Edu Komputika 9 (2) (2022)

Edu Komputika Journal

http://journal.unnes.ac.id/sju/index.php/edukom

Penanganan *Imbalanced Dataset* untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter

Cindy Magnolia✉, Ade Nurhopipah, dan Bagus Adhi Kusuma

Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Indonesia

Info Artikel

Sejarah Artikel:
Diterima: November 2022
Direvisi: Januari 2023
Disetujui: Januari 2023

Keywords:
ADASYN, *Imbalanced Dataset, Random Combination Sampling*, SMOTE, *Under Sampling,*

Abstrak

Imbalanced dataset merupakan hal yang sering ditemukan secara alami dalam proses penambangan data. Kondisi ini sangat mempengaruhi keakuratan klasifikasi data seperti yang terjadi dalam klasifikasi komentar program Kampus Merdeka yang peneliti lakukan. Penelitian ini akan fokus pada penanganan *Imbalanced dataset* untuk meningkatkan kinerja klasifikasi komentar yang berasal dari aplikasi Twitter. Data diklasifikasikan ke dalam empat kelas yaitu kelas 0 (untuk informasi), kelas 1 (untuk opini), kelas 2 (untuk pertanyaan), dan kelas 3 (untuk *out of topic*). Metode yang digunakan untuk *balancing dataset* adalah *Undersampling, Oversampling* menggunakan SMOTE dan ADASYN, serta *Random Combination Sampling*. Evaluasi performa dilakukan menggunakan algoritma *Support Vector Machine (SVM)* dengan perbandingan komposisi data *training* dan *testing* 80:20. Metode pembobotan data yang digunakan adalah *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan nilai *max_features* 3000, 5000, dan 7000. Hasil pengujian awal menunjukan bahwa nilai akurasi dan *F1-score* pada *Imbalanced dataset* secara berurut-urut adalah 0,7 dan 0,7. Sedangkan metode penanganan *Imbalanced dataset* dapat meningkatkan nilai *F1-score*, kecuali pada penerapan metode *Undersampling*. Metode terbaik ditunjukan oleh penerapan ADASYN dengan nilai akurasi dan *F1-score* berurut-urut sebesar 0,9 dan 0,9. Penggunaan *max_features* pada TF-IDF juga mempengaruhi hasil performa klasifikasi, dengan *max_features* terbaik ditunjukan pada jumlah 5000.

Abstract

Imbalanced dataset is often found naturally in the data mining process. This condition significantly affects the accuracy of data classification as happened in the comment classification on the Kampus Merdeka program that the researchers did. This study will focus on handling imbalanced dataset to improve the performance of the comment classification from Twitter. Data is classified into four classes, namely class 0 (for informations), class 1 (for opinions), class 2 (for questions), and class 3 (for out-of-topics). The methods used for dataset balancing are Undersampling, Oversampling using SMOTE and ADASYN, and Random Combination Sampling. Performance evaluation is using the Support Vector Machine with a composition of 80:20. The data weighting method used is Term Frequency-Inverse Document Frequency (TF-IDF) with max_features values of 3000, 5000, and 7000. The initial test results show that the accuracy and F1-score on the imbalanced dataset is 0,7 and 0.7. While the method of handling imbalanced dataset can increase the F1-score, except for the application of the Undersampling method. The best method is shown by the application of ADASYN with an accuracy and F1-score 0,9 and 0.9. Using max_features in TF-IDF also affects the classification performance results, with the best max_features shown at 5000.

✉ Alamat korespondensi:
Jurusan Informatika, Universitas Amikom Purwokerto
Jl. Letjend Pol. Soemarto No.127, Purwokerto, Banyumas 53127
E-mail: 19sa1164@mhs.amikompurwokerto.ac.id

ISSN 2252-6811
E-ISSN 2599-297X

PENDAHULUAN

Salah satu upaya untuk mengikuti dinamika perkembangan jaman adalah perlu adanya pembaharuan sistem pendidikan. Sejak tahun 1947, kurikulum dalam dunia pendidikan telah mengalami 11 (sebelas) kali perubahan. Kurikulum terbaru saat ini adalah Kurikulum Merdeka Belajar bagi Pendidikan Dasar hingga Menengah yang diluncurkan pada bulan Februari 2022. Konsep ini diawali dengan peluncuran Program Kampus Merdeka bagi Perguruan Tinggi. Merdeka Belajar Kampus Merdeka (MBKM) merupakan kurikulum yang diterapkan pada jenjang perguruan tinggi. Dalam penyesuaiannya kurikulum ini menjawab kebutuhan dalam melatih kreativitas, inovatisi, serta mengasah bakat dan minat peserta didik untuk mempersiapkan mereka menghadapi tantangan jaman.

