ABSTRACT

The growth of vehicles in Indonesia by 9.56% in a year is not proportional to the growth of roads by 1.46% in a year. Besides infrastructure development, a traffic monitoring system is needed to enable traffic officers to adjust traffic flow immediately when congestion occurs.

Harnessing of social media data for traffic monitoring system is widely used. The monitoring system often use training set containing irrelevant data which burdens the system during the process of training the classification model. Prior researches in dataset reduction by Fang et al. [1] and Yang et al. [2] has high time complexity and are specialized their reduction method for kNN classification, while research by Yu et al. [3] is not suitable for text data with numerous features. Those prior researches cannot reduce training set for text classification which used SVM algorithm.

This research aims to develop a method to reduce training set of text classification which uses SVM algorithm to handle irrelevant data so that it can decrease the required computing resources to train the classification model.

In this research, training set was reduced by using text similarity coefficient Cosine, Dice, Jaccard, and Overlap with threshold varied between 0.1 to 1.0. Data that were considered similar were omitted to obtain a smaller training set.

Training set reduction by using text similarity coefficient calculation that has been developed in this research worked well. The system could reduce dataset significantly and produced better accuracy than before reduction. In order to achieve equal or better accuracy than before reduction, Dice calculation with threshold 0.4 had the most significant reduction. Using that configuration, the data were reduced by 78.06% and the accuracy increased by 0.07% to become 98.88%. The system that has been developed in this research could outperform random subset selection method which could not achieve even the same accuracy as before reduction. With the smaller training set used to train the classification model, the traffic monitoring system by harnessing social media data will be cheaper and easier to implement in limited computing resources.

Keywords: *traffic, Twitter, data set reduction, classification, machine learning*

INTISARI

Pertumbuhan jumlah kendaraan di Indonesia sebesar 9,56% setiap tahun tidak sebanding dengan pertumbuhan jalan yang hanya sebesar 1,46% setiap tahunnya. Selain pembangunan infrastruktur, diperlukan juga sistem pemantauan lalu lintas agar pihak berwenang dapat segera melakukan rekayasa lalu lintas ketika terjadi kemacetan.

Pemantauan keadaan lalu lintas dengan menggunakan data media sosial saat ini mulai banyak digunakan. Kumpulan data latih yang digunakan dalam pemantauan tersebut sering kali mengandung data yang kurang penting yang membebani sistem ketika pembuatan model klasifikasi. Penelitian sebelumnya mengenai reduksi kumpulan data latih oleh Yuan dkk. [1] dan Yang dkk. [2] mempunyai kompleksitas waktu yang tinggi dan dikhususkan untuk algoritme klasifikasi kNN, sementara penelitian oleh Yu dkk. [3] tidak sesuai dengan data teks yang mempunyai fitur yang sangat banyak. Penelitian-penelitian sebelumnya tersebut belum dapat melakukan reduksi kumpulan data latih yang sesuai dengan klasifikasi teks yang menggunakan algoritme SVM.

Penelitian ini bertujuan untuk mengembangkan metode reduksi kumpulan data latih klasifikasi teks yang sesuai untuk *Support Vector Machine* (SVM) untuk mengatasi data yang kurang penting sehingga dapat mengurangi sumber daya komputasi yang dibutuhkan ketika pembuatan model klasifikasi.

Pada penelitian ini, kumpulan data latih akan direduksi dengan menggunakan perhitungan koefisien kemiripan teks *Cosine*, Dice, Jaccard, dan *Overlap* dengan batas ambang yang divariasikan dari 0,1 hingga 1. Data yang dianggap mirip akan dihilangkan sehingga menghasilkan kumpulan data baru yang lebih kecil.

Reduksi kumpulan data dengan perhitungan kemiripan antar teks yang dikembangkan pada penelitian ini dapat bekerja dengan baik. Sistem dapat mereduksi kumpulan data dengan cukup signifikan dan akurasi yang lebih baik daripada sebelum direduksi. Untuk mencapai hasil akurasi yang lebih baik atau sama dengan sebelum direduksi, reduksi data paling signifikan ditunjukkan oleh perhitungan Dice dengan batas ambang 0,4. Dengan konfigurasi tersebut, data berkurang sebanyak 78,06% dan akurasi meningkat 0,07% menjadi 98,88%. Sistem yang dikembangkan dalam penelitian ini dapat mengungguli metode pemilihan sebagian data secara acak yang tidak dapat mencapai angka akurasi yang sama dengan atau lebih dari akurasi sebelum direduksi. Dengan berkurangnya jumlah data yang dibutuhkan untuk melatih model klasifikasi, sistem pemantauan kondisi lalu lintas menggunakan data media sosial akan menjadi lebih murah dan lebih mudah diimplementasikan dengan sumber daya komputasi yang terbatas.

Kata kunci: lalu lintas, Twitter, reduksi kumpulan data, klasifikasi, *machine learning*