ANALISIS PENGGUNAAN METODE *UNDERSAMPLING CLUSTER CENTROIDS* UNTUK MODEL KLASIFIKASI SVM DAN EKSTRAKSI FITUR BOW PADA ULASAN APLIKASI SIREKAP 2024

Novia Hidayati Ramadhani, Naufal Azmi Verdikha, Taghfirul Azhima Yoga Siswa
Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur
Jl. Ir. H. Juanda No.15, Samarinda
2011102441233@umkt.ac.id

ABSTRAK

Sistem Informasi Rekapitulasi (SIREKAP) digunakan sebagai aplikasi perhitungan suara pemilu. Namun, pada awal bulan februari 2024 aplikasi ini mendapat banyak ulasan negatif di *platform google play store*, dengan rata-rata peringkat 2,7 dari 5. Hal ini menunjukkan adanya ketidakseimbangan pada peringkat ulasan pengguna. Ketidakseimbangan ini dapat mempengaruhi hasil analisis dan evaluasi aplikasi secara keseluruhan. Untuk mengatasi ketidakseimbangan data pada ulasan penelitian in menggunakan algoritma *machine learning*. Penelitian ini bertujuan untuk mengklasifikasikan teks ulasan aplikasi SIREKAP 2024 menggunakan algoritma SVM dengan ekstraksi fitur BoW serta metode *undersampling Cluster Centroids* (CC) untuk menangani ketidakseimbangan data. Dataset yang digunakan berjumlah 8235 ulasan. Proses validasi dilakukan menggunakan teknik *k-fold cross-validation* dengan k=10 untuk memastikan performa secara konsisten pada berbagai subset data. Hasil penelitian menunjukkan bahwa undersampling CC efektif dalam menyeimbangkan distribusi kelas, dengan *Imbalanced Ratio* (IR) awal sebesar 21,02 berhasil menjadi 1. Selain itu nilai entropi meningkat dari 1,079 menjadi 1,609 yang menunjukkan peningkatan keragaman ketidakpastian data. Namun, meskipun distribusi kelas menjadi seimbang, rata rata nilai IBA mengalami penurunan dari 0,274 pada Skenario I menjadi 0,271 pada Skenario II. Penurunan ini mengindikasikan bahwa teknik CC dapat menyebabkan hilangnya informasi penting pada kelas mayoritas selama proses *undersampling*.

Kata kunci : *Ketidakseimbangan Data, Klasifikasi Teks, Undersampling, BoW, SVM*

1. PENDAHULUAN

Sistem Informasi Rekapitulasi (SIREKAP) merupakan aplikasi yang digunakan untuk publikasi hasil penghitungan suara pada pemilihan umum yang dilaksanakan pada bulan Februari 2024. Aplikasi ini dirancang untuk mempermudah kelompok penyelenggara pemungutan suara (KPPS) di Tempat Pemungutan Suara (TPS) dalam mencatat hasil penghitungan suara dan meneruskannya ke pusat pengumpulan data [1].

Ulasan dan opini mengenai SIREKAP banyak bermunculan di internet menggambarkan pengalaman pengguna termasuk kelebihan, kekurangan, dan kendala yang dihadapi. Namun, ulasan pengguna di *Google Play Store* menunjukkan bahwa aplikasi ini memiliki distribusi data ulasan yang tidak merata di antara kelas rating [2].

Terdapat 8232 data ulasan pengguna dengan rata-rata peringkat 2,7 dari 5 bintang. Hal ini menunjukkan banyak pengguna yang memiliki pengalaman negatif dengan aplikasi SIREKAP 2024. Distribusi ulasan yang tidak merata dan rendah pada ulasan aplikasi seperti pada penelitian sebelumnya. Salah satunya adalah penelitian [3] mengenai ulasan pengguna aplikasi IPUSNAS di *Google Play Store*, yang menggunakan 2430 ulasan dari Januari hingga Desember 2021. Dari beberapa algoritma yang diuji, seperti *Naïve Bayes*, *Logistic Regression*, dan SVM, algoritma *Support Vector Machine* (SVM) menunjukkan performa terbaik dengan *precision*, *recall*, dan *f1-score* masing-masing sebesar 87%.

Ketidakseimbangan data (*imbalanced data*) terjadi ketika jumlah data antar kelas tidak seimbang, yang menyebabkan model cenderung memihak kelas mayoritas dan mengabaikan kelas minoritas. Salah satu metode yang digunakan untuk mengatasi masalah ini adalah teknik undersampling. Teknik ini dilakukan dengan menghapus sebagian data dari kelas mayoritas hingga jumlahnya seimbang dengan kelas minoritas [4].

Perkembangan awal Februari 2024 pada *platform* SIREKAP menunjukkan ketidakseimbangan signifikan dalam distribusi ulasan antar peringkat, dengan peringkat satu yang dominan. Situasi ini dapat menimbulkan bias yang memengaruhi performa analisis dan proses pengambilan keputusan berbasis data [5].

Untuk mengatasi masalah ini, penelitian ini mengusulkan kombinasi algoritma SVM dengan ekstraksi fitur BoW serta metode undersampling CC untuk menangani ketidakseimbangan data. Evaluasi akhir menggunakan IBA bertujuan untuk memastikan performa model yang optimal pada data yang tidak seimbang [6].