Dilansir melalui laman Kampus Merdeka oleh Kemdikbud, terdapat 9 (sembilan) macam program yang dapat disesuaikan dengan kemampuan dan minat mahasiswa seperti, Magang dan Studi Independen Bersertifikat (MSIB), Wirausaha Merdeka, Indonesian International Student Mobility Awards (IISMA), Pertukaran Mahasiswa Merdeka (PMM), Kampus Mengajar, Praktisi Mengajar, Bangkit by GOTO (Google, Goto, Traveloka), Kementerian ESDM-GERILYA. Sedangkan bagi internal kampus ditambahkan tiga program yaitu Riset atau Penelitian, Membangun Desa (KKN Tematik - KKNT), dan Proyek Kemanusiaan.

Tidak semua pihak dapat menyambut baik perubahan kurikulum ini. Pro dan kontra dari berbagai pihak diekspesikan melalui berbagai bentuk. Media sosial digunakan sebagai wadah utama untuk menampung aspirasi, mendorong percepatan dan pemerataan informasi. Sebagai program baru yang telah diluncurkan, merdeka belajar kampus merdeka masih menjadi topik perbincangan yang cukup hangat di masyarakat. Kesimpangsiuran informasi dan provokasi tidak dapat terhindarkan dari situasi tersebut. Oleh karena itu, klasifikasi komentar diperlukan untuk membantu penyaringan komentar dari masyarakat.

Imbalanced dataset sangat umum ditemukan dalam pengambilan data secara langsung. Penelitian tentang program MBKM telah dilakukukan oleh beberapa peneliti seperti Zhafira, Rahayudi, & Indriati (2021), Kholila (2021), serta Pipin & Kurniawan (2022). Namun pada penelitian tersebut belum ada yang terfokus pada proses klasifikasi sentimennya, terutama pada kondisi ketidakseimbangan data yang mempengaruhi proses klasifikasi. Tantangan yang terjadi dari kondisi ini salah satunya dapat terlihat ketika kita melakukan klasifikasi data. Pada proses ini, data minoritas akan didominasi oleh data mayoritas yang menyebabkan algoritma yang digunakan tidak dapat memisahkan data dengan baik karena cenderung lebih banyak mempelajari data mayoritas. Pembelajaran mesin cenderung memberikan label hanya pada kelas mayoritas untuk data yang diprediksi dengan mengabaikan kelas minoritas, sehingga kelas mayoritas cenderung menunjukkan nilai akurasi yang lebih baik (Fitriani, Yasin, & Tarno, 2021). Studi tentang penanganan *imbalanced dataset* telah dilakukan pada penelitian dengan judul 'Seleksi Fitur dan Penanganan *Imbalanced Data* menggunakan RFECV dan ADASYN'. Pada penelitian tersebut, kondisi data sangat tidak seimbang sangat ekstrim. Metode ADASYN berhasil meningkatkan nilai akurasi hingga mencapai 88% (Pratama, Chandra, & Prasetyaningrum, 2021). Pada kondisi ini, metode ADASYN sangat diperlukan untuk mengurangi bias yang ditimbulkan dari hasil klasifikasi data.

Begitu pula pada penelitian (Khasanah, Muladi, & Pujianto, 2019) dengan judul 'Penerapan Teknik SMOTE untuk Mengatasi *Imbalanced Class* dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN' dengan menggunakan data yang berasal dari Kompas.com. Melalui metode SMOTE yang diterapkan dapat disimpulkan bahwa metode ini berhasil meningkatkan nilai akurasi yang mencapai 87,5%. Evaluasi juga menunjukan performa yang baik, kecuali pada kelas k=5, k=7, dan k=9 yang menunjukan penurunan nilai akurasi. Dalam kondisi *imbalanced,* perlu dilakukan *balancing* agar dapat membantu dalam pemilahan dan pengklasifikasian berita berdasarkan pada objektivitas berita itu sendiri.

Penelitian lainnya dengan judul 'Penanganan *Imbalance Data* pada Klasifikasi Kemungkinan Penyakit Stroke' berkaitan dengan perbandingan antara teknik *oversampling* dan teknik *undersampling* pada *imbalanced* data yang memperoleh nilai akurasi masing-masing sebesar 95% dan 76%. Pentingnya melakukan *balancing* pada penelitian ini adalah untuk meminimalisir kekeliruan dalam pengklasifikasian data (Mutmainah, 2021).

