JURNAL PUBLIKASI ILMU KOMPUTER DAN MULTIMEDIA
Halaman Jurnal: http://ejurnal.stie-trianandra.ac.id/index.php/jupkom
Halaman UTAMA Jurnal : http://ejurnal.stie-trianandra.ac.id/index.php

PERBANDINGAN METODE *RESAMPLING* PADA *IMBALANCED DATASET* UNTUK KLASIFIKASI KOMENTAR PROGRAM MBKM

Ade Nurhopipah[a], Cindy Magnolia[b]
[a] Fakultas Ilmu Komputer/ Program Studi Informatika,
ade_nurhopipah@amikompurwokerto.ac.id, Universitas Amikom Purwokerto
[b] Fakultas Ilmu Komputer/ Program Studi Informatika,
gabriellamagnolia641@gmail.com, Universitas Amikom Purwokerto

ABSTRAK

An imbalanced dataset, a condition where the dataset is dominated by one class, is a problem commonly found in real-world applications. In this study, the problem occurs in the collected dataset to classify four types of public comments on the Merdeka Belajar Kampus Merdeka (MBKM) program. The dataset has a high Imbalanced Ratio of 5:1 and a low classification performance with an F-Measure between 0.6209 and 0.6672. The problem underlies this research objective to explore several resampling techniques to see their effect on the classification model performance. The resampling methods studied were undersampling using Near Miss and Tomek Links, oversampling using SMOTE and ADASYN, and a combination of undersampling and oversampling using Random Combination Sampling (RCS). This research used four classifiers: Random Forest, Logistic Regression, SVM and MLP, to see the stability of the resampling method effect. Based on the analysis performed, it can be concluded that the Near Miss method in this study did not have a positive effect on improving model performance. In contrast, other methods could improve classifier model performance by increasing the *F-Measure* value. The best performance was obtained in the SVM model classification on the dataset using SMOTE resampling method. After analyzing the optimization of the model and the resampling method, the maximum F-Measure value is 0.9524.

Keywords: Classification, Imbalanced Dataset, MBKM, Resampling.

Abstrak

Imbalanced dataset yaitu kondisi di mana dataset didominasi oleh salah satu kelas adalah permasalahan yang umum ditemukan dalam aplikasi di dunia nyata. Pada penelitian ini, permasalahan tersebut terjadi pada dataset yang dikumpulkan untuk klasifikasi empat jenis komentar publik terhadap program Merdeka Belajar Kampus Merdeka (MBKM). Dataset tersebut memiliki *Imbalanced Rasio* yang tinggi sebesar 5:1 dan kinerja klasifikasi yang rendah dengan *F-Measure* di antara 0,6209 sampai 0,6672. Masalah ini mendasari tujuan penelitian, yaitu mencoba mengeksplorasi beberapa teknik *resampling* untuk melihat pengaruhnya terhadap kinerja model klasifikasi. Metode *resampling* yang diteliti adalah *undersampling* dengan Near Miss dan Tomek Links, *oversampling* dengan SMOTE dan ADASYN, dan kombinasi *undersampling* dan *oversampling* dengan *Random Combination Sampling* (RCS). Penelitian ini menggunakan empat *classifier* yaitu *Random Forest*, *Logistic Regression*, SVM dan MLP untuk melihat stabilitas efek metode *resampling*. Berdasarkan analisis yang dilakukan, dapat disimpulkan bahwa metode Near Miss pada penelitian ini tidak memberikan efek positif dalam peningkatan kinerja model. Sebaliknya, metode lainnya dapat memperbaiki kinerja model *classiifier* dengan meningkatkan nilai *F-Measure*. Kinerja terbaik diperoleh pada model klasifikasi SVM dengan dataset hasil *resampling* metode SMOTE. Setelah melalui analisis optimasi model dan metode *resampling* diperoleh nilai *F-Measure* maksimal sebesar 0.9524.

Kata Kunci: Klasifikasi, *Imbalanced Dataset*, MBKM, *Resampling*.

Received Agustus 30, 2022; Revised Oktober 2, 2022; Accepted Desember 22, 2022

1. PENDAHULUAN

Program Merdeka Belajar Kampus Merdeka (MBKM) yang digagas Kementrian Pendidikan, Kebudayaan, Riset dan Teknologi Republik Indonesia adalah gagasan inovatif metode pembelajaran dengan pendekatan *active learning* yang memungkinkan mahasiswa menyesuaikan jenis dan konten pendidikan dengan minat dan bakat serta kebutuhan industri. Program ini memberikan peluang bagi mahasiswa untuk mengembangkan kemandirian dalam menemukan pengetahuan melalui realitas dan dinamika lapangan, seperti bagaimana cara menghadapi permasalahan di dunia nyata, kebutuhan peningkatan *skill*, interaksi sosial, kolaborasi, manajemen diri, tuntutan kinerja, pemenuhan target dan prestasi [1]. Namun begitu, sebagai program yang baru digagas, terdapat berbagai tantangan dalam berbagai hal, misalnya dalam aspek penerimaan mahasiswa, pengelolaan oleh penyelenggara dan juga kesiapan mitra.

Salah satu metode untuk mengevaluasi kesiapan, penerimaan, ataupun tanggapan atas penyelenggaraan MBKM adalah menggali sebanyak mungkin aspirasi publik melalui berbagai metode. Penjaringan aspirasi ini dapat dilakukan secara langsung maupun tidak langsung. Sebagai contoh, penelitian berbasis survey yang dilakukan pada [2] dan [3] bertujuan untuk melihat perspektif mahasiswa secara langsung terhadap program MBKM pada aspek-aspek yang ditentukan penyelenggara. Contoh lainnya dapat ditemukan pada penelitian [4] yang menyajikan studi kasus tentang persepsi, pemahaman, dan minat mahasiswa terhadap program MBKM dengan pendekatan *Focus Group Discussion* (FGD). Dengan metode pengumpulan data seperti ini, dapat diperoleh tanggapan langsung dan terstruktur sesuai dengan tujuan spesifik yang dirancang.

Di sisi lain, sosial media sebagai salah satu *platform* yang saat ini banyak digunakan, menjanjikan tersedianya data opini publik secara tidak langsung dalam jumlah yang besar. Oleh karena itu, banyak penelitian yang memanfaatkan data dari sosial media untuk menganalisis tanggapan publik terkait program MBKM. Sebagai contoh penelitian [5] melakukan analisis sentimen terhadap delapan jenis program MBKM dengan sumber data yang berasal dari 591 komentar pada grup telegram pengawas. Penelitian tersebut menggunakan metode *Naive Bayes* dan *K-Nearest Neighbors*. Studi lain pada penelitian [6], melakukan analisis sentimen berupa enam kelas emosi dengan sumber data dari 968 komentar Twitter menggunakan algoritma LSTM. Sedangkan penelitian [7] yang juga menggunakan komentar Twitter sebanyak 475 data, mengelompokan sentimen positif, negatif dan netral dengan menggunakan metode *Naïve Bayes*, *K-Nearest Neighboars* dan *Decision Tree*. *Platform* lain yaitu Youtube juga dimanfaatkan dalam penelitian [8] dengan mengambil 900 komentar. Penelitian ini bertujuan untuk menganalisis sentimen positif dan negatif penonton pada video unggahan Mendikbud di kanalnya.

Permasalahan utama menyangkut data yang diperoleh pada sosial media adalah data yang tidak terstruktur dan memiliki banyak *noise*. Oleh karena itu, pemrosesan lebih lanjut diperlukan sehingga data yang melimpah dapat dimanfaatkan dengan baik. Permasalahan lain adalah komentar di meda sosial juga memiliki berbagai jenis dan variasi sehingga mempersulit proses penjaringan tanggapan terhadap program. Komentar tersebut tidak hanya berupa opini, namun juga kadang berupa pertanyaan, dan informasi. Selain itu terdapat juga banyak komentar yang sama sekali tidak berhubungan dengan program tersebut namun menggunakan *tag* atau *keyword* nama program untuk menarik perhatian misalnya komentar berupa iklan atau *spam*. Tertarik dengan permasalahan tersebut, penelitian ini mencoba melakukan klasifikasi terhadap jenis komentar pada Twitter berkaitan dengan program MBKM untuk empat kelas yaitu informasi, opini, pertanyaan dan komentar lainnya yang tidak berhubungan (*out of topic*). Namun pada proses pengolahan data, kami menemukan bahwa dataset yang kami peroleh memiliki kelas yang tidak seimbang (*imbalanced dataset*).

