P-ISSN: 1979-5661
E-ISSN: 2622-321X

Balancing Dataset Untuk Klasifikasi Komentar Program Kampus Merdeka Menggunakan *Synonym Replacement*

Soleh Nifanto[a1], Ade Nurhopipah[a2]

[a] Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
Jl. Letjend Pol. Soemarto No.127 Purwokerto Utara, Banyumas, Jawa Tengah, Indonesia
[1]solehnifanto86@email.com (Corresponding author)
[2]ade_nurhopipah@amikompurwokerto.ac.id

Abstrak

Klasifikasi komentar dalam program Kampus Merdeka merupakan langkah penting dalam menganalisis sentimen pengguna terhadap berbagai fitur dan layanan yang ditawarkan oleh program tersebut. Namun demikian pada dataset yang diolah dalam penelitian ini, terdapat masalah yang dihadapi yaitu ketidakseimbangan jumlah data pada masing-masing kelas. Rasio ketidakseimbangan pada dataset tersebut cukup tinggi yaitu sebesar 5:1. Ketidakseimbangan ini umumnya mengakibatkan model klasifikasi cenderung memprioritaskan kelas mayoritas dan menghasilkan kinerja yang rendah pada kelas minoritas. Oleh karena itu, suatu pendekatan *augmentasi data* digunakan dalam penelitian ini dengan metode *Synonym Replacement* untuk menghasilkan variasi data dalam kelas minoritas, sehingga mengurangi ketidakseimbangan dan meningkatkan kinerja klasifikasi. Metode ini memanfaatkan teknik penggantian sinonim dalam kalimat-kalimat pada komentar dengan harapan dapat memperkaya dataset dan meningkatkan representasi fitur. Hasil dari penelitian menunjukan peningkatan nilai *F-Measure* dari 0,6672 menjadi 0,7875. Evaluasi menggunakan ROC menunjukan nilai maksimum sebesar 0,96. Sedangkan kelas yang tidak mendapatkan *augmentasi* memiliki kecenderungan nilai ROC yang rendah di antara 0,81 sampai 0,88.

Kata Kunci: Analisis Sentimen, Augmentasi Data, Kampus Merdeka, Klasifikasi, *Synonym Replacement.*

Abstract

The classification of comments in the Merdeka Campus program is an essential step in analyzing user sentiment towards the various features and services offered by the program. However, in the dataset processed in this study, problems are encountered, namely the imbalance of the amount of data in each class. The Imbalanced Ratio in this dataset is relatively high by 5:1. This generally leads to a classification model that prioritizes the majority class and results in low performance in the minority class. Therefore, a data augmentation approach is used in this study with the Synonym Replacement method to produce data variations in minority classes, thereby reducing the imbalance and improving classification performance. This method utilizes the technique of replacing synonyms in sentences in comments to enrich the dataset and increase the representational features. The study's results showed an increase in the F-Measure value from 0.6672 to 0.7875. Evaluation using ROC shows a maximum value of 0.96. In contrast, the class that did not get augmentation tended to have low ROC values between 0.81 to 0.88.

Keywords: *Classification, Data Augmentation, Kampus Merdeka, Sentiment Analysis, Synonym Replacement.*

1. Pendahuluan

Dalam era digital yang semakin maju, *platform* media sosial seperti Twitter telah menjadi sarana penting bagi pengguna untuk berbagi pandangan dan pendapat mereka secara langsung. Dalam

konteks tersebut, analisis sentimen atau klasifikasi komentar menjadi topik yang menarik perhatian para peneliti dan praktisi[1]. Meskipun ada banyak algoritma klasifikasi yang telah dikembangkan, namun tetap ada tantangan dalam menghadapi masalah yang kompleks seperti variabilitas bahasa dan skala dataset yang tidak seimbang.

Salah satu pendekatan yang umum digunakan untuk memproses data teks adalah pemrosesan bahasa alami atau *Natural Language Processing* (NLP). Pendekatan berbasis NLP telah menjadi salah satu bidang penelitian yang kritis dalam analisis sentimen dan klasifikasi teks. NLP memungkinkan komputer untuk memahami, memproses, dan memanipulasi bahasa manusia dengan cara yang serupa dengan manusia[2], [3]. Dalam konteks klasifikasi komentar pada Program Kampus Merdeka di Twitter, pendekatan NLP dapat memberikan kerangka kerja yang kuat untuk memahami dan memproses teks secara otomatis.

Dalam bidang NLP, terdapat metode augmentasi data atau teknik pembuatan dataset tiruan. Augmentasi data berpotensi mengurangi *overfitting* dengan cara meningkatkan keragaman data pelatihan sehingga meningkatkan kemampuan generalisasi model yang dilatih. Menurut Madukwe dkk, augmentasi data untuk jenis data teks dapat dibagi menjadi dua yaitu *local data augmentation* di mana augmentasi dilakukan dengan perubahan di wilayah lokal dalam kalimat aslinya dan *global data augmentation* di mana perubahan akan secara global memengaruhi seluruh kalimat. Contoh *local data augmentation* adalah *synonym replacement, random swap, insertion* dan *deletion*. Sedangkan contoh *global data augmentation* diantaranya menggunakan teknik *paraphrasing, back translation* dan *text generation* dengan model Bahasa[4].

Synonym replacement merupakan salah satu metode yang efisien dan metode yang mudah diakses untuk augmentasi data teks. Dalam teknik ini kata-kata dalam teks diperbarui dengan sinonimnya untuk memperluas keragaman dan representasi data yang lebih baik. Contoh penggunaan *synonym replacement* telah dilakukan pada penelitian [5] untuk pengolahan *essay scoring* dan pada penelitian [6] untuk meningkatkan performa model yang diuji pada beberapa dataset. Dalam penelitian ini fokus pembahasan adalah pada penerapan teknik *synonym replacement* terhadap dataset komentar Program Kampus Merdeka di Twitter untuk mengatasi masalah ketidakseimbangan dataset.

