Intisari

Penggunaan media sosial di Indonesia semakin meningkat. Salah satu media sosial yang banyak digunakan hingga kini adalah Twitter. Pengguna Twitter dapat melakukan *tweet*, saling membalas komentar, dan mendapatkan berita terkini secara *realtime*.Opini yang dikemukakan oleh masing-masing individu dapat bernilai baik atau buruk. Sehingga hal ini menjadi hal yang penting bagi instansi atau pihak tertentu. Banyaknya opini masyarakat yang berupa *tweet* pada Twitter dapat dimanfaatkan dalam sebuah analisis sentimen.

Analisis sentimen merupakan salah satu cabang pengetahuan dari *text mining* untuk menggali informasi secara lebih mendalam dan mengelompokkannya ke dalam sentimen positif, negatif, atau netral. Tahapan dalam analisis sentimen antara lain pengambilan dan pengumpulan data, *preprocessing,* pembobotan fitur, klasifikasi, dan evaluasi. Dari berbagai penelitian analisis sentimen berbasis teks berbahasa Indonesia yang telah dilakukan belum mempertimbangkan pentingnya sentimen netral dan data yang kaya akan kata atau frasa yang mengandung sentimen. Oleh karena itu, penelitian ini akan membandingkan pengaruh keberadaan sentimen netral serta penggunaan *dataset* yang kaya akan *tweet* mengandung sentimen terhadap kinerja algoritme klasifikasi *Support Vector Machine* (SVM) dan *Random Forest* (RF).Dalam penelitian ini digunakan*dataset* berdomain spesifik dengan topik ojek *online, dataset* berdomain umum yang terdiri dari data berbagai topik, serta *dataset* gabungan dari *dataset* spesifik dan *dataset* umum.TF-IDF digunakan sebagai metode pembobotan fitur yang dikombinasikan dengan ragam tokenisasi N-Gram(1,1), N-Gram(1,2), dan N-Gram(1,3). Klasifikasi dilakukan dengan menyilangkan data latih dan data uji dari masing-masing domain.

Berdasarkan berbagai kombinasi pengujian yang digunakan, algoritme SVM memiliki rata-rata performa tertinggi pada model klasifikasi dengan data latih dan data uji berdomain umum untuk klasifikasi dua kelas dengan nilai akurasi rerata sebesar 85,36%, *recall* rerata sebesar 85,70%, dan *f1-score* rerata sebesar 85,70%.

Kata kunci: Analisis Sentimen, Twitter, Domain Silang, *Support Vector Machine, Random Forest,* Pembobotan Fitur, N-Gram

Abstract

The use in Indonesia is increasing by time. One of the most widely used social media is Twitter. Through Twitter, users can tweet, reply to comments, and get the latest news in realtime. The opinions expressed by each user can be of good or bad value. This could be an important matter for certain agencies or parties. The number of public opinions in the form of tweets on Twitter can be used in a sentiment analysis.

Sentiment analysis is a branch of text mining knowledge to explore information more deeply and classify it into positive, negative, or neutral sentiments. Stages in sentiment analysis include data collection, preprocessing, term weighting, classification, and evaluation. From various studies on analysis of Indonesian text-based sentiments that have been conducted, it has not considered the importance of neutral sentiments and data that is rich in words or phrases that contain sentiments. Therefore, this study will compare the effect of the existence of neutral sentiments and the use of dataset that will contain tweets with rich sentiment terms towards the performance of the Support Vector Machine (SVM) and Random Forest (RF) classification algorithms. This study used datasets with specific domains on the topic of online transportation, a general domain dataset consists of data from various topics, as well as a combined dataset of specific datasets and general datasets. TF-IDF is used as a term weighting method combined with various tokenisation of N-Gram(1,1), N-Gram(1,2), and N-Gram(1,3).

Classification is done by crossing training data and test data from each domain. Based on various test combinations that have been conducted, the SVM algorithm has the highest performance average in the classification model with training data and test data with a general domain for the classification of two classes with mean accuracy of 85.36%, average recall of 85.70%, and f1 Average score of 85.70%.

Keywords : *Sentiment Analysis, Twitter, Cross Domain, Support Vector Machine, Random Forest, Feature Weighting, N-Gram*