Penelitian ini diharapkan dapat memperkenalkan kombinasi metode tersebut sebagai pendekatan yang inovatif dan memberikan kontribusi ilmiah baru dalam analisis data teks berbasis *machine learning*. Oleh karena itu, penelitian ini bertujuan untuk menganalisis

hasil evaluasi IBA dalam mengukur performa metode undersampling CC terhadap klasifikasi teks ulasan aplikasi SIREKAP 2024 menggunakan SVM dan teknik ekstraksi fitur BoW.

2. TINJAUAN PUSTAKA

Penelitian mengenai klasifikasi teks pada sistem *Auto Essay Scoring* (AES) untuk menilai jawaban esai secara otomatis telah dilakukan oleh [7] dan [8].

Penelitian ini menggunakan teknik ekstraksi fitur *Bag of Words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF), serta algoritma *Support Vector Machine* (SVM), *Latent Semantic Analysis* (LSA), dan *K-Nearest Neighbors* (KNN). Hasil pengujian menunjukkan bahwa kombinasi SVM dengan BoW menghasilkan evaluasi *Root Mean Square Error* (RMSE) sebesar 1,993, yang lebih rendah dibandingkan penggunaan SVM dengan TF-IDF. Nilai RMSE yang lebih kecil menunjukkan tingkat akurasi yang lebih baik, sehingga kombinasi BoW dan SVM terbukti menjadi metode yang efektif untuk klasifikasi teks.

Pada penelitian [9] juga menggunakan algoritma SVM dalam penelitian klasifikasi data stunting di Kota Samarinda, dengan hasil akurasi mencapai 96,6% dan nilai *precision*, *recall*, dan *f1-score* masing-masing sebesar 97%. Penelitian ini menunjukkan bahwa SVM efektif untuk menangani data berdimensi tinggi dan mampu memberikan hasil yang optimal dalam klasifikasi data.

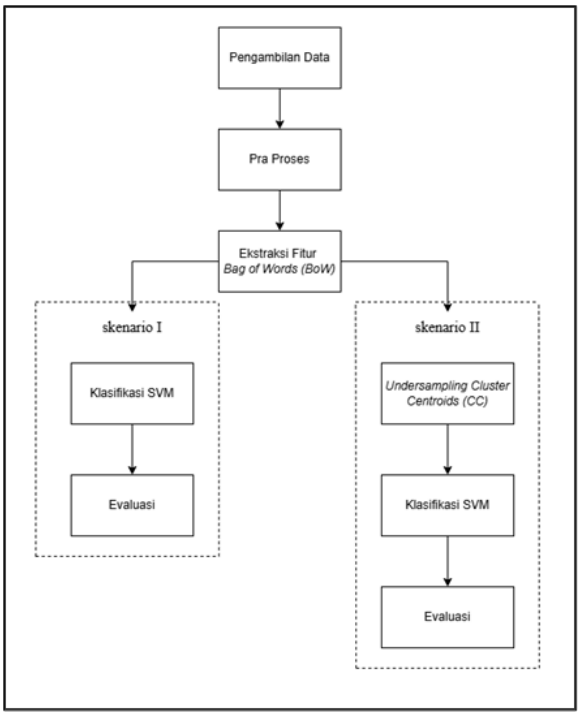
Penelitian terkait teknik *undersampling* dilakukan oleh [10] dalam mengklasifikasikan data cuaca Bangladesh dengan sampel 20,544 dan menggunakan teknik undersampling serta cross-validation untuk mencegah *overfitting*. Beberapa algoritma yang dibandingkan termasuk KNN, SVM, RFC, *XGBoost*, serta metode undersampling seperti *Cluster Centroids* (CC), RUS, dan *NearMiss* (NM). Hasilnya, penggunaan CC dengan SVM menghasilkan akurasi 91,17%, *precision* 91,30%, dan *recall* 91,94%, yang menunjukkan kinerja yang baik dan konsisten pada berbagai metrik evaluasi.

3. METODE PENELITIAN

3.1. Prosedur Penelitian

Alur penelitian yang akan dibahas dalam penelitian ini terdapat pada Gambar 1 tahap awal hingga akhir adalah sebagai berikut.

Tahapan penelitian pada Gambar 1 diawali dengan pengambilan data, diikuti oleh pra-pemrosesan. Data yang telah diproses diekstraksi menggunakan fitur BoW, lalu dilanjutkan ke dua skenario: Skenario I, klasifikasi langsung menggunakan algoritma SVM dengan evaluasi IBA dan Skenario II, dilakukan *undersampling* menggunakan *cluster centroids* (CC) sebelum klasifikasi dengan SVM, kemudian diakhiri evaluasi IBA.



Gambar 1. Alur Penelitian

3.2. Pengambilan Data

Data yang digunakan dalam penelitian ini adalah ulasan aplikasi SIREKAP 2024 pada *google play store*. Pengembalian data diambil pada 6 Februari 2024. Data diambil dengan teknik scraping data ulasan berjumlah 8358 dengan 5 kategori Nama, Rating, Waktu, Komentar, dan *Like*.

3.3. Pra Proses

Berikut adalah tahapan yang dilakukan dalam pra pemrosesan data teks :

a. *Insert Columns ID*, menambahkan kolom baru "*ID*" pada tabel, dengan nomor urut pada setiap baris data.
b. *Lower Case*, teknik ini mengubah huruf pada teks menjadi huruf kecil, dilakukan untuk menghilangkan makna yang muncul dari penggunaan huruf besar ke kecil.
c. *Remove Character*, menghapus karakter-karakter tertentu pada teks seperti tanda baca, simbol, dan angka.
d. *Spell Checker*, mengidentifikasi dan memperbaiki kata-kata yang salah eja dalam teks menjadi ejaan yang benar, membantu mengoreksi kesalahan ketik dalam teks.
e. *Stemming*, proses penghapusan akhiran kata untuk menghasilkan kata dasar.
f. Visualisasi Kata, membuat representasi grafis dari kata-kata, seperti *wordcloud* frekuensi kata untuk memahami pola dalam data teks.
g. Hapus Data Kosong, menghapus baris atau kolom yang mengandung nilai kosong dari dataset untuk memastikan data lebih bersih dan data kosong yang berisi spasi.

