Prediksi Retweet Berdasarkan Konten dan Pengguna dengan Metode Classifier Selection

Muhamad Febiansyah[1], Jondri[2], Indwiarti[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]febiansyah@students.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id,
[3]indwiarti@telkomuniversity.ac.id

Abstrak

Perkembangan media sosial telah merubah cara penyebaran informasi, dengan Twitter memainkan peran utama. Penelitian ini bertujuan mengembangkan model prediksi retweet di Twitter menggunakan fitur content-based dan user-based, serta teknik oversampling untuk meningkatkan kinerja model. Hasil eksperimen menunjukkan bahwa meta learner tanpa oversampling pada fitur content-based memiliki macro average F1-score sebesar 0.52, namun dengan recall yang sangat rendah untuk kelas retweet (6%) dan F1-score 0.11. Sebaliknya, meta learner dengan oversampling pada fitur content-based memperbaiki performa dengan presisi 0.86, recall 0.77, dan F1-score 0.80 untuk retweet, dengan nilai macro average F1-score sebesar 0.82 yang menunjukan kenaikan dibandingkan dengan meta learner tanpa oversampling. Untuk model user-based, tanpa oversampling, macro average F1-score memiliki nilai 0.75 dengan keseimbangan baik antara presisi dan recall pada kelas non retweet. Setelah oversampling, model user-based mempertahankan keseimbangan yang baik dengan presisi, recall, F1-score, dan macro average F1-score masing-masing sebesar 0.88 pada kelas retweet dan non retweet. Secara keseluruhan, oversampling meningkatkan kinerja model, terutama pada fitur content-based, dengan model user-based menunjukkan performa yang paling konsisten dan baik.

Kata kunci : twitter, pemilihan pengklasifikasi, berbasis pengguna, berbasis konten

Abstract

The development of social media has changed the way information is disseminated, with Twitter playing a major role. This study aims to develop a retweet prediction model on Twitter using content-based and user-based features, as well as oversampling techniques to improve model performance. The experimental results show that the meta learner without oversampling on the content-based feature has an average macro F1-score of 0.52, but with a very low recall for the retweet class (6%) and an F1-score of 0.11. In contrast, the meta learner with oversampling on the content-based feature improves performance with a precision of 0.86, a recall of 0.77, and an F1-score of 0.80 for retweets, with a macro average F1-score of 0.82 showing an improvement compared to the meta learner without oversampling. For the user-based model, without oversampling, the average macro F1-score has a value of 0.75 with a balance of both precision and recall in the non-retweet class. After oversampling, the user-based model maintains a good balance with precision, recall, F1-score, and macro average F1-score of 0.88 in the retweet and non-retweet classes, respectively. Overall, oversampling improves the model performance, especially on content-based features, with the user-based model showing the most consistent and good performance.

Keywords: twitter, classifier selection, user based, content based

1. Pendahuluan

Latar Belakang

Perkembangan media sosial mempercepat penyebaran informasi, termasuk informasi terkait COVID-19. Pada Januari 2022, pengguna aktif media sosial di Indonesia mencapai 191 juta, meningkat 12,35% dari tahun sebelumnya. Twitter menjadi salah satu platform populer dengan lebih dari 500 juta pengguna global dan 340 juta retweet setiap hari [1][2]. Melalui tweet, pengguna dapat berbagi informasi berupa foto, teks, video, dan suara secara real-time, serta memposting ulang konten dari pengguna lain [4].

Namun, tidak semua tweet terkait COVID-19 mendapatkan retweet. Membuat model prediksi untuk retweet sangat penting karena retweet berperan signifikan dalam memperluas jangkauan dan dampak dari sebuah pesan, terutama selama pandemi di mana informasi yang tepat waktu dan akurat sangat dibutuhkan. Memahami pola retweet dapat membantu dalam menyebarkan informasi kesehatan yang kritis lebih efektif, sehingga dapat mendukung upaya penanggulangan pandemi dan pengambilan keputusan oleh masyarakat. Dengan adanya model prediksi retweet yang akurat, dapat diidentifikasi faktor-faktor yang membuat suatu informasi lebih mungkin untuk tersebar luas [5].

Penelitian sebelumnya telah menggunakan berbagai metode machine learning seperti Naïve Bayes, Fuzzy, SVM, dan Decision Tree untuk prediksi retweet, namun hasilnya masih kurang memadai [6]. Kelemahan dari metode tersebut terletak pada kemampuannya yang terbatas dalam menangani kompleksitas dan variasi data, seperti interaksi pengguna, waktu posting, dan konten tweet. Misalnya, penelitian yang menggunakan Naïve Bayes sering kali mengasumsikan bahwa fitur-fitur independen, yang tidak selalu mencerminkan kenyataan. Metode