Ketidakseimbangan dataset adalah masalah yang sering terjadi dalam tugas-tugas klasifikasi di mana jumlah sampel dalam satu kelas jauh melebihi jumlah kelas yang lain. Sayangnya, seringkali data kelas minoritas adalah kelas yang penting untuk diteliti atau dideteksi [9]. Ketidakseimbangan dataset adalah hal yang perlu diatasi karena nilai akurasi yang tinggi tidak hanya bergantung pada algoritma yang digunakan, namun juga tergantung pada faktor karakter dataset [10]. Oleh karena itu, ketidakseimbangan dataset akan menyebabkan performa klasifikasi tidak optimal karena model lebih banyak belajar dengan data pada kelas mayoritas. Pada penelitian ini, akurasi model pada beberapa algoritma hanya mampu menghasilkan *F-Measure* di antara 0,6209 sampai 0,6672. Oleh sebab itu, penelitian ini mencoba mengekplorasi beberapa

metode untuk memodifikasi *imbalanced dataset* dengan harapan dapat meningkatkan kinerja model klasifikasi.

Terdapat beberapa teknik untuk mengatasi permasalahan *imbalanced dataset* yang secara umum terbagi menjadi dua yaitu metode berbasis algoritma (*algorithm-level methods*) dan metode berbasis modifikasi data (*data-level methods*). Pendekatan berbasis modifikasi data dilakukan dengan cara menyeimbangkan distribusi kelas, sedangkan pendekatan berbasis algoritma mencoba untuk memperbaiki algoritma pembelajaran atau *classifier* tanpa mengubah dataset pelatihan [11]. Teknik *data-level methods* dapat dilakukan dengan melakukan *resampling* dengan mengurangi jumlah data pada kelas mayoritas (*undersampling*), menambah jumlah data sintetis (*augmented dataset*) pada kelas minoritas (*oversampling*), atau dapat juga dilakukan dengan kombinasi antara *undersampling* dan *oversampling*.

Penelitian tentang perbandingan metode *resampling* dapat kita temukan dalam berbagai referensi. Pada penelitian [12], metode Near Miss dibandingkan dengan *Synthetic Minority Over-sampling Technique* (SMOTE) pada data riwayat medis pasien dan menyimpulkan bahwa Near Miss dengan nilai *F-Measure* sebesar 0,8623 lebih unggul daripada SMOTE. Metode SMOTE juga dibandingkan dengan *Adaptive Synthetic* (ADASYN) dan digunakan pada penelitian [13] untuk data klasifikasi penyakit diabetes. Pada penelitian ini disimpulkan bahwa ADASYN mampu mengatasi masalah *imbalanced data* lebih baik dengan nilai akurasi 87,3%. Pada penelitian untuk mengklasifikasikan ujaran kebencian, penelitian [14] menyimpulkan bahwa SMOTE memberikan hasil yang baik dengan nilai Geo sebesar 0,778. Perbandingan berbagai metode *undersampling*, *oversampling* dan kombinasinya juga dilakukan pada [15] pada klasifikasi jurnal yang menyimpulkan bahwa SMOTE adalah metode *oversampling* terbaik dan Tomek-Link adalah metode *undersampling* terbaik.

Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa terdapat perbedaan hasil penelitian tentang metode *resampling* mana yang menghasilkan kinerja yang lebih baik. Oleh karena itu, penelitian ini fokus mencari metode yang sesuai dengan dataset yang dimiliki dan dapat meningkatkan kinerja model menjadi lebih baik. Penelitian ini menyajikan perbandingan dari dua teknik *undersampling* yaitu Near Miss dan Tomek Links, dua metode *oversampling* yaitu SMOTE dan ADASYN dan satu metode kombinasi antara *Undersampling* dan *Oversampling* yaitu *Random Combination Sampling* (RCS). Untuk memastikan bahwa metode yang digunakan memberikan hasil yang stabil, dataset hasil modifikasi kelima metode ini digunakan sebagai input untuk empat algoritma klasifikasi yaitu *Random Forest, Logistic Regression*, SVM dan *Multi Layer Perceptron* (MLP). Kami mengevaluasi kinerja metode tersebut, terutama dengan meninjau nilai akurasi, *F-Measure* dan kurva *Receiver Operating Characteristic* (ROC*)*.

2. TINJAUAN PUSTAKA

2.1. Metode *Resampling*

Dalam dunia statistik, istilah *resampling* didefinisikan sebagai pembuatan sampel baru berdasarkan sampel yang telah ada. Teknik yang paling sederhana dalam *resampling* adalah memilih sampel secara acak atau *random sampling*. Pada teknik *oversampling* hal ini dilakukan dengan memilih sampel data dari kelas mayoritas dan menghapusnya. Meskipun sederhana, teknik ini berpotensi menghapus data yang penting dalam menentukan batas antar kelas. Pada teknik *oversampling* pemilihan sampel acak dilakukan pada kelas minoritas lalu kemudian di duplikasi. Metode ini menyebabkan model dilatih dengan data sampel yang sama berulang kali, sehingga metode ini tidak efektif dalam mengatasi *overfitting*. Oleh karena itu, munculah beberapa konsep untuk membangun teknik *resampling* yang lebih representatif. Pada bagian ini akan diuraikan beberapa metode *resampling* untuk mengatasi efek *imbalanced dataset* yang digunakan pada penelitian ini. Metode yang digunakan adalah teknik dengan pendekatan *oversampling*, *undersampling* dan kombinasinya.

2.1.1 Near Miss

Metode Near Miss bekerja dengan memilih sampel berdasarkan jarak sampel pada kelas mayoritas ke sampel kelas minoritas. Jarak tersebut ditentukan dalam ruang fitur menggunakan *Euclidean Distance*. Terdapat tiga pendekatan yang dapat dilakukan dalam memilih sampel dari kelas mayoritas, yaitu Near Miss-1 dengan menemukan sampel dengan jarak rata-rata terkecil ke sampel terdekat dari kelas lain, Near Miss-2 yaitu dengan memilih sampel dengan jarak rata-rata terkecil ke sampel terjauh kelas lain, dan Near

Miss-3 yaitu dengan mempertahankan tetangga terdekat dari setiap sampel kelas lain, dan sampel kelas mayoritas dipilih berdasarkan jarak rata-rata terkecil antara sampel-sampel tersebut dengan tetangga terdekatnya [16].

2.1.2 Tomek Links

Berbeda dengan Near Miss yang memilih sampel pada kelas mayoritas untuk dipertahankan, sebaliknya metode Tomek Links memilih sampel untuk dihapus. Metode ini memilih sampel dengan menemukan pasangan antar kelas (*cross-class pairs*) yang memiliki *Euclidean Distance* terkecil satu sama lain dalam ruang fitur. Metode ini dapat juga diterapkan untuk pembersihan data pasca-pemrosesan untuk menghapus sampel baik dari kelas mayoritas maupun minoritas, karena wilayah batas yang kurang terdefinisi dengan baik [17]. Jika sampel di kelas minoritas dianggap konstan, maka sampel tersebut dapat digunakan untuk menemukan semua contoh di kelas mayoritas yang paling dekat dengan kelas minoritas. Oleh karena itu, sampel yang dihasilkan oleh Tomek Links adalah sampel batas atau *noise*. Hal ini disebabkan fakta bahwa hanya sampel pada batas kelas dan *noise* yang memiliki tetangga terdekat yang berasal dari kelas lawan.

2.1.3 *Synthetic Minority Over-sampling Technique* (SMOTE)

SMOTE merupakan teknik *oversampling* yang merupakan metode manipulasi data berbasis interpolasi. Metode ini bekerja dengan cara mengambil sampel dari kelas minoritas dan menyisipkan sampel sintetis sepanjang garis yang menghubungkan satu sampel dengan sampel lainnya, yang merupakan tetangga terdekat di kelas tersebut. Teknik SMOTE terdiri dari tiga tahap yaitu memilih sampel dari kelas minoritas secara acak, memilih sampel dari *k* sampel tetangganya, dan membuat sampel baru menggunakan interpolasi linier dari kedua sampel tersebut. Interpolasi linier adalah teknik untuk memperkirakan suatu titik dari dua titik yang ada dengan menggambar garis linier dari titik-titik tersebut [18].