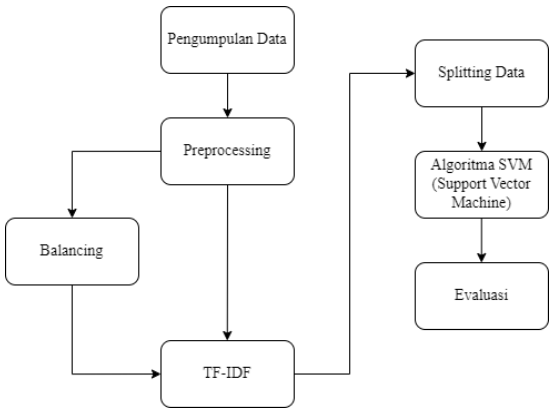
Dari uraian yang telah dijelaskan sebelumnya, penelitian ini akan fokus pada penanganan *imbalanced* data yang berasal dari komentar *Twitter* terhadap program Kampus Merdeka untuk jenjang Perguruan Tinggi. Metode *balancing* yang akan digunakan untuk penelitian ini adalah *Undersampling*, SMOTE, ADASYN, dan *Random Combination Sampling*. Untuk mengetahui performa terbaik, peneliti menerapkan metode *Term Frequency - Inverse Document Frequency* untuk pembobotan dan algoritma SVM (*Support Vector Machine*) untuk klasifikasi data. Data diperoleh dengan proses *scarping* menggunakan Twitter API dan diproses secata penuh menggunakan bahasa pemrograman *python* pada Google Colab. Beberapa *library* pendukung seperti NLTK (*Natural Language Toolkit*) dan Sastrawi juga digunakan dalam tahap *pre-processing.*

METODE PENELITIAN

Klasifikasi teks adalah suatu proses untuk mengkategorikan dokumen ke dalam kelas yang telah didefinisikan sebelumnya (Mutawali, Zaen, & Bagye, 2019). Dalam tugas pengklasifikasian kelas, sangat umum ditemukan kondisi data yang bersifat *imbalanced. Imbalanced data* atau data tidak seimbang adalah sebuah keadaan dimana distribusi kelas data tidak seimbang, jumlah satu kelas dengan jumlah kelas lainnya lebih sedikit atau lebih banyak. Menurut (Fithriasari, Hariastuti, & Wening, 2020), penanganan *imbalanced data* sangat diperlukan karena hal ini menyebabkan mesin akan sulit untuk memprediksikan dengan benar. Dalam *balancing* data terdapat 3 kategori yaitu *undersampling methods, oversampling methods,* dan *hybrid methods* (yang menggabungkan metode sampling data) (Fernandez, et al., 2018). Saat model algoritma dibangun, model akan cenderung mengarah kepada kelas mayoritas sehingga kelas minoritaspun akan diprediksi menjadi kelas mayoritas.

Metode yang dilakukan untuk penelitian ini adalah pengumpulan teks (*scarping*), *pre-processing*, pembobotan kata (*feature extraction*), *balancing data*, *splitting data*, memeriksa performa dari metode *balancing* dengan algoritma SVM, dan evaluasi. Berikut pada gambar 1 menunjukan diagram alir untuk pemrosesan data.

Penelitian ini diproses menggunakan bahasa pemrograman Python pada semua langkah, kecuali pada proses klasifikasi atau *labelling* yang dilakukan secara manual.



Gambar 1. Diagram alir pemrosesan data

A. *Pre-processing*

Setelah melakukan proses pengumpulan data, proses dilanjutkan dengan tahap *pre-processing*. Pada tahap *pre-processing* dilakukan beberapa langkah antara lain *delete punct, remove duplicating*, *delete short instance*, *replacing*, *case folding*, *labelling*, *remove stopwords*, *stemming,* dan tokenisasi.

1. *Delete punct*

Pada tahap pertama, untuk memperoleh data komentar yang lebih representatif diperlukan penghapusan *punctuation* atau tanda baca, hastag, dan *mention*. Hasil dari tahap ini berupa kalimat komentar yang sudah bersih.

2. *Remove duplicating*

Proses ini dilakukan untuk menghapus duplikasi untuk memastikan bahwa tidak ada baris komentar yang sama.

3. *Delete short instance*

Pemrosesan selanjutnya bertujuan untuk menjaga agar komentar memiliki makna yang lengkap sehingga baris yang hanya memiliki kalimat kurang dari 20 karakter atau huruf akan dihapus.

4. *Replacing*

Pada tahap ini, kata - kata yang berupa singkatan dan menggunakan kata pengganti diganti dengan kata yang sesuai dengan KBBI (Kamus Besar Bahasa Indonesia).

5. *Case folding*

Pada tahap selanjutnya, data komentar yang memiliki inkonsistensi penggunaan huruf kapital dan huruf kecil dalam penulisan kalimat akan diubah dan disamaratakan menjadi huruf kecil.

6. *Labelling*

Pada tahap ini, data komentar yang telah dalam kondisi cukup bersih diklasifikasikan ke dalam kelas yang ditentukan yaitu kelas 0 (untuk informasi), kelas 1 (untuk opini), kelas 2 (untuk pertanyaan), dan kelas 3 (untuk *out of topic*).

7. *Remove Stopwords*

Setelah diklasifikasikan, pada tahap ini, data kembali dibersihkan untuk menghapus kata-kata yang tidak diperlukan atau tidak penting pada setiap baris.

8. *Stemming*

Pada tahapan ini, setiap kata dalam data yang memiliki imbuhan awalan maupun akhiran akan dihilangkan dan dikembalikan ke bentuk kata dasarnya.

9. Tokenisasi

Pada tahap terakhir proses *pre-processing*, kalimat akan dipecah ke dalam bagian–bagian kata.

