Smart Comp: Jurnalnya Orang Pintar Komputer
Volume 14, Nomor 1, Januari 2025, hlm. 123-129
Terakreditasi Sinta, Peringkat 4, SK No. 105/E/KPT/2022
DOI : 10.30591/smartcomp.v14i1.7166

P-ISSN : 2089-676X
E-ISSN : 2549-0796

Prediksi Retweet Berdasarkan Konten dan Pengguna dengan Metode Pemilihan Klasifikator

Muhamad Febiansyah[1], Jondri[2*], Indwiarti[3]
[1,2,3)]Program Studi Informatika, Fakultas Informatika, Universitas Telkom
Email: [1]febiansyah@student.telkomuniversity.ac.id, [*2]jondri@telkomuniversity.ac.id, [3]indwiarti@telkomuniversity.ac.id

(Naskah masuk: *3 Agustus 2024*, diterima untuk diterbitkan: *10 Januari 2025*)

Abstrak: *Perkembangan media sosial telah mengubah cara penyebaran informasi secara drastis. Twitter, sebagai salah satu platform utama, memiliki peran penting dalam proses ini dengan jutaan pengguna dan retweet yang terjadi setiap hari. Memahami faktor-faktor yang memengaruhi retweet sangat penting untuk berbagai keperluan, seperti pemasaran digital, penyebaran informasi, dan analisis sentimen. Penelitian ini mengembangkan model prediksi retweet pada Twitter dengan memanfaatkan fitur content-based dan user-based. Fitur content-based mencakup elemen-elemen dalam konten tweet, seperti panjang teks, penggunaan hashtag, dan keberadaan tautan. Sementara itu, fitur user-based melibatkan karakteristik pengguna, seperti jumlah pengikut, tingkat aktivitas, dan pengaruh akun. Metode classifier selection digunakan untuk memilih model terbaik yang memberikan prediksi akurat. Teknik seperti oversampling diterapkan untuk menangani ketidakseimbangan data, tantangan umum dalam analisis data media sosial. Oversampling pada data content-based memastikan representasi merata dari data minoritas, sedangkan eksplorasi pada fitur user-based meningkatkan kinerja model. Hasil eksperimen menunjukkan bahwa kombinasi meta learner dengan oversampling pada data content-based memberikan hasil yang sangat baik. Secara keseluruhan, penggunaan meta learner dan oversampling memberikan dampak signifikan terhadap akurasi model. Penelitian ini menyimpulkan bahwa pendekatan terpadu yang melibatkan fitur content-based dan user-based, didukung oleh teknik pembelajaran mesin yang tepat, dapat meningkatkan prediksi retweet secara substansial.*

Kata Kunci – *Twitter; Classifier Selection; Oversampling; Predicting; Meta Learner*

Retweet Prediction Based on Content and User Based with Classifier Selection Method

Abstract: *The evolution of social media has drastically changed the way information is disseminated. Twitter, as one of the leading platforms, plays a significant role in this process, with millions of users and retweets occurring daily. Understanding the factors influencing retweets is crucial for various purposes, such as digital marketing, information dissemination, and sentiment analysis. This study develops a retweet prediction model on Twitter by utilizing content-based and user-based features. Content-based features include elements within tweet content, such as text length, hashtag usage, and the presence of links. Meanwhile, user-based features involve user characteristics, such as the number of followers, activity levels, and account influence. The classifier selection method is used to identify the best model that provides accurate predictions. Techniques such as oversampling are applied to address data imbalance, a common challenge in social media data analysis. Oversampling on content-based data ensures more balanced representation of minority data, while exploring user-based features enhances overall model performance. Experimental results show that combining meta learners with oversampling on content-based data yields excellent outcomes. Overall, the use of meta learners and oversampling significantly improves model accuracy. This study concludes that an integrated approach involving content-based and user-based features, supported by appropriate machine learning techniques, can substantially enhance retweet prediction.*

Keywords – *Twitter; Classifier Selection; Oversampling; Predicting; Meta Learner*

1. PENDAHULUAN

Perkembangan media sosial mempercepat penyebaran informasi. Pada Januari 2022, pengguna aktif media sosial di Indonesia mencapai 191 juta, meningkat 12.35% dari tahun

sebelumnya. Twitter adalah 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 [3].

