JASIE "Jurnal Aplikasi Sistem Informasi dan Elektronika"
Vol. 5, No. 2, Desember 2023, Hal: 65-77
p-ISSN: 2714-612X, e-ISSN: 2798-1509

Analisis Sentimen Pasca Pertandingan Sepak Bola Indonesia Melawan Argentina Pada Unggahan Media Sosial *Twitter* Menggunakan Metode *Multinomial Naïve Bayes* Dan *Gaussian Naïve Bayes*

Moh Machrus Alfani[1*], Qurrota A'yun[2]
Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember[1,2,3]
Email: Fannypandut@gmail.com[1*], qurrota.ayun@unmuhjember.ac.id[2]

ABSTRAK

Timnas Indonesia baru saja menjadi tuan rumah pertandingan persahabatan melawan Argentina yang digelar di Gelora Bung Karno Jakarta yang menjadi perbincangan hangat masyarakat terutama di *twitter*. Pada penelitian ini akan dilakukan analisis sentimen pasca pertandingan sepak bola Indonesia melawan Argentina pada *twitter* dengan sejumlah 700 *tweet* data. Metode yang digunakan adalah *Multinomial Naïve Bayes* dan *Gaussian Naïve Bayes.* Namun dalam proses klasifikasi, seringkali terdapat masalah yang ditemukan oleh peneliti yaitu ketidakseimbangan data. Maka dari itu, pada penelitian ini menambahkan teknik *balancing* data, *Random Oversampling*, *Undersampling* dan *SMOTE*. Teknik *balancing* diharapkan dapat meningkatkan hasil pada klasifikasi. Seluruh data akan diproses menggunakan metode *K-Fold Cross Validation* dengan variasi nilai K 2, 5, 7 dan 10. Hasil pengujian metode *Multinomial Naïve Bayes* tanpa menggunakan teknik *balancing* mendapatkan nilai akurasi tertinggi sebesar 69%, presisi 69% dan recall 69%. Metode *Gaussian Naïve Bayes* tanpa teknik *balancing* mendapatkan nilai akurasi tertinggi sebesar 58%, presisi 58% dan *recall* 57%. Sedangkan hasil uji menggunakan metode *Multinomial Naïve Bayes* dengan menambahkan teknik *balancing* dapat diketahui jika menggunakan Random *Oversampling* mendapatkan nilai akurasi tertinggi sebesar 80%, presisi 81% dan *recall* 80%. Untuk metode *Gaussian Naïve Bayes* dengan menambahkan teknik *balancing* dapat diketahui jika menggunakan *SMOTE* mendapatkan nilai akurasi tertinggi sebesar 79%, presisi 79% dan *recall* 79%. Dapat disimpulkan bahwa metode *Multinomial Naïve Bayes* pada penelitian ini lebih efektif dibandingkan dengan *Gaussian Naïve Bayes.*

Kata Kunci: Sepak Bola, Analisis Sentimen, *Multinomial Naïve Bayes*, *Gaussian Naïve Bayes*

ABSTRACT

The Indonesian national football team recently hosted a friendly match against Argentina at Gelora Bung Karno Stadium in Jakarta, sparking widespread discussions, especially on Twitter. This research focuses on sentiment analysis post the Indonesia vs. Argentina football match on Twitter, utilizing a dataset of 700 tweets. The methods employed include Multinomial Baive Bayes and Gaussian Naive Bayes. However, during the classification process, researchers encountered data imbalance issues. Therefore, this study incorporates data balancing techniques such as Random Oversampling, Undersampling, and SMOTE to enhance classification results. The entire dataset will be processed using K-Fold Cross Validation with varying K values of 2, 5, 7, and 10. Testing the Multinomial Naive Bayes method without balancing techniques yielded the highest accuracy of 69%, precision of 69%, and recall of 69%. The Gaussian Naive Bayes method without balancing techniques achieved the highest accuracy of 58%, precision of 58%, and recall of 57%. In contrast, testing the Multinomial Naive Bayes method with the addition of balancing techniques revealed that using Random Oversampling resulted in the highest accuracy of 80%, precision of 81%, and recall of 80%. For the Gaussian Naive Bayes method with balancing techniques, using SMOTE achieved the highest accuracy of 79%, precision of 79%, and recall of 79%. In conclusion, the Multinomial Naive Bayes method in this study proved to be more effective than the Gaussian Naive Bayes method.

Keywords: *Football, Sentimen Analysis, Multinomial Naïve Bayes, Gaussian Naïve Bayes*

1. PENDAHULUAN

Sepak bola adalah jenis permainan berkelompok yang memerlukan kolaborasi tim. Prestasi suatu tim tidak hanya bergantung pada kontribusi individu, melainkan bergantung pada sinergi seluruh anggota dalam satu tim (Tarju & Wahidi, 2017). Sepak bola juga bisa dibilang salah satu jenis olahraga yang sangat populer di seluruh dunia, terutama di Indonesia (Prasetyo & Pradana, 2021).

Timnas Indonesia baru saja menjadi tuan rumah pertandingan persahabatan melawan Argentina yang digelar di Gelora Bung Karno Jakarta. Pertandingan timnas Indonesia melawan timnas Argentina menjadi perbincangan hangat masyarakat terutama di *Twitter*. Hal ini telah memicu berbagai respon

mengenai pertandingan tersebut. Mengamati hal ini, penulis ingin mengetahui sentimen masyarakat terutama pada media sosial *twitter* terhadap adanya pertandingan timnas Indonesia melawan timnas Argentina. Maka dari itu diperlukanlah analisis sentimen.

Analisis sentimen merupakan pengolahan data untuk memperoleh arti atau ungkapan pada sebuah kata atau kalimat (Astari, dkk., 2020). Salah satu cara untuk mengukur dan menganalisis suatu kasus atau objek tertentu adalah melalui analisis sentimen, di mana kesimpulan dan keputusan dapat ditarik berdasarkan teks dalam bentuk kalimat atau dokumen. Salah satu teknik yang dapat digunakan untuk melakukan analisis sentimen adalah menggunakan algoritma *Naïve Bayes*. Dalam algoritma ini, sistem melakukan klasifikasi berdasarkan probabilitas data yang diperoleh.

