ABSTRAK

Twitter memiliki data yang besar dan mudah didapatkan yang dapat dimanfaatkan untuk menambang informasi**.** Banyak peneliti yang menggunakan Twitter untuk mengambil pesan pendek (*tweet*) karena memiliki pengguna yang bervariasi dan jumlah data yang banyak. Salah satu pemanfaatan data Twitter tersebut adalah sistem *clustering* keluhan yang ada di kota Bandung. *Tweet* mengenai keluhan kota Bandung akan dikelompokkan berdasarkan kemiripannya sehingga lebih memudahkan pengguna untuk melihat kumpulan keluhan yang sama beserta jumlahnya. Pada tahap awal, *tweet* akan dipraproses agar dapat diolah di proses klasifikasi dan *clustering*. Selanjutnya, *tweet* akan dipilah berdasarkan kelas keluhan dan bukan keluhan menggunakan algoritma kNN untuk menunjang tahap praproses. Untuk *tweet* pada kelas keluhan akan diolah pada proses *clustering* menggunakan algoritma kMeans agar menghasilkan kelompok keluhan yang terbentuk. Pada proses *clustering*, *tweet* akan diolah sesuai kategori keluhan yaitu bukan macet dan macet serta berdasarkan rentang *tweet* diterbitkan yaitu per minggu dan per bulan. Pada proses klasifikasi didapatkan akurasi tertinggi sebesar 75,06% pada k=1. Sedangkan pada proses *clustering*, pada rentang waktu per minggu, data keluhan bukan macet menghasilkan *purity* tertinggi sebesar 0,8064 pada k=6. Sedangkan untuk data keluhan macet menghasilkan *purity* tertinggi sebesar 0,8464 pada k=13. Sementara itu, pada rentang waktu per bulan, data keluhan bukan macet menghasilkan *purity* tertinggi sebesar 0,6422 pada k=13. Sedangkan untuk data keluhan macet menghasilkan *purity* tertinggi sebesar 0,6089 pada k=29.

Kata kunci: klasifikasi, *clustering*, kNN, kMeans, *purity*

ABSTRACT

Twitter has a huge data and readily available that can be used to mine information. Many researchers are using Twitter to take a short text (tweet) because it has users are varied and vast amounts of data. One of these is the use of Twitter data clustering existing complaints system in the city of Bandung. Tweet about complaints of Bandung will be grouped by similarity that making it easier for users to see the same set of complaints and their number. In the early stages, will tweet dipraproses to be processed in the process of classification and clustering. Furthermore, the tweet will be sorted by class 'complaint' and 'not a complaint' using kNN algorithm to support the preprocessing stage. For tweet on the class 'complaint' will be processed in k-Means clustering algorithm to produce a group complaint form. In the process of clustering, the tweet will be processed according to the category of 'non-traffic jam complaint' and 'traffic jam complaint', and under a range of published tweets are per week and per month. In the classification process obtained the highest accuracy of 75.06% at k = 1. While in the process of clustering, in the span of a week, 'non-traffic jam complaint' produce the highest purity of 0.8064 at k = 6. As for the 'traffic jam complaint' produce the highest purity of 0.8464 at k = 13. Meanwhile, in the span of a month, the data 'non-traffic jam complaint' produce the highest purity of 0.6422 at k = 13. As for the 'traffic jam complaint' produce the highest purity of 0.6089 at k = 29.

Keyword: classification, *clustering*, kNN, kMeans, *purity*