Tidak semua postingan di Twitter mendapatkan retweet. Oleh karena itu, penting untuk membangun model prediksi apakah suatu postingan akan di-retweet atau tidak, dengan menggunakan fitur seperti content-based, time-based, dan user-based [4].

Penelitian sebelumnya menggunakan metode machine learning seperti Naïve Bayes, Fuzzy, SVM, dan Decision Tree untuk prediksi retweet, namun hasilnya masih lemah [5]. Oleh karena itu, diperlukan model baru untuk prediksi yang lebih baik.

Penelitian ini akan menggunakan metode classifier selection, yang menumpuk beberapa metode machine learning untuk prediksi yang lebih akurat. Tujuannya adalah untuk menentukan apakah observasi terhadap pengguna Twitter dapat menjadi fitur untuk memprediksi penyebaran informasi dan mencari algoritma terbaik dengan memanfaatkan fitur content-based dan user-based [6].

A. Twitter

Twitter merupakan salah satu media sosial yang memungkinkan berbagai aktivitas seperti memposting foto, video, suara, dan teks, serta memposting ulang informasi dari pengguna lain [7]. Informasi di Twitter dapat menyebar dengan cepat karena adanya fitur retweet yang sering digunakan oleh pengguna. Secara struktur, fitur retweet mirip dengan penggunaan email, di mana pengguna dapat mengirim ulang email yang diterima dari orang lain. Oleh karena itu, fitur retweet memungkinkan penyebaran informasi yang lebih luas dan dapat dipahami oleh pengguna lain [8].

B. Selection Feature

Untuk melakukan penelitian, perlu dilakukan pemilihan fitur yang bertujuan untuk mendapatkan hasil prediksi yang diinginkan. Pemilihan fitur dapat dilakukan setelah pengumpulan data dan preprocessing. Fitur yang akan digunakan dalam penelitian ini adalah content-based dan user-based.

1) Content Based

Content based merupakan fitur untuk memfilter konten, dimana system akan memberikan rekomendasi kepada pengguna berdasarkan aktivitas pengguna tersebut [9]. Pemilihan yang dilakukan dengan fitur ini antara lain:

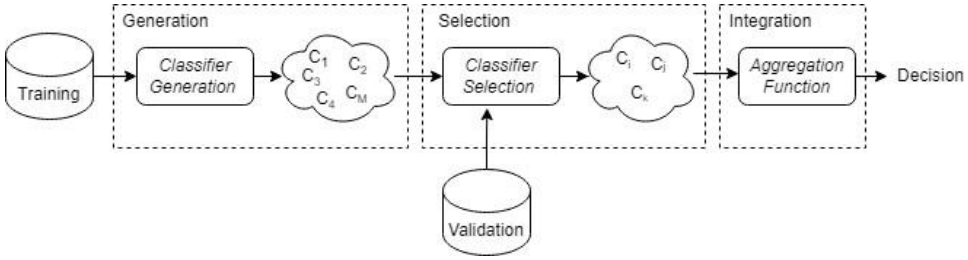
1. Contain_picture, merupakan tweet yang mengandung gambar.
2. Contain_hastag, merupakan tweet yang mengandung tag.
3. Contain_video, merupakan tweet yang mengandung video.
4. Contain_mentioned, merupakan tweet yang mengandung penyebutan pengguna lain.
5. Contain_url, merupakan tweet yang mengandung tautan url.
6. Len_of_tweet, merupakan Panjang karakter dari tweet tersebut.