Proses klasifikasi seringkali menghadapi masalah ketidakseimbangan data, di mana distribusi kelas tidak seimbang karena jumlah data yang lebih banyak atau lebih sedikit untuk beberapa kelas. Ketidakseimbangan data dapat menyebabkan tingkat akurasi yang tidak seimbang antara mayoritas dan minoritas (Kasanah, dkk., 2021). Oleh karena itu, diperlukan metode seperti *Random Oversampling*, *Random Undersampling*, dan *SMOTE* untuk menyeimbangkan distribusi data. Dalam melakukan klasifikasi terdapat beberapa metode diantaranya adalah *Multinomial Naïve Bayes* dan *Gaussian Naïve Bayes.*

Algoritma *Multinomial Naïve Bayes* adalah pengembangan dari metode *Naïve Bayes* yang efektif dalam analisis sentimen. Beberapa penelitian menunjukkan bahwa algoritma ini terkenal karena tingkat akurasi yang tinggi (Verawati & Audit, 2022).

Metode *Multinomial Naïve Bayes* dipilih karena mampu menghitung dengan efisien dalam proses klasifikasi data dan efektif dalam menangani masalah yang melibatkan banyak kategori atau lebih dari dua kelas data (Fariz, dkk., 2021). Perbedaan terletak dalam pemilihan jenis data, jika *Naïve Bayes* menggunakan model *gaussian* cocok untuk data *continue*, sedangkan *Multinomial Naïve Bayes* cocok digunakan untuk data diskrit seperti jumlah kata dalam dokumen (Ernayanti, dkk., 2023).

2. KAJIAN PUSTAKA

A. *Twitter*

Twitter adalah bentuk media sosial yang populer di seluruh dunia. Bukti kepopuleran ini tampak dari beragam konten yang diunggah di *platform* tersebut, terutama melalui *tweet*. *Tweet-twee*t ini mencerminkan opini-opini yang berkisar pada topik-topik seperti politik, sosial, pendidikan, dan bahkan olahraga (Astiningrum, dkk., 2020).

B. *Text Minning*

Text mining adalah proses penggalian data yang mengambil data dalam bentuk teks, yang sering kali tidak terstruktur atau disebut sebagai data tak terstruktur. *Text mining* merupakan komponen dari data *mining* yang fokus pada ekstraksi pengetahuan dan informasi dari pola-pola yang terdapat dalam koleksi dokumen teks memakai alat analisis tertentu (Ghiffarie, dkk., 2019).

C. Metode Klasifikasi

Klasifikasi adalah metode pengelompokan objek berdasarkan atribut yang dimilikinya. Proses ini bisa dilakukan dengan cara manual atau dengan dukungan teknologi. Klasifikasi manual melibatkan campur tangan manusia tanpa memanfaatkan komputer, sedangkan klasifikasi yang dibantu teknologi melibatkan penggunaan berbagai algoritma, seperti *Support Vector Machine, Decision Tree,* dan *Naïve Bayes* (Wibawa, 2018).

D. *Balancing* Data

Melakukan pengklasifikasian pada data yang memiliki ketidakseimbangan kelas adalah permasalahan utama yang perlu diatasi dalam ranah machine learning dan data mining. ketidakseimbangan data (*imbalanced* data) merujuk pada distribusi data yang memiliki perbedaan signifikan dalam jumlah antara kelas-kelas data yang ada (Diantika, 2023). Terdapat beberapa metode yang bisa digunakan untuk mengatasi isu ketidakseimbangan jumlah kelas data yaitu *over sampling*, *undersampling*, dan *SMOTE*.

1) *Oversampling*
   Oversampling adalah sebuah teknik dalam analisis data dan *machine learning* yang digunakan untuk mengatasi masalah ketidakseimbangan dalam jumlah kelas data. Teknik *oversampling* dilakukan dengan maksud untuk meningkatkan jumlah sampel pada kelas minoritas hingga mencapai tingkat yang setara dengan jumlah sampel pada kelas mayoritas, dengan cara menggandakan sampel kelas minoritas secara acak (Saputro & Rosiyadi, 2022).
2) *Undersampling*
   Undersampling merupakan metode yang digunakan untuk menangani ketidakseimbangan data dengan mengurangi jumlah contoh dari kelas mayoritas hingga setara dengan jumlah contoh kelas minoritas. Ini bertujuan untuk menyelaraskan jumlah contoh di setiap kelas, mencegah model pembelajaran mesin dari *overfitting* pada kelas mayoritas, dan meningkatkan kemampuan model untuk mengenali kelas minoritas (Syukron & Subekti, 2018).
3) *SMOTE*
   SMOTE adalah pendekatan *oversampling* yang secara sintetis menghasilkan *instance* dengan memilih secara acak *instance* dari kelas minoritas dan menggunakan metode interpolasi untuk menghasilkan *instance* antara titik yang dipilih dan *instance* yang berdekatan (Kaur & Gosain, 2018).

E. TF-IDF

TF-IDF adalah metode yang digunakan untuk memberikan nilai bobot pada hubungan antara sebuah kata (*term*) dan sebuah dokumen. Data hasil *preprocessing* yang berupa kata akan diubah dalam bentuk angka dengan dilakukan pembobotan kata yang bermaksud untuk menghitung bobot pada masing-masing kata yang akan dipakai untuk fitur (Siregar, dkk., 2017).

$$tf_{t,d} = \frac{n_{t,d}}{N} \quad (1)$$

dengan,
$n_{t,d}$ = nilai istilah yang muncul
$tf_{t,d}$ = frekuensi kemunculan kata pada sebuah dokumen
N = semua *term* dalam dokumen

$$idf_d = log\ log\ (\frac{Nd}{df}) \quad (2)$$

dengan,
Nd = total dokumen
d*f* = banyak dokumen yang mengandung *term* tersebut

$$tfidf_{t,d} = tf_{t,d}\ \text{x}\ idf_d \quad (3)$$

dengan,
$tfidf_{t,d}$ = *Term Frequency-Inverse Document Frequency*
$tf_{t,d}$ = nilai *TF*
idf_d = nilai *IDF*

F. *Multinomial Naïve Bayes*

Algoritma *Multinomial Naïve Bayes* adalah salah satu varian dari algoritma *Naive Bayes* yang khusus digunakan dalam konteks pemrosesan teks, terutama untuk tugas klasifikasi teks. Algoritma ini sering digunakan dalam analisis sentimen, klasifikasi dokumen, dan tugas lain yang melibatkan data teks (Sanrilla, dkk., 2022).

$$P(c) = \frac{Tct+1}{\left(\sum \quad t'\epsilon V\, Tct\right)+B'} \tag{4}$$

dengan,
$P(c)$ = Probabilitas kata ke-n dengan diketahui kelas c
$Tct + 1$ = Jumlah kemunculan kata pada kelas c + 1
$\left(\sum \quad t'\epsilon V\, Tct\right)$ = Jumlah semua kata pada kelas c
B' = Semua kosakata (*Vocabulary*) yang muncul pada dokumen