2.1.4 *Adaptive Synthetic* (ADASYN)

ADASYN adalah salah satu teknik *oversampling* untuk meningkatkan kinerja model berdasarkan pada distribusi data. Metode ini bekerja dengan mengurangi bias yang disebabkan oleh ketidakseimbangan kelas, dan secara adaptif menggeser batas keputusan klasifikasi ke arah sampel yang sulit. Ide penting dari ADASYN adalah menggunakan pembobotan yang memperhatikan distribusi untuk sampel kelas minoritas menurut tingkat kesulitan untuk diklasifikasikan. Metode ini lebih banyak menghasilkan data sintetis untuk kelas minoritas yang lebih sulit dipelajari dibandingkan dengan contoh-contoh minoritas yang lebih mudah mempelajari. ADASYN dirancang untuk membuat sampel sintetis di wilayah dengan kepadatan sampel minoritas rendah [19].

2.1.5 *Random Combination Sampling* (RCS)

Sementara *undersampling* berpotensi menghilangkan informasi yang penting dan *oversampling* berpotensi membangkitkan terlalu banyak sampel yang tidak representatif, metode kombinasi dari keduanya dapat dipertimbangkan untuk mendapatkan dataset yang lebih representatif. Pada penelitian ini, metode RCS dilakukan dengan dua langkah yaitu pertama, menghapus sebagian data pada kelas mayoritas secara acak, dan kedua menambah sampel untuk kelas minoritas dengan cara menggabungkan dua buah sampel random pada kelas tersebut. Dengan cara demikian, diharapkan sampel baru memiliki variasi baru dan mampu membuat sebuah titik di ruang vektor pada wilayah yang masih representatif.

2.2. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Agar dapat menjadi input bagi algoritma *classifier*, kata-kata dalam data teks biasanya direpresentasikan sebagai fitur kategori diskrit. Mengekstrak satu set fitur menggunakan algoritma yang efektif tidak hanya akan mengurangi dimensi ruang fitur, tetapi juga berguna untuk menghapus fitur yang berlebihan [20]. Bobot yang diberikan pada kata menunjukkan pentingnya kata tersebut dalam dokumen [21]. Metode ini menggabungkan dua konsep penghitungan bobot yaitu frekuensi kemunculan kata dalam dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut. Berdasarkan eksperimen pada penelitian [22], representasi teks TF-IDF berfungsi lebih baik untuk dataset kecil dengan ketidakseimbangan data yang besar, daripada pembobotan yang lebih kompleks seperti Glove, doc2vec, atau FastText.

2.3. Algoritma Klasifikasi

Support Vector Machine (SVM) adalah teknik klasifikasi yang biasanya digunakan dalam kasus *imbalanced dataset* [23]. Namun begitu, untuk menganalisis efek dari metode *resampling* terhadap kinerja

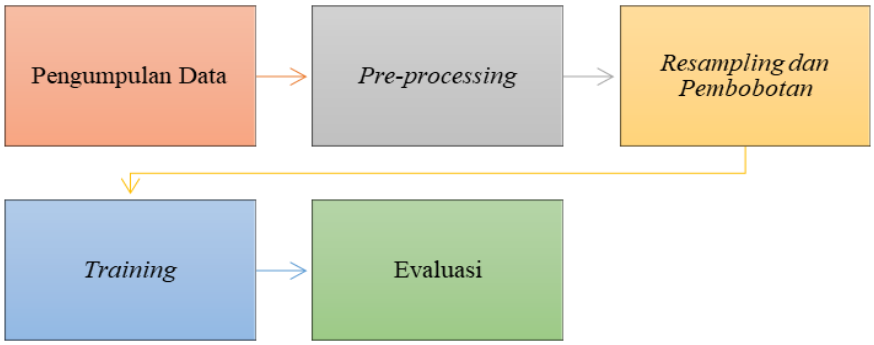
klasifikasi, beberapa macam algoritma klasifikasi digunakan untuk melihat konsistensi hasil masing-masing metode. Pada penelitian ini algoritma yang digunakan diharapkan mewakili beberapa variasi metode klasifikasi. Penelitian ini menggunakan empat algoritma yaitu *Random Forest, Logistic Regression, Multi Layer Perceptron* (MLP) dan *Support Vector Machine* (SVM). *Random Forest* adalah salah satu *classifier* yang mengkombinasikan sejumlah *Decision Trees*, sedangkan *Logistic Regression* adalah *classifier* dengan pendekatan probabilitas kondisional. Sementara *Logistic Regression* dinilai kurang stabil, *Random Forest* dinilai lebih tidak terpengaruh oleh *imbalanced dataset* [24]. Algoritma selanjutnya adalah algoritma-algoritma yang lebih kompleks yaitu MLP dengan menggunakan pendekatan *Artificial Neural Network*, dan *Support Vector Machine* (SVM) yang menggunakan pendekatan pembuatan batas berupa *hyperplane* antar kelas di ruang multi dimensi.

2.4. Evaluasi

Untuk menganalis hasil *resampling* kami melihat jumlah, sebaran data dan nilai *Imlanced Ratio* (IR). Kami akan mencoba memvisualisasikan hasilnya dengan terlebih dahulu mereduksi dimensi dengan *Principal Component Analysis* (PCA). Selanjutnya untuk mengukur keberhasilan klasifikasi, kami menganalisis *confusion matrix* dan menghitung akurasi pada data *training* dan *testing*. Ketika bekerja dengan *imbalanced dataset*, evaluasi dengan nilai akurasi saja tidak cukup karena nilai yang dihasilkan akurasi didominasi oleh kelas mayoritas. Oleh karena itu, penelitian ini juga meninjau nilai *recall* (perbandingan *true positive* dengan keseluruhan kelas positif), *precission* (perbandingan *true positive* dengan keseluruhan prediksi positif), dan skor *F-Measure* yang merupakan perbandingan harmonik antara keduanya. Selain itu kami menganalisa grafik ROC-AUC. *Receiver Operating Characteristic (*ROC*)* adalah grafik *true positive* terhadap *false positive*, sedangkan nilai *Area Under the Curve* (AUC) sering digunakan untuk membandingkan kinerja antar model [25].

3. METODOLOGI PENELITIAN

Langkah-langkah penelitian ini pada dasarnya serupa dengan tugas klasifikasi teks secara umum. Namun demikian terdapat satu tahapan tambahan yaitu *resampling dataset* yang kami lakukan sebelum atau sesudah pembobotan. Penelitian ini menggunakan bahasa pemrograman Python dengan *platform* Google Colaboratory. Gambaran umum proses penelitian ditunjukan pada Gambar 1.



Gambar 1. Desain Umum Penelitian

3.1. Dataset

Tahap pertama dalam penelitian ini adalah pengumpulan data yang dilakukan dengan menggunakan Twitter API. Data diambil dalam rentang waktu bulan Juni 2022 hingga Agustus 2022. Kata kunci yang digunakan pada pengumpulan data adalah "Kampus Merdeka", "Merdeka Belajar" dan "MBKM". Dari proses pengumpulan data diperoleh sejumlah 16.946 baris data.

3.2. *Pre-processing*

Proses selanjutnya adalah *pre-processing* atau normalisasi data. Proses ini melibatkan beberapa langkah diantaranya penghapusan tanda baca, penghapusan data duplikat, penghapusan contoh kalimat yang terlalu pendek, penggantian beberapa daftar kata dengan kata baku, *case folding* dengan mengubah semua huruf

menjadi kecil, menghapus *stopword*, melakukan *stemming* atau mengekstrak kata menjadi kata dasarnya dan yang terakhir adalah tokenisasi yaitu memisahkan kalimat menjadi kata-kata. *Library* yang digunakan dalam proses ini adalah *Natural Language Toolkit* (NLTK) dan Sastrawi.