2) User Based

User based merupakan fitur untuk memproses pemberian rating oleh pengguna lain terhadap suatu informasi dengan menggunakan cosine similarity antar pengguna [10]. Fitur ini didasarkan pada interaksi antar satu pengguna dengan pengguna lainnya sehingga menjadi penting untuk diperhatikan dalam penelitian. Pemilihan yang dilakukan dengan fitur ini antara lain:

1. Age_account merupakan umur dari akun pengguna twitter.
2. Avg_tweets_per_day, merupakan jumlah rata-rata tweet yang di unggah per harinya. Perhitungan ini dengan membagi total_tweets dengan age_account.
3. Total_tweets, merupakan jumlah keseluruhan tweet yang telah di unggah.
4. No_followers, merupakan jumlah pengguna lain yang mengikuti pengguna tersebut.
5. No_following, merupakan jumlah pengguna lain yang di ikuti.

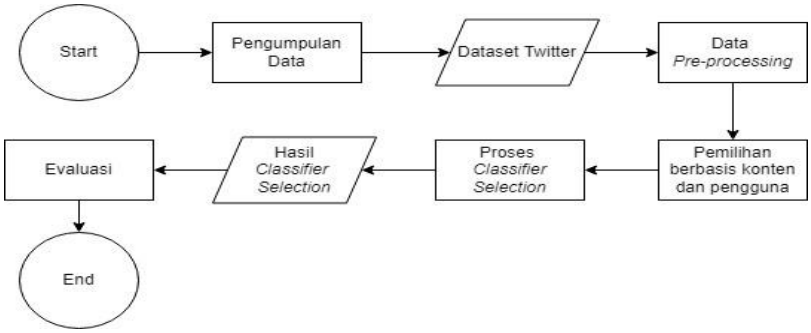
C. Clasifier Selection



Gambar 1. Arsitektur Classifier Selection

Classifier selection adalah cara untuk memilih model terbaik dalam menyelesaikan masalah. Dalam proses ini, mesin pembelajaran akan mencoba memprediksi data uji seakurat mungkin untuk hasil terbaik. Prosesnya terdiri dari tiga tahap: generasi, seleksi, dan integrasi. Pada tahap generasi, beberapa model dasar dibangun dengan berbagai strategi. Tahap seleksi melibatkan pemilihan model terbaik berdasarkan kriteria yang ditentukan menggunakan data validasi. Tahap terakhir adalah integrasi, di mana output dari model terpilih digabungkan sesuai dengan aturan yang telah ditetapkan [11].

2. METODE PENELITIAN



Gambar 2. Flowchart

Gambar 2 merupakan rancangan system dari prediksi retweet berbasis konten dan pengguna dengan metode classifier selection. Tahap ini meliputi pengumpulan data, preprocessing, pemilihan berbasis konten dan pengguna, classifier selection, dan berakhir pada evaluasi.

2.1. Dataset

Dataset berasal dari pengumpulan data didalam twitter, data yang dikumpulkan dengan menggunakan bantuan twitter API (Application Program Interface) yang telah tersedia untuk pengguna yang telah terdaftar sebagai developer twitter.

2.2. Preprocessing

Preprocessing merupakan tahap yang dilakukan setelah pengumpulan data, pada data yang akan di gunakan, perlu dilakukan penyeleksian kata yang ada pada tweets sehingga menghasilkan kata-kata yang lebih terstruktur. Preprocessing dilakukan dengan berbagai tahap, yaitu:

1. Mengatasi missing value yang ada pada dataset yang digunakan.
2. Mengecek kembali apakah data yang ada mempunyai nilai duplicate.
3. Menghapus data yang memiliki nilai duplicate
4. Mengecek apakah ada outliner pada dataset.
5. Menghapus outliner yang ada pada dataset.
6. Mengecek imbalance class, dimana 0 jmerupakan kelas yang tidak mendapatkan retweet, sedangkan 1 merupakan kelas yang mendapatkan retweet.

