Prediksi Retweet Berdasarkan Konten dan Pengguna dengan Metode Classifier Selection

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi S1 Informatika

Fakultas Informatika

Universitas Telkom

1301180293

Muhamad Febiansyah

Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2024

LEMBAR PENGESAHAN

Prediksi Retweet Berdasarkan Konten dan Pengguna dengan Metode Classifier Selection

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

NIM : 1301180293

Muhamad Febiansyah

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Informatika
Fakultas Informatika
Universitas Telkom

Bandung, 1 Agustus 2024
Menyetujui

Pembimbing I,	Pembimbing II,
Jondri, S.Si., M.Si.	Dra. Indwiarti, M.Si.
95700035	98690022

Ketua Program Studi
Sarjana Informatika,

Dr. Erwin Budi Setiawan, S.Si, M.T.
NIP: 007600045

LEMBAR PERNYATAAN

Dengan ini saya, Muhamad Febiansyah, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul Prediksi Retweet Berdasarkan Konten dan Pengguna dengan Metode Classifier Selection beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 1 Agustus 2024

Yang Menyatakan

Muhamad Febiansyah

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

Fuzzy dan SVM juga menunjukkan keterbatasan dalam menangkap pola non-linear dalam data, sementara Decision Tree rentan terhadap overfitting ketika digunakan pada dataset yang lebih besar dan kompleks. Oleh karena itu, diperlukan pendekatan baru yang lebih mampu menangani kompleksitas ini untuk meningkatkan akurasi prediksi retweet.

Penelitian ini akan menggunakan metode classifier selection, yang menggabungkan beberapa algoritma machine learning untuk mencapai prediksi yang lebih akurat. Base model yang dipilih adalah Support Vector Machine (SVM), Decision Tree (DT), dan Logistic Regression (LR) karena ketiganya memiliki kekuatan komplementer: SVM efektif dalam menangani data non-linear dan tinggi dimensi, Decision Tree baik dalam menangkap hubungan non-linear serta interpretasi model, dan Logistic Regression cocok untuk situasi di mana interpretabilitas model dibutuhkan serta mampu memberikan probabilitas output. Meta learner yang dipilih adalah SVM, karena kemampuannya yang kuat dalam mengklasifikasikan data kompleks setelah menerima input dari base models, sehingga dapat menggabungkan kekuatan dari ketiga model dasar tersebut untuk mencapai hasil prediksi yang lebih akurat dan andal. Dengan classifier selection ini, diharapkan dapat ditemukan algoritma yang paling efektif dalam memanfaatkan fitur content-based dan user-based, sehingga dapat meningkatkan akurasi prediksi retweet, khususnya pada tweet terkait COVID-19 [7].

Topik dan Batasannya

Penelitian ini berfokus pada prediksi retweet di Twitter dengan menggunakan fitur berbasis konten (content-based) dan berbasis pengguna (user-based). Model prediksi dibangun dengan metode classifier selection, yang mengombinasikan beberapa algoritma pembelajaran mesin. Batasan penelitian termasuk penggunaan teknik oversampling untuk menangani ketidakseimbangan kelas dan evaluasi model dengan metrik klasifikasi biner. Penelitian ini terbatas pada data yang dikumpulkan dari Twitter melalui API dan tidak mencakup faktor-faktor lain seperti sentimen atau waktu pengiriman tweet yang mungkin memengaruhi prediksi retweet.

Tujuan

Penelitian ini bertujuan untuk mengembangkan model prediksi retweet yang lebih akurat dengan memanfaatkan fitur content-based dan user-based. Penelitian ini juga bertujuan untuk mengeksplorasi efektivitas teknik oversampling dalam meningkatkan kinerja model, terutama dalam menghadapi ketidakseimbangan kelas. Dengan menggunakan pendekatan meta learner dan *classifier selection*, penelitian ini bertujuan untuk menemukan kombinasi algoritma yang optimal untuk memprediksi retweet, serta mengevaluasi dampak teknik oversampling terhadap kinerja prediksi.

Organisasi Tulisan

Penulisan dimulai dengan melakukan tinjauan literatur yang mencakup berbagai topik. Selanjutnya, metodologi yang digunakan dalam penelitian ini akan dijelaskan. Pada tahap berikutnya, hasil penelitian akan dievaluasi dan dibahas. Terakhir, kesimpulan dan saran akan disampaikan.