Tujuan utama dari penelitian ini adalah untuk membuat dataset komentar Program Kampus Merdeka yang lebih seimbang dengan harapan bahwa teknik ini dapat meningkatkan performa klasifikasi komentar secara keseluruhan. Dalam penelitian ini, penggantian kata-kata dalam komentar dengan sinonim yang relevan dapat menciptakan variasi sintaksis dan semantik yang lebih kaya, sehingga meningkatkan representasi data dan meminimalkan bias yang mungkin terjadi. Dalam konteks klasifikasi komentar pada Program Kampus Merdeka dari Twitter, penggantian sinonim dapat membantu menghadapi perbedaan gaya penulisan, kosakata yang berbeda, atau bahkan perubahan makna kata yang umum terjadi dalam konteks sosial media.

Penelitian sebelumnya dengan dataset yang sama telah dilakukan oleh Magnolia [7] dan Nurhopipah [8], namun penelitian ini menggunakan metode augmentasi data untuk data numerik seperti Near Miss, Tomek Links, SMOTE dan ADASYN. Pada penelitian ini dilakukan augmentasi langsung pada level kalimat, sehingga data yang dihasilkan lebih real dan relevan. Demikian juga dengan menggunakan metode *synonym replacement* diharapkan tidak ada bias dalam pelabelan karena kelas atau label yang dibangkatkan dipastikan sesuai.

Manfaat utama dari teknik penggantian sinonim adalah peningkatan variasi dalam teks, yang dapat membantu model klasifikasi untuk mempelajari dan mengenali pola yang lebih banyak dalam dataset. Dengan memperkenalkan variasi kata-kata melalui penggantian sinonim, diharapkan dapat meningkatkan kemampuan model untuk mengenali sentimen atau kategori komentar dengan lebih baik. Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang berharga tentang penggunaan teknik *synonym replacement* dalam mengatasi ketidakseimbangan dataset dalam klasifikasi komentar di Twitter. Selain itu, penelitian ini juga diharapkan dapat memberikan kontribusi pada pengembangan metode dalam analisis sentimen dalam bidang NLP secara umum[9].

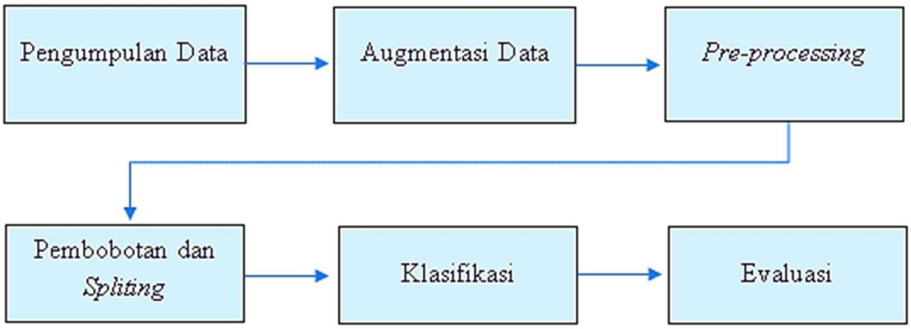
Dalam jurnal ini, dijelaskan metodologi yang digunakan, termasuk deskripsi dataset, teknik penggantian sinonim yang diterapkan, dan algoritma klasifikasi yang digunakan. Terdapat juga laporan dan analisis hasil dari eksperimen yang dilakukan, serta temuan dan implikasi dari penelitian ini.

P-ISSN: 1979-5661
E-ISSN: 2622-321X

2. Metode Penelitian

2.1. Diagram Alur Penelitian

Alur penelitian ini terdiri dari enam langkah dengan fokus penelitian pada langkah augmentasi data. Penelitian ini terbatas pada ruang lingkup pengolahan data teks dan klasifikasi data teks. Diagram alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

2.2. Tahapan Penelitian

1. Pengumpulan Data
   Tahap pertama penelitian adalah pengumpulan data di mana dataset diperoleh dengan meminta izin pada peneliti sebelumnya. Dataset berisi komentar terhadap program Merdeka Belajar Kampus Merdeka (MBKM) yang diperoleh menggunakan Twitter API dalam kurun waktu Juni 2022 hingga Agustus 2022. Dataset terdiri dari 7883 data teks berisi komentar terhadap program MBKM[7].
2. Augmentasi Data
   Tahap selanjutnya adalah augmentasi data. Sebelum melakukan augmentasi data terlebih dahulu dilakukan pemisahan data berdasar kelas. Setelah pemisahan data, maka augmentasi data dilakukan dengan menyamanyakan jumlah data pada setiap kelas. Metode yang digunakan dalam augmentasi data adalah *synonym replacement* seperti dilakukan oleh penelitian [10]. Metode ini dilakukan dengan mengganti beberapa kata dengan kata sinonim dan menjadikan teks hasil pergantian sinonim menjadi data baru. Modul kamus sinonim bahasa indonesia digunakan sebagai acuan untuk mengganti kata dengan sinonimnya. Kamus ini dibangun berdasarkan kesamaan makna kata yang terdapat dalam Kamus Besar Bahasa Indonesia(KBBI). Metode *synonym replacement* dilakukan dengan mengambil kalimat dari dataset yang akan di augmentasi, kemudian 20% kata dari kalimat yang dipilih diambil secara acak dan digantikan dengan sinonim kata yang ada pada kamus yang dibangun.
3. *Pre-processing*
   Tahap selanjutnya adalah *pre-processing* dengan melakukan normalisasi *dataset*. Tahap ini terdiri dari tahapan penghapusan tanda baca, penghapusan data ganda, penggantian kata menjadi kata baku, mengubah semua huruf pada dataset ke bentuk huruf kecil, penghapusan *stopword* dan mengekstrak kata dasar dari kalimat atau disebut *stemming*. Proses ini dilakukan dengan menggunakan *library* Natural Language Toolkit(NLTK) dan sastrawi.
4. Pembobotan dan *Spitting*
   Tahap pembobotan dan *splitting* data adalah tahap di mana data yang sebelumnya dalam bentuk teks kemudian diberi bobot menjadi bentuk *vector*. Pembobotan dilakukan dengan *Term Frequency-Inverse Document Frequency* (TF-IDF). Parameter yang digunakan pada