2.3. Classification

a. Base-Learner
Pada tingkat satu classifier selection yang nantinya akan disusun memiliki tiga metode classifikasi untuk bagian base-learner yaitu:
1) SVM
SVM (Support Vector Machine) adalah proses tipe supervisi dalam pembelajaran mesin yang menganalisis dan mengidentifikasi pola dalam data input untuk melakukan klasifikasi atau analisis regresi. SVM digunakan dalam berbagai aplikasi, seperti pengenalan angka, pengenalan tulisan tangan, deteksi wajah, klasifikasi kanker, peramalan deret waktu, dan lain-lain [12].
2) Decision Tree
Decision tree adalah model prediktif dalam pembelajaran mesin yang digunakan untuk klasifikasi atau regresi. Model ini membagi data ke dalam subset berdasarkan fitur tertentu hingga mencapai hasil akhir. Decision tree diilustrasikan sebagai struktur pohon, di mana setiap simpul internal adalah fitur, setiap cabang adalah aturan keputusan, dan setiap simpul daun adalah hasil atau label [13].
3) Logistic Regression
Logistic Regression adalah salah satu metode klasifikasi yang umum digunakan dalam analisis data. Dalam konteks Machine Learning, Logistic Regression adalah salah satu algoritma yang sering digunakan untuk masalah klasifikasi biner, di mana tujuan utamanya adalah untuk memprediksi probabilitas bahwa suatu instance tertentu termasuk dalam kelas tertentu [14].
4) Meta Learner
Meta learner adalah pendekatan dalam Machine Learning di mana algoritma mempelajari dari berbagai tugas atau dataset untuk mempercepat pembelajaran pada tugas atau dataset baru. Dalam penelitian ini, kami akan menggunakan meta-learning untuk mengoptimalkan proses klasifikasi dengan menggunakan Support Vector Machine (SVM) sebagai classifier pada tingkat kedua. Meta-learner kami akan dilatih dengan hasil prediksi dari beberapa metode dasar, sehingga dapat menghasilkan akurasi yang baik [15].

b. Eksperimen
Pada penelitian ini dilakukan eksperimen menggunakan teknik oversampling pada dataset yang mengalami ketidakseimbangan kelas. Oversampling merupakan strategi yang umum digunakan dalam pengolahan data untuk menangani masalah ketidakseimbangan kelas dengan meningkatkan jumlah sampel dari kelas minoritas [16].

c. Evaluation
Evaluasi model digunakan untuk mengevaluasi kinerja system yang telah dirancang, yang nantinya akan menggunakan binary classification metrics. Di dalam binary classification metrics terdapat berbagai macam perhitungan performasi, salah satunya adalah confusion metrix. Confusion metrix akan menghasilkan perhitungan berupa akurasi, presisi, recall, dan F1-Measure. Berikut adalah table dari confusion matrix:

Tabel 1. Confosuion Matrix

		Actual Values	
		Positive	**Negative**
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Akurasi merupakan hasil yang menunjukkan seberapa akurat sebuah system yang telah dibuat dalam melakukan klasifikasi dengan benar. Berikut rumus dari akurasi:

$$Accuracy = \frac{TP+TN}{Total\ Number\ of\ Data} \quad (1)$$

Presisi merupakan hasil yang menunjukkan perbandingan jumlah sampel yang diprediksi berada dikelas yang benar dan jumlah sampel yang diprediksi oleh sistem klasifikasi. Berikut rumus dari presisi:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall merupakan hasil yang menunjukkan rasio jumlah sampel yang diprediksi dengan benar dengan jumlah yang seharusnya diprediksi. Berikut rumus dari recall:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Measure merupakan hasil yang menunjukkan pengukuran untuk analisis kinerja klasifikasi. Berikut rumus dari F1-Measure

$$F1 - Measure = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, dibuat beberapa fungsi untuk melatih dataset, seperti SVM, Decision Tree, dan Logistic Regresion menggunakan base learner. Base learner merupakan fungsi cross validation untuk menghasilkan probabilitas dari masing masing model. Kemudian hasil prediksi akan digunakan untuk melatih fungsi meta learner, kemudian akan menghasilkan model meta learner yang telah dilatih. Lalu, terdapat fungsi classifier yang akan melatih base learner dan meta learner yang mana nanti akan dihitung hasil pemodelan dengan confusion matrix. Di percobaan awal, dilakukan beberapa eksperimen untuk mencari model terbaik diantaranya meta learner tanpa oversampling data, dan meta learner dengan oversampling data.

