

VISUALISASI SEBARAN PENGGUNA BAHASA DAERAH DI INDONESIA BERBASIS DATA CROWDSOURCED TWITTER

Didit Setiawan

Email: diditsetiawan192@mail.ugm.ac.id

INTISARI

Geovisualisasi merupakan metode dalam mempresentasikan data geospasial. Salah contoh bentuk geovisualisasi yaitu peta. Peta dibuat untuk memudahkan pengguna dalam melakukan analisis hubungan spasial antar fenomena di permukaan bumi. Sumber pembuatan peta terdiri atas 6 jenis, meliputi survei terestris, data statistik, fotogrametri, citra satelit, digitasi peta, serta sensus. Seiring perkembangan zaman, kini muncul sumber data baru yaitu *big data*. *Big data*, terutama *big data* media sosial dapat dimanfaatkan untuk pemetaan berbagai fenomena, salah satunya adalah pemetaan pengguna bahasa daerah. Kelebihan penggunaan *big data* untuk pemetaan ini yaitu lebih efisien serta hemat waktu, biaya, dan tenaga. Berdasarkan hal tersebut, kemudian dilakukan penelitian ini dengan tujuan untuk mengekstraksi data pengguna bahasa daerah dari *big data* media sosial Twitter dan memvisualisasikannya secara kartografis dalam bentuk peta statis guna menjadi gambaran sebaran penduduk menurut penggunaan bahasa di Indonesia. Selain itu, juga dilakukan uji perbandingan data terhadap data sebaran suku sebagai data referensi.

Tahapan penelitian terdiri atas tahap ekstraksi dan visualisasi data, serta tahap analisis data. Ekstaksi data dilakukan menggunakan *library snsrape* pada *Python* sedangkan visualisasi dilakukan menggunakan peta statis. Analisis dilakukan dengan uji kesesuaian secara visual serta uji statistik menggunakan tabel kontingensi.

Hasil ekstraksi dan pengolahan data diperoleh 119.380 pengguna bahasa daerah dengan pengguna terbanyak adalah pengguna Bahasa Jawa. Visualisasi dapat dilakukan dengan metode *choropleth map*, *proportional symbol map*, dan *dot map*. Hasil analisis menunjukkan bahwa data pengguna bahasa daerah yang diekstrak dari media sosial Twitter mempunyai kesesuaian tinggi dengan data populasi suku hasil sensus penduduk tahun 2010 menurut perhitungan *overall accuracy* dan mempunyai kesesuaian sedang menurut perhitungan kappa.

Kata kunci: Geovisualisasi, *Big data*, Twitter, Bahasa Daerah

VISUALIZATION OF THE DISTRIBUTION OF LOCAL LANGUAGE USERS IN INDONESIA BASED ON CROWDSOURCED TWITTER DATA

Didit Setiawan

Email: diditsetiawan192@mail.ugm.ac.id

ABSTRACT

Geovisualization is a method of presenting geospatial data. One example of a form of geovisualization is a map. Maps are made to make it easier for users to analyze spatial relationships between phenomena on the earth's surface. The sources for making maps consist of six types: terrestrial surveys, statistical data, photogrammetry, satellite imagery, map digitization, and censuses. Along with the times, a new data source has appeared, namely big data. Big data, especially social media big data, can be used to map various phenomena, one of which is the mapping of local language users. The advantages of using big data for this mapping is that it's more efficient and saves time, cost, and effort. Based on this, this research was then carried out with the aim of extracting local language users data from Twitter's social media big data and visualizing it cartographically in a static map to illustrate the distribution of the population according to language use in Indonesia. In addition, a data comparison test was also carried out on ethnic distribution data as reference data.

The research stages consist of the extraction and data visualization stages, and the data analysis stages. Data extraction uses the snsrape library in Python, while visualization uses static maps. Analysis was carried out by visual comparison test and statistical test using contingency table.

The results of data extraction and processing obtained 119.380 local language users with the most users being Javanese Users. Visualization can be done by the method of choropleth map, symbol proportional map, and dot map. The results of the analysis show that local language user data extracted from social media Twitter has high compatibility with ethnic population data from the 2010 population census according to overall accuracy calculations and has moderate suitability according to kappa calculations.

Keywords: Geovisualization, Big Data, Twitter, Local Languages