B. *Balancing Data*

Setelah melalui tahap *pre-processing,* data yang dihasilkan dilanjutkan ke tahap *balancing*. *Balancing* perlu dilakukan apabila data mengalami ketidakseimbangan kelas. Hal ini dapat diketahui melalui visualisasi dari data yang telah diklasifikasikan ke dalam kelas sebelumnya. *Balancing* sendiri merupakan sebuah metode untuk menyamakan jumlah data pada setiap kelas. Pada penelitian ini, metode yang digunakan untuk *balancing* adalah *Undersampling*, SMOTE (*Synthetic Minority Over-Sampling Technique*), ADASYN (*Adaptive Synthetic Sampling Approach*), dan *Random Combination Sampling*.

1. *Undersampling*

Undersampling adalah metode untuk mengurangi atau mengeliminasi jumlah data pada kelas mayoritas hingga jumlahnya sama dengan kelas minoritas. Metode yang digunakan dalam teknik ini adalah metode *NearMiss*. *NearMiss* didasarkan pada heuristic informasi. Proses kerja dari *NearMiss* adalah dengan melihat distribusi kelas dan menghilangkan sampel kelas mayoritas secara acak. Apabila jarak kedua kelas mengalami perbedaan yang ekstrim, *NearMiss* secara otomatis akan menghilangkan jarak tersebut dan mencoba untuk menyeimbangkan distribusi kelas (Fernandez, et al., 2018).

2. SMOTE (*Synthetic Minority Over-Sampling Technique*)

SMOTE merupakan salah satu turunan dari metode *oversampling*. SMOTE melakukan replikasi dari data minoritas (yang dikenal sebagai data sintetis/*synthetic*). Menurut (Siringoringo, 2018), metode ini mencari *k nearest neighbors* atau ketetanggaan terdekat data sebanyak nilai *k* untuk setiap data dalam kelas minoritas. Setelah itu, secara acak, data sintetis akan dibuat sebanyak *n* persentase duplikasi sesuai dengan yang diinginkan oleh data minor dan *k-nearest neighbors.*

3. ADASYN (*Adaptive Synthetic Sampling Approach*)

ADASYN juga merupakan salah satu turunan dari metode *oversampling*. ADASYN (Hidayat, Ardiansyah, & Setyanto, 2021) secara adaptif menghasilkan sampel data sintetis untuk kelas minoritas yang dibentuk dari distribusi data yang tidak merata pada kelas mayoritas untuk mengurangi bias. Tujuan dari ADASYN adalah memberikan bobot untuk kelas minoritas (diperoleh dari replikasi lebih banyak pada bagian yang sulit dipelajari) (Fithriasari, Hariastuti, & Wening, 2020).

4. *Random Combination Sampling*

Random Combination Sampling merupakan metode *sampling* yang berasal dari modifikasi peneliti. Metode ini mengkombinasikan antara pengurangan kelas dan replikasi data. Cara kerja dari metode ini adalah mengurangi jumlah kelas mayoritas dan mereplikasi secara acak pada kelas minoritas dengan menggabungkan baris sebanyak jumlah yang diinginkan pada setiap kelas. Untuk penelitian ini, jumlah yang ditentukan dalam pembuatan kelas berkisar diantara 1.950 data hingga 2.100 data.

C. Pembobotan Kata

Setelah dilakukan *balancing*, pada tahapan ini akan dilakukan pembobotan untuk kata yang telah diperoleh pada tahap

pre-processing. Dalam penelitian ini, teknik yang akan digunakan untuk pembobotan kata atau *feature extraction* adalah TF-IDF (*Term Frequency - Inverse Document Frequency*). TF-IDF digunakan untuk mencari representasi nilai dari setiap dokumen dari kumpulan data training, dimana akan dibentuk sebuah vector antara dokumen dengan kata (Widyasanti, dkk., 2018).

D. Pembuatan Model dengan algoritma SVM

Pembuatan model diperlukan untuk memeriksa performa dari metode balancing yang digunakan sebelumnya. Algoritma SVM (Support Vector Machine) digunakan pada penelitian ini karena algoritma ini memiliki performa yang cukup bagus dalam mengolah data dalam kondisi imbalanced. Algoritma ini juga sangat popular untuk menemukan pola pada klasifikasi berdasarkan pada kemampuannya dan fleksibilitasnya dalam beradaptasi di pembelajaran yang sulit (Fernandez, et al., 2018). Karena penelitian ini menggunakan empat kelas, maka algoritma SVM yang digunakan adalah multiclass. Pada algoritma SVM terdapat dua jenis teknik untuk menangani klasifikasi multiclass diantaranya OVO (One Vs One) dan OVA (One Vs All) (Delimayanti, et al., 2021). Perbedaan mendasar pada keduanya terletak pada perbandingannya. OVO perbandingan antar kelas, sedangkan OVA perbandingan satu kelas dengan keseluruhan kelas. Algoritma SVM yang digunakan merupakan bentuk baseline tanpa memodifikasi apapun.