P-ISSN: 1979-5661
E-ISSN: 2622-321X

saat dilakukan pembobotan adalah *max feature* yang merupakan jumlah kata yang sering muncul dari daftar kata pada dataset [9]. Parameter *max feature* yang digunakan pada penelitian ini sebesar 5000. Hasil pembobotan menghasilkan *vektor* yang merupakan gambaran bobot setiap kata berdasarkan frekuensi kemunculannya dan frekuensi dokumen di mana kata tersebut muncul. Teks dengan nilai bobot besar menggambarkan bahwa dalam teks tersebut memiliki banyak kata yang unik, sedangkan teks dengan nilai bobot kecil menggambarkan bahwa kata dalam teks sering muncul di dokumen lain. Selanjutnya dilakukan tahap *splitting* di mana dataset yang sudah dinormalisasi dipecah menjadi data *training* dan data *testing.* Data dibagi dengan perbandingan 80:20 masing-masing untuk *data training* dan *data testing.*

5. Klasifikasi
   Tahapan ini dilakukan dengan menggunakan *classifier Support Vector Machines* (SVM). Kernel yang digunakan adalah *linear, poly, rbf* dan *sigmoid*. Klasifikasi dilakukan empat kali berdasarkan jumlah kernel sehingga menghasilkan empat hasil klasifikasi. Proses klasifikasi ini merupakan sarana untuk memvalidasi pengaruh *balancing dataset* karena hasil augmentasi data dapat dinilai berdasar dari hasil klasifikasi[11].
6. Evaluasi
   Tahap ini menunjukan hasil augmentasi data berdasar hasil dari klasifikasi dengan empat kernel yang digunakan pada SVM. Hasill klasifikasi akan ditampilan dengan *confusion matrix* dan grafik ROC-AUC. *Confusion matrix* adalah matriks yang menggambarkan hasil prediksi. Dengan matriks ini dapat dihitung nilai *recall*, *precision, F1-score* dan akurasi dari *classifier*. ROC-AUC adalah grafik yang menunjukan perbandingan kinerja tiap kernel *classifier* terhadap data pada tiap kelas.

3. Hasil dan Pembahasan

3.1. Pengolahan Data

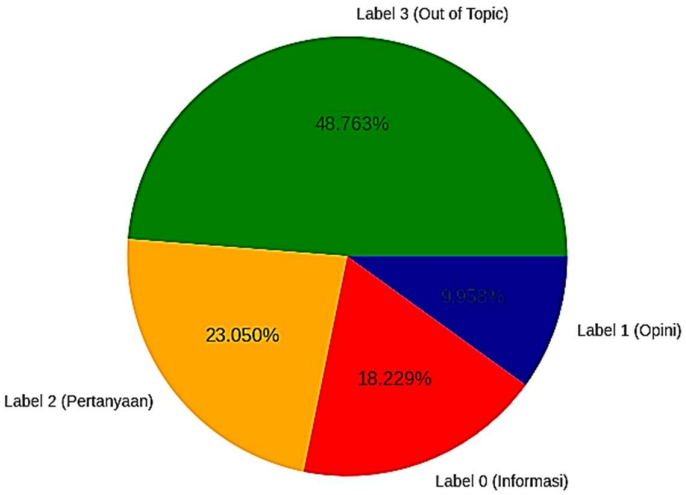
Dataset awal yang digunakan adalah sebanyak 7.883. Kemudian dataset dibagi menjadi empat kelas dan diberi label untuk setiap kelas yaitu label 0 untuk kelas informasi, label 1 untuk kelas opini, label 2 untuk kelas pertanyaan dan label 3 untuk kelas *out of topic* atau komentar yang tidak berkaitan. Sampel hasil pelabelan dataset dapat dilihat pada Tabel 1.

Tabel 1. Sampel Dataset

Isi	Label
penyampaian materi merdeka belajar teknik jaya muda sriwijaya oleh ibu leily nurul komariah sangat memotivasi dan seru	0
Kamis telah dilaksanakan pelepasan mahasiswa jurusan teknik mesin untuk magang industri mandiri kampus merdeka batch ii di pt imip	0
sebenarnya sudah merdeka belajar atau baru belajar merdeka wahai	1
capek sama kurikulum merdeka belajar	1
ada yang ketrima mbkm di paragon apa ada grupnya ya buat yang uda ketrima makasih	2
adakah cewe yang ketrima mbkm di paragon terus penempatan di jakarta aku mau ajakin kos bareng	2
aloo ini aku belajar bio sama ngerjain lkpnya juga mayan seru iya kalian bio udah sampe materi mana kalian masih lanjut kurikulum yang lama or sudah ganti yang merdeka	3
hari merdeka aku belajar kimia plus bikin fanadadakan pas hari h walau ada gempuran tugas pjok	3

Dataset tersebut terdiri dari 1.437 untuk kelas informasi, 785 untuk kelas opini, 1.817 untuk kelas pertanyaan dan 3.844 untuk kelas *out of topic.* Dapat dilihat bahwa hampir separuh dataset terdiri dari kelas ke 3 atau kelas *out of topic* dengan persentase 48,763%. Pembagian kelas ditunjukan pada pada Gambar 2.

