribadian seseorang. Pemanfaatan data linguistik tersebut untuk keperluan *screening* kepribadian telah banyak dilakukan, namun terbatas pada penutur bahasa Inggris. Padahal, identifikasi awal kepribadian seseorang menjadi hal yang penting untuk bidang-bidang tertentu seperti rekrutmen karyawan maupun seleksi masuk sekolah. Untuk itu dibutuhkan suatu mekanisme yang dapat melakukan klasifikasi kepribadian berdasarkan data linguistik media sosial untuk penutur bahasa Indonesia.

Penelitian yang dilakukan adalah membangun model klasifikasi kepribadian pengguna Twitter berdasarkan data *tweet* berbahasa Indonesia menggunakan model arsitektur *deep learning* CNN dan LSTM. Klasifikasi yang dilakukan adalah *multi-label classification* ke dalam lima kelas kepribadian Big-Five. Teks linguistik direpresentasikan ke dalam bentuk vektor menggunakan dua *word embedding* yaitu Word2Vec dan fastText yang dilatih menggunakan korpus berbahasa Indonesia. Hasil klasifikasi CNN dan LSTM dengan masing-masing Word2Vec dan fastText dibandingkan hasilnya untuk melihat kombinasi terbaik dalam melakukan klasifikasi kepribadian.

Pengujian dilakukan dengan menggunakan 15 data pengguna Twitter, masing-masing dengan data *tweet* sebanyak 5 *tweet* terakhir dalam bahasa Indonesia. Hasil pengujian pada kedua model menunjukkan bahwa kombinasi model LSTM dan fastText dengan dimensi 50 mengungguli CNN dan Word2Vec maupun fastText. LSTM mendapat hasil dengan nilai akurasi sebesar 70%, *alpha evaluation* sebesar 82% dan *f-measure* sebesar 81%.

Kata Kunci: *Deep learning*, Klasifikasi teks, *word embedding*, Kepribadian, Big-Five, Twitter, linguistik

ABSTRACT

TWITTER USERS PERSONALITY CLASSIFICATION BY TWEET DATA USING DEEP LEARNING

by

Rian Apriansyah

17/418662/PPA/05446

The number of social media active users reaches almost half of the global population. Linguistic data is one of the most generated contents of social media usages. The study said that there is a correlation between linguistic aspect with personality. The linguistic data utilization for personality screening purpose has been widely used but limited only to English speakers. In Indonesia itself, early personality identification is an important thing for several purposes such as employee recruitment or student entry selection. Therefore, a certain mechanism is needed to perform a personality classification by linguistic data from social media specifically for Indonesian speaker.

This study aims to address the problem by building a twitter users personality classification model by tweet data using deep learning architectures CNN and LSTM. Word2Vec and fastText were trained using the Indonesian language corpus to be able to represent words into vector. The classification result for CNN and LSTM with Word2Vec and fastText respectively were compared to see the best combination in terms of classifying the personality.

The testing was done by utilizing 15 data of Twitter users, each with their 5 latest Indonesian-tweets along with the big-5 personality labels. The testing result for both models shows that the LSTM and fastText 50 combinations outperformed all the CNN combinations. The LSTM with fastText 50 has the accuracy of 70%, alpha-evaluation of 82% and f-measure of 81%.

Keywords: *deep learning, text classification, word embedding, personality, twitter, linguistic data*