3.4. Ekstraksi Fitur BoW

Bag of Words (BoW) merupakan teknik pembelajaran *machine learning* dengan model mengekstraksi fitur teks pemrosesan bahasa alami, mengubah teks menjadi representasi *numerik* yang dapat dipahami oleh mesin. Tujuan utama BoW adalah untuk mengidentifikasi dan menghitung nilai frekuensi kemunculan kata kata kunci [11].

Proses ini dilakukan dengan menggunakan fungsi *CountVectorizer* dari *library sklearn*.

3.5. Cross Validation

Teknik *k-fold cross validation* digunakan untuk menilai keakuratan model secara objektif dengan membagi data pra-pemrosesan menjadi 10 subset (*fold*) yang sama besar. Proses dilakukan 10 kali, di mana satu *fold* digunakan sebagai data uji dan 9 *fold* sebagai data latih serta diuji pada *fold* lainnya secara bergantian. Metode ini membantu mendapatkan estimasi performa yang lebih tepat dan mencegah *overfitting* [12].

3.6. Skenario I

3.6.1. Klasifikasi SVM

Model klasifikasi yang digunakan adalah algoritma klasifikasi *Support Vector Machine* (SVM). Pada konsep SVM mencari fungsi pemisah atau *hyperplane* untuk menemukan bidang pemisah terbaik diantara fungsi yang tidak terbatas jumlahnya. Dengan mengukur dari margin *hyperplane* maka dapat dicari titik maksimalnya, menemukan hyperplane perbedaan terbaik antara dua kelas [13].

Proses ini menggunakan *library sklearn.svm* dengan fungsi *LinearSVC*.

3.6.2. Evaluasi

Menggunakan evaluasi untuk menganalisis performa model pada data tidak seimbang disajikan sebagai berikut.

a. Nilai IR

Imbalanced Ratio (IR) adalah rasio antara jumlah data kelas mayoritas dan minoritas dalam suatu dataset. IR yang tinggi menunjukkan ketidakseimbangan data, yang dapat menyebabkan bias model terhadap kelas mayoritas dan menurunkan performa prediksi pada kelas minoritas. Penggunaan IR diperlukan untuk menjaga performa model pada dataset tidak seimbang [14]; [15].

$$Imbalanced\ Ratio\ (IR) = \frac{Kelas\ Mayoritas}{Kelas\ Minoritas} \quad \ldots (1)$$

b. IBA

Index Balanced Accuracy (IBA) adalah metrik evaluasi yang digunakan untuk mengukur performa keseluruhan classifier pada kelas mayoritas dan minoritas. IBA berguna untuk mengidentifikasi kelas dominan serta hubungan dominansinya, terutama pada data tidak seimbang, di mana model cenderung bias terhadap kelas mayoritas dengan tingkat kesalahan tinggi pada kelas minoritas [16].

Tabel 1. Confusion Matrix

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

Confusion matrix tabel 1 untuk klasifikasi binary yaitu (a)*true Positive* (TP) yaitu, sampel yang sebenarnya positif diklasifikasikan dengan benar sebagai positif. (b)*true Negative* (TN) yaitu, sampel yang sebenarnya negatif diklasifikasikan dengan benar sebagai negatif (c)*false Negative* (FN) yaitu, sampel yang sebenarnya positif tetapi salah diklasifikasikan sebagai negatif. (d)*false Positive* (FP) yaitu, sampel yang sebenarnya negatif tetapi salah diklasifikasikan sebagai positif [17].

Metrik evaluasi meliputi :

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN} \quad \ldots (2)$$

Mengukur proporsi prediksi yang benar.

$$\text{Sensitivitas} = \frac{TP}{TP + FN} \quad \ldots (3)$$

Mengukur kemampuan model dalam mengidentifikasi kelas positif yang benar.

$$\text{Spesifisitas} = \frac{TN}{TN + FP} \quad \ldots (4)$$

Mengukur kemampuan model dalam mengidentifikasi kelas negatif yang benar.

$$\text{Gmean} = \sqrt{Sensitivitas\ x\ Spesifisitas} \quad \ldots (5)$$

Menggambarkan keseimbangan antara *sensitivitas* dan *spesifisitas*.

$$\text{Dominance} = (Sensitivitas - Spesifisitas) \ldots (6)$$

Mengukur perbedaan antara *sensitivitas* dan *spesifitas*.

$$\text{IBA } (\alpha) = (1 + \alpha\ x(Dominance))\ x\ Gmean^2 \ldots (7)$$

IBA mengukur keseimbangan antara kinerja keseluruhan model (*G-mean*) dan keseimbangan kelas (*Dominance*). Nilai a mengatur pengaruh *sensitivitas* dan *spesifisitas*, di mana jika a = 0, IBA hanya bergantung pada pada *G-Mean* [18]. IBA memastikan performa baik pada kedua kelas, terutama pada kelas minoritas, sehingga cocok untuk data tidak seimbang.

c. Entropi

Entropi mengukur tingkat ketidakpastian dalam data. Rentang nilai entropi adalah 0 hingga 1, di mana 0 berarti data murni tanpa ketidakpastian, dan 1 menunjukkan data sepenuhnya acak dengan ketidakpastian tertinggi [19]. Dalam *python* entropi dihitung menggunakan *library SciPy*.