3.3. *Resampling*
Setelah melalui tahap *pre-processing*, data yang lebih bersih melalui proses *resampling*. Tujuan *resampling* adalah memodifikasi dataset sehingga lebih proporsional atau lebih representatif. Dalam hal ini kami akan meninjau nilai *Imbalance Ratio* dan visualisasi dari hasil *resampling*. *Resampling* dengan menggunakan teknik *undersampling* dan *oversampling* dilakukan setelah proses pembobotan menggunakan TF-IDF, namun untuk teknik kombinasi dilakukan sebelum pembobotan. Hal ini dikarenakan *resampling* dengna RCS dilakukan pada tingkat kalimat, sedangkan metode *resampling* lain dilakukan pada ruang vektor. Dalam implementasinya, *imblearn library* digunakan dengan tanpa memodifikasi parameter apapun atau menggunakan *default setting.*

3.4. Pembobotan
Pembobotan menggunakan TF-IDF dilakukan untuk merubah dataset pada proses sebelumnya dari bentuk kata dalam dokumen menjadi bobot numerik dalam vektor. Parameter *max feature* pada penelitian ini adalah 5000. *Max feature* menggambarkan jumlah kata yang diambil dari daftar kata yang sering muncul pada dokumen. Hasil dari proses ini berupa vektor *sparse* yang menggambarkan bobot setiap kata berdasarkan frekuensi kemunculannya dan frekuensi dokumen di mana kata tersebut muncul. Bobot yang besar menggambarkan bahwa kata tersebut unik pada dokumen tersebut, sedangkan bobot yang kecil menggambarkan bahwa kata-kata tersebut berada hampir di setiap dokumen.

3.5. *Training*
Sebelum melakukan *training*, *dataset* dibagi menjadi dua, yaitu untuk proses *training* dan untuk proses *testing* dengan proporsi 80:20. Selanjutnya *training dataset* menjadi masukan dalam algoritma klasifikasi *Random Forest*, *Logistic Regression*, SVM dan MLP. Sehingga dalam proses ini kami memiliki lima jenis dataset yang diinput ke dalam empat algoritma klasifikasi. Performa setiap algoritma klasifikasi ditinjau dengan melihat hasil akurasi *training*-nya.

3.6. Evaluasi
Pada proses ini kami akan meninjau seperti apa hasil setiap metode *resampling* dan sebaran datanya. Selanjutnya kami juga menganalisis nilai akurasi untuk mengukur rasio prediksi yang benar dari keseluruhan prediksi yang dilakukan. Kami juga akan meninjau secara umum nilai *recall* dan *precission* dan lebih spesifik menyajikan skor *F-Measure* untuk melihat apakah model memiliki hasil klasifikasi yang seimbang antar kelas. Selain itu kami akan melihat grafik ROC-AUC pada semua hasil klasifikasi.

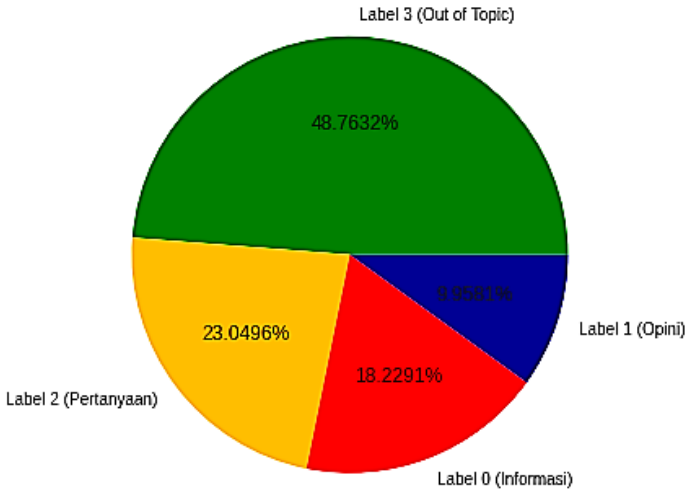
4. HASIL DAN PEMBAHASAN

4.1. Pengolahan Dataset
Dataset yang terkumpul pada penelitian ini adalah sebanyak 16.946 baris data. Setelah *pre-processing*, terdapat pengurangan jumlah data secara signifikan terutama setelah dilakukan proses penghapusan duplikasi menjadi sebesar 7883. Adapun contoh komentar publik yang dikumpulkan ditunjukkan melalui Tabel 1. Nama kelas untuk komentar diberi label 0 untuk kelas informasi, 1 untuk kelas opini, 2 untuk kelas pertanyaan dan 3 untuk komentar yang tidak berkaitan (*out of topic*). Dapat ditunjukan pada Tabel 1 bahwa data hasil *pre-processing* lebih bersih, tanpa *tag*, simbol, alamat *website*, tanda baca dan lain-lain.

Jumlah kelas setelah pelabelan adalah 1.437 untuk kelas 0, 785 untuk kelas 1, 1.817 untuk kelas 2 dan 3.844 untuk kelas 3. Distribusi kelas ditunjukan pada Gambar 2 yang menunjukan bahwa proporsi data untuk kelas 3 atau *out of topic* adalah sebanyak 48,7632% atau hampir setengahnya dari total data. Dengan kondisi seperti ini, jika dilakukan penelitian berkaitan dengan opini program MBKM tentunya data pada kelas tersebut akan sangat mengganggu. Oleh karena itu, salah satu manfaat penelitian ini adalah sebagai dasar filterisasi jenis komentar yang selanjutnya dapat menjadi input pada proses analisis lain seperti analisis sejauh mana informasi MBKM diketahui publik, analisis opini publik dan analisis tentang pertanyaan publik mengenai program ini.

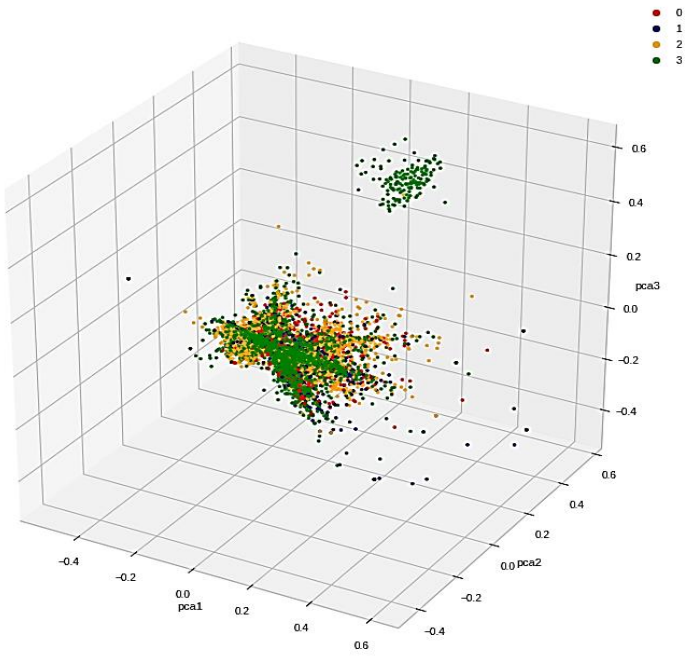
Tabel 1. Sampel dataset sebelum dan setelah *pre-processing*

Label	Sebelum pre-processing	Setelah *pre-processing*
0	b'@naaboahctay @collegemenfess Ga bisa kak, uin dibawah kemenag, kampus merdeka dibawah kemendikbud'	ga bisa kak uin bawah kemenag kampus merdeka bawah kemendikbud
0	b'RT @unqity: Mendikbudristek\xc2\xa0Nadiem Makarim mendorong Mahasiswa jd pengusaha dgn meluncurkan program wirausaha merdeka pada Jumat 15 Juli 20\xe2\x80\xa6'	mendikbudristek nadiem makarim dorong mahasiswa usaha luncur program wirausaha merdeka jumat juli
1	b'kampus merdeka is a very interesting plan'	kampus merdeka sangat asik
1	b'ya allah ini tes buat kampus merdeka cepet bgt anjai. gw bsk ad kuliah pula'	ya allah tes kampus merdeka cepet banget anjai saya besok ada kuliah
2	b'knp pas udah lulus, program kampus merdeka br ada'	kenapa pas lulus program kampus merdeka baru ada
2	b'kenapa paul dipanggil ui sih emangnya dia kampus merdeka https://t.co/B7IzfMBsDx'	kenapa paul panggil ui dia kampus merdeka
3	b'RT @OposisiCerdas: Breaking News: Istri TNI Ditembak di Depan Rumah\nhttps://t.co/rQoanqs70a'	breaking news istri tni tembak di depan rumah
3	b'@convomf ikut kampus merdeka?'	ikut kampus merdeka