E. Evaluasi

Tahap terakhir dari pemrosesan data merupakan evaluasi. Tahap evaluasi digunakan untuk melihat F1-*score* dari hasil pengujian sebelumnya. F1-*score* digunakan untk melihat indikasi dari model klasifikasi apakah *precision* dan *recall* dari model memiliki hasil yang baik.

HASIL DAN PEMBAHASAN

A. Dataset

Data yang digunakan pada penelitian ini berasal dari *platform* Twitter dengan periode Juni hingga Agustus 2022. Tabel 1 menunjukan sampel tampilan data mentah yang menjadi input pada tahap *pre-processing*.

Dataset yang dikumpulkan berasal dari komentar publik dengan kata kunci "kampus merdeka", "mbkm", dan "merdeka belajar". Pengambilan data dilakukan dengan Twitter API. Data yang diperoleh dalam proses *scarping* sejumlah 16946 baris.

B. *Pre-processing*

Pada tahapan *pre-processing* terdapat beberapa tahapan antara lain *delete punct, remove duplicating*, *delete short instance*, *replacing*, *case folding*, *labelling*, *remove stopwords*, *stemming,* dan tokenisasi. Tahap ini menggunakan *library* python yaitu *Natural Language Toolkit* (NLTK) dan Sastrawi.

Bagian *pre-processing* dibagi menjadi 2 (dua) tahap dalam kondisi sebelum dan setelah *labelling*. Untuk 5 (lima) tahap pertama bagian *pre-processing* akan dilakukan *delete punct, remove duplicating, delete short instance*, *replacing,* dan *case folding*. Dari tahap ini akan memperoleh data yang cukup bersih dan bisa digunakan untuk proses labelling. Sampel dari hasil *pre-processing* pertama ditunjukan pada Tabel 2.

Tabel 1. Data Mentah Komentar *Twitter*

No	Tweet
1	b'knp pas udah lulus, program kampus merdeka br ada'
2	b'kenapa paul dipanggil ui sih emangnya dia kampus merdeka https://t.co/B7IzfMBsDx'
3	b'@naaboahctay @collegemenfess Ga bisa kak, uin dibawah kemenag, kampus merdeka dibawah kemendikbud'
4	b'kampus merdeka is a very interesting plan'
5	b'RT @OposisiCerdas: Breaking News: Istri TNI Ditembak di Depan Rumah\nhttps://t.co/rQoanqs70a'
6	b'@convomf ikut kampus merdeka?'
7	b'guyss ada yang ikutan kampus merdeka ga yaa, aku mau tanya tanya heuheu'
8	b'ya allah ini tes buat kampus merdeka cepet bgt anjai. gw bsk ad kuliah pula'
9	b'RT @unqity: Mendikbudristek\xc2\xa0Nadiem Makarim mendorong Mahasiswa jd pengusaha dgn meluncurkan program wirausaha merdeka pada Jumat 15 Juli 20\xe2\x80\xa6'
10	b'- ada yg udh keterima magang kampus merdeka? kasitau company nya dong manatau samaa'

Selanjutnya dilakukan pelabelan data secara manual. Data dari proses sebelumnya diklasifikasikan ke dalam kelas yang ditentukan yaitu kelas 0 (untuk informasi), kelas 1 (untuk opini), kelas 2 (untuk pertanyaan), dan kelas 3 (untuk *out of topic*). Yang termasuk dalam kelas 3 (*out of* topic) adalah baris yang membahas di luar topik pembicaraan, baris yang tidak menggunakan bahasa Indonesia, dan yang tidak memiliki makna yang jelas. Tabel 3 menunjukan sampel hasil proses *labelling*,

Tabel 2. Data Hasil *Pre-Processing* tahap Pertama

No	Tweet
1	kenapa pas udah lulus program kampus merdeka baru ada
2	kenapa paul dipanggil ui sih emangnya dia kampus merdeka
3	ga bisa kak uin dibawah kemenag kampus merdeka dibawah kemendikbud
4	kampus merdeka sangat asik
5	breaking news istri tni ditembak di depan rumah
6	ikut kampus merdeka
7	guys ada yang ikutan kampus merdeka ga ya aku mau tanya tanya
8	ya allah ini tes buat kampus merdeka cepet bgt anjai saya besok ada kuliah pula
9	mendikbudristek nadiem makarim mendorong mahasiswa jadi pengusaha dengan meluncurkan program wirausaha merdeka pada jumat juli
10	ada yang udah keterima magang kampus merdeka kasitau company nya dong manatau sama

Setelah melalui proses *labelling,* dataset kembali dilakukan pemrosesan dalam python untuk dilanjutkan proses *stopwords*, *stemming*, dan tokenisasi. Pada akhir tahap ini, diperoleh data bersih sejumlah 7883 baris, dimana setiap sampel komentar direpresentasikan dalam bentuk *list* kata dasar dan telah dipecah menjadi kumpulan dari kata - kata. Tabel 4 menunjukan contoh hasil dari *pre-processing* tahap akhir.