G. *Gaussian Naïve Bayes*

Metode *Gaussian Naïve Bayes* merupakan teknik pengklasifikasian data yang berfokus pada perhitungan probabilitas berdasarkan Teorema *Bayes*. *Gaussian Naive Bayes* adalah salah satu turunan dari algoritma *Naïve Bayes* yang terdapat pada data mining dimana penggunaannya tergolong mudah dan pemrosesannya juga memiliki waktu yang cepat (Mujahidin, dkk., 2022).

$$P(X_i = x_i\,|Y = \ y_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{\frac{(x_i-\mu_{ij})^2}{2\sigma^2 ij}} \tag{5}$$

dengan,
P = peluang
Xi = atribut ke *i*
xi = nilai atribut ke *i*
Y = kelas yang dicari
yi = sub-kelas yang dicari
σ = standart deviasi
μ = nilai rata-rata / *mean*

H. *Confusion Matrix*

Confusion matrix merupakan adalah tabel yang memberikan data perbandingan antara klasifikasi yang diprediksi oleh sistem dengan hasil klasifikasi yang sebenarnya. Pengukuran yang diterapkan *confusion matrix* yaitu dengan menghitung *accuracy, precision, recall*, yang mengacu pada nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *dan False Negative* (FN) yang merupakan keluaran dari *confusion matrix*.

I. *K-Fold Cross Validation*

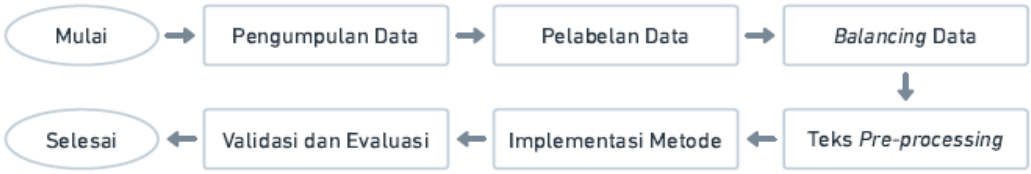
K-Fold Cross-Validation adalah suatu teknik dalam *machine learning* yang digunakan untuk mengukur sejauh mana model atau algoritma yang telah dibuat mampu menggeneralisasi data dengan baik. Teknik ini membagi *dataset* menjadi sejumlah K bagian, yang disebut "*Fold*". Di sisi lain, teknik ini mengukur performa model dan memperkirakan seberapa akurat suatu model prediktif ketika digunakan dalam situasi praktis (Ridwansyah, 2022).

J. Phyton

Python merupakan bahasa pemrograman tingkat tinggi yang diracik oleh Guido van Rossum. Bahasa pemrograman yang bersifat interaktif dan dapat dipakai di berbagai macam platform dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode dan merupakan salah satu bahasa popular yang ada kaitannya dengan data *science*, *machine learning*, dan *internet of things* (*IoT*) (Rahmadhika & Thantawi, 2021).

3. METODE PENELITIAN

Metode penelitian merupakan suatu langkah yang digunakan untuk mengatasi suatu problem yang logis. Dalam penelitian ini digunakan metode penelitian deskriptif. Dalam konteks penelitian deskriptif ini, bertujuan untuk menggambarkan fakta-fakta dan informasi dengan akurat, dengan mengambil data yang sesuai dengan kenyataan. Tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan penelitian

A. Pengumpulan Data dan Pelabelan Data

Data yang dipakai pada penelitian ini adalah komentar dari pengguna media sosial *Twitter* tentang pertandingan antara timnas Indonesia melawan timnas Argentina. Data dikumpulkan melalui teknik *crawling* sejumlah 700 komentar.

Proses selanjutnya akan dilakukan pelabelan data. Pada proses ini data akan dibagi menjadi 3 sentimen yaitu, positif, negatif dan netral 0. Pelabelan data dilakukan secara manual dengan divalidasi oleh ahli bahasa.

B. Teks *Pre-processing*

Selanjutnya akan dilakukan *preprocessing text* untuk meningkatkan akurasi klasifikasi data dengan mengubah bentuk dokumen menjadi data yang terstruktur. Tahapan *preprocessing* ditunjukkan pada Tabel 1, Tabel 2, dan Tabel 3.

1. *Cleansing*

Tabel 1. *Cleansing*

Data *tweet*	*cleansing*
@Argentina Messi ga mau datang ke Indonesia karena takut sama timnas? @Kwkwkw https://t.co/Vh7DrAB8Xf	Argentina Messi ga mau datang ke Indonesia karena takut sama timnas yak

2. *Case Folding* dan *tokenizing*

Tabel 2. *Case Folding dan Tokenizing*

Case Folding	*Tokenizing*
argentina messi ga mau datang ke indonesia karena takut sama timnas	[‘argentina’, ‘messi’, ‘ga’, ‘mau’, ‘datang’, ‘ke’, ‘indonesia’, ‘karena’, ‘takut, ‘sama’, ‘timnas’]

3. *Stop removal* dan *Stemming*

Tabel 3. *Stop removal dan Stemming*

Stop removal	*stemming*
[‘argentina’, ‘messi’, ‘mau’, ‘datang’, ‘indonesia’, ‘karena’, ‘takut, ‘sama’, ‘timnas’]	[‘argentina’, ‘messi’, ‘mau’, ‘datang’, ‘indonesia’, ‘karena’, ‘takut, ‘sama’, ‘timnas’]

C. Implementasi Metode

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan pembobotan pada kata (*term*) terhadap teks dokumen. Pada tahap ini output nilai *continue* yang sudah diproses melalui *TF-IDF* akan diproses menggunakan metode *Gaussian Naïve Bayes*. Sedangkan untuk *Multinomial Naïve Bayes* ini membutuhkan sebuah tabel yang bernilai diskrit seperti kemunculan kata pada dokumen yang digunakan untuk acuan perhitungannya.

D. Validasi dan evaluasi

Tahap terakhir yaitu, setelah dilakukannya pengklasifikasian menggunakan metode *Multinomial Naïve Bayes* dan *Gaussian Naïve Bayes* akan dilakukan pengukuran dan pengujian hasil dengan persamaan *confusion matrix*. *Confusion Matrix* ini berfungsi untuk mengetahui nilai kriteria yang

didapat sehingga hasil akhirnya yaitu berupa nilai akurasi, nilai presisi dan nilai *recall*, dengan membandingkan nilai aktual dan hasil klasifikasi.