Gambar 2. Distribusi kelas pada dataset komentar program MBKM

Terkait dengan kondisi *imbalanced dataset*, perhitungan *Imbalanced Ratio* (IR) yaitu perbandingan jumlah mayoritas dengan kelas minoritas adalah 3844/785=4,8968. Dengan demikian dapat disimpulkan bahwa perbandingan datanya adalah 5:1. Untuk kebutuhan visualisasi dataset, reduksi dimensi dengan PCA dilakukan dan menghasilkan visualisasi dataset *original* yang ditunjukan pada Gambar 3. Selain memilki jumlah data mayoritas, Pada Gambar 3 dapat ditunjukan bahwa sebagian kelas 3 (*out of topic*) memiliki data-data yang memiliki karakteristik jauh berbeda dengan data lainnya. Hal ini dapat diindikasikan dengan kumpulan titik sampel berwarna hijau yang berada jauh di wilayah kumpulan data lainnya.



Gambar 3. Visualisasi dataset komentar program MBKM

4.2. Hasil *Resampling*

Metode *resampling* yang dilakukan pada penelitian ini menghasilkan jumlah dataset yang berbeda-beda. Tabel 2 menunjukan jumlah dataset yang dihasilkan beserta sebaran data per kelasnya.

Tabel 2. Hasil *resampling* dataset pada berbagai metode

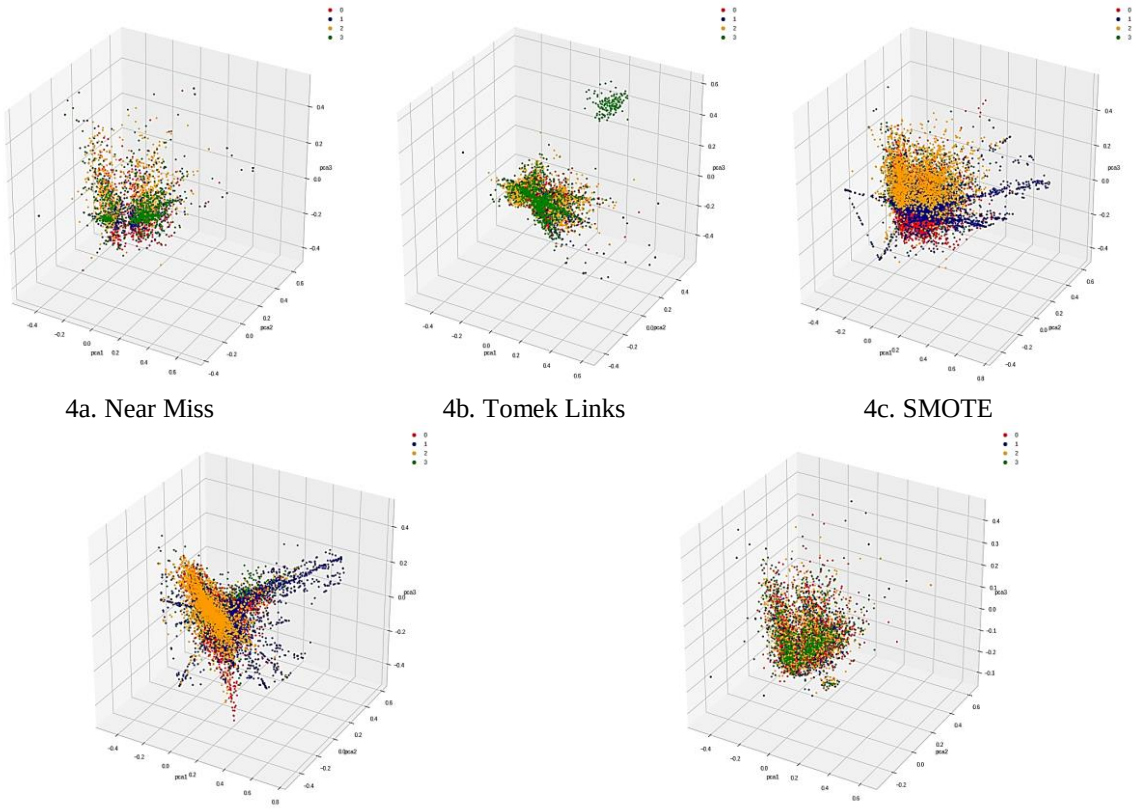
	Near Miss		Tomek Links		SMOTE		ADASYN		RCS	
Data	**Jumlah**	**%**	**Jumlah**	**%**	**Jumlah**	**%**	**Jumlah**	**%**	**Jumlah**	**%**
Kelas 0	785	25%	1.258	17%	3.844	25%	3.876	25%	2.000	25%
Kelas 1	785	25%	785	11%	3.844	25%	3.733	24%	2.000	25%
Kelas 2	785	25%	1657	23%	3.844	25%	4.095	26%	2.000	25%
Kelas 3	785	25%	3.523	49%	3.844	25%	3.844	25%	2.000	25%
Total	3.140		7.223		15.376		15.548		8.000	

Pada teknik *oversampling*, metode Near Miss menghasilkan sampel dengan jumlah masing-masing kelas sebanyak kelas minoritas yaitu 785 sampel, sehingga jumlah total dataset berkurang menjadi 3140. Versi Near Miss yang dipanggil adalah versi Near Miss-1 dengan menemukan sampel dengan jarak rata-rata terkecil tetangga ke tiga buah sampel terdekat dari kelas lain. Visualisasi dataset hasil Near Miss ditunjukan pada Gambar 4a. Dari gambar tersebut dapat ditunjukant bahwa Near Miss menghilangkan dominasi kelas 3. Informasi lain yang juga dapat dicatat adalah bahwa kumpulan data kelas 3 yang berada terpisah dengan wilayah data lain menjadi hilang.

Oversampling dengan metode Tomek Links dilakukan dengan memilih data yang baik untuk membantu menentukan batas daerah keputusan. Tidak seperti metode lain, Tomek Links tidak menghapus data mayoritas berdasarkan *Imbalanced Ratio*. Pada penelitian ini, data yang dihapus pada kelas non-minoritas hanya sebanyak 660 yaitu 179 pada kelas 0, 160 pada kelas 2 dan 321 pada kelas 3. Total dihasilkan data hasil *resampling* Tomek Link sejumlah 7223 dengan visualisasi data ditunjukan pada Gambar 4b. Dengan metode ini, kondisi *imbalanced dataset* masih tinggi dengan nilai IR 0,4878. Jika metode ini memberikan hasil yang baik, metode ini dapat menjadi dasar bahwa rendahnya kinerja klasifikasi bukan berasal dari pengaruh *imbalanced dataset.*

Teknik *oversampling* menggunakan SMOTE menghasilkan data dengan jumlah masing-masing kelas sama dengan jumlah kelas mayoritas yaitu 3844, sehingga jumlah total data adalah 15.376. Jumlah tetangga yang dihitung jaraknya untuk proses interpolasi adalah 5 buah sampel. Dalam hasil visualisasinya pada Gambar 4c, dapat ditunjukan bahwa hasil interpolasi linear SMOTE yang kadang kala menghasilkan titik sampel yang sejajar dengan sampel lain membentuk garis lurus. Metode ADASYN membangkitkan sampel dengan metode serupa dengan SMOTE, namun ADASYN membangkitkan jumlah sampel yang berbeda tergantung pada distribusi kelas atau kepadatannya. Namun begitu persentasi kelas yang dihasilkan pada metode ini tetap seimbang. Pada gambar 4d divisualisaikan hasil *resampling* dengan ADASYN dengan distribusi data yang lebih alami dari pada SMOTE. Jumlah data yang dihasilkan pada metode ini adalah 15.548.