P-ISSN: 1979-5661
E-ISSN: 2622-321X



Gambar 2. Pembagian Kelas pada Dataset

Berdasarkan diagram pembagian kelas terdapat ketidakseimbangan dataset. Ketidakseimbagan dataset ini dapat diindikasikan dengan suatu rasio yang disebut *Imbalanced Ratio* (IR). Nilai IR dapat dihitung dengan membandingkan jumlah kelas mayoritas dengan kelas minoritas. Nilai IR pada dataset ini adalah 3844/785=4,8968. Dengan kata lain rasio ketidakseimbangan data sebesar 5:1, sehingga perlu dilakukan augmentasi data pada kelas 0, 1 dan 2.

3.2. Augmentasi Data

Berdasarkan proses pengolahan data, maka metode augmentasi diperlukan. Metode augmentasi yang digunakan adalah metode pergantian sinonim (*synonym replacement*). Sebelum melakukan augmentasi, dataset dipecah menjadi empat kelas. Karena jumlah data di kelas 0, 1 dan 2 terlalu sedikit, namun jumlah data di kelas 3 sangat banyak, maka pada penelitian ini, jumlah data pada semua kelas akan diseragamkan sebanyak 2.000. Hal ini dilakukan agar augmentasi data tidak dilakukan secara berlebihan. Oleh karena itu, pada kelas 3 dengan jumlah data sebanyak 3.844 tidak dilakukan augmentasi melainkan diambil 2.000 data secara acak. Sedangkan kelas dengan data kurang dari 2.000 maka akan dilakukan augmentasi hingga jumlah data mencapai 2.000. Data pada kelas 3 yang diambil dapat dilihat pada Gambar 3.

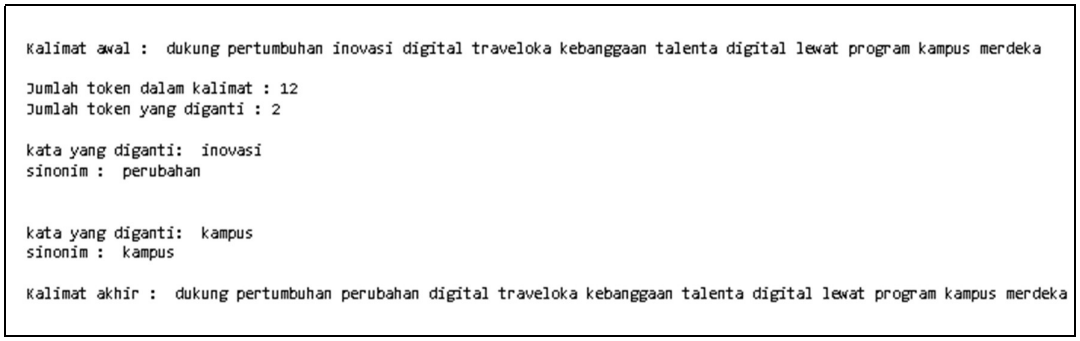
```
                                                   isi  label
863   believe me sistur di kelas uda jarang anak kar...      3
4     hari merdeka aku belajar kimia plus bikin fana...      3
473   aku ikutt mbkm tahun kemarin dan di konversi k...      3
7269   dulu aku mbkm darisampaimeskipun nilainya bar...      3
7645   kadariun goblog  kan dah gua bilang rendang a...      3
...                                                  ...    ...
7797   program kampus merdeka caa tapi ehange ke ln ...      3
7464  gabung tempat magang internship online samak s...      3
7330  gabung tempat magang internship online kampus ...      3
7671   pas thn lalu masih semester  cv ala kadarnya ...      3
7201  kegagalan kita mengekalkan budaya membaca buka...      3

[2000 rows x 2 columns]
```

Gambar 3. Data Kelas 3

Metode augmentasi yang dilakukan adalah mengambil kalimat dari kelas yang akan ditambah datanya, kemudian secara acak mengambil 20% kata pada kalimat tersebut dan menggantinya dengan sinonimnya yang terdapat pada kamus sinonim. Kamus sinonim tersebut terdiri dari kata-kata berbahasa Indonesia beserta sinonimnya yang dibuat berdasarkan pada kesamaan makna kata dalam KBBI. Kamus sinonim yang digunakan berasal dari akun github *victoriasovereigne* yang merupakan versi json kamus sinonim Tesaurus bahasa Indonesia. Selanjutnya kamus

P-ISSN: 1979-5661
E-ISSN: 2622-321X

tersebut ditambahkan dengan kata-kata baru yang sering muncul pada dataset namun tidak ada di kamus. Contoh hasil dari pergantian sinonim dapat dilihat pada Gambar 4.



Gambar 4. Hasil Proses *Synonym Replacement*

Pada tampilan tersebut kalimat awal yang diambil memiliki dua belas token. Penelitian ini mengambil batas *replacement* sebanyak 20 persen per kalimat, sehingga pada kalimat tersebut diambil dua kata untuk diganti dengan sinonimnya. Selain itu ada beberapa kata yang tidak diganti, contohnya kata “merdeka”, “belajar” dan “kampus” untuk menjaga konteks kalimat “merdeka belajar kampus merdeka”. Pada contoh tersebut kata yang diganti adalah “inovasi” dengan sinonimnya “perubahan” sedangkan kata “kampus” tidak diganti. Prosedur lainya yang kami lakukan adalah jika kata yang diambil dari kalimat tidak memiliki sinonim pada kamus sinonim maka akan dikembalikan ke kata asli.