4. HASIL DAN PEMBAHASAN

Tahap Berikutnya adalah proses pengujian klasifikasi data dengan menggunakan teknik *K-Fold Cross Validation*.

A. *Multinomial Naïve Bayes*

Tabel 4 dan Tabel 5 adalah tabel yang menunjukkan hasil pengujian klasifikasi data dengan menggunakan *K-Fold Cross Validation* 2,5,7,10 menggunakan metode *Multinomial Naïve Bayes*.

Tabel 4. Hasil *Multinomial Naïve Bayes* Normal Dataset dan *Oversampling*

K-Fold	Langkah Uji	*Multinomial Naïve Bayes*					
		Normal Dataset			*Random Oversampling*		
		Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
2-*Fold*	Uji 1	66%	67%	66%	76%	76%	72%
	Uji 2	64%	67%	67%	75%	69%	74%
5-*Fold*	Uji 1	73%	62%	72%	80%	80%	78%
	Uji 2	71%	71%	70%	77%	78%	75%
	Uji 3	66%	70%	75%	77%	82%	81%
	Uji 4	70%	69%	69%	82%	81%	79%
	Uji 5	63%	69%	60%	79%	79%	82%
7-*Fold*	Uji 1	65%	65%	69%	81%	79%	81%
	Uji 2	67%	67%	69%	73%	80%	81%
	Uji 3	77%	68%	79%	82%	79%	75%
	Uji 4	73%	73%	75%	79%	76%	76%
	Uji 5	65%	74%	68%	73%	80%	73%
	Uji 6	65%	63%	60%	73%	79%	81%
	Uji 7	69%	70%	64%	81%	85%	83%
10-*Fold*	Uji 1	64%	65%	60%	83%	78%	86%
	Uji 2	71%	62%	71%	76%	77%	72%
	Uji 3	67%	65%	78%	80%	82%	83%
	Uji 4	68%	75%	70%	89%	75%	85%
	Uji 5	65%	60%	72%	75%	86%	79%
	Uji 6	74%	71%	65%	76%	80%	77%
	Uji 7	72%	80%	52%	78%	82%	81%
	Uji 8	70%	70%	65%	89%	86%	83%
	Uji 9	72%	75%	71%	79%	81%	75%
	Uji 10	62%	68%	75%	81%	81%	79%

Tabel 5. Hasil *Multinomial Naïve Bayes Undersampling* dan *SMOTE*

K-Fold	Langkah Uji	*Multinomial Naïve Bayes*					
		Random UnderSampling			*SMOTE*		
		Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
2-*Fold*	Uji 1	63 %	66%	56%	66%	62%	64%
	Uji 2	65%	65%	60%	66%	63%	62%
5-*Fold*	Uji 1	59 %	71%	59%	65%	70%	74%
	Uji 2	62%	60%	67%	70%	74%	70%
	Uji 3	73%	67%	69%	77%	71%	66%
	Uji 4	65%	64%	67%	68%	65%	69%
	Uji 5	53%	64%	71%	62%	71%	67%
7-*Fold*	Uji 1	68%	63%	52%	67%	68%	77%
	Uji 2	53%	74%	74%	68%	74%	71%
	Uji 3	73%	63%	52%	69%	72%	64%

K-Fold	Langkah Uji	*Multinomial Naïve Bayes*					
		Random UnderSampling			*SMOTE*		
		Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
	Uji 4	77%	66%	73%	68%	75%	72%
	Uji 5	66%	65%	61%	71%	74%	71%
	Uji 6	65%	71%	69%	77%	69%	79%
	Uji 7	66%	57%	76%	61%	64%	66%
10-*Fold*	Uji 1	71%	60%	75%	72%	75%	65%
	Uji 2	52%	68%	65%	70%	78%	68%
	Uji 3	65%	72%	65%	71%	65%	68%
	Uji 4	63%	65%	72%	71%	80%	70%
	Uji 5	70%	81%	59%	81%	80%	80%
	Uji 6	79%	50%	68%	72%	68%	78%
	Uji 7	72%	61%	75%	75%	69%	69%
	Uji 8	63%	72%	68%	65%	62%	68%
	Uji 9	63%	70%	61%	62%	66%	66%
	Uji 10	72%	70%	70%	66%	65%	72%

Berdasarkan Tabel 4 dan 5 di atas dapat diperoleh jika menggunakan metode *Multinomial Naïve Bayes* nilai akurasi tertinggi sebesar 77%, presisi sebesar 80% dan nilai *recall* sebesar 79%. Pada *Multinomial Naïve Bayes random oversampling* memperoleh nilai akurasi tertinggi sebesar 89%, presisi sebesar 86% dan nilai *recall* sebesar 86%, lalu *Multinomial Naïve Bayes random undersampling* memperoleh nilai akurasi tertinggi sebesar 79%, presisi sebesar 81% dan nilai *recall* sebesar 76%, dan *Multinomial Naïve Bayes SMOTE* memperoleh nilai akurasi tertinggi sebesar 81%, presisi sebesar 80% dan nilai *recall* sebesar 80%.

B. *Gaussian Naïve Bayes*

Tabel 6 dan Tabel 7 adalah tabel yang menunjukkan hasil pengujian klasifikasi data dengan menggunakan *K-Fold Cross Validation* 2,5,7,10 menggunakan metode *Gaussian Naïve Bayes*.

Tabel 6. Hasil *Gaussian Naïve Bayes* Normal Dataset dan *Oversampling*

K-Fold	Langkah Uji	*Gaussian Naïve Bayes*					
		Normal Dataset			*Random Oversampling*		
		Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
2-*Fold*	Uji 1	58%	58%	56%	72%	70%	66%
	Uji 2	62%	56%	54%	69%	69%	72%
5-*Fold*	Uji 1	58%	59%	52%	78%	74%	73%
	Uji 2	57%	59%	57%	74%	75%	72%
	Uji 3	59%	56%	55%	74%	79%	78%
	Uji 4	59%	54%	56%	73%	69%	76%
	Uji 5	52%	57%	61%	70%	75%	73%
7-*Fold*	Uji 1	61%	50%	52%	71 %	78%	80%
	Uji 2	60%	52%	60%	72%	79%	78%
	Uji 3	48%	70%	57%	78%	71%	68%
	Uji 4	57%	53%	48%	80%	73%	76%
	Uji 5	50%	60%	66%	79%	74%	80%
	Uji 6	58%	62%	56%	73%	73%	73%
	Uji 7	52%	62%	54%	74%	74%	76%
10-*Fold*	Uji 1	61%	61%	55%	79%	79%	72%
	Uji 2	57%	62%	54%	73%	68%	76%
	Uji 3	57%	51%	60%	71%	68%	78%
	Uji 4	67%	58%	58%	78%	78%	77%
	Uji 5	54%	67%	60%	79%	79%	72%
	Uji 6	58%	44%	57%	73%	71%	74%
	Uji 7	62%	51%	62%	82%	82%	80%