Medode RCS yang dilakukan pada tingkat kalimat menghasilkan dataset yang divisualisasikan oleh Gambar 4e. Hasil *resampling* dengan metode ini lebih natural karena cukup merepresentasikan dataset di dunia nyata. Pada metode ini kami mengurangi jumlah sampel kelas mayoritas dan menambah kelas minoritas, sehingga jumlah setiap kelas sama banyak yaitu 2.000. Total dataset yang dihasilkan pada metode RCS adalah 8.000.



4a. Near Miss 4b. Tomek Links 4c. SMOTE

4d. ADASYN 4e. RCS

Gambar 4 Visualisasi dataset hasil *resampling* dengan berbagai metode

4.3. Hasil Klasifikasi

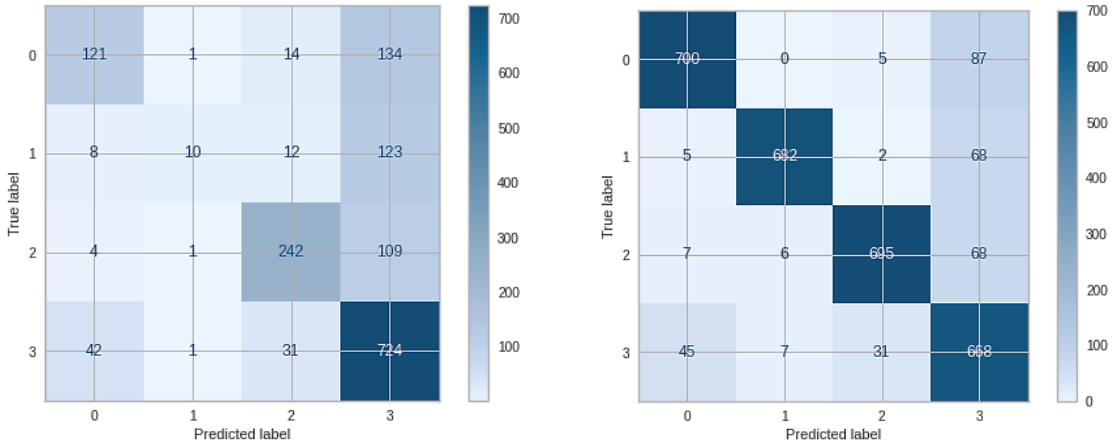
Pengaruh *resampling* dari berbagai metode diuji dengan beberapa metode klasifikasi yang hasilnya ditunjukan pada Tabel 3. Berdasarkan nilai akurasi *training* dan *testing* dapat disimpulkan bahwa kecenderungan *overfiting* di mana akurasi pada data *training* tinggi, namun akurasi pada data *testing* rendah terjadi pada dataset *original*. Sayangnya fenomena ini malah meningkat ketika klasifikasi dilakukan pada dataset hasil *resampling* dengan metode Near Miss. Hasil akurasi *testing* terendah ditunjukan pada dataset Near Miss dengan model MLP sebesar 0,5366. Walaupun kondisi data sudah seimbang, namun dataset ini mengalami penurunan jumlah data yang signifikan. Hal tersebut kemungkinan besar menyebabkan model

tidak cukup belajar dari sampel yang ada. Namun demikian untuk dataset hasil *resampling* metode lainnya, walaupun *overfiting* masih terjadi, akurasi *testing* mengalami peningkatan. Hasil akurasi *training* tertinggi diberikan oleh dataset hasil *resampling* metode ADASYN dengan menggunkan model *Random Forest* sebesar 0,9986, sedangkan hasil *testing* tertinggi ditunjukan oleh dataset hasil metode SMOTE dengan model SVM sebesar 0,8901.

Untuk *F-Measure* pada dataset *original* menunjukan nilai yang lebih rendah dari pada akurasinya. Oleh karena itu kami menganalisis nilai *precision* dan *recall*. Ketika memperhatikan polanya, ternyata nilai *recall* untuk kelas 1 pada semua model sangat rendah. Sebagai contoh, pada Gambar 5 ditunjukan *confussion matrix* untuk dataset *original* pada metode *SVM*. Pada Gambar 5a ditunjukan bahwa *True Positive* untuk kelas 1 hanya 10 sampel dari total 153 sampel kelas 1. Nilai *recall* pada kasus ini adalah sebesar 0,0654. *False Negative* terbanyak terjadi karena model mengklasifikasikan sebanyak 123 kelas 1 sebagai kelas 3. Hal ini sangat mungkin terjadi karena dataset di kelas 3 mendominasi proses *training*. Pada Gambar 5b ditampilkan juga hasil *confusion matrix* dengan nilai *recall* kelas 1 tertinggi menggunakan model SVM pada dataset ADASYN.

Tabel 3. Hasil akurasi dan *F-Measure* berbagai model klasifikasi pada berbagai dataset hasil *resampling*

Classifier	**Evaluasi**	**Original**	**Near Miss**	**Tomek Links**	**SMOTE**	**ADASYN**	**RCS**
Random Forest	*Training*	0,9974	0,9948	0,9967	0,9983	0,9986	0,9985
	Testing	0,6766	0,5843	0,6899	0,8576	0,8694	0,7331
	F-Measure	0,6411	0,5822	0,6548	0,8574	0,8684	0,7332
Logistic Regression	*Training*	0,8065	0,8387	0,8205	0,8973	0,8928	0,8893
	Testing	0,6968	0,6178	0,7231	0,8299	0,8144	0,7675
	F-Measure	0,6672	0,6132	0,6972	0,8275	0,8110	0,7653
SVM	*Training*	0,9305	0,9621	0,9415	0,9839	0,9869	0,9812
	Testing	0,6956	0,6130	0,7107	0,8901	0,8871	0,8143
	F-Measure	0,6597	0,6109	0,6762	0,8928	0,8896	0,8151
MLP	*Training*	0,9968	0,9948	0,9958	0,9973	0,9967	0,9984
	Testing	0,6227	0,5366	0,6532	0,8670	0,8729	0,7381
	F-Measure	0,6209	0,5347	0,6502	0,8610	0,8654	0,7375

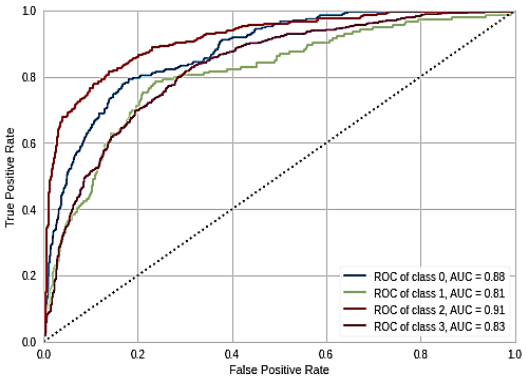


5a. Dataset *original* 5b. Dataset ADASYN

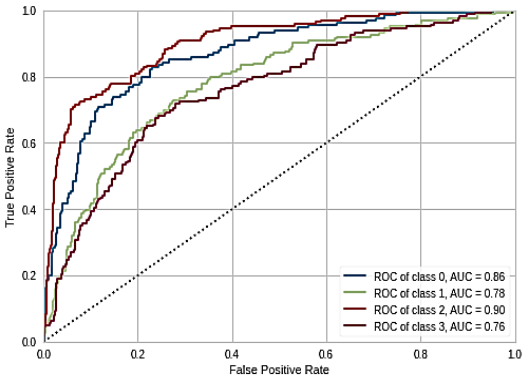
Gambar 5 Contoh *confussion matrix* hasil klasifikasi dengan model SVM

Pada dataset dari metode Near Miss nilai *F-Measure* juga lebih kecil dari nilai akurasinya. Setelah hasil *recall* dianalisis, ditemukan bahwa secara umum metode ini cukup besar meningkatkan *recall* untuk kelas 1, namun sayangnya nilai *recall* untuk kelas 3 berkurang secara signifikan. Selanjutnya pada dataset hasil *resampling* Tomek Link hasil *recall* kelas 1 tidak mengalami peningkatan yang signifikan, sehingga nilai