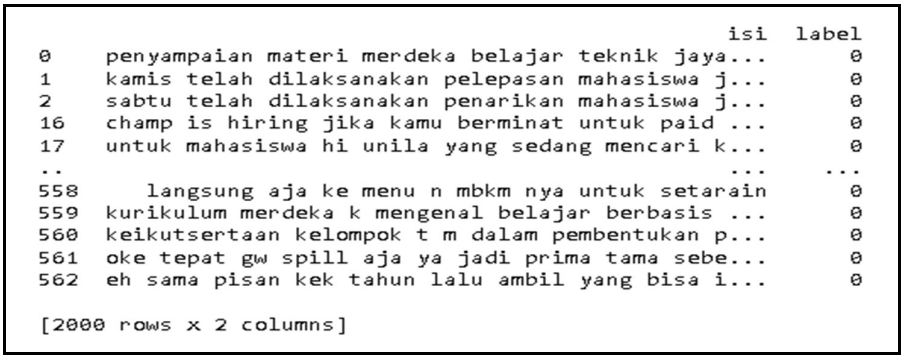
Berdasarkan proses augmentasi, berikut adalah hasil augmentasi dari beberapa kalimat pada kelas 0, 1 dan 2 yang ditunjukan pada Tabel 2.

Tabel 2. Perbandingan Kalimat Asli Dengan Hasil Proses *Synonym Replacement*

kalimat asli	kalimat augmentasi	kelas
narasumber darisandi budi iriawan mpdlandasan filosofis dan teoritis kurikulum merdeka belajar sabtujuni	narasumber darisandi budi iriawan mpdlandasan filosofis dan teoritis kompendium merdeka belajar sabtujuni	0
tahun ajaran baru kurikulum merdeka belajar mulai diterapka nselengkapinnya	warsa ajaran baru kompendium merdeka belajar mulai diterapka nselengkapinnya	0
tanpa bermaksud d menyinggung tetapi diterima mbkm magang merdeka batch tahun semuanya menyebalkan	tanpa bermaksud d menyinggung tetapi diterima mbkm magang merdeka batch warsa semuanya menyebalkan	1
mbkm tf bener menarik sekali	mbkm tf bener merentangkan sekali	1
gais kalian yang lolos magang kampus merdeka ipeka nya pada diatas	gais kamu yang lolos magang kampus merdeka ipeka nya pada diatas	2
hi minskill setelah test online mbkm juli kmrn kapan pengumuman lanjuata ya ya terima kasih	hi minskill setelah test online mbkm juli kmrn bilamana penyiaran lanjuata ya ya songsong kasih	2

Berdasarkan tabel hasil augmentasi, berikut adalah gambar hasil augmentasi data untuk setiap kelas yaitu kelas 0, 1 dan 2. Kelas informasi atau kelas 0 dengan jumlah data 1.437 memerlukan 563 data agar menjadi 2.000 data. Augmentasi dilakukan dengan mengambil 563 data dari kelas 0 kemudian mengganti beberapa kata dengan sinonimnya. Selanjutnya data ditambahkan dengan data asli sehingga data pada kelas 0 menjadi 2.000 data. Dataset hasil augmentasi ditunjukan pada Gambar 5.

P-ISSN: 1979-5661
E-ISSN: 2622-321X

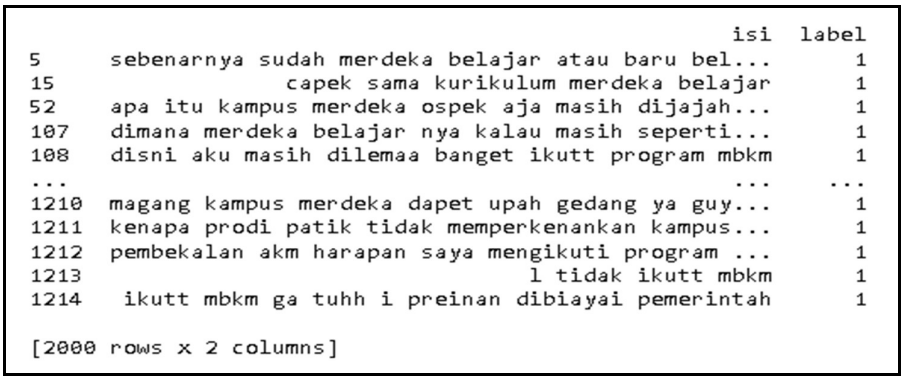


	isi	label
0	penyampaian materi merdeka belajar teknik jaya...	0
1	kamis telah dilaksanakan pelepasan mahasiswa j...	0
2	sabtu telah dilaksanakan penarikan mahasiswa j...	0
16	champ is hiring jika kamu berminat untuk paid ...	0
17	untuk mahasiswa hi unila yang sedang mencari k...	0
..	...	...
558	langsung aja ke menu n mbkm nya untuk setarain	0
559	kurikulum merdeka k mengenal belajar berbasis ...	0
560	keikutsertaan kelompok t m dalam pembentukan p...	0
561	oke tepat gw spill aja ya jadi prima tama sebe...	0
562	eh sama pisan kek tahun lalu ambil yang bisa i...	0

[2000 rows x 2 columns]

Gambar 5. Dataset Kelas 0 Hasil Augmentasi

Kelas opini atau kelas 1 dengan jumlah data 785, memerlukan 1.215 data agar menjadi 2.000 data. Augmentasi dilakukan dengan mengambil data dari kelas 1 kemudian mengganti beberapa kata dengan sinonimnya. Selanjutnya data ditambahkan dengan data asli sehingga data pada kelas 1 menjadi 2.000 data. Dataset hasil augmentasi dapat dilihat pada Gambar 6.