3.7. Skenario II

Undersampling cluster centroids (CC) yaitu pada kelas mayoritas dengan mengganti sebuah *cluster* sampel mayoritas dengan *cluster centroid* dari

algoritma *K-means*. Metode ini memisahkan kelas mayoritas menjadi *cluster* K dan mengganti kelas mayoritas dengan *centroid* dari *cluster* ini. Untuk mengolah data *machine* K maka algoritma dalam data dimulai dengan kelompok pertama dari *centroids* yang dipilih secara acak, lalu digunakan sebagai titik awal untuk setiap *cluster*, dan kemudian melakukan iteratif perhitungan berulang untuk mengoptimalkan posisi *centroids* [20]. Penelitian ini menggunakan *library imblearn.under_sampling*. Kemudian pada Skenario II juga menggunakan model SVM dan IBA sebagai evaluasi akhir. Setelah proses *undersampling* CC, langkah berikutnya adalah prioritas dan eliminasi data teks. Prioritas memilih kata-kata yang relevan dan sering muncul, sementara eliminasi menghapus kata-kata dengan frekuensi rendah atau tidak sesuai konteks.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Input Data

Data setelah *scraping* penelitian ini berbentuk file CSV yang diimpor ke dalam *python*. Data ini berisi 8232 ulasan SIREKAP 2024 yang dikelompokkan berdasarkan beberapa atribut yaitu, Nama, Rating, Waktu, Komentar, dan *Like* seperti pada Gambar 2.

	Nama	Rating	Waktu	Komentar	Like
0	Nency Miranda	5	05/02/2024 14:07	Mantap	0.0
1	ichal sabiel	1	05/02/2024 14:06	Sudah tau pemakainya masyarakat biasa Malah sp...	0.0
2	Evandra Aditya	1	05/02/2024 14:06	Apk nya nggk bisa buat log in	0.0
3	Ganesh insann	1	05/02/2024 14:05	Susah masuk dih	0.0
4	Yohana Frediana	1	05/02/2024 14:04	Tidak bisa masuk inisiliasi	0.0
5	Hendra Hirasawa	1	05/02/2024 14:04	Internal server error!	0.0
6	windi sari	2	05/02/2024 14:04	Baik	0.0
7	Desy Yasmine	1	05/02/2024 14:03	Susah bgtz inialisasi .. sumpah.. dari pagi mu...	0.0
8	nuril rani	5	05/02/2024 14:03	Susah buat login	0.0
9	setyo aji	1	05/02/2024 14:03	kamera ter-reduce sangat parah, untuk scan tul...	0.0

Gambar 2. Hasil Tampilan 10 Data Teratas Dataset

4.2. Hasil Pra Proses

Selanjutnya, kolom ID ditambahkan untuk setiap baris agar lebih mudah diidentifikasi dan dianalisis. Pada ID dimulai dari d00001 hingga baris terakhir yaitu d08232. Hasilnya adalah sebagai berikut pada Gambar 3.

	ID	Rating	komentar_clean
0	d00001	5	mantap
1	d00002	1	sudah tahu maka masyarakat biasa malah spesifi...
2	d00003	1	aplikasi nya tidak bisa buat masuk ini
3	d00004	1	susah masuk dih
4	d00005	1	tidak bisa masuk inisial

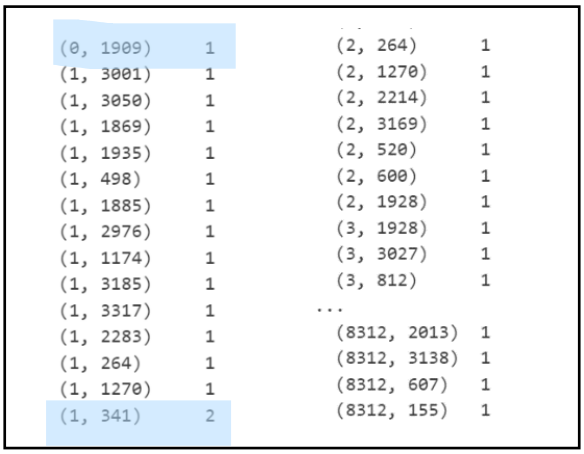
Gambar 3. Hasil Colum ID Dataset

Kata-kata yang sering muncul divisualisasikan dalam bentuk *wordcloud*, di mana ukuran kata mencerminkan frekuensi kemunculannya. Visualisasi ini membantu memahami focus utama komentar pengguna dan mengidentifikasi masalah utama, seperti Gambar 4 “tidak bisa”, “kata sandi”, dan “bisa masuk”.

Gambar 4. Hasil Visualisasi Kata

4.3. Hasil Ekstraksi Fitur BoW

Hasil representasi *Bag of Words* (BoW) dapat dilihat pada Gambar 6 pasangan dokumen dan indeks kata, diikuti oleh nilai frekuensi kata dalam dokumen. Misalnya, (0, 1909) 1 menunjukkan bahwa pada dokumen pertama (*indeks* 0), kata dengan indeks 1909 muncul 1 kali. Contoh lain, (1, 341) 2, berarti pada dokumen kedua (*indeks* 1), kata dengan indeks 341 muncul sebanyak 2 kali.