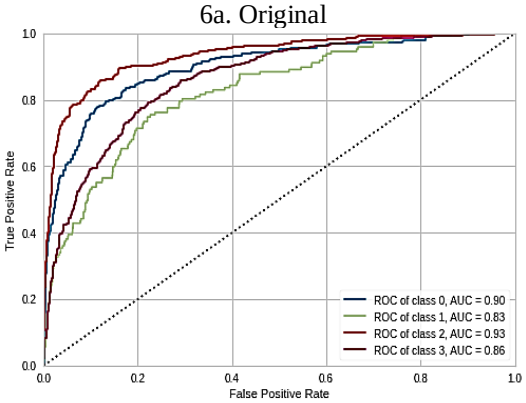
F-Measure klasifikasi dataset ini rendah. Untuk dataset yang dibangkitkan oleh SMOTE, ADASYN dan juga RCS, hasil klasifikasi memiliki kesamaan yaitu nilai *F-Measure* yang rata-rata setara dengan hasil akurasinya. Pada ketiga dataset ini, nilai *recall* kelas 1 meningkat secara signifikan. Hasil *recall* tertinggi kelas 1 adalah 0,9815 pada dataset ADASYN dengan model MLP. Namun demikian nilai *recall* kelas 3 mengalami penurunan walaupun tidak begitu drastis seperti pada dataset Near Miss. Untuk mendapatkan intuisi tentang pengaruh metode *resampling* terhadap hasil klasifikasi tiap kelas, Gambar 6 menunjukan grafik ROC-AUC hasil klasifikasi menggunakan SVM pada berbagai dataset hasil *rasampling*.



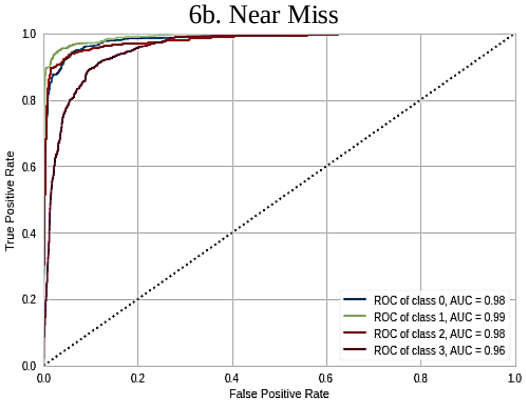
6a. Original



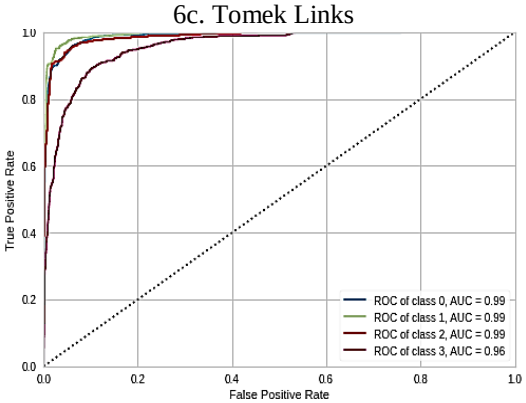
6b. Near Miss



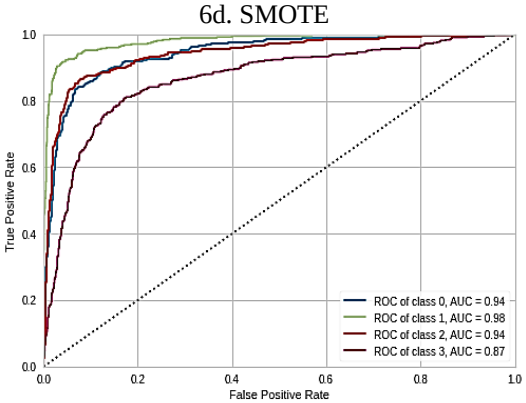
6c. Tomek Links



6d. SMOTE



6e. ADASYN



6f. RCS

Gambar 5 Grafik ROC dengan model SVM

Berdasarkan analisis yang dilakukan, dapat disimpulkan bahwa metode *undersampling* Near Miss pada penelitian ini tidak memberikan efek positif dalam peningkatan kinerja model klasifikasi, sedangkan model lainnya dapat membantu meningkatkan kinerja dengan meningkatkan nilai *F-Measure* dan nilai AUC pada setiap kelas. *Undersampling* dengan Tomek Links hanya menghasilkan sedikit peningkatan kinerja. Sedangkan metode *oversampling* menggunakan SMOTE dan ADASYN menghasilkan peningkatan kinerja

yang signifikan. Meskipun demikian, pengaruh positif ini perlu diteliti lebih jauh karena peningkatan kinerja ini mungkin bukan hanya disebabkan karena dataset yang lebih seimbang, namun karena jumlah data latih yang besar. Pada metode kombinasi yang disusun pada penelitian ini yaitu RCS, terdapat peningkatan kinerja yang baik walaupun tidak lebih tinggi dari metode *oversampling*. Mengingat jumlah data pada dataset RCS juga hanya sekitar setengahnya dari dataset pada metode *oversampling*, penelitian tentang pengaruh penambahan jumlah data menggunakan metode RCS layak untuk dilakukan pada penelitian selanjutnya.

Selanjutnya kami mengambil dataset dan model yang menghasilkan evaluasi *F-Measure* tertinggi yaitu SMOTE dan SVM untuk dikembangkan lebih lanjut. Pada model SVM kami mencari pengaruh kernel terhadap hasil klasifikasi dan memperoleh nilai terbaik menggunakan kernel *polynomial* dengan nilai *F-Measure* sebesar 0,9508. Hasil penelitian terhadap pengaruh jenis kernel yang digunakan ditunjukan pada Tabel 4. Selanjutnya dengan menggunakan model SVM dengan kernel *polynomial*, kami mencari jumlah tetangga yang memberikan hasil lebih baik dalam pembangkitan dataset dengan menggunakan SMOTE. Hasil penelitian terhadap pengaruh jumlah tetangga ditunjukan pada Tabel 5. Meskipun hanya memberikan sedikit peningkatan, dataset terbaik ditunjukan pada SMOTE menggunakan jumlah tetangga = 3 dengan skor *F-Measure* 0,9524.

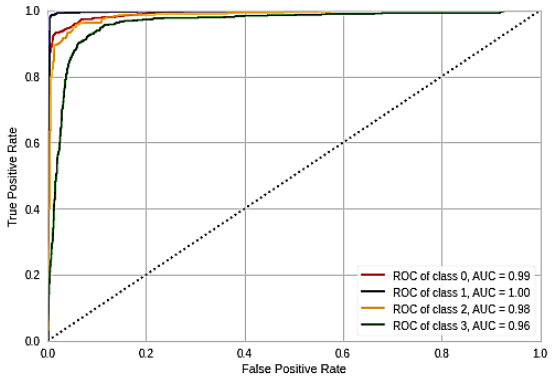
Tabel 4 Pengaruh jenis kernel pada model SVM

Evaluasi	**Linier**	**Polynomial**	**RBF**	**Sigmoid**
Training	0,9202	0,9883	0,9839	0,8301
Testing	0,8442	0,9512	0,8901	0,7834
F-Measure	0,8409	0,9508	0,8928	0,7797

Tabel 5 Pengaruh jumlah tetangga pada *resampling* SMOTE

Evaluasi	**1**	**3**	**5**	**7**	**9**
Training	0,9884	0,9884	0,9883	0,9884	0,9877
Testing	0,9489	0,9528	0,9512	0,9525	0,9512
F-Measure	0,9485	0,9524	0,9508	0,9521	0,9508

Hasil dataset yang dibangkitkan dengan *resampling* SMOTE dengan jumlah tetangga=3 dan model klasifikasi SVM dengan kernel *polynomial* menghasilkan grafik ROC-AUC yang ditunjukan pada Gambar 7. Pada grafik tersebut dapat dilihat bahwa nilai klasifikasi pada setiap kelas menunjukan nilai optimal dengan AUC terbaik di kelas 1 mencapai angka 100% dan nilai AUC terrendah pada kelas 3 yaitu 96%.



Gambar 6 Grafik ROC pada dataset SMOTE menggunakan model SVM dengan paramater terbaik.