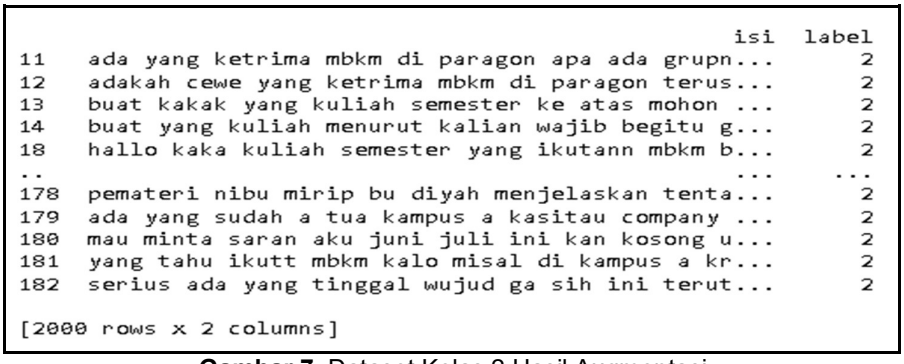


	isi	label
5	sebenarnya sudah merdeka belajar atau baru bel...	1
15	capek sama kurikulum merdeka belajar	1
52	apa itu kampus merdeka ospek aja masih dijajah...	1
107	dimana merdeka belajar nya kalau masih seperti...	1
108	disni aku masih dilemaa banget ikutt program mbkm	1
...	...	...
1210	magang kampus merdeka dapet upah gedang ya guy...	1
1211	kenapa prodi patik tidak memperkenankan kampus...	1
1212	pembekalan akm harapan saya mengikuti program ...	1
1213	l tidak ikutt mbkm	1
1214	ikutt mbkm ga tuhh i preinan dibiayai pemerintah	1

[2000 rows x 2 columns]

Gambar 6. Dataset Kelas 1 Hasil Augmentasi

Kelas informasi atau kelas 2 dengan jumlah data 1.817 memerlukan 183 data agar menjadi 2.000 data. Augmentasi dilakukan dengan mengambil 183 data dari kelas 2 kemudian mengganti beberapa kata dengan sinonimnya. Selanjutnya data ditambahkan dengan data asli sehingga data pada kelas 2 menjadi 2.000 data. Dataset hasil augmentasi dapat dilihat pada Gambar 7.



	isi	label
11	ada yang ketrima mbkm di paragon apa ada grupn...	2
12	adakah cewe yang ketrima mbkm di paragon terus...	2
13	buat kakak yang kuliah semester ke atas mohon ...	2
14	buat yang kuliah menurut kalian wajib begitu g...	2
18	hallo kaka kuliah semester yang ikutann mbkm b...	2
..	...	...
178	pemateri nibu mirip bu diyah menjelaskan tenta...	2
179	ada yang sudah a tua kampus a kasitau company ...	2
180	mau minta saran aku juni juli ini kan kosong u...	2
181	yang tahu ikutt mbkm kalo misal di kampus a kr...	2
182	serius ada yang tinggal wujud ga sih ini terut...	2

[2000 rows x 2 columns]

Gambar 7. Dataset Kelas 2 Hasil Augmentasi

Kemudian dataset pada setiap kelas digabungkan kembali menjadi satu dataset utuh. Dataset baru memiliki jumlah 8.000. Dataset hasil augmentasi dapat dilihat pada Gambar 8.

P-ISSN: 1979-5661
E-ISSN: 2622-321X

	isi	label
463	kabar gembira gensoepreneur buka peluang mahas...	0
1429	mahasissayaa informatika univ yarsi milik semp...	0
633	lapor akmn npelaksanan sharing session diikutt...	0
256	fakultas ekonomi unair menyeletidakrakan rapat...	0
1165	hadir kampus merdeka kedaireka kabar gembira f...	0

	isi	label
2246	kampus merdeka sayang otak panitia ospek untir...	1
2176	ikutt mbkm cpt kkn	1
2989	imissu anak mj ikutann mbkm ga kamu bakal nemp...	1
2512	tuju kampus merdeka nder keluar dr jurus asal	1
2199	inti sosialisasi hari mbkm anak emas kkn regul...	1

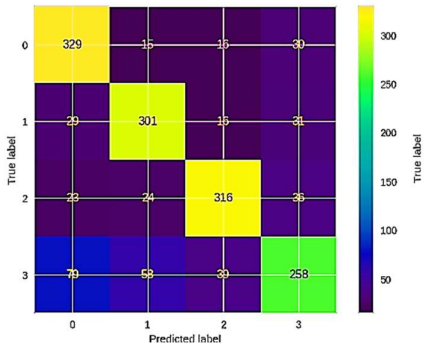
	isi	label
5891	mahasissayaa fkip ikutt mbkm tah	2
4758	kemarin acc program magang kampus merdeka digi...	2
5219	ugm gaes daftar msib kampus merdeka tu kalay g...	2
4207	askrl nunggu berapa lama aku mau tes mbkm tlap...	2
4289	cara magang kampus merdeka gimanaa nder	2

	isi	label
7712	sayaarta beasissayaa beasissayaa merdeka ajar ...	3
6782	gantung ga kerja ga ngelarang mbkm ngelarang g...	3
6278	aku harap ada akm sini temen temen munetidakin...	3
7032	guys aku terima program kampus merdeka sekiann...	3
7328	masuk buktiin bahsayaa masuk universitas neger...	3

