ABSTRACT

Make the world a better place is the goal of Sustainable Development Goals (SDGs), consisting of 17 goals as policy guidance from 2015 to 2030. Every country desperately needs supporting data in monitoring and evaluation of the achievement of SDGs, including Indonesia. The challenges faced with the monitoring and evaluation of SDGs achievements that have to monitor SDGs in all Indonesia regions, so the surveillance is quite difficult. In addition, the period of every activity monitoring and evaluation of SDGs is every year (365 days), making it quite difficult to conduct daily monitoring without involving a sufficiently large number of teams.

One way to get supporting data in the monitoring and evaluation of SDGs achievements is to retrieve data from the news on the official online news site. However, the process of acquisition and classification of news stories into 17 categories of SDGs manually by humans takes a very long time regarding the amount of data. For example, on an official online news site, there are about 600 news stories released every day. The data source used in This research came from the official online news site Detik.com, the data is text-shaped. Detik.com News Indexing Service presents news from 29 March 2004 to this day (about 16 years), this is a surplus that is not available on other Indonesian online news sites. With the availability of data in a long window will certainly help the analysis process of Indonesia SDGs in the past (Trace Back). This research focuses on the need to process data categorizing quickly and automatically using a computer.

To realize this goal, the architecture of a distributed classification system was developed in this research. The Naive Bayes classification algorithm is used for categorizing Indonesian-language online news headlines into 17 Goals SDGs and one non-SDGs news category as Noise Data. The distributed classification system in this study was implemented with RabbitMQ Message Broker. Furthermore, to accommodate all the results of the classification data and as a data analysis tool, this research uses the ELK Stack. As a result, this architecture achieve faster processing time of 13% and more effective CPU utilization rate of 35% when compared to centralized classification systems. Also, it was proven that the utilization of online news sources can complement SDGs Indonesia's surveillance and evaluation support data across categories (17 goals).

Keywords: Distributed Classification System, Sustainable Development Goals, Machine Learning, Naive Bayes Classifier, RabbitMQ Message Broker, Big Data Analytics

INTISARI

Mewujudkan kehidupan dunia yang lebih baik adalah tujuan dari *Sustainable Development Goals (SDGs)* yang terdiri dari *17 Goals* sebagai tuntunan kebijakan mulai tahun 2015 sampai tahun 2030. Setiap negara sangat membutuhkan data pendukung dalam melakukan monitoring dan evaluasi terhadap capaian SDGs, termasuk Indonesia. Tantangan yang dihadapi pada monitoring dan evaluasi capaian SDGs antara lain lokasi pemantauan capaian SDGs mencakup seluruh daerah di Indonesia sehingga *Surveilance* cukup sulit. Selain itu, rentang waktu setiap kegiatan monitoring dan evaluasi terhadap capaian SDGs umumnya adalah periodik setiap satu tahun sekali (365 hari), sehingga cukup sulit untuk melakukan pemantauan harian tanpa melibatkan tim dengan jumlah yang cukup besar.

Salah satu cara untuk mendapatkan data pendukung dalam monitoring dan evaluasi capaian SDGs adalah dengan mengambil data dari berita pada situs berita daring resmi. Namun, proses akuisisi dan proses klasifikasi berita-berita ke dalam 17 kategori SDGs secara manual oleh manusia membutuhkan waktu yang sangat lama terkait banyaknya data. Misalnya, pada suatu situs berita daring resmi, terdapat sekitar 600 berita yang dirilis setiap hari. Sumber data yang digunakan pada penelitian ini berasal dari situs berita daring resmi Detik.com, data tersebut berbentuk teks. Layanan indeks berita Detik.com menyajikan berita mulai dari 29 Maret 2004 sampai dengan hari ini (sekitar 16 tahun), ini merupakan kelebihan yang tidak didapatkan pada situs berita daring Indonesia yang lain. Dengan ketersediaan data dalam rentang waktu yang panjang tentunya akan sangat membantu proses analisis SDGs Indonesia di masa lampau *(Trace Back)*. Penelitian ini berfokus pada kebutuhan pemrosesan pengkategorian data secara cepat dan otomatis menggunakan komputer.

Untuk mewujudkan tujuan tersebut, suatu arsitektur sistem klasifikasi terdistribusi dikembangkan pada penelitian ini. Algoritme klasifikasi *Naive Bayes* dipakai untuk pengkategorian judul berita daring berbahasa Indonesia ke dalam *17 Goals SDGs* dan satu kategori berita *non-SDGs* sebagai *Noise Data*. Sistem komputasi terdistribusi pada penelitian ini diimplementasikan dengan *RabbitMQ Message Broker*. Selanjutnya, untuk menampung semua hasil data klasifikasi dan sebagai alat analisis data, penelitian ini menggunakan *ELK Stack*. Hasil yang dicapai oleh arsitektur sistem klasifikasi terdistribusi pada penelitian ini adalah waktu proses yang lebih cepat 13% dan tingkat pemanfaatan CPU yang lebih efektif 35% jika dibandingkan sistem klasifikasi terpusat. Selain itu, terbukti bahwa pemanfaatan sumber berita daring dapat melengkapi data pendukung monitoring dan evaluasi SDGs Indonesia di seluruh kategori (17 kategori).

Kata Kunci: Sistem Klasifikasi Terdistribusi, *Sustainable Development Goals*, *Machine Learning, Naive Bayes Classifier, Big Data Analytics*