Walaupun dalam penelitian ini SMOTE memberikan hasil peningkatan kinerja model secara signifikan, namun perlu dianalisis lebih jauh apakah data yang dibangkitkan sudah cukup representatif atau tidak. Hal ini karena kita membangkitkan sampel data di ruang vektor dan tidak mendapatkan intuisi yang jelas tentang kalimat sebenarnya yang dibangkitkan. Tidak seperti data numerik atau gambar, teks sintetis sangat

berpeluang kehilangan informasi kontekstual dan semantik. Sampel yang dihasilkan seringkali berakhir dengan tata bahasa dan struktur teks yang buruk sehingga kehilangan makna [9]. Oleh karena itu pembangkitan teks dalam level kalimat seringkali lebih rasional. Selain itu, *algorithm-level methods* yang tidak memerlukan modifikasi dataset juga dapat dipertimbangkan, misalnya menggunakan model *Deep Learning* untuk klasifikasi text seperti yang dilakukan pada penelitian [26] dan [27].

5. KESIMPULAN DAN SARAN

Penelitian ini mencoba mengeksplorasi beberapa teknik *resampling* untuk melihat pengaruhnya terhadap kinerja model klasifikasi komentar publik terhadap program MBKM. Hal ini, dikarenakan dataset yang diperoleh memiliki ketidaseimbangan kelas dengan *Imbalanced Rasio* (IR) sebesar 5:1 dan kinerja klasifikasi yang rendah dengan nilai *F-Measure* di antara 0,6209 sampai 0,6672. Metode *resampling* yang digunakan adalah *undersampling* dengan Near Miss dan Tomek Links, *oversampling* dengan SMOTE dan ADASYN, dan kombinasi *undersampling* dan *oversampling* dengan RCS. Untuk melihat stabilitas efek metode *resampling*, empat *classifier* digunakan yaitu *Random Forest*, *Logistic Regression*, SVM dan MLP.

Berdasarkan analisis yang dilakukan, dapat disimpulkan bahwa metode Near Miss pada penelitian ini tidak memberikan efek positif dalam peningkatan kinerja model klasifikasi, sedangkan model lainnya dapat membantu meningkatkan kinerja dengan meningkatkan nilai *F-Measure* dan nilai AUC pada setiap kelas. *undersampling* dengan Tomek Links hanya menghasilkan sedikit peningkatan kinerja, sedangkan metode *oversampling* menggunakan SMOTE dan ADASYN menghasilkan peningkatan kinerja yang signifikan. Pada metode kombinasi yang disusun pada penelitian ini yaitu RCS, terdapat peningkatan kinerja yang baik walaupun tidak lebih tinggi dari metode *oversampling*. Nilai terbaik diperoleh pada model klasifikasi SVM dengan dataset hasil *resampling* metode SMOTE. Setelah melalui optimasi model dan metode *resampling* diperoleh nilai *F-Measure* optimal sebesar 0,9524

DAFTAR PUSTAKA

[1] C. M. Toquero, “Challenges and Opportunities for Higher Education amid the COVID-19 Pandemic: The Philippine Context,” *Pedagog. Res.*, vol. 5, no. 4, 2020.

[2] P. Arjanto, W. F. Antariksa, and A. Timan, “Persepsi Mahasiswa Terhadap Implementasi Merdeka Belajar Kampus Merdeka (MBKM),” *J. Adminitrasi dan Manaj. Pendidik.*, vol. 5, no. 3, pp. 247–257, 2022.

[3] K. D. P. Meke, R. B. Astro, and M. H. Daud, “Dampak Kebijakan Merdeka Belajar Kampus Merdeka (MBKM) pada Perguruan Tinggi Swasta di Indonesia,” *Edukatif J. Ilmu Pendidik.*, vol. 4, no. 1, pp. 675–685, 2021.

[4] C. E. C. Citraningtyas, A. A. Setiawan, and E. Purwanto, “Students ’ Perception toward the Merdeka Belajar Kampus Merdeka Policy (Case Study at a Private University in South Tangerang),” *Cent. ASIA CAUCASUS*, vol. 22, no. 5, pp. 1157–1164, 2021.

[5] A. Rozaq, Y. Yunitasari, K. Sussolaikah, E. Resty, and N. Sari, “Sentiment Analysis of Kampus Mengajar 2 Toward the Implementation of Merdeka Belajar Kampus Merdeka Using Naïve Bayes and Euclidean Distence Methods,” vol. 3, no. 1, pp. 30–37, 2022.

[6] S. J. Pipin and H. Kurniawan, “Analisis Sentimen Kebijakan MBKM Berdasarkan Opini Masyarakat di Twitter Menggunakan LSTM,” vol. 23, no. 2, pp. 197–208, 2022.

[7] A. Rozaq, Y. Yunitasari, K. Sussolaikah, E. Resty, N. Sari, and R. I. Syahputra, “Analisis Sentimen Terhadap Implementasi Program Merdeka Belajar Kampus Merdeka Menggunakan Naïve Bayes , K-Nearest Neighboars Dan Decision Tree,” *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 2, pp. 746–750, 2022.

[8] D. F. Zhafira, B. Rahayudi, and I. Indriati, “Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF Berdasarkan Komentar pada Youtube,” *J. Sist. Informasi, Teknol. Informasi, dan Edukasi Sist. Inf.*, vol. 2, no. 1, pp. 55–63, 2021.

[9] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, “Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models,” *Appl. Sci.*, vol. 11, no. 869, pp. 1–20, 2021.

[10] M. Hayaty, S. Muthmainah, and S. M. Ghufran, “Random and Synthetic Over-Sampling Approach

to Resolve Data Imbalance in Classification," *Int. J. Artif. Intell. Res.*, vol. 4, no. 2, pp. 86–94, 2021.

[11] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data:Review of Methods and Applications," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, pp. 1–8, 2021.

[12] A. R. B. Alamsyah, S. Rahma, N. S. Belinda, and A. Setiawan, "SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data," in *International Conference On Data Science and Official Statistics*, 2021, pp. 305–314.

[13] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," vol. 8, no. 2, pp. 276–282, 2021.

[14] N. A. Verdikha *et al.*, "KOMPARASI METODE OVERSAMPLING UNTUK KLASIFIKASI TEKS UJARAN KEBENCIAN," pp. 85–90, 2018.

[15] A. Indrawati, H. Subagyo, A. Sihombing, W. Wagiyah, and S. Afandi, "Analyzing the Impact of Resampling Method for Imbalanced Data Text in Indonesian Scientific Articles Categorization," *Baca J. Dokumentasi Dan Inf.*, vol. 41, no. 2, p. 133, 2020.

[16] N. M. Mqadi, N. Naicker, and T. Adeliyi, "Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss," *Hindaawi Math. Probl. Eng.*, no. 7194728, pp. 1–16, 2021.

[17] E. F. Swana and W. Doorsamy, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," 2022.

[18] A. Nurhopipah, Y. Ceasar, and A. Priadana, "Improving Machine Leaming Accuracy using Data Augmentation in Recruitment Recommendation Process," in *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, 2021, pp. 203–208.

[19] A. B. Ebenezer, O. K. Boyinbode, and O. M. Idowu, "A Comprehensive Analysis of Handling Imbalanced Dataset," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 2, pp. 454–463, 2021.

[20] H. Z. Mauni, T. Hossain, and R. Rab, "Classification of Underrepresented Text Data in an Imbalanced Dataset Using Deep Neural Network," in *2020 IEEE Region 10 Symposium, TENSYMP 2020*, 2020, no. June, pp. 997–1000.

[21] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, A. A. Saputra, and H. D. Fahma, "Implementasi Term-Frequency Inverse Document Frequency (TF- IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 1, no. 2, pp. 58–64, 2019.

[22] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," *Procedia Comput. Sci.*, vol. 159, pp. 736–745, 2019.

[23] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1560–1571, 2019.

[24] M. Zheng, F. Wang, X. Hu, Y. Miao, H. Cao, and M. Tang, "A Method for Analyzing the Performance Impact of Imbalanced Binary Data on Machine Learning Models," *Axioms*, vol. 11, no. 607, pp. 1–19, 2022.

[25] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, 2019.

[26] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, no. 181, pp. 1–12, 2022.

[27] L. Jiao *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.