2. Studi Terkait

2.1 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 [3]. 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].

2.2 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.

2.2.1 Atribut Data

a. author: author merupakan penulis tweet.
b. deskripsi: merupakan isi dari konten.
c. lokasi: merupakan lokasi yang disebutkan pada isi tweet.
d. favorit: merupakan tweet yang dimasukan kedalam favorit.
e. jumlah_retweet: penandaan apakah tweet telah di retweet atau belum.
f. sentiment: penandaan apakan tweet memiliki sentiment negative atau positive.

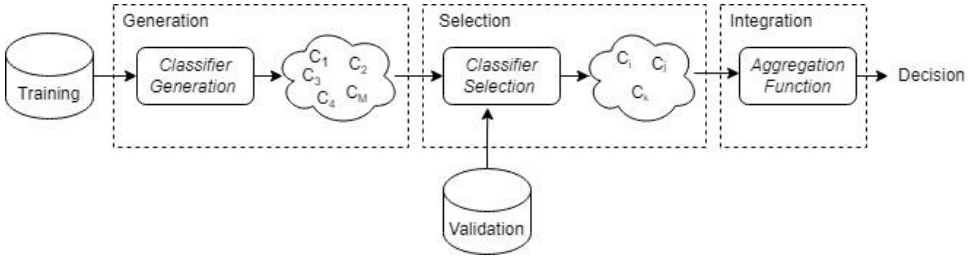
2.2.2 Content Based

Content based merupakan fitur untuk memfilter konten, dimana system akan memberikan rekomendasi kepada pengguna berdasarkan aktivitas pengguna tersebut[9]

2.2.3 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.

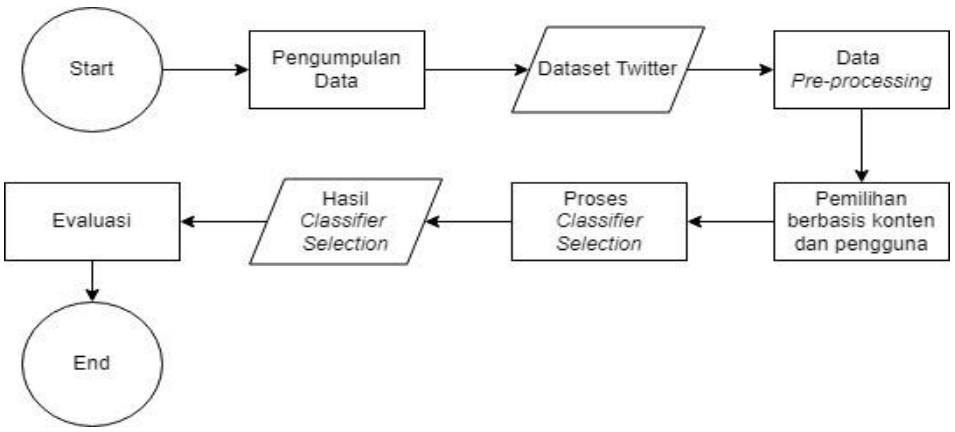
2.3 Classifier Selection



Gambar 2. 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].

3. Sistem yang Dibangun



Gambar 3. 1 Flowchart

Gambar 3.1 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.

a. Dataset

Dataset ini dikumpulkan dari Twitter menggunakan Twitter API, yang tersedia untuk pengguna terdaftar sebagai developer. Dataset ini terdiri dari 1.275 baris dan 34 kolom, mencakup berbagai informasi seperti akun pengguna, konten tweet, dan sentimen. Untuk analisis ini, dataset dibagi menjadi dua subset: 60% digunakan sebagai data train dan 40% sebagai data test. Pembagian ini memungkinkan pengujian model dengan data yang tidak terlihat sebelumnya untuk mengevaluasi kinerjanya.

b. 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.

c. Classification

1) Base Learner

Pada tingkat satu classifier selection yang nantinya akan disusun memiliki tiga metode classifikasi untuk bagian base-learner yaitu:

a) 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].

b) 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].

c) 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].