K-Fold	Langkah Uji	*Gaussian Naïve Bayes*					
		Normal Dataset			*Random Oversampling*		
		Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
	Uji 8	54%	64%	58%	73%	79%	86%
	Uji 9	54%	64%	60%	74%	78%	85%
	Uji 10	58%	61%	51%	76%	77%	75%

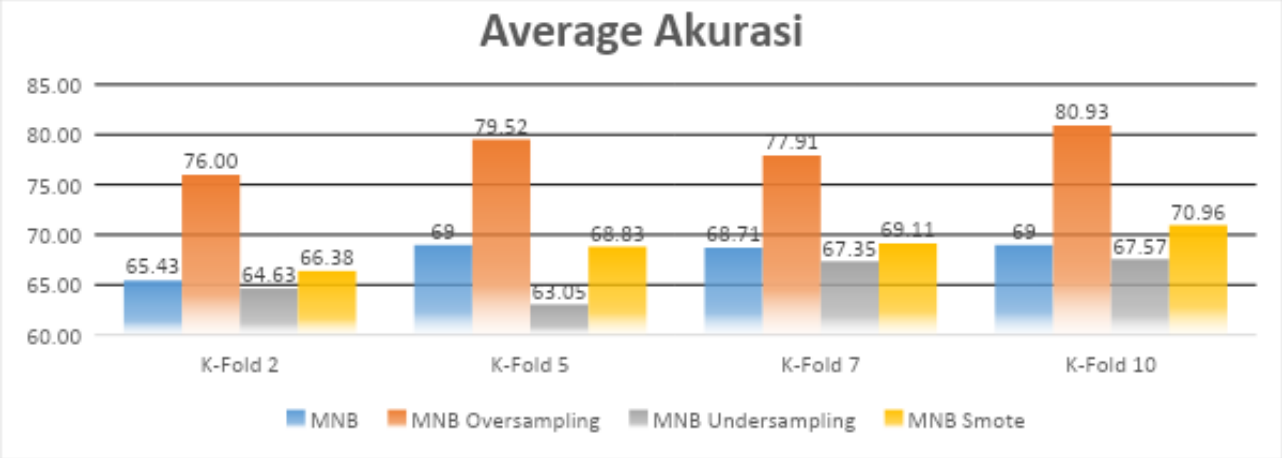
Tabel 7. Hasil *Gaussian Naïve Bayes* Normal *Undersampling* dan *SMOTE*

K-Fold	Langkah Uji	*Gaussian Naïve Bayes*					
		Random Undersampling			*SMOTE*		
		Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
2-*Fold*	Uji 1	54%	54%	56%	74%	75%	76%
	Uji 2	56%	55%	60%	77%	77%	72%
5-*Fold*	Uji 1	68%	64%	60%	72%	71%	79%
	Uji 2	53%	51%	59%	74%	76%	76%
	Uji 3	61%	67%	62%	75%	84%	79%
	Uji 4	64%	60%	53%	84%	71%	74%
	Uji 5	55%	60%	68%	79%	85%	74%
7-*Fold*	Uji 1	66%	57%	63%	84 %	79%	77%
	Uji 2	58%	73%	66%	80%	84%	80%
	Uji 3	61%	63%	58%	74%	81%	84%
	Uji 4	55%	57%	61%	83%	74%	81%
	Uji 5	68%	65%	52%	80%	79%	76%
	Uji 6	58%	61%	63%	78%	82%	78%
	Uji 7	65%	66%	65%	76%	77%	78%
10-*Fold*	Uji 1	48%	68%	62%	81%	68%	78%
	Uji 2	68%	65%	65%	82%	81%	71%
	Uji 3	68%	52%	63%	78%	85%	91%
	Uji 4	56%	52%	56%	78%	85%	90%
	Uji 5	65%	68%	59%	78%	80%	81%
	Uji 6	68%	68%	61%	72%	78%	82%
	Uji 7	72%	50%	63%	86%	78%	78%
	Uji 8	61%	54%	68%	82%	82%	65%
	Uji 9	56%	59%	45%	78%	84%	72%
	Uji 10	56%	61%	63%	78%	79%	84%

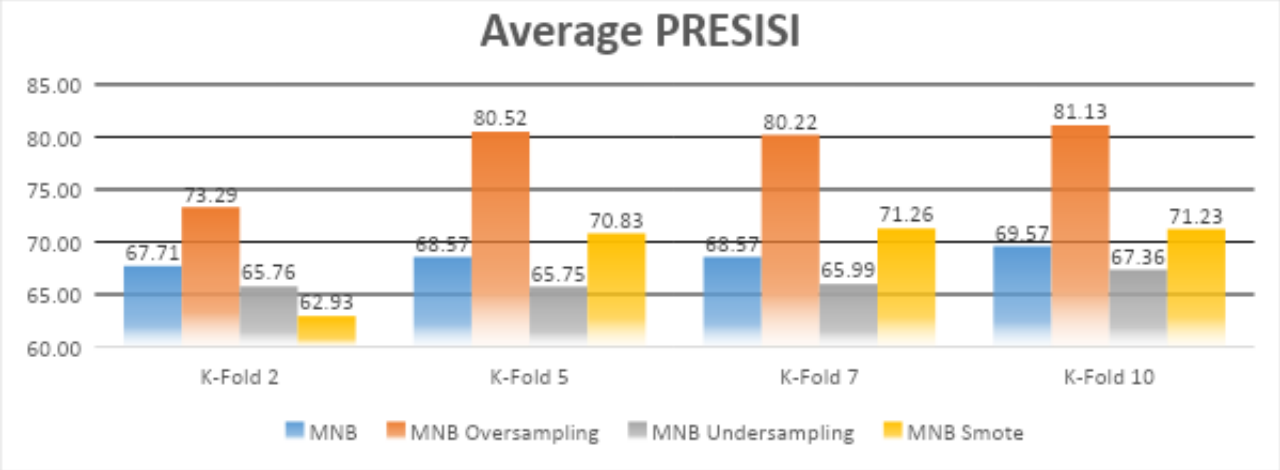
Berdasarkan Tabel 6 dan 7 di atas dapat diperoleh jika menggunakan metode *Gaussian Naïve Bayes* memperoleh nilai akurasi tertinggi sebesar 67%, presisi sebesar 70% dan nilai *recall* sebesar 62%. Pada *Gaussian Naïve Bayes random oversampling* memperoleh nilai akurasi tertinggi sebesar 82%, presisi sebesar 82% dan *recall* 86%, lalu *Gaussian Naïve Bayes random undersampling* memperoleh nilai akurasi tertinggi sebesar 72%, presisi 73% dan *recall* sebesar 68%, dan *Gaussian Naïve Bayes SMOTE* memperoleh nilai akurasi tertinggi sebesar 86%, presisi 85% dan *recall* 91%.