Tabel 3. Data yang Telah Diberikan Label

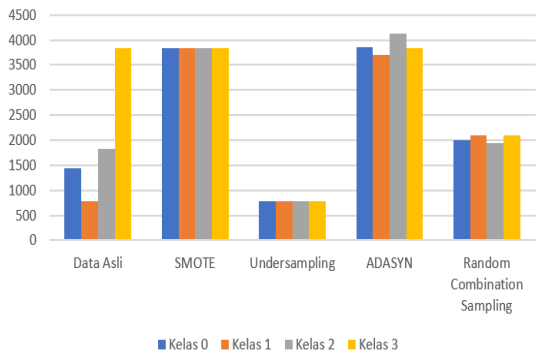
No	Tweet	Label
1	kenapa pas udah lulus program kampus merdeka baru ada	2
2	kenapa paul dipanggil ui sih emangnya dia kampus merdeka	2
3	ga bisa kak uin dibawah kemenag kampus merdeka dibawah kemendikbud	0
4	kampus merdeka sangat asik	1
5	breaking news istri tni ditembak di depan rumah	3
6	ikut kampus merdeka	3
7	guys ada yang ikutan kampus merdeka ga ya aku mau tanya tanya	2
8	ya allah ini tes buat kampus merdeka cepet bgt anjai saya besok ada kuliah pula	3
9	mendikbudristek nadiem makarim mendorong mahasiswa jadi pengusaha dengan meluncurkan program wirausaha merdeka pada jumat juli	0
10	ada yang udah keterima magang kampus merdeka kasitau company nya dong manatau sama	2

Tabel 4. Data Hasil *Pre-Processing* tahap Kedua

No	Tweet	Label
1	[‘kenapa’, ‘pas’, ‘lulus’, ‘program’, ‘kampus’, ‘merdeka’, ‘baru’, ‘ada’]	2
2	[‘kenapa’, ‘paul’, ‘panggil’, ‘ui’, ‘dia’, ‘kampus’, ‘merdeka’]	2
3	[‘ga’, ‘bisa’, ‘kak’, ‘uin’, ‘bawah’, ‘kemenag’, ‘kampus’, ‘merdeka’, ‘bawah’, ‘kemendikbud’]	0
4	[‘kampus’, ‘merdeka’, ‘sangat’, ‘asik’]	1
5	[‘breaking’, ‘news’, ‘istri’, ‘tni’, ‘tembak’, ‘di’, ‘depan’, ‘rumah’]	3
6	[‘ikut’, ‘kampus’, ‘merdeka’]	3
7	[‘guys’, ‘ada’, ‘ikut’, ‘kampus’, ‘merdeka’, ‘engga’, ‘ya’, ‘aku’, ‘mau’, ‘tanya’, ‘tanya’]	2
8	[‘ya’, ‘allah’, ‘tes’, ‘kampus’, ‘merdeka’, ‘cepet’, ‘banget’, ‘anjai’, ‘saya’, ‘besok’, ‘ada’, ‘kuliah’]	3
9	[‘mendikbudristek’, ‘nadiem’, ‘makarim’, ‘dorong’, ‘mahasiswa’, ‘usaha’, ‘luncur’, ‘program’, ‘wirausaha’, ‘merdeka’, ‘jumat’, ‘juli’]	0
10	[‘ada’, ‘sudah’, ‘terima’ ‘magang’, ‘kampus’, ‘merdeka’, ‘kasih’, ‘company, ‘manatau’, ‘sama’]	2

C. *Balancing Data*

Berdasarkan visualisasi persebaran data disetiap kelas, dapat disimpulkan bahwa data bersifat *Imbalanced* sehingga, setelah melalui tahap *pre-processing,* data yang dihasilkan dilanjutkan ke tahap *balancing*. Pada tahap *balancing*, data pada setiap kelas akan diolah untuk diseimbangkan jumlah datanya. Gambar 2 menunjukan visualisasi perbandingan jumlah data dari data asli dan data yang telah dilakukan proses *balancing*.



Gambar 2. Visualisasi jumlah data asli (tanpa *balancing*) dan dengan teknik *balancing* (SMOTE, *Undersampling*, ADASYN, dan RCS)

Dari metode yang telah diterapkan, dari Gambar 2 dapat dilihat untuk metode SMOTE dan ADASYN, dataset di replikasi hingga mencapai jumlah data pada kelas mayoritas. Sedangkan pada metode *undersampling* data kelas mayoritas dikurangi hingga mencapai jumlah data pada kelas minoritas terrendah. Lain halnya dengan metode *random combination sampling* yang mengambil nilai tengah antara kelas minoritas dan kelas mayoritas.

D. Pembobotan Kata

Setelah dilakukan *balancing*, data melalui proses pembobotan menggunakan TF-IDF (*Term Frequency – Inverse Document Frequency*). Proses ini bertujuan mengekstrak informasi pada data, sehingga klasifikasi dapat dilakukan berdasarkan informasi tersebut. Dari dataset yang diproses sebelumnya, bobot angka akan diberikan sesuai dengan jumlah kemunculannya pada setiap dokumen (Amira, Utama, & Fahmi, 2020). Tabel 5 menunjukan sampel pembobotan menggunakan TF-IDF.