2) 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].

d. Eksperimen

Dalam penelitian ini, eksperimen dilakukan dengan menerapkan teknik oversampling pada dataset yang mengalami ketidakseimbangan kelas. Oversampling adalah strategi yang sering digunakan dalam pengolahan data untuk mengatasi masalah ketidakseimbangan kelas dengan meningkatkan jumlah sampel dari kelas minoritas[16]. Pada kode yang digunakan, teknik SMOTE (Synthetic Minority Over-sampling Technique) diterapkan untuk menghasilkan sampel sintetis dari kelas minoritas, sehingga menciptakan keseimbangan yang lebih baik antara kelas-kelas dalam dataset. Setelah proses oversampling, dataset dibagi menjadi set pelatihan dan set pengujian dengan proporsi 60% untuk pelatihan dan 40% untuk pengujian. Pembagian ini memastikan bahwa model yang dilatih dapat belajar dari data yang lebih seimbang antara kelas minoritas dan kelas mayoritas. Distribusi label setelah oversampling ditampilkan untuk memverifikasi bahwa jumlah instance dari setiap kelas telah diperbaiki sesuai dengan tujuan oversampling.

e. Evaluasi

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.

4. Evaluasi

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. Confusion Matrix

		Actual Value	
		Positive	Negatif
Predicted Value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

a. Akurasi

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

$$Accuracy = \frac{TP + TN}{Total\ Number\ of\ Data}$$

b. Presisi

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 :

$$Precision = \frac{TP}{TP + FP}$$

c. Recall

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

$$Recall = \frac{TP}{TP + FN}$$

d. F1-Measure

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

$$F1 - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

4.1 Hasil Pengujian

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.

a. Meta Learner Result

i. Content Based

Hasil evaluasi dari meta learner terhadap metode berbasis konten terbagi menjadi dua kategori: 0 untuk non-retweet dan 1 untuk retweet. Evaluasi terhadap model berbasis konten menunjukkan kinerja yang cukup baik dengan macro average F1-score 0.52. Meskipun model ini memiliki nilai precision yang tinggi untuk kelas non-retweet (0.88), model ini sangat kurang efektif dalam mendeteksi retweet (kelas 1), dengan recall yang sangat rendah yaitu hanya 0.06. Artinya, model hanya mampu mendeteksi 6% dari seluruh instance retweet yang ada, sementara 94% retweet tidak terdeteksi dengan benar. Selain itu, nilai F1-score yang tinggi (0.93) hanya tercapai pada kelas non-retweet, sedangkan F1-score untuk kelas retweet hanya sebesar 0.11. Nilai macro average menunjukkan bahwa meskipun

model ini efektif dalam mendeteksi kelas non-retweet, performa pada kelas retweet sangat buruk, menghasilkan ketidakseimbangan yang signifikan antara presisi dan recall. Macro average memberikan gambaran yang lebih komprehensif tentang kinerja model pada kedua kelas secara keseluruhan, dan dalam hal ini, menunjukkan bahwa model mengalami kesulitan yang signifikan dalam mendeteksi retweet. Oleh karena itu, meskipun akurasi keseluruhan tampak baik, macro average menyoroti ketidakseimbangan yang signifikan dalam kinerja model, terutama dalam mendeteksi retweet. Berikut adalah hasil penelitian meta learner terhadap content based:

Tabel 2. Content Based Meta Learner

	Precision	Recall	F1-Score	Support
0	0.88	1.00	0.93	3448
1	0.71	0.06	0.11	503
Macro Avg F1-Score			0.52	3951

ii. User Based

Hasil evaluasi dari meta learner terhadap metode berbasis user-based terbagi menjadi dua kategori: 0 untuk non-retweet dan 1 untuk retweet. Evaluasi terhadap model user-based menunjukkan kinerja yang sangat baik dengan macro average F1-Score sebesar 0.75. Model ini berhasil mencapai nilai precision 0.94, f1-score dengan nilai 0.96, dan recall 0.99 untuk kelas non-retweet, yang menunjukkan bahwa model sangat efektif dalam mendeteksi hampir semua instance non-retweet dengan tingkat kesalahan prediksi yang sangat rendah, yaitu hanya sekitar 1%. Namun, untuk kelas retweet, model menunjukkan performa yang kurang optimal, dengan precision 0.84, recall 0.40, dan F1-score 0.54. Meskipun precision untuk kelas retweet relatif baik, recall yang rendah menunjukkan bahwa model hanya mampu mendeteksi 40% dari seluruh instance retweet yang ada, sementara 60% retweet tidak terdeteksi dengan benar. Macro average memberikan gambaran menyeluruh tentang kinerja model di kedua kelas, mencerminkan keseimbangan performa yang lebih baik meskipun terdapat perbedaan signifikan antara deteksi kelas non-retweet dan retweet. Secara keseluruhan, model user-based sangat andal dalam mengidentifikasi non-retweet, tetapi masih membutuhkan perbaikan dalam mendeteksi retweet. Berikut adalah hasil penelitian meta learner terhadap user-based:

Tabel 3. User Based Meta Learner

	Precision	Recall	F1-Score	Support
0	0.94	0.99	0.96	1815
1	0.84	0.40	0.54	1286
Macro Avg F1-Score			0.97	738

b. Meta Learner With Oversampling Data Result

Setelah dilakukan oversampling, jumlah data untuk content based menjadi 6914 data dan user based menjadi 3617 data.

i. Content Based

Hasil evaluasi dari meta learner dengan oversampling data terhadap metode berbasis konten menunjukkan perbaikan kinerja yang signifikan. Untuk kelas non-retweet, model ini menghasilkan precision sebesar 0.79, yang menunjukkan bahwa 79% dari prediksi yang diklasifikasikan sebagai non-retweet adalah benar. Nilai recall sebesar 0.87 menunjukkan bahwa model mampu mendeteksi 87% dari instance non-retweet yang sebenarnya. Meskipun F1-score sebesar 0.83 pada kelas non-retweet menunjukkan kinerja yang baik, masih terdapat ketidakseimbangan antara presisi dan recall, seperti yang tercermin dari nilai precision dan recall. Untuk kelas retweet, model menunjukkan performa yang lebih baik dengan precision sebesar 0.86 dan recall sebesar 0.77. Ini mengindikasikan bahwa model lebih efektif dalam mendeteksi retweet dibandingkan dengan non-retweet, dengan F1-score sebesar 0.81 yang menunjukkan keseimbangan yang lebih baik antara presisi dan recall.

Penerapan oversampling data telah membantu meningkatkan kinerja model dalam mendeteksi retweet dibandingkan dengan pendekatan sebelumnya. Macro average F1-score sebesar 0.82 mencerminkan keseimbangan yang baik antara performa model pada kedua kelas, menandakan bahwa oversampling data telah memberikan kontribusi positif terhadap keseimbangan antara deteksi non-retweet dan retweet. Berikut adalah hasil penelitian meta learner with oversampling data terhadap content-based:

Tabel 4. Content Based Meta Learner dengan Oversampling Data

	Precision	Recall	F1-Score	Support
0	0.79	0.87	0.83	3454
1	0.86	0.77	0.81	3460
Macro Avg F1-Score			0.82	6914

ii. User Based

Hasil penelitian yang menggunakan meta learner dengan oversampling data pada model user-based menunjukkan kinerja yang sangat konsisten dan seimbang dalam mendeteksi baik retweet (kelas 1) maupun non-retweet (kelas 0). Model ini mencapai nilai precision, recall, dan F1-score yang seragam, yaitu sebesar 0.88 untuk kedua kelas. Untuk kelas non-retweet (kelas 0), model ini memiliki precision dan recall masing-masing sebesar 0.88, menunjukkan bahwa 88% dari prediksi non-retweet adalah benar dan model berhasil mendeteksi 88% dari seluruh instance non-retweet yang ada. F1-score yang juga sebesar 0.88 menunjukkan keseimbangan yang baik antara precision dan recall, yang menunjukkan kinerja model yang konsisten dalam mendeteksi non-retweet. Untuk kelas retweet (kelas 1), model ini juga menunjukkan nilai precision dan recall sebesar 0.88, yang berarti model dapat mendeteksi 88% dari instance retweet dengan tingkat akurasi yang sama dengan kelas non-retweet. F1-score sebesar 0.88 pada kelas retweet menegaskan bahwa model ini mencapai keseimbangan yang efektif antara precision dan recall. Dengan macro average F1-score sebesar 0.88, hasil ini mencerminkan performa yang sangat baik di kedua kelas, tanpa kecenderungan bias terhadap salah satu kelas. Berikut adalah hasil penelitian meta learner dengan oversampling data terhadap user-based:

Tabel 5. User Based Meta Learner dengan Oversampling Data

	Precision	Recall	F1-Score	Support
0	0.88	0. 88	0. 88	1804
1	0. 88	0. 88	0. 88	1813
Macro Avg F1-Score			0. 88	3617