C. *Average Multinomial Naive Bayes*

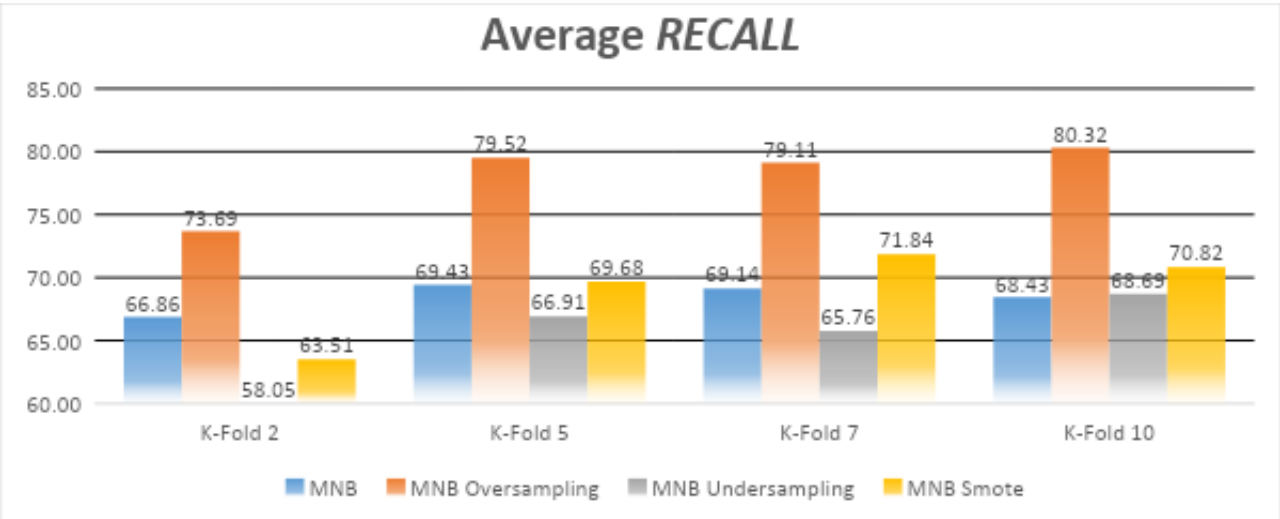
Hasil pengujian validasi dan evaluasi menggunakan 2,5,7,10 *cross validation* menggunakan *Multinomial Naïve Bayes* dengan teknik random *oversampling*, *undersampling*, dan *SMOTE*. Diketahui rata-rata akurasi, presisi, *recall* yang ditunjukkan pada Gambar 2, Gambar 3, dan Gambar 4.



Gambar 2. Hasil rata-rata akurasi *Multinomial Naive Bayes*



Gambar 3. Hasil rata-rata Presisi *Multinomial Naïve Bayes*



Gambar 4. Hasil rata-rata *Recall Multinomial Naïve Bayes*

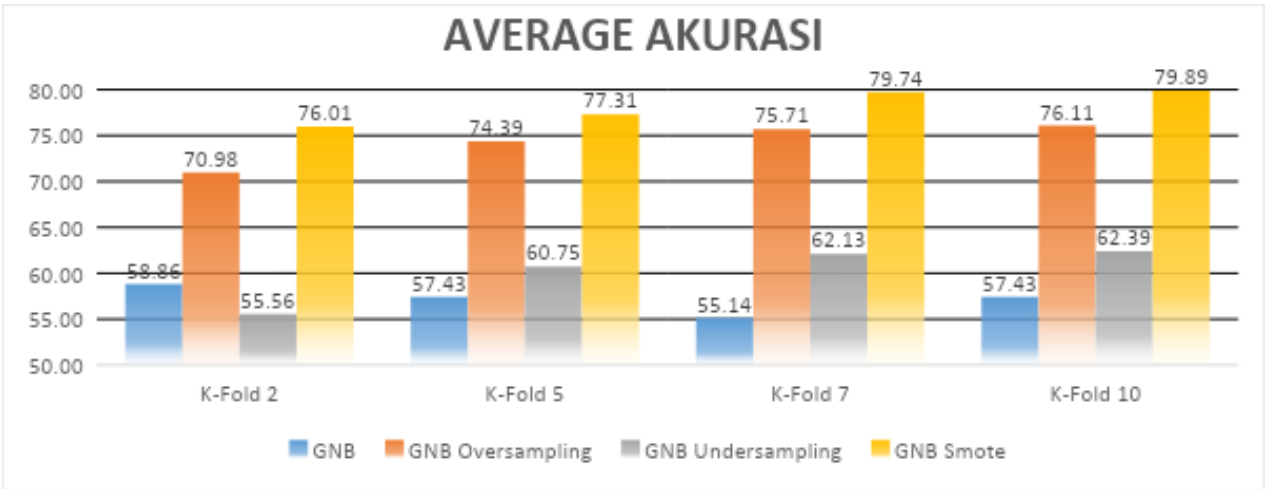
Berdasarkan Gambar 2, Gambar 3 dan Gambar 4 adalah hasil rata-rata perhitungan *K-Fold* dengan metode *Multinomial Naïve Bayes* dan *Multinomial Naïve Bayes* menambahkan teknik *balancing* yaitu *random oversampling, random undersampling* dan *SMOTE*. Hasil perhitungan *K-Fold* terdapat peningkatan pada tingkat akurasi, presisi, dan *recall*, di mana awalnya metode *Multinomial*

Naïve Bayes tanpa menggunakan teknik *balancing* diketahui nilai rata-rata akurasi 69%, presisi 69%, dan *recall* 69%. Namun, ketika menambahkan dengan teknik *balancing*, diketahui nilai rata-rata akurasi, presisi dan *recall* tertinggi jika menggunakan metode *Multinomial Naïve Bayes random oversampling* memperoleh rata-rata akurasi sebesar 80%, nilai presisi 81% dan nilai *recall* 80%.