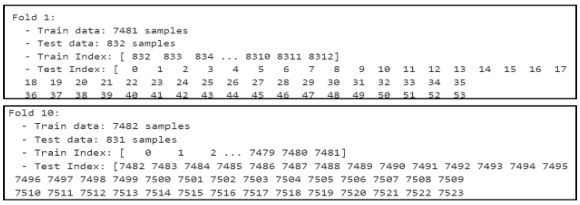


Gambar 5. Hasil BoW

4.4. Hasil Skenario I

4.4.1. Cross Validation

Pada *fold* 1 data latih terdiri dari 7481 sampel, sedangkan data uji terdiri dari 832 sampel (*indeks* 832 hingga 8312). Proses ini berlangsung dalam waktu kurang dari 0,00 detik. Sedangkan pada *fold* 10 terdapat Data latih terdiri dari 7482 sampel, dengan 83 sampel sebagai data uji (*indeks* 7482 hingga 8312). Proses ini juga memakan waktu kurang dari 0,00 detik terdapat pada Gambar 7.



Gambar 6. Hasil fold 1 dan 10 Skenario I

4.4.2. Klasifikasi dan Evaluasi

Untuk mencari nilai IRpada Skenario I (sebelum dilakukan *undersampling*) sebagai berikut.

$$\text{IR} = \frac{kelas\ mayoritas}{kelas\ minoritas} = \frac{5341}{254} = 21{,}02$$

Nilai *imbalanced Ratio* (IR) pada Skenario I sebesar 21,02 menunjukkan bahwa dataset masih termasuk dalam kategori *imbalanced data*, karena nilai IR jauh lebih besar dari 1. Menunjukkan dominasi jumlah sampel pada kelas mayoritas rating 1 sebanyak 5341 sampel dibandingkan kelas minoritas rating 4 yang hanya 254 sampel.

Berikut adalah hasil yang diperoleh Skenario I klasifikasi pada Tabel 2.

Tabel 2. Evaluasi Setiap Fold SVM

Fold	SVM				
	Accuracy	Sensitivity	Specificity	G-mean	IBA
1	0,704	0,338	0,875	0,544	0,280
2	0,698	0,317	0,867	0,525	0,260
3	0,681	0,308	0,866	0,516	0,252
4	0,708	0,352	0,878	0,556	0,293
5	0,724	0,353	0,885	0,559	0,296
6	0,687	0,339	0,872	0,544	0,280
7	0,684	0,335	0,867	0,539	0,275
8	0,704	0,340	0,875	0,546	0,282
9	0,701	0,350	0,872	0,553	0,289
10	0,708	0,365	0,880	0,567	0,305
Rata-Rata	0,696	0,333	0,871	0,539	0,274

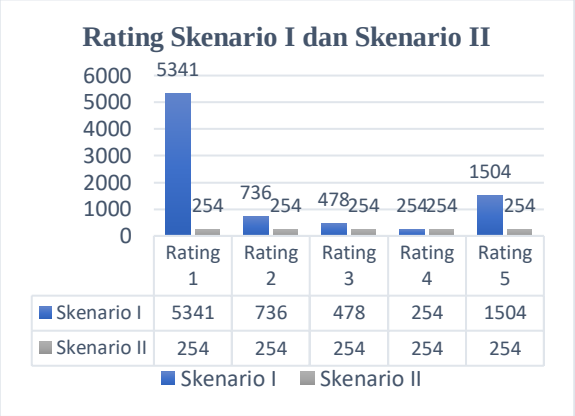
Dari Tabel 1 Hasil Evaluasi Setiap Fold SVM, terlihat bahwa nilai *accuracy* tertinggi terdapat pada *Fold* 5 dengan nilai 0,724. Selain itu, IBA tertinggi juga terdapat pada *Fold* 10 dengan nilai 0,305. Untuk *sensitivity*, nilai tertinggi terdapat pada *fold* 10 dengan nilai 0,365 dan nilai *specificity* tertinggi terdapat pada *fold* 5 dengan 0,885. Semantara itu untuk *G-Mean* tertinggi pada *fold* 10 dengan nilai 0,567.

Secara keseluruhan, rata-rata *accuracy* dari seluruh *fold* adalah 0,696, sedangkan *sensitivity* rata-rata adalah 0,333 menunjukkan kemampuan yang rendah dalam mengidentifikasi kelas positif. *specificity* rata-rata adalah 0,871 menunjukkan kemampuan yang baik dalam mengidentifikasi kelas negatif. Dengan rata-rata *G-mean* sebesar 0,539 menunjukkan keseimbangan antara *sensitivitas* dan *spesifisitas*. Selain itu,rata-rata IBA sebesar 0,274 menunjukkan performa keseluruhan model dalam menangani ketidakseimbangan. Hasil ini menunjukkan bahwa performa model SVM cukup baik secara umum, meskipun terdapat variasi performa antar *fold* dan masih belum optimal dalam mengklasifikasikan karena belum seimbang.

4.5. Hasil Skenario II

4.5.1. Undersampling Cluster Centroids

Pada bagian ini disajikan perbandingan distribusi kelas untuk Skenario I dan Skenario II. Grafik berikut memperhatikan perubahan jumlah sampel pada setiap kategori rating sebelum dan setelah dilakukan *undersampling* menggunakan CC terdapat pada Gambar 7.