5. Kesimpulan

Berdasarkan hasil pada Bab IV, penelitian ini menunjukkan bahwa penggunaan meta learner dan teknik oversampling memiliki dampak signifikan terhadap kinerja model prediksi retweet. Pada tahap awal, meta learner tanpa oversampling menunjukkan kinerja baik pada model content-based dengan macro average F1-score sebesar 0.52. Namun, model ini menghadapi kesulitan dalam mendeteksi retweet, dengan recall yang sangat rendah dan F1-score yang tidak seimbang antara kelas retweet dan non-retweet. Sebaliknya, model user-based tanpa oversampling menunjukkan hasil lebih baik, dengan macro average F1-score sebesar 0.75 dan keseimbangan yang lebih baik antara presisi dan recall, terutama dalam mendeteksi kelas retweet.

Setelah penerapan teknik oversampling, terjadi peningkatan signifikan pada beberapa metrik, terutama dalam mendeteksi retweet. Pada model content-based dengan oversampling, presisi dan recall untuk kelas retweet meningkat menjadi 0.86 dan 0.77, dengan F1-score 0.80, menunjukkan peningkatan jelas dalam kemampuan model untuk mendeteksi retweet. Meskipun ada perbaikan, model ini masih menunjukkan ketidakseimbangan antara presisi dan recall pada kelas non-retweet.

Model user-based dengan oversampling menunjukkan hasil sangat baik, dengan presisi, recall, F1-score, dan macro average F1-score seragam sebesar 0.88 untuk kedua kelas, mencerminkan kinerja yang lebih seimbang dan andal. Secara keseluruhan, penerapan teknik oversampling pada meta learner terbukti meningkatkan kinerja model, terutama dalam mendeteksi retweet. Namun, tantangan dalam menjaga keseimbangan antara presisi dan recall masih perlu diatasi, khususnya pada model content-based.

Daftar Pustaka

[1] D. Indonesia, "DataIndonesia.id," 25 February 2022. [Online]. Available: https://dataindonesia.id/Digital/detail/pengguna-media-sosial-di-indonesia-capai-191- juta-pada-2022. [Accessed 22 April 2022].

[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.

[3] S. N. Firdaus, C. Ding and A. Sadeghian, "Retweet: A popular information diffusion mechanism--A survey paper," Online Social Networks and Media, vol. 6, pp. 26--40, 2018.

[4] 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.

[5] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter--Analysis of predictive features," Journal of computational science, vol. 28, pp. 257--264, 2018.

[6] X. Dong, Z. Yu, W. Cao, Y. Shi and Q. Ma, "A survey on ensemble learning," Frontiers of Computer Science, vol. 14, no. 2, pp. 241--258, 2020.

[7] 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.

[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.

[9] F. A. Utami, "Warta Ekonomi," 9 May 2022. [Online]. Available: https://wartaekonomi.co.id/read412507/apa-itu-content-based-filtering. [Accessed 18 May 2022].

[10] D. Nugraha, T. W. Purboyo and R. A. Nugrahaeni, "Sistem Rekomendasi Film

Menggunakan Metode User Based Collaborative Filtering," eProceedings of Engineering, vol. 8, no. 5, 2021.

[11] Suyanto, A. Arifianto, R. Rismala and A. Sunyoto, Evolutionary Machine Learning (Pembelajaran Mesin Otonom Berbasis Komputasi Evolusioner), INFORMATIKA, 2020.

[12] Behera, M. P., Sarangi, A., Mishra, D., & Sarangi, S. K. (2023). A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine. Procedia Computer Science, 218, 818-827.

[13] Abdulazeez, A. M., Brifcani, A., & Issa, A. S. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(1), 21-46.

[14] Smith, J., Johnson, M., & Williams, R. (2018). Application of Logistic Regression in Health Data Classification: A Machine Learning Approach. Journal of Health Informatics, 10(2), 87-95.

[15] Vanschoren, J., et al. (2018). Meta-Learning: A Survey. arXiv preprint arXiv:1810.03548.

[16] Islam, M. A. K., Islam, M. M., Shahriar, M. S., & Alam, M. R. (2021). A Comprehensive Review on Class Imbalance Problem: Dataset Characteristics, Oversampling Methods, and Their Effects. Journal of Machine Learning Research, 22(3), 567-589.