Dalam tahap pembobotan, untuk mencari hasil terbaik dilakukan beberapa percobaan dengan mengubah *max features* yang digunakan untuk membatasi jumlah jumlah fitur atau kata dari dataset yang akan dihitung skor TF-IDF-nya. Semakin tinggi frekuensi kata muncul dalam dokumen maka akan semakin tinggi pula nilai TF-IDFnya. Pada penelitian ini, *max_features* yang digunakan adalah 3000, 5000, dan 7000.

Tabel 5. Hasil Pembobotan Kata TF-IDF

Hasil TF-IDF	
(0, 24)	0.33179397764698754
(0, 15)	0.2820551936906091
(1, 48)	0.3685786183397816
(1, 36)	0.3685786183397816
(1, 11)	0.3685786183397816
.	
.	
.	
(8, 39)	0.16418226474950082
(8, 25)	0.17997535094081832
(9, 30)	0.3268419060182692
(9, 0)	0.3268419060182692
(9, 8)	0.3268419060182692

E. Klasifikasi dengan algoritma SVM

Sebelum digunakan dalam proses klasifikasi dalam algoritma, data dibagi menjadi dua bagian yaitu untuk pelatihan (*training)* dan untuk pengujian (*testing*). Perbandingan komposisi data *training* dan *testing* yang digunakan 80:20. Untuk pengacakan data pada penelitian ini menggunakan parameter *random_state* = 75.

F. Evaluasi

Tahap terakhir dari pemrosesan data merupakan evaluasi. Dalam tahap ini dapat dilihat kinerja masing-masing metode *balancing* dengan menggunakan nilai akurasi dan nilai F1-*score*. Nilai akurasi digunakan untuk melihat prediksi benar dari dataset secara keseluruhan. Evaluasi F1-*score* yang melibatkan perbandingan *precision* dan *recall* dinilai lebih tepat untuk mengukur kinerja klasifikasi pada dataset yang tidak seimbang.

Tabel 6 menunjukan bahwa hasil akurasi klasifikasi dengan nilai *max_features* yang berbeda, konsisten menunjukan bahwa metode ADASYN merupakan metode paling baik dengan nilai akurasi sebesar 0,903.

Tabel 7 menunjukan bahwa hasil *F-1 Score* dengan nilai *max_features* yang berbeda juga

konsisten menunjukan bahwa metode ADASYN merupakan metode paling baik dengan nilai akurasi sebesar 0,9. Pada Tabel 6 dan Tabel 7 kita dapat melihat bahwa nilai *max-features* yang berbeda pada TF-IDF tidak begitu menunjukan perbedaan berarti, namun secara umum *max_feature*=5000 lebih banyak menunjukan hasil klasifikasi yang lebih tinggi. Nilai akurasi dan *F1-Score* pun menunjukkan hasil yang serupa.

Tabel 6. Perbandingan Nilai Akurasi dari Berbagai Metode dan Nilai *Max Features*

Metode	Max features		
	3000	5000	7000
Data Asli	0,702	0,707	0,706
Undersampling	0,599	0,619	0,623
SMOTE	0,888	0,896	0,890
ADASYN	**0,895**	**0,903**	**0,896**
Random Combination Sampling	0,801	0,794	0,785

Tabel 7. Perbandingan Nilai *F-1 Score* dari Berbagai Metode dan Nilai *Max Features*

Metode	Max features		
	3000	5000	7000
Data Asli	0,7	0,7	0,7
Undersampling	0,6	0,62	0,62
SMOTE	0,89	0,9	0,89
ADASYN	**0,9**	**0,9**	**0,9**
Random Combination Sampling	0,8	0,79	0,79

Salah satu hal yang menarik pada hasil yang diperoleh adalah bahwa metode *undersamping* yang diterapkan pada dataset peneliti tidak mampu meningkatkan kinerja klasifikasi. Walaupun pada beberapa referensi seperti pada penelitian dengan judul 'Penerapan Teknik Sampling untuk Mengatasi Imbalance Class pada Klasifikasi Online Shoppers Intention' (Ardiyansyah & Rahayuningsih, 2020) metode *undersampling* ini dapat menunjukkan hasil yang lebih baik dengan adanya peningkatan hingga mencapai 4% dari nilai akurasi data awal dan percobaan pada beberapa algoritma seperti KNN, Naïve Bayes, dan Random Forest. Namun, pada dataset peneliti *undersampling* menunjukan nilai akurasi dan nilai F-1 Score yang lebih buruk dari klasifikasi yang dilakukan terhadap data asli. Hasil *undersampling* yang lebih buruk juga pernah ditemukan pada penelitian 'O*versampling, Undersampling, Smote SVM* dan *Random Forest* pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017' dimana hasil akurasi dari data asli 0,9672, sedangkan untuk hasil *undersampling* 0,032 dengan komposisi pembagian data *training* 75% : *testing* 25% dan algoritma SVM (*Support Vector Machine*) (Qadrini, Hikmah, & Megasari, 2022).