3.1. Meta Learner Result

Hasil dari meta learner terhadap content based menunjukkan kinerja baik dengan presisi 0.74, recall 1.00, F1 score 0.85, dan akurasi 0.76. Model ini efektif dalam mendeteksi semua retweet, meskipun ada 26% prediksi salah. F1 score tinggi menunjukkan keseimbangan baik antara presisi dan recall. Berikut adalah hasil penelitian meta learner terhadap content based:

Tabel 2. Content Based Meta Learner

	Precision	**Recall**	**F1-score**	**Support**
0	0.74	1.00	0.85	1354
1	0.00	0.00	0.00	437
Accuracy			0.74	1791

Hasil dari meta learner terhadap content based terbagi menjadi dua bagian, 0 untuk yang tidak retweet, dan 1 untuk retweet. Dalam kelas tidak retweet menghasilkan kinerja yang sangat baik, menghasilkan presisi, recall, dan f1-score dengan nilai 1.00. Tetapi pada kelas retweet menghasilkan presisi, recall, dan f1-score 0.91. hasil ini sangat efektif dalam mendeteksi semua value meskipun terdapat 3% prediksi salah. berikut adalah hasil penelitian meta learner terhadap user based:

Tabel 3. User Based Meta Learner

	Precision	**Recall**	**F1-score**	**Support**
0	1.00	1.00	1.00	715
1	0.91	0.91	0.91	22
Accuracy			0.97	738

3.2. Meta Learner With Oversampling Data Result

Setelah dilakukan oversampling, jumlah data untuk content based menjadi 3347 data dan user based menjadi 1793 data. Hasil dari meta learner with oversampling data berdasarkan content based menunjukan bahwa pada hasil tidak retweet menghasilkan presisi sebesar 0.77, recall 0.38, f1-score 0.51, dapat diartikan bahwa model ini menghasilkan presisi yang baik tetapi masih menghasilkan deteksi semua data yang sebenarnya tidak di retweet pada dataset dan tidak menghasilkan keseimbangan terhadap recall dan presisi. Pada hasil retweet, model ini menghasilkan presisi 0.57, recall 0.88, f1-score 0.69, dapat di artikan bahwa terhadap retweet, menghasilkan presisi yang kurang baik tetapi dapat mendeteksi semua data yang sebenarnya di retweet pada dataset ini, menghasilkan keseimbangan yang cukup baik. Secara keseluruhan, model ini menghasilkan akurasi 62% yang mana model ini mendeteksi value yang salah sebesar 38%. Berikut hasil penelitian meta learner with oversampling data terhadap content based:

Tabel 4. Content Based Meta Learner With Oversampling Data

	Precision	**Recall**	**F1-score**	**Support**
0	0.77	0.38	0.51	1392
1	0.57	0.88	0.69	1286
Accuracy			0.62	2678

Hasil penelitian pada meta learner with oversampling data berdasarkan user based, menunjukkan presisi, recall, f1-score, dan akurasi 1.00 terhadap konten retweet maupun tidak retweet. dapat disimpulkan bahwa model ini menghasilkan presisi, recall, f1-score, dan akurasi yang sangat baik. Berikut hasil penelitian meta learner with oversampling data terhadap user based:

Tabel 5. User Based Meta Learner With Oversampling Data

	Precision	**Recall**	**F1-score**	**Support**
0	1.00	1.00	1.00	700
1	1.00	1.00	1.00	735
Accuracy			1.00	1435

