# ABSTRACT

Geolocation information from social media data like Twitter has opened many opportunities for geolocation-based application development such as location-based sentiment analysis, tourism analysis and disaster or accident location identification. However, the availability of data with geolocation information (geotagged data) is still very limited. Geolocation prediction on non-geotagged data becomes the solution to solve the problem. Unfortunately, the availability of named entity extraction tools to support the geolocation prediction process for Indonesian text data is still not publicly available yet.

In this research, a geolocation prediction model with named entity extraction approach is proposed to process text data in Indonesian language. There are three stages of process that is used in the proposed geolocation prediction model. The first stage is Part-of-speech Tagging (PoS Tagging) to extract the unique word from the input text, the second stage is Named Entity Recognition (NER) to recognize the entity type of each unique word and the third stage is Geocoding to convert location indicative word into coordinates (latitude and longitude).

Based on the experiments and evaluations conducted by using Haversine Formula, the proposed model can predict the geolocation information of Tweet with accuracy value of 35.80% within 5KM from the actual location and 49.58% within 160KM from the actual location.

**Keywords :** geolocation prediction, geotagged data, named entity extraction, named entity recognition, location indicative word extraction, social media.

# INTISARI

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 atau kecelakaan. Walau demikian, ketersedian data dengan informasi geolokasi (*geotagged data*) masih sangat terbatas. Prediksi geolokasi pada data *non-geotagged* menjadi solusi untuk masalah tersebut. Namun sayangnya, ketersedian sumber daya *named entity extraction* untuk mendukung proses prediksi geolokasi untuk data teks Bahasa Indonesia masih belum tersedia secara umum.

Pada penelitian ini, sebuah model prediksi geolokasi dengan pendekatan *named entity extraction* diajukan untuk mengolah data teks berbahasa Indonesia. Terdapat tiga tahap proses yang digunakan dalam model prediksi geolokasi yang diajukan. 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 dilakukan dengan menggunakan *Haversine Formula*, teknik yang diajukan dapat memprediksi informasi geolokasi *Tweet* dengan nilai akurasi sebesar 35.80% pada kisaran *error tolerance* 5KM dari lokasi aktual dan 49.58% pada kisaran *error tolerance* 160KM dari lokasi aktual.

**Kata kunci –** prediksi geolokasi, data *geotagged*, *named entity extraction, named entity recognition*, ekstraksi *location indicative word*, media sosial.