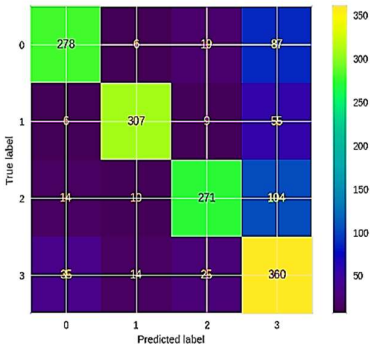
Gambar 8. Dataset Baru Hasil Augmentasi

3.3. Hasil

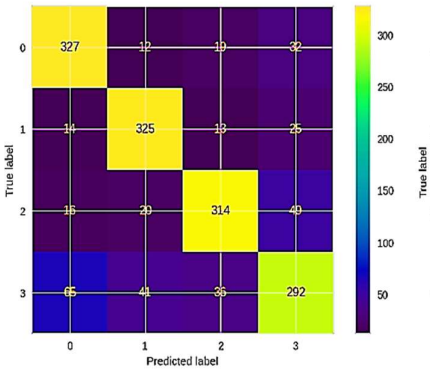
Pengaruh augmentasi data menggunakan metode *Synonym Replacement* dapat dilihat pada hasil akurasi *classifier* menggunakan empat kernel yang berbeda. Klasifikasi dilakukan dengan terlebih dahulu melakukan *pre-processing* kemudian pembobotan dan dilanjutkan dengan pembagian dataset. Klasifikasi dilakukan menggunakan metode *Support Vector Machines*(SVM), di mana klasifikasi dilakukan empat kali dengan kernel yang diubah-ubah. Kernel yang digunakan pada metode *Support Vector Machines*(SVM), terdiri dari karnel *linear, poly, rbf* dan *sigmoid*. Hasil klasifikasi ditunjukan dengan tampilan *confusion matrix* pada Gambar 9a untuk hasil kernel *linear*, 9b untuk hasil karnel *poly*, 9c untuk kernel *rbf,* dan 9d untuk hasil kernel *sigmoid*.



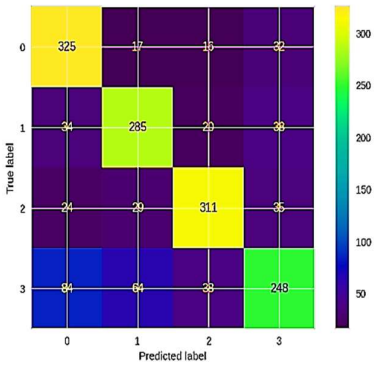
9a. Hasil Kernel *Linear*



9b. Hasil Kernel *Poly*



9c. Hasil Kernel *RBF*



9d. Hasil Kernel *Sigmoid*

Gambar 9. Grafik *Confusion Matrix* Menggunakan SVM

P-ISSN: 1979-5661
E-ISSN: 2622-321X

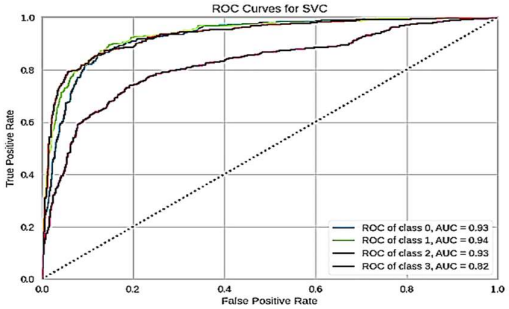
Dengan *confusion matrix* dapat dihitung nilai akurasi, *recall*, *precision* dan *F1-score.* Tabel hasil evaluasi dapat dilihat pada Tabel 3.

Tabel 3. Hasil Akurasi dan *F1-score* Empat Kernel *Support Vector Machines* (SVM)

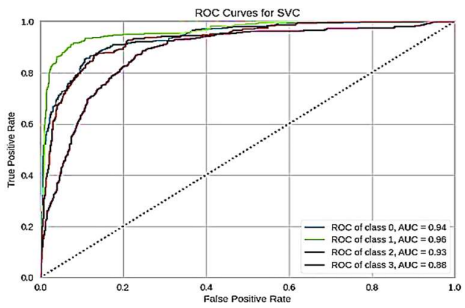
Karnel	Akurasi Training	Akurasi Testing	*Precission*	*Recall*	*F1-score*
linear	0,8923	0,7525	0,7537	0,7571	0,7522
poly	0,9951	0,7600	0,7940	0,7894	0,7677
rbf	0,9732	0,7862	0,7868	0,7901	0,7875
sigmoid	0,8228	0,7306	0,7319	0,7350	0,7301

Pada Tabel 3 tersebut ditunjukan nilai tertinggi pada *precision* sebesar 0,7940 dengan kernel *poly*, nilai *recall* sebesar 0,7901 dan nilai *F1-score* 0,7875 dengan kernel *rbf*. Begitu juga akurasi *training* tertinggi didapatkan dengan menggunakan kernel *poly*, namun akurasi *testing* tertinggi pada adalah sebesar 0,7862 dengan kernel *rbf*. Dapat disimpulkan bahwa dua kernel terbaik adalah *poly* dan *rbf*. Namun begitu, kernel *poly* memiliki kecenderungan *overfitting* lebih besar, karena selisih nilai akurasi *training* dan akurasi *testing* yang lebih besar. Dengan demikian pada penelitian ini dapat disimpulkan bahwa kernel terbaik adalah kernel *rbf*. Secara umum, penggunaan metode *synonym replacement* dapat meningkatkan nilai *F-Measure* dari 0,6672 pada klasifikasi data asli menjadi 0,7875 setelah dilakukan augmentasi.

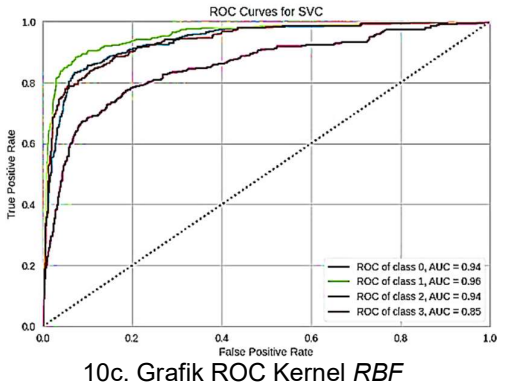
Evaluasi juga dilakukan dengan melihat grafik ROC yang ditunjukan pada Gambar 10. Gambar 10a adalah grafik ROC untuk kernel *linear*, Gambar 10b adalah grafik ROC untuk kernel *poly*, Gambar 10c adalah grafik ROC untuk kernel *rbf*, Gambar 10d adalah grafik ROC untuk kernel *sigmoid*.