SIMPULAN

Berdasarkan penelitian yang dilakukan, dapat disimpulkan bahwa hasil performa terbaik dari metode penanganan *imbalanced dataset* pada data yang peneliti kumpulkan adalah dengan metode *oversampling* menggunakan ADASYN. Terdapat penurunan akurasi dan nilai *F-1 Score* yang terjadi pada metode *undersampling* yang dapat terjadi karena mesin belajar pada jumlah dataset yang lebih sedikit. Penggunaan nilai *max_features* pada pembobotan TF-IDF tidak begitu menunjukan perbedaan berarti, namun secara umum nilai *max_features*=5000 lebih banyak menunjukan hasil klasifikasi yang lebih tinggi.

DAFTAR PUSTAKA

Amira, S. A., Utama, S., & Fahmi, M. H. (2020). Penerapan Metode Support Vector Machine untuk Analisis Sentimen pada Review Pelanggan Hotel. *Edu Komputika Journal, 7*(2), 40-48. doi:

Ardiyansyah, & Rahayuningsih, P. A. (2020). Penerapan Teknik Sampling untuk Mengatasi Imbalance Class pada Klasifikasi Online Shoppers Intention. *Jurnal Teknik Informatika Kaputama (JTIK), 4*(1), 7-15. doi:

Delimayanti, M.K., Sari, R., Laya, M., Faisal, M. R., & Pahrul. (2021). Pemanfaatan Metode Multiclass-SVM pada Model Klasifikasi Pesan Bencana Banjir di Twitter. *Edu Komputika Journal 8*(1), 39-47.

Fernandez, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalance Data Sets.* Cham, Switzerland: Springer. doi:10.1007/978-3-319-98074-4

Fithriasari, K., Hariastuti, I., & Wening, K. S. (2020). Handling Imbalance Data in Classification Model with Nominal Predictors. *International Journal of Computing Science and Applied Mathematics, 6*(1), 33-37.

Fitriani, R. D., Yasin, H., & Tarno. (2021). Penanganan Klasifikasi Kelas Data Tidak Seimbang dengan Random Oversampling pada Naive Bayes. *Jurnal Gaussian, 10*(1), 11-20.

Hidayat, W., Ardiansyah, M., & Setyanto, A. (2021). Pengaruh Algoritma ADASYN dan SMOTE terhadap performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb. *Edumatic: Jurnal Pendidikan Informatika, 5*(1), 11-20.

Kampus Merdeka. (2021). *Program Kampus Merdeka*. Retrieved from Kampus Merdeka: https://kampusmerdeka.kemdikbud.go.id/program

Kasanah, A. N., Muladi, & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 3*(2), 196-201. doi: 10.29207/resti.v3i2.945

Kholila, N. (2021). Analisis Sentimen terhadap Program Merdeka Belajar - Kampus Merdeka pada Twitter menggunakan Support Vector Machine (SVM). ANTIVIRUS: Jurnal Ilmiah Teknik Informatika, 15(2), 252-261. doi:10.35457/antivirus.v15i2.1866

Mutawali, L., Zaen, M. T., & Bagye, W. (2019). Klasifikasi Teks Sosial Media Twitter menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto). *Jurnal Informatika & Rekayasa Elektronika (JIRE), 2*(2), 43-51.

Mutmainah, S. (2021). Penanganan Imbalance Data pada Klasifikasi Kemungkinan Penyakit Stroke. *Jurnal SNATi, 1*(1), 10-16.

Pipin, S.J., Kurniawan, H. (2022). Analisis Sentimen Kebijakan MBKM berdasarkan Opini Masyarakat di Twitter menggunakan LSTM. Jurnal SIFO Mikroskil, 23(2), 197-208.

Pratama, I., Chandra, A. Y., & Prasetyaningrum, P. T. (2021). Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN. *Jurnal Eksplora Informatika, 11*(1), 38-49. doi:10.30864/eksplora.v11i1.578

Qadrini, L., Hikmah, & Megasari. (2022). Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017. *Journal of Computer System and Informatics (JoSYC), 3*(4), 386-391. doi:10.47065/josyc.v3i4.2154

Siringoringo, R. (2018). Klasifikasi Data Tidak Seimbang menggunakan Algoritma SMOTE dan k-Nearest Neighbor. *Jurnal ISD, 3*(1), 44-49.

Zhafira, D.F., Rahayudi, B., Indriati. (2021). Analisis Sentimen Kebijakan Kampus Merdeka menggunakan Naïve Bayes dan Pembobotan TF-IDF berdasarkan Komentar pada Youtube. Jurnal Sistem Informasi, Teknologi Informasi, dan Edukasi Sistem Informasi (JUST-SI), 2(1), 55-63.