Gambar 7. Grafik Rating Skenario I dan II

Menunjukkan perbandingan distribusi kelas antara Skenario I dan Skenario II. Pada Skenario I, kelas Rating 1 dan Rating 5 memiliki jumlah sampel yang jauh lebih banyak dibandingkan kelas lainnya, yaitu 5341 dan 1504 sampel. Sedangkan pada Skenario II, teknik *undersampling* diterapkan untuk menyeimbangkan semua kelas, sehingga setiap rating memiliki 254 sampel. Selisih terbesar terlihat pada Rating 1 dan Rating 5, yang menurun drastis untuk mencapai keseimbangan antar kelas pada Skenario II.

Dari grafik ini dapat dilihat bahwa Skenario II menggunakan *undersampling* untuk menyeimbangkan jumlah sampel di setiap rating menjadi 254. Selisih yang signifikan pada Rating 1 dan Rating 5 menunjukkan ketidakseimbangan yang ada pada Skenario I, yang diatasi dalam Skenario II. Untuk mencari nilai (IR) pada Skenario II (setelah dilakukan *undersampling*) sebagai berikut.

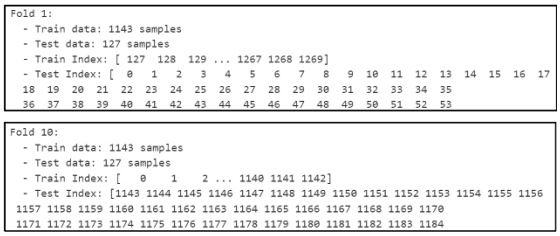
$$\text{IR} = \frac{kelas\ mayoritas}{kelas\ minoritas} = \frac{254}{254} = 1$$

Dalam Skenario II teknik *undersampling* menggunakan *cluster centroids* telah berhasil menyamakan jumlah sampel antara mayoritas dan minoritas, sehingga nilai *imbalanced ratio* (IR) menjadi 1. Ini yang berarti jumlah sampel untuk setiap kelas, tanpa memandang rating kini telah sama. Ketika nilai IR mencapai 1, ini menandakan bawa distribusi sampel antara kedua kelas telah seimbang. Hal ini sangat penting dalam pembuatan model prediktif karena dengan sampel yang seimbang, model dapat dengan lebih efektif dan tidak akan bias hanya ke satu kelas saja. Tahap berikutnya kata-kata yang merupakan prioritas dan eliminasi pada Tabel 3 berikut.

Tabel 3. Prioritas dan Eliminasi

Rating 1		Rating 2		Rating 3		Rating 4		Rating 5	
Prioritas	**Eliminasi**	**Prioritas**	**Eliminasi**	**Prioritas**	**Eliminasi**	**Prioritas**	**Eliminasi**	**Prioritas**	**Eliminasi**
aplikasi	amat	aplikasi	iphone	tidak	aktif	bisa	-	aplikasi	ayo
tidak	bentar	tidak	oke	masuk	yah	masuk	-	bisa	top
yang	apk	bisa	log	aplikasi	1999	tidak	-	tidak	betul
ini	log	masuk	lumayan	bisa	abu	aplikasi	-	masuk	allah
bisa	rusak	di	bang	ini	bena	di	-	di	damai
masuk	benah	sudah	gmn	sudah	bismillah	untuk	-	ini	daring
di	bobrok	ini	konfirmasi	di	bodoh	ini	-	yang	digitalisasi
sudah	operasi	yang	liat	guna	detail	dan	-	dan	informasi
dan	valid	dan	mumet	yang	deteksi	yang	-	saya	jiwa
guna	ubah	kata	palsu	sandi	download	sudah	-	kata	kualitas

4.5.2. Cross Validation



Gambar 8. Hasil Fold 1 dan 10

Pada *fold* 1 memiliki 1143 sampel untuk data latih dan 127 sampel untuk data uji dengan indeks 0 hingga 126, sementara *fold* 10 memiliki 1143 sampel untuk data latih dan 127 sampel untuk data uji dengan indeks 1143 hingga 1269. Proses ini berlangsung sangat cepat, kurang dari 0,00 detik terdapat pada Gambar 8.

4.5.3. Klasifikasi dan Evaluasi

Tahap berikutnya adalah Skenario II menggunakan metode *undersampling* dengan *cluster centroids* terdapat pada Tabel 4.

Tabel 4. Hasil Evaluasi setiap fold CC

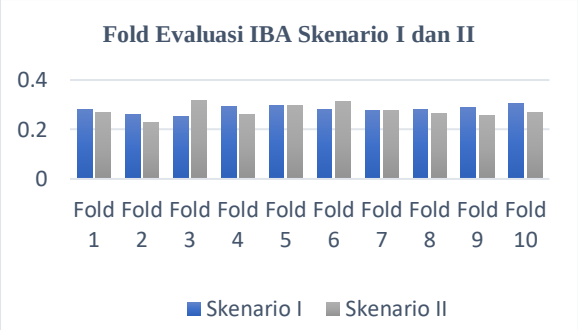
Fold	***Undersampling Cluster Centroids***				
	Accuracy	***Sensitivity***	***Specificity***	***G-Mean***	***IBA***
1	0,338	0,338	0,833	0,531	0,268
2	0,291	0,292	0,822	0,490	0,227
3	0,409	0,392	0,851	0,578	0,318
4	0,322	0,332	0,830	0,525	0,262
5	0,393	0,368	0,848	0,558	0,297
6	0,370	0,388	0,842	0,571	0,312
7	0,370	0,347	0,842	0,541	0,278
8	0,322	0,338	0,829	0,529	0,266
9	0,322	0,326	0,830	0,520	0,257
10	0,346	0,339	0,836	0,533	0,270
Rata-Rata	0,341	0,341	0,835	0,534	0,271

Dari Tabel 3 hasil Evaluasi Setiap Fold CC, terlihat bahwa nilai *accuracy* tertinggi diperoleh pada *fold* 3 dengan nilai 0,409. Selain itu, nilai IBA tertinggi juga terdapat pada *fold* 3 dengan nilai 0,318. Untuk *sensitivity* nilai tertinggi tercatat pada *fold* 3 dengan nilai 0,392 dan nilai *specificity* tertinggi terdapat pada *fold* 3 dengan nilai 0,851. Sementara itu, *G-Mean* tertinggi pada *fold* 3 dengan nilai 0,578.