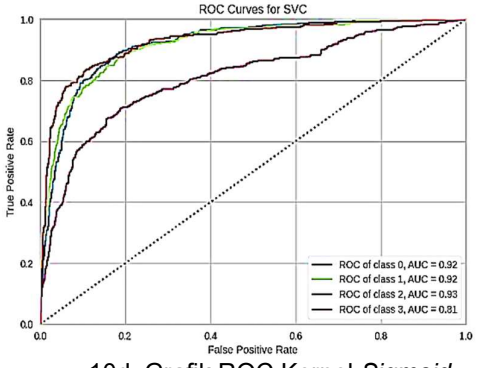
10a. Grafik ROC Kernel *Linear*



10b. Grafik ROC Kernel *Poly*



10c. Grafik ROC Kernel *RBF*



10d. Grafik ROC Kernel *Sigmoid*

Gambar 10. Grafik ROC Menggunakan SVM

Pada Gambar 10 secara umum dapat dilihat bahwa nilai ROC dengan kernel *poly* dan *rbf* menunjukan nilai yang paling baik dari kernel lain. Untuk kelas 0,1 dan 2, nilai ROC cukup baik yaitu berkisar dari nilai 0,92 sampai 0,96. Namun untuk kelas 3 yaitu kelas mayoritas yang tidak dilakukan augmentasi, nilai ROC hanya berkisar 0,81 sampai dengan 0,88. Hal ini dapat terjadi karena walaupun jumlah data sudah seimbang, namun variasi data pada kelas 3 yang merupakan

kelas *out of topic* lebih besar. Sebagai contoh komentar pada kelas 3 dapat terdiri dari iklan, spam, dan topik lainnya. Karena hasil maksimal tidak menunjukan nilai yang memuaska, pada penelitian selanjutnya dapat dilakukan augmentasi pada level kalimat dengan menggunakan metode lain seperti *back translation, random swap, random delete, random insert,* dan *paraphrase.*

4. Kesimpulan

Pada penelitian ini teknik augmentasi data digunakan untuk menyeimbangkan dataset komentar masyarakat terhadap program MBKM. Hal ini, karena pada dataset terdapat ketidakseimbangan dataset yang besar. Ketidakseimbangan ini ditunjukan dengan nilai *Imbalanced Rasio*(IR) sebesar 5:1. Teknik augmentasi data yang dilakukan adalah metode *synonym replacement* di mana data diambil dari data asli kemudian diganti beberapa kata dengan kata sinonimnya. Hasil penggantian sinonim kemudian ditambahkan ke dataset asli hingga memenuhi jumlah dataset yang diinginkan.

Berdasarkan hasil analisis yang dilakukan, dapat disimpulkan bahwa pengaruh augmentasi dengan *synonym replacement* kurang menunjukan hasil optimal karena terjadinya *overfitting.* Hal ini dapat dilihat dari nilai akurasi yang turun saat dilakukan *testing*. Namun saat melihat kembali grafik ROC maka diketahui kelas 3 atau kelas *out of topic* menjadi kelas yang menyebabkan *overfitting.* Kelas 3 sebelumnya merupakan kelas mayoritas dan tidak menerima augmentasi data. Nilai akurasi tertinggi pada saat pengujian model SVM adalah 0,7875 dengan menggunakan kernel *rbf*.

References

[1] Prianto C, Isti Rahayu W, and Izza Hamka N, “Sentimen Analisis Terhadap Pembelajaran Jarak Jauh,” *Jurnal Ilmu Komputer* , vol. 14, 2021.

[2] S. Y. Feng *et al.*, “A Survey of Data Augmentation Approaches for NLP,” in *Finding of the Association for Computational Linguistics*, 2021. [Online]. Available: https://github.com/styfeng/DataAug4NLP.

[3] S. Ren, Y. Deng, K. He, and W. Che, “Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency,” in *57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019, pp. 1085–1097. [Online]. Available: https://wordnet.princeton.edu/

[4] K. J. Madukwe, X. Gao, and B. Xue, “Token replacement-based data augmentation methods for hate speech detection,” *World Wide Web*, vol. 25, no. 3, pp. 1129–1150, May 2022, doi: 10.1007/s11280-022-01025-2.

[5] N. Fadilah and S. Priyanta, “Automatic Essay Scoring Using Data Augmentation in Bahasa Indonesia,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 16, no. 4, p. 401, Oct. 2022, doi: 10.22146/ijccs.76396.

[6] J. Wei and K. Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” in *9th International Joint Conference on Natural Language Processing*, 2019. [Online]. Available: http://github.

[7] C. Magnolia, A. Nurhopipah, D. Bagus, and A. Kusuma, “Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter,” 2022. [Online]. Available: http://journal.unnes.ac.id/sju/index.php/edukom

[8] A. Nurhopipah and C. Magnolia, “Perbandingan Metode Resampling Pada Imbalanced Dataset Untuk Klasifikasi Komentar Program MBKM,” *JUPIKOM*, vol. 1, no. 2, 2022.

[9] T. Fazar Tri Hidayat and dan Azhari Ali Ridha, “Analisis Sentimen Pemindahan Ibu Kota Pada Twitter Dengan Metode Support Vector Machine,” *Jurnal Ilmu Komputer*, vol. 14, 2021.

[10] S. Garg and G. Ramakrishnan, “BAE: BERT-based Adversarial Examples for Text Classification,” in *The 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[11] C. Shorten, T. M. Khoshgoftaar, and B. Furht, “Text Data Augmentation for Deep Learning,” *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00492-0.