INTISARI

PERBANDINGAN METODE *NAMED ENTITY RECOGNITION* UNTUK PREDIKSI GEOLOKASI INFORMASI LALU LINTAS DI MEDIA SOSIAL

Oleh:

Firnanda Akmal Subarkah

19/448700/PPA/05783

Informasi geolokasi dari data media sosial seperti Twitter telah membuka banyak peluang pengembangan aplikasi berbasis geolokasi seperti *location-based sentiment analysis*, *tourism analysis* dan identifikasi lokasi bencana. Walau demikian, ketersedian data informasi keadaan lalu lintas dengan informasi geolokasi (*geotagged data*) masih sangat terbatas. Prediksi geolokasi pada data *non-geotagged* menjadi solusi untuk masalah tersebut.

Pada penelitian ini, sebuah model prediksi geolokasi informasi lalu lintas dengan pendekatan *named entity extraction* untuk mengolah data teks berbahasa Indonesia. Terdapat tiga tahap proses yang digunakan dalam model prediksi geolokasi. Tahap pertama yaitu *Part-of-Speech Tagging* (PoS Tagging) untuk mengekstrak kata unik dari input teks, tahap kedua yaitu *Named Entity Recognition* (NER) untuk mengenali tipe entitas setiap kata unik dan tahap ketiga yaitu *Geocoding* untuk mengkonversi *location indicative word* menjadi koordinat (*latitude* dan *longitude*).

Berdasarkan eksperimen dan evaluasi yang telah dilakukan dengan menggunakan metode *Cosine Similarity* untuk menghitung tingkat kemiripan antara entitas 3 (tiga) model NER dengan nama lokasi aktual (*gold standard*) ditunjukkan model *Naïve Bayes* lebih unggul dari model lainnya, dan evaluasi kedua dengan *Haversine Distance* untuk menghitung selisih jarak antara hasil geolokasi dari entitas 3 (tiga) model NER dengan geolokasi lokasi aktual (*gold standard*) menunjukkan nilai total dan rata-rata selisih jarak model *Logistic Regression* mendapatkan nilai lebih kecil dibandingkan Naïve Bayes dan SVM.

Kata Kunci: media sosial, lalu lintas, prediksi geolokasi, *named entity recognition.*

ABSTRACT

COMPARISON OF NAMED ENTITY RECOGNITION METHOD FOR GEOLOCATION PREDICTION OF TRAFFIC INFORMATION ON SOCIAL MEDIA

By:

Firnanda Akmal Subarkah

19/448700/PPA/05783

Geolocation information from social media such as Twitter has opened up many opportunities for geolocation-based application development, such as location-based sentiment analysis, tourism analysis, and identification of disaster locations. However, traffic condition information data with geotagged data still needs to be improved. Geolocation prediction on non-geotagged data is a solution to this problem.

This study proposes a geolocation prediction model for traffic state information using a named entity extraction approach to process Indonesian language text data. Three stages of the process are used in the proposed geolocation prediction model. The first stage is Part-of-Speech Tagging (PoS Tagging) to extract unique words from text input, and the second stage is Named Entity Recognition (NER) to recognize the entity type for each unique word. The third stage is Geocoding to convert location-indicative words into coordinates (latitude and longitude).

Based on experiments and evaluations conducted using Cosine Similarity method to calculate the similarity level between the entities of the 3 (three) NER models and the actual location names (gold standard), Naïve Bayes model demonstrated superiority over the other models. Additionally, in the second evaluation using Haversine Distance, Logistic Regression model showed smaller total and average distance differences compared to Naïve Bayes and SVM models.

Keywords: social media, traffic, geolocation prediction, *named entity recognition.*