Secara keseluruhan, rata-rata *accuracy* dari seluruh *fold* adalah 0,341 sebelumnya 0,696 dengan perbandingan 35,5% performa mengalami penurunan pada Skenario II. Sedangkan rata-rata *sensitivity* sebesar 0,341 sebelumnya 0,333 dengan peningkatan 0,8% dalam mengidentifikasi kelas positif pada Skenario I. Rata-rata *specificity* sebesar 0,835 sebelumnya 0,871 dengan perbandingan sebesar 3,6% yang mengindikasikan adanya penurunan, meskipun kemampuan yang baik dalam mengidentifikasi kelas negatif meski masih lebih tinggi mengidentifikasi dalam kelas negatif pada Skenario I. Dengan rata-rata *G-Mean* 0,534 sebelumnya 0,539 perbandingan sebesar 0,5% yang mengalami penurunan, tetapi tetap menjaga keseimbangan antara *sensitivitas* dan *spesifisitas*. Selain itu, rata-rata IBA sebesar 0,271 sebelumnya 0,274 dengan perbandingan 0,3% yang menunjukkan adanya penurunan, namun performa keseluruhan model dalam menangani ketidakseimbangan kelas cukup baik. Hasil ini menunjukkan bahwa teknik *undersampling cluster centroids* tetap menghasilkan performa yang stabil, dengan nilai terbaik tercatat pada *fold* 3 untuk semua metrik. Hal ini menunjukkan kemampuan model untuk mencapai performa yang konsisten.

4.6. Pembahasan dan Hasil Evaluasi

Hasil evaluasi untuk kedua Skenario I dan II berikut hasil analisis dari performa di setiap Skenario seperti pada Gambar 9 sebagai berikut.



Gambar 9. Grafik Fold IBA Skenario I dan I

Nilai IBA pada Skenario I lebih signifikan dibandingkan dengan Skenario II terdapat pada *fold* 1, *fold* 2, *fold* 4, *fold* 8, *fold* 9 dan *fold* 10 di mana Skenario I menunjukkan keunggulan pada setiap *fold* tersebut. Sebaliknya, Skenario II hanya unggul pada *fold* 3, *fold* 5, dan *fold* 6 serta memiliki keunggulan sebesar 1,8% dibandingkan Skenario I pada *fold* 7. Berdasarkan nilai IBA pada setiap *fold*, teknik *undersampling Cluster Centroids* menunjukkan tidak meningkatkan performa secara keseluruhan, namun mengalami penurunan performa.

Terdapat nilai entropi Skenario I (sebelum *undersampling*) dan Skenario II (setelah *undersampling*) sebagai berikut.

```
Entropi Skenario I (sebelum undersampling): 1.079003928885893
Entropi Skenario II (setelah undersampling): 1.6094379124341005
```

Gambar 10. Hasil Nilai Entropi

Nilai entropi pada Skenario I dan II ditampilkan pada Gambar 10. Pada Skenario I nilai entropi sebesar 1,078. Sementara itu, pada Skenario II setelah dilakukan *undersampling* CC nilai entropi mengalami peningkatan menjadi 1,609. Peningkatan ini membuktikan bahwa teknik ini menambah nilai keragaman dan ketidakpastian dalam dataset.

5. KESIMPULAN DAN SARAN

Hasil penelitian ini menunjukkan pada Skenario I, nilai IR mencapai 21,02. Setelah penerapan CC Skenario II, IR berhasil seimbang menajdi 1. Nilai entropi meningkat dari 1,079 menjadi 1,609 yang menunjukkan peningkatan keragaman ketidakpastian data. Berdasarkan rata-rata IBA Skenario I sebesar 0,274 lebih tinggi dibandingkan Skenario II sebesar 0,271 dengan penurunan sebesar 0,3%. Secara keseluruhan, meskipun teknik CC efektif dalam menyeimbangkan distribusi kelas, hasil penelitian ini menunjukkan bahwa metode ini tidak cukup efektif dalam mempertahankan atau meningkatkan performa model SVM. Hal ini disebabkan oleh hilangnya informasi penting pada kelas mayoritas selama proses *undersampling*.

Untuk penelitian selanjutnya, disarankan untuk melakukan tahapan *preprocessing* data yang lebih mendalam agar kata-kata yang tidak relevan dapat dibersihkan ulang. Dengan demikian, data yang digunakan menjadi lebih optimal sebelum memasuki tahapan pemodelan. Selain itu, eksplorasi lebih lanjut terhadap parameter *Cluster Centroids* (CC) dan *Support Vector Machine* (SVM) juga dapat dilakukan dengan berbagai Skenario tambahan. Hal ini bertujuan untuk menemukan kombinasi parameter yang lebih baik sehingga dapat meningkatkan performa model. Selanjutnya, penelitian mendatang juga dapat mempertimbangkan perbandingan dengan metode ekstraksi fitur dan klasifikasi lain untuk mengetahui hasil terbaik dari setiap metode lainnya.

