

Analisis dan Implementasi Algoritma C4.5 dan Pembobotan TF-IDF untuk Menentukan *Trending Topic* pada Media Sosial Twitter

Ridwan Rafif¹, Erwin Budi Setiawan², Isman Kurniawan³

^{1,2,3}Fakultas Informatika, Universitas Telkom, Bandung

¹ridwanrafif@students.telkomuniversity.ac.id, ²erwinbudisetiawan@telkomuniversity.ac.id,

³ismankrn@telkomuniversity.ac.id,

Abstrak

Twitter merupakan media sosial *microblogging* dengan perkembangan tercepat diantara semua media sosial. Pesan yang disematkan pada *twitter* merupakan kejadian yang sedang terjadi. Dengan banyaknya penggunaan *twitter* saat ini, banyaknya jumlah *tweet* yang disematkan setiap harinya dapat dikelompokkan menjadi sebuah *trending topik* yang menggambarkan berita yang banyak dibicarakan pada saat ini. Permasalahannya adalah bagaimana *tweet* yang disematkan ini dapat menjadi sebuah berita yang dibicarakan banyak penggunanya dengan *mapping* atau pembagian kelas untuk setiap *tweet* yang disematkan dan bagaimana sebuah *tweet* dapat dijadikan kedalam satu kelas tertentu. Pada penelitian ini, penulis membangun sistem untuk mengklasifikasi *trending topik* dan menganalisa *trending topik twitter* apa yang kemungkinan muncul menggunakan algoritma C4.5 dengan metode pembobotan TF-IDF. Hasil terbaik yang diperoleh dengan akurasi sebesar 56 %.

Kata Kunci: TF-IDF, *Twitter*, C4.5, *Trending Topik*

Abstract

Twitter is the fastest growing microblogging social media among all social media. Messages embedded in Twitter are events that are happening. With the current number of Twitter uses, the number of tweets embedded with each payment can be recognized as a trending topic that illustrates the news that is widely discussed at the moment. The problem is how these embedded tweets can be the news that many users talk about by mapping or class divisions for each tweet embedded and how tweets can be used for one specific class. In this final project research, the author makes a system to classify trending topics and analyze what Twitter trending topics appear using the C4.5 algorithm with the TF-IDF weighting method. The best results with an accuracy of 56%.

Keywords: TF-IDF, *Twitter*, C4.5, *Trending Topic*

1. Pendahuluan

Teknologi informasi yang dengan cepat berkembang membuat media sosial menjadi konsumsi yang banyak digunakan pengguna internet seperti *Facebook*, *Twitter*, *Instagram*, *path*, dan lainnya. *Twitter* saat ini menjadi media sosial dengan pengguna terbanyak yaitu sebanyak 140 juta pengguna dan memproduksi 400 juta *tweet* setiap harinya[1]. Berdasarkan dari data yang dirilis oleh *world of tweets dot com* pada tahun 2010, Indonesia menjadi peringkat 3 dunia dengan *tweet* sebanyak 13.39% dari total pengguna *twitter* dunia dan menjadi yang terbanyak di asia dengan aktifitas *twitter* 59.97%[2].

Twitter sangat cocok untuk penerbitan dan penyebaran informasi berita. Teks *tweet* pendek, sehingga mudah bagi pengguna untuk memposting *tweet* mereka dengan ponsel atau alat lainnya. Akibatnya, pengguna dapat mempublikasikan apa yang mereka lihat secara *real-time*, terutama untuk informasi berita. Fungsi "*following*" membuatnya sangat tepat untuk difusi informasi berita. Saat pengguna berpartisipasi dalam diskusi topik berita, *tweet* terkait akan didorong ke pengikutnya. Jika para pengikutnya juga tertarik dengan topik tersebut, mereka dapat mengambil bagian dalam diskusi juga. Dengan demikian, topik berita menyebar dengan sangat cepat.[3]

Trending topik adalah fitur yang unik pada media sosial *twitter* yang dapat dijadikan parameter mengenai berita apa yang sedang terjadi saat ini. Penggunaan fitur unik *twitter* ini menjadi alasan penggunaan data *trending* untuk melihat apakah data yang berupa *trending topik* ini dapat memberikan sebuah informasi.

Pada penelitian sebelumnya [4], masalah yang diangkat, *tweet* yang menjadi *trending topik* harus merujuk dari beberapa *tweet* sebelumnya, pada kasus *boone logan*, harus ada beberapa *tweet* yang merujuk kepada *pitcher boone logan* sehingga kasus tersebut menjadi *trending*, penentuan *trending topik* hanya dilakukan menggunakan kata kunci. klasifikasi *tweet* dapat dibagi menjadi beberapa *class* diantaranya opini, kejadian, berita, Algoritma untuk mengklasifikasikan teks yang dapat digunakan pada penelitian ini terdapat pada *data mining*. *Data mining* adalah proses menemukan data dan mengetahui *big data* [5],. dengan menggunakan algoritma C4.5