INTISARI

KLASIFIKASI TWEET SPAM DAN VALID MENGGUNAKAN SELEKSI FITUR CHI SQUARE DAN ALGORITMA NAÏVE BAYES CLASSIFIER PADA TWEET BERBAHASA INDONESIA

Oleh :

Derta Isyajora Rakhman

13/356734/PA/15740

Pengguna Twitter bebas untuk mengirimkan tweet dengan isi konten yang beragam, termasuk tweet spam. Salah satu aplikasi yang memanfaatkan Twitter API adalah JalananYogya. JalananYogya adalah platform *crowdsourcing* yang mampu mengumpulkan laporan jalan rusak di Yogyakarta dari masyarakat melalui Twitter. Untuk menjaga integritas data JalananYogya, diperlukan sebuah sistem yang dapat melakukan *filtering* terhadap tweet valid dan tweet spam.

Sebagai langkah awal dalam pembuatan sistem *spam filtering*, akan dilakukan pengujian terhadap algoritma klasifikasi yang potensial. Fokus penelitian ini adalah untuk melakukan evaluasi performa algoritma Naïve Bayes Classifier untuk mengklasifikasikan tweet yang dipadukan dengan seleksi fitur Chi Square untuk meningkatkan performa. Metode yang digunakan dalam penelitian ini adalah Multinomial Naïve Bayes Classifier dan Bernoulli Naïve Bayes Classifier.

Berdasarkan pengujian yang dilakukan, akurasi terbaik dihasilkan oleh sistem yang menggunakan Naïve Bayes Classifier model Multinomial yang dipadukan dengan seleksi fitur Chi Square, yakni 95 %.

Kata Kunci : Twitter, JalananYogya, Tweet, Spam, Naïve Bayes Classifier, Chi Square

ABSTRACT

TWEET SPAM AND VALID CLASSIFICATION USING CHI SQUARE SELECTION FEATURE AND NAÏVE BAYES CLASSIFIER ALGORYTHM FOR INDONESIAN TWEETS

By :

Derta Isyajora Rakhman

13/356734/PA/15740

Twitter users have no limitation to send any tweet that contain diversified information, including spam. JalananYogya is one of the application that use that feature. JalananYogya is crowdsourcing platform that leverages community participation to report damaged roads in Yogyakarta using Twitter. To maintain JalananYogya's data integrity, system that can perform filtering on valid report and spam report is required

As a first step in making spam filtering system, potential classification algorithms will be tested. The focus of this research was to evaluate the performance of Naïve Bayes Classifier algorithm to classify tweet combined with Chi Square feature selection to improve performance. The method used by this reasearch is Multinomial Naive Bayes Classifier and Bernoulli Naïve Bayes Classifier.

Based on the tests performed, 95 % is the best accuracy produced by systems that have been built on this research. That system using Multinomial Naïve Bayes Classifier methods combined with Chi Square feature selection

Keyword: Tweet, JalananYogya, Tweet, Spam, Naïve Bayes Classifier, Chi Square