DAFTAR PUSTAKA

[1] I. A. Pradesa, "Analisis Penggunaan Sistem Rekapitulasi Suara (Sirekap) Dalam Menghadapi Problematika Pemilu 2024," *Triwikrama J. Multidisiplin Ilmu Sos.*, vol. 03, no. 04, pp. 47–57, 2024, [Online]. Available: https://ejournal.warunayama.org/index.php/triwikrama/article/view/2578

[2] M. Fajar, Y. Herjanto, U. S. Karawang, and T. Timur, "SIREKAP PADA PLAY STORE MENGGUNAKAN ALGORITMA RANDOM FOREST CLASSIFER," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 12, no. 2, pp. 1204–1210, 2024, doi: http://dx.doi.org/10.23960/jitet.v12i2.4192.

[3] A. Septiani and I. Budi, "Klasifikasi Ulasan Pengguna Aplikasi: Studi Kasus Aplikasi Ipusnas Perpustakaan Nasional Republik Indonesia (PNRI)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 4, pp. 1110–1120, 2022, doi: 10.29100/jipi.v7i4.3216.

[4] M. Koziarski, "Radial-Based Undersampling for imbalanced data classification," *Pattern Recognit.*, vol. 102, 2020, doi: 10.1016/j.patcog.2020.107262.

[5] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, *A survey on imbalanced learning: latest research, applications and future directions*, vol. 57, no. 6. 2024. doi: 10.1007/s10462-024-10759-6.

[6] G. AlMahadin, A. Lotfi, M. M. Carthy, and P. Breedon, "Enhanced Parkinson's Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques," *SN Comput. Sci.*, vol. 3, no. 1, pp. 1–21, 2022, doi: 10.1007/s42979-021-00953-6.

[7] N. A. Verdikha, J. H. Dwiagam, and R. Hasudungan, "Indonesian Automated Essay Scoring with Bag of Word and Support Vector Regression," *JSE J. Sci. Eng.*, vol. 1, no. 2, pp. 95–100, 2024, doi: 10.30650/jse.v1i2.3841.

[8] H. Thamrin, N. A. Verdikha, and A. Triyono, "Text Classification and Similarity Algorithms in Essay Grading," *2021 4th Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2021*, pp. 201–205, 2021, doi: 10.1109/ISRITI54043.2021.9702808.

[9] T. A. Y. Siswa and N. A. Verdikha, "Optimasi Algoritma SVM dengan Chi-Square dan SMOTE untuk High Dimension Data Stunting Kota

Samarinda," *Ilk. J. Ilm.*, vol. xx, no. 200, pp. 1–10, 2021, doi: https://doi.org/10.33096/ilkom.v1xix.xxx.x-x.

[10] M. A. Rahman, A. Akter, F. S. Richi, A. Shoud, and T. Ahmed, "A Comparative Study of Undersampling and Oversampling Methods for Flood Forecasting in Bangladesh using Machine Learning," *2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023*, no. December, 2023, doi: 10.1109/ICCCNT56998.2023.10306368.

[11] Y. S. Kencana, "Klasifikasi Sentiment Analysis pada Review Buku Novel Berbahasa Inggris dengan Menggunakan Metode Support Vector Machine (SVM)," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 6, no. 3, pp. 10451–10462, 2019, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/6171%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/viewFile/6171/6150

[12] Y. C. D. A. N. Khotimah, "PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[13] Z. Rakhmawati, S. Basuki, and G. W. Wicaksono, "Klasifikasi Kalimat Tanya Berdasarkan Taksonomi Bloom Menggunakan Support Vector Machine," *J. Repos.*, vol. 2, no. 4, pp. 427–436, 2020, doi: 10.22219/repositor.v2i4.69.

[14] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *Sistemasi*, vol. 10, no. 2, p. 445, 2021, doi: 10.32520/stmsi.v10i2.1303.

[15] K. Akbar and M. Hayaty, "Data Balancing untuk Mengatasi Imbalance Dataset pada Prediksi Produksi Padi Balancing Data to Overcome Imbalance Dataset on Rice Production Prediction," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 2, no. 02, pp. 1–14, 2020, [Online]. Available: http://jurnal.umus.ac.id/index.php/intech

[16] K. Dwivedi, R. Lakshmanan, and R. Regunathan, "K-means under-sampling for hypertension prediction using NHANES dataset," *Int. J. Performability Eng.*, vol. 17, no. 8, pp. 733–740, 2021, doi: 10.23940/ijpe.21.08.p9.733740.

[17] L. S. Lin, C. H. Kao, Y. J. Li, H. H. Chen, and H. Y. Chen, "Improved support vector machine classification for imbalanced medical datasets by novel hybrid sampling combining modified mega-trend-diffusion and bagging extreme learning machine model," *Math. Biosci. Eng.*, vol. 20, no. 10, pp. 17672–17701, 2023, doi: 10.3934/mbe.2023786.

[18] V. García, R. A. Mollineda, and J. S. Sánchez, "Index of balanced accuracy: A performance measure for skewed class distributions," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5524 LNCS, pp. 441–448, 2009, doi: 10.1007/978-3-642-02172-5_57.

[19] C. . Shannon, "A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379-423," 1948, doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[20] J. POLROB, "A MODIFIED WHALE OPTIMIZATION ALGORITHM FOR IMPROVING DATA BALANCE BASED ON UNDERSAMPLING TECHNIQUES," *Sch. Math. Inst. Sci. Suranaree Univ. Technol.*, 2021.