4. KESIMPULAN

Berdasarkan hasil pada chapter IV, penelitian ini menunjukkan bahwa penggunaan meta learner dan teknik oversampling dalam model prediksi retweet memiliki dampak yang signifikan terhadap kinerja model. Pada tahap awal, hasil dari meta learner tanpa oversampling menunjukkan kinerja yang baik untuk model content-based dengan presisi 0.74, recall 1.00, F1-score 0.85, dan akurasi 0.76, meskipun terdapat kekurangan dalam mendeteksi kelas retweet dengan nilai presisi, recall, dan F1-score yang nol. Untuk model user-based, hasilnya sangat memuaskan dengan akurasi 0.97 dan kinerja yang sangat baik pada kedua kelas. Setelah dilakukan oversampling, jumlah data meningkat dan hasilnya menunjukkan perbaikan dalam beberapa metrik namun juga beberapa penurunan. Pada model content-based dengan oversampling, presisi untuk kelas tidak retweet meningkat menjadi 0.77 tetapi recall menurun menjadi 0.38, dengan F1-score 0.51, menunjukkan ketidakseimbangan yang masih ada antara presisi dan recall. Namun, untuk kelas retweet, presisi 0.57, recall 0.88, dan F1-score 0.69 menunjukkan peningkatan kemampuan dalam mendeteksi retweet meskipun masih ada kekurangan dalam presisi. Model user-based dengan oversampling menunjukkan hasil yang sangat baik dengan presisi, recall, F1-score, dan akurasi 1.00 untuk kedua kelas, menandakan kinerja yang sempurna setelah oversampling. Secara keseluruhan, teknik oversampling meningkatkan jumlah data dan memperbaiki beberapa aspek kinerja model, tetapi keseimbangan antara presisi dan recall masih menjadi tantangan pada model content-based.

DAFTAR PUSTAKA

[1] M Ivan Mahdi, "Pengguna Media Sosial di Indonesia Capai 191 Juta pada 2022," Dataindonesia.ID.
[2] Z. Luo, M. Osborne, J. Tang, and T. Wang, "Who will retweet me? Finding retweeters in Twitter," in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 869–872. doi: 10.1145/2484028.2484158.
[3] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet prediction considering user's difference as an author and retweeter," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 852–859. doi: 10.1016/j.osnem.2018.04.001.
[4] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter–Analysis of predictive features," J Comput Sci, vol. 28, pp. 257–264, 2018, doi: 10.1016/j.jocs.2017.10.010.
[5] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," Front Comput Sci, vol. 14, pp. 241–258, 2020, doi: 10.1007/s11704-019-8208-z.
[6] I. Khan, X. Zhang, M. Rehman, and R. Ali, "A literature survey and empirical study of meta-learning for classifier selection," IEEE Access, vol. 8, pp. 10262–10281, 2020, Accessed: Jul. 15, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8951014
[7] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet: A popular information diffusion mechanism–A survey paper," Online Soc Netw Media, vol. 6, pp. 26–40, 2018, doi: 10.1109/ASONAM.2016.7752337.
[8] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in 2010 43rd Hawaii international conference on system sciences, IEEE, 2010, pp. 1–10. doi: 10.1109/HICSS.2010.412.
[9] F. A. Utami, "Apa Itu Content-based Filtering?," warnaekonomi.
[10] D. Nugraha, "Sistem Rekomendasi Film Menggunakan Metode User Based Collaborative Filtering," 2021.
[11] A. Sunyoto, A. Arifianto, and R. Rismala, "Evolutionary Machine Learning: Pembelajaran Mesin Otonom Berbasis Komputasi Evolusioner," 2023.
[12] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, "A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine," Procedia Comput Sci, vol. 218, pp. 818–827, 2023, doi: 10.1016/j.procs.2023.01.062.
[13] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.
[14] A. Holzinger, Machine learning for health informatics. Springer, 2016.
[15] J. Vanschoren, "Meta-learning: A survey," arXiv preprint arXiv:1810.03548, 2018.
[16] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," Appl Soft Comput, vol. 143, p. 110415, 2023.