Ahmad Alif Robit Al Hazmi. 2019. Klasterisasi dan Geovisualisasi *Tweet* dengan Algoritma K-Means untuk Kasus Penyebaran Penyakit Menular Langsung (Studi Kasus Tuberkulosis dan Diare). Skripsi ini di bawah bimbingan Ira Puspitasari, S.T., M.T., Ph.D. dan Taufik, S.T., M.Kom. Program Studi S1 Sistem Informasi. Fakultas Sains dan Teknologi, Universitas Airlangga.

ABSTRAK

Berkembangnya media sosial mendorong munculnya situs *microblogging* seperti Twitter yang dapat berperan sebagai sistem sensor dengan menjadikan penggunanya sebagai sensor dan *tweet* yang memiliki informasi tentang suatu lokasi. Volume dan kecepatan *tweet* pada saat suatu kejadian berlangsung sangat tinggi sehingga masyarakat yang terdampak dan petugas profesional mengalami kesulitan dalam pemrosesan informasi. Pada penelitian ini dilakukan klasterisasi *tweet* dengan algoritma K-Means untuk kasus penyebaran penyakit menular langsung (studi kasus tuberkulosis dan diare). Penelitian ini menggunakan data teks dari media sosial Twitter dengan kata kunci tentang penyebaran penyakit menular langsung (studi kasus tuberkulosis dan diare). Data hasil klasterisasi divisualisasikan untuk menerapkan geovisualisasi *tweet*. Sebelum proses klasterisasi, dilakukan tahap praproses seperti *tokenizing*, normalisasi kata, penghapusan *stopwords*, dan *stemming* sehingga mengubah data *tweet* menjadi *Term Document Matix* (TDM). Proses geovisualisasi dilakukan menggunakan Plotly, sebuah *tools* analitik dan visualisasi data secara daring**.** Tahap klasterisasi dengan K-Means pada TDM memperoleh klaster terbaiknya pada klaster dengan jumlah *term* sebesar 20 dan nilai *k* sebesar 6 sehingga menghasilkan nilai *Sum Squared Error* sebesar 7.652,71 dan koefisien silhouette sekitar 0,41. Geovisualisasi dari hasil klasterisasi terkait penyebaran penyakit menular langsung (studi kasus tuberkulosis dan diare) menyebar hampir di seluruh Indonesia**.** Berdasarkan kemunculan *term* pada setiap klaster, *term* terkait diare tersebar pada semua klaster karena memiliki jumlah *tweet* yang lebih dominan dibandingkan dengan *tweet* terkait tuberkulosis.

Kata kunci: klasterisasi, *tweet*, geovisualisasi

Ahmad Alif Robit Al Hazmi. 2019. Clustering and Geovisualization of Tweet using K-Means Algorithm for Dissemination Issues of Infectious Disease with Direct Contact (A Case Study of Tuberculosis and Diarrhea). This thesis supervised by Ira Puspitasari, S.T., M.T., Ph.D. and Taufik, S.T., M.Kom. Bachelor of Computer Science (Information System). Faculty of Science and Technology, Universitas Airlangga.

ABSTRACT

The proliferation of social media fosters the microblogging sites such as Twitter as a sensor system with their users acting as sensors and their tweet conveying information with a geographic location. Volume and velocity of tweets posted during events today tend to be extremely high, making it hard for event-affected communities and professional responders to process the information in a timely manner. This study aims to cluster tweet data using the K-Means algorithm for dissemination issues of infectious disease with direct contact (a case study of tuberculosis and diarrhea). This study uses text data from Twitter with keywords related to dissemination issues of infectious disease with direct contact (a case study of tuberculosis and diarrhea). Clustering result data is visualized to implements tweet geovisualization. Before the clustering process, preprocessing steps will be performed such as tokenizing, word normalization, stopwords removal, and stemming so that transform tweet data into Term Document Matrix (TDM). Geovisualization process is performed by using Plotly, an online data analytics and visualization tool. Clustering step with K-Means algorithm on TDM uses 20 as the best number of term and 6 as the best *k* value so that Sum Squared Error (SSE) value is 7,652.71 and silhouette coefficient is 0.41. Geovisualization from clustering result for dissemination issues of infectious disease with direct contact (a case study of tuberculosis and diarrhea) spreads out to all over Indonesia. Based on term frequency from each cluster, terms which related to diarrhea spread to all cluster because they have more dominant tweet frequency than tweets which related to tuberculosis.

Keywords: clustering, tweet, geovisualization