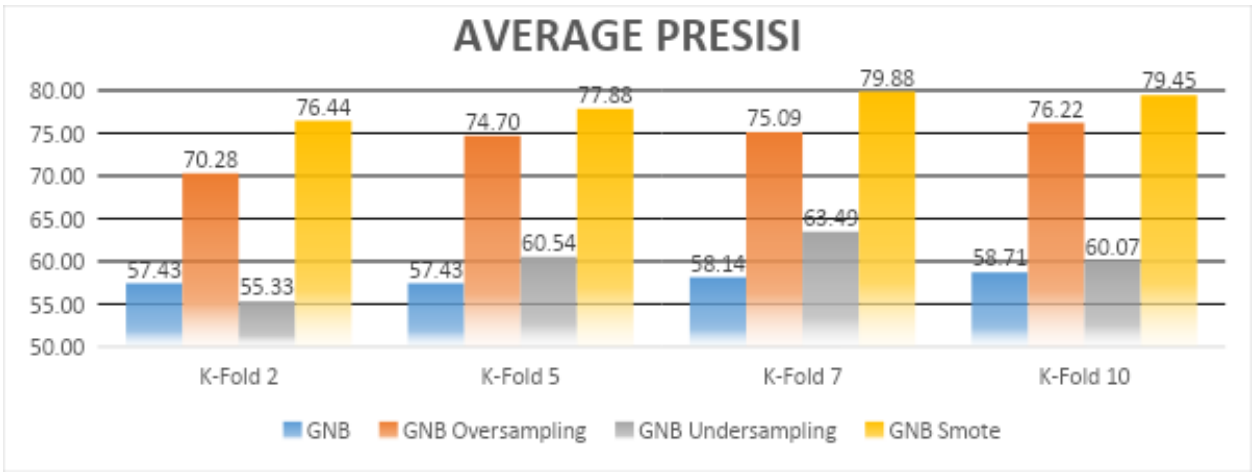
D. *Average Gaussian Naive Bayes*

Hasil pengujian validasi dan evaluasi menggunakan 2,5,7,10 *cross validation* menggunakan *Gaussian Naïve Bayes* dengan teknik *random oversampling*, *undersampling*, *SMOTE*. Diketahui rata-rata akurasi, presisi, *recall* sebagai berikut.

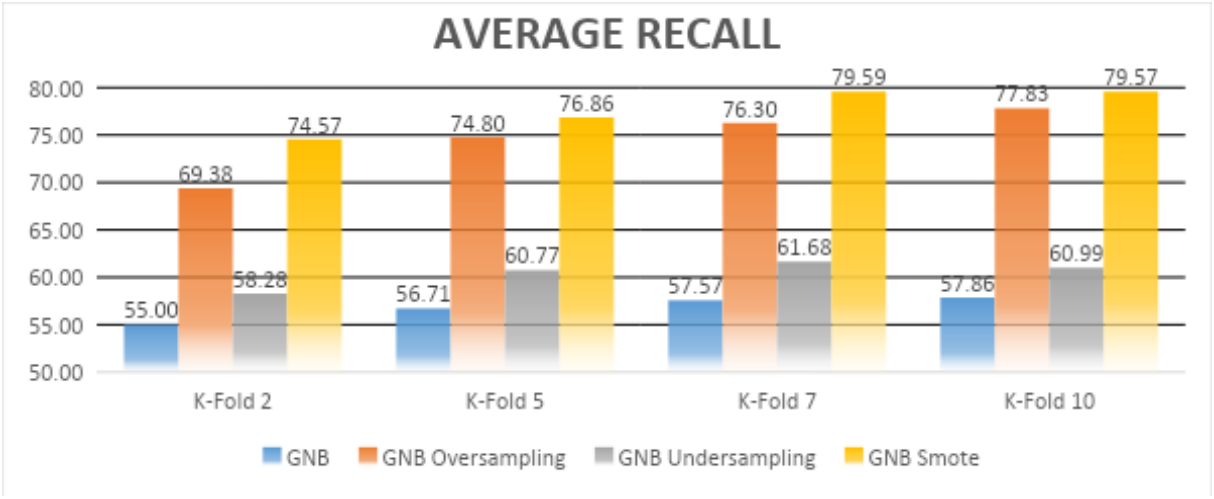
Dari hasil rata-rata perhitungan *K-Fold* dengan *Gaussian Naïve Bayes* dan *Gaussian Naïve Bayes* menambahkan teknik *balancing* yaitu *random oversampling*, *random undersampling* dan *SMOTE*. Hasil perhitungan *K-Fold* terdapat peningkatan pada tingkat akurasi, presisi, dan *recall*, di mana awalnya metode *Gaussian Naïve Bayes* tanpa menggunakan teknik *balancing* diketahui nilai rata-rata akurasi 58%, presisi 58%, dan *recall* 57%. Namun, ketika menambahkan dengan teknik *balancing*, diketahui nilai rata-rata akurasi, presisi dan *recall* tertinggi jika menggunakan metode *Gaussian Naïve Bayes SMOTE* memperoleh rata-rata akurasi sebesar 79%, nilai presisi 79% dan nilai *recall* 79%. Berikut adalah hasil rata-rata akurasi, presisi dan *recall* jika menggunakan metode *Gaussian Naïve Bayes* dengan menambahkan teknik *balancing*, *Random Oversampling*, *undersampling*, *SMOTE* yang terdapat pada Gambar 5, Gambar 6 dan Gambar 7.



Gambar 5. Hasil rata-rata akurasi *Gaussian Naïve Bayes*



Gambar 6. Hasil rata-rata presisi *Gaussian Naïve Bayes*



Gambar 7. Hasil rata-rata *recall Gaussian Naïve Bayes*

5. KESIMPULAN

Setelah dilakukan pengujian validasi dan evaluasi hasil pada metode *Multinomial Naïve Bayes* dan *Gaussian Naïve Bayes* dengan menambahkan teknik *balancing*, maka dapat diambil kesimpulan sebagai berikut:

a. Pada penelitian ini, metode *Multinomial Naïve Bayes* tanpa menggunakan teknik *balancing* diketahui mencapai akurasi tertinggi sebesar 77% pada K-*Fold* 7 langkah ketiga, presisi 80% pada K-*Fold* 10 langkah ketujuh dan *recall* 79% pada K-*Fold* 7 langkah ketiga. Sedangkan untuk metode *Gaussian Naïve Bayes* tanpa menggunakan teknik balancing diketahui mencapai akurasi tertinggi sebesar 67% pada K-*Fold* 10 langkah keempat, presisi 70% pada K-*Fold* 7 langkah ketiga dan *recall* 62% pada K-*Fold* 10 langkah ketujuh.
b. Pada penelitian ini, metode *Multinomial Naïve Bayes* menggunakan teknik *balancing* random *oversampling* diketahui mencapai akurasi tertinggi sebesar 89% pada K-*Fold* 10 langkah uji 4 dan 8, presisi 86% pada K-*Fold* 10 langkah uji 5 dan 8, untuk *recall* 86% pada K-*Fold* 10 langkah uji 1. Random *undersampling* diketahui mencapai akurasi tertinggi sebesar 79% pada K-*Fold* 10 langkah uji 5, presisi 81% pada K-*Fold* 10 langkah uji 5, dan *recall* 76% pada K-*Fold* 7 langkah uji 7. *SMOTE* diketahui mencapai akurasi tertinggi sebesar 81% pada K-*Fold* 10 langkah uji 5, presisi 80% pada K-*Fold* 10 langkah uji 4 dan 5, untuk *recall* 80% pada K-*Fold* 10 langkah uji 5. Sedangkan untuk metode *Gaussian Naïve Bayes* menggunakan teknik *balancing random oversampling* diketahui mencapai akurasi tertinggi sebesar 82% pada K-*Fold* 10 langkah ketujuh, presisi 82% pada K-*Fold* 10 langkah ketujuh, untuk *recall* 86% pada K-*Fold* 10 langkah kedelapan. random *undersampling* diketahui mencapai akurasi tertinggi sebesar 72% pada K-*Fold* 10 langkah uji 7, presisi 73% pada K-*Fold* 7 langkah uji 2, dan *recall* 68% pada K-*Fold* 5 langkah uji 5 dan K-*Fold* 10 langkah uji 8. *SMOTE* diketahui mencapai akurasi tertinggi sebesar 86% pada K-*Fold* 10 langkah uji 7, presisi 85% pada K-*Fold* 10 langkah uji 3 dan 4, untuk *recall* 91% pada K-*Fold* 10 langkah uji 3.

Pada penelitian ini terdapat beberapa kekurangan yaitu proses normalisasi, metode validasi, teknik pembagian data, dan pembobotan kata yang dijelaskan secara detail sebagai berikut:

a. Kalimat tidak terstruktur dan kata yang tidak formal dapat menambahkan metode normalisasi kata seperti Levenshtein distance untuk memperoleh hasil yang maksimal pada proses normalisasi kata dan memperoleh tingkat akurasi yang optimal.
b. Pelabelan data dapat menggunakan metode pelabelan otomatis menggunakan sistem seperti *lexicon-based* mungkin dapat memperoleh akurasi lebih optimal.

c. Teknik pembagian data dapat menggunakan metode lain seperti *Holdout Validation*, sehingga kita dapat mengetahui teknik pembagian data mana yang lebih efektif dalam mengklasifikasikan sebuah data.
d. Dapat menggunakan metode pembobotan kata selain *TF-IDF* seperti *Binary weighting*, *chi-square* dan lain-lain. Sehingga kita dapat mengetahui pembobotan kata mana yang lebih cocok digunakan dalam metode yang akan dipakai.

6. DAFTAR PUSTAKA

Astari, N. M. A. J., Divayana, D. G. H. & Indrawan, G. (2020). Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode *Naive Bayes* Classifier. *Jurnal Sistem Dan Informatika (JSI)*, *15*(1), 27–29. https://doi.org/10.30864/jsi.v15i1.332

Astiningrum, M., Haniah, M., & Pradana, Y. R. Y. (2020). Analisis Sentimen Tentang Opini Terhadap Performa Timnas Sepak Bola Indonesia Pada Twitter. *Seminar Informatika Aplikatif Polinema (Siap)*, 35— 39.

Diantika, S. (2023). Penerapan Teknik Random Oversampling Untuk Mengatasi Imbalance Class Dalam Klasifikasi Website Phishing Menggunakan Algoritma Lightgbm. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 19–25. https://doi.org/10.36040/jati.v7i1.6006

Ernayanti, T., Mustafid, M., Rusgiyono, A., & Hakim, A. R. (2023). Penggunaan Seleksi Fitur Chi-Square Dan Algoritma *Multinomial Naïve Bayes* Untuk Analisis Sentimen Pelanggan Tokopedia. *Jurnal Gaussian*, *11*(4), 562–571. https://doi.org/10.14710/j.gauss.11.4.562-571

Fariz, M. I., Arifianto, D., & Rahayu, Y. D. (2021). *OPTIMASI METODE MULTINOMIAL NAÏVE BAYES*. *2*(2), 84–91.

Ghiffarie, A., Salsabila, K. D. A., Baistama, R. P., Variadi, M. I., & Rhajendra, M. D. (2019). Analisis Sentimen Terhadap Produk The Body Shop Tea Tree Oil. *JTMI: Jurnal Teknologi & Manajemen Informatika* -Vol.5, No.1.

Indrawati, A. (2021). Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset. *JIKO (Jurnal Informatika Dan Komputer)*, *4*(1), 38–43. https://doi.org/10.33387/jiko.v4i1.2561

Kaur, P., & Gosain, A. (2018). Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. *Advances in Intelligent Systems and Computing*, *653*(January), 23–30. https://doi.org/10.1007/978-981-10-6602-3_3

Mujahidin, S., Prasetio, B., & Utomo, M. C. C. (2022). Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM Pada Komentar Youtube Dengan Metode *Gaussian Naïve Bayes*. *Voteteknika (Vocational Teknik Elektronika Dan Informatika)*, *10*(3), 17. https://doi.org/10.24036/voteteknika.v10i3.118299

Prasetyo, D., & Pradana, M. (2021). Analisis Sentimen Media Sosial Terhadap Bank Rakyat Indonesia (bri) Sebagai Sponsor Resmi Liga Sepak Bola Indonesia (liga 1). *EProceedings …*, *8*(6), 8556–8561. https://openlibrarypublications.telkomuniversity.ac.id/index.php/management/article/view/17081%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/management/article/view/17081/16793

Rahmadhika, M. K., & Thantawi, A. M. (2021). Rancang Bangun Aplikasi Face Recognition Pada Pendekatan CRM Menggunakan Opencv Dan Algoritma Haar Cascade. *IKRA-ITH INFORMATIKA: Jurnal Komputer Dan Informatika*, *5*(1), 109–118.

Ridwansyah, T. (2022). Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-*Fold* Cross Validation Dan Naïve Bayes Classifier. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *2*(5), 178–185. https://doi.org/10.30865/klik.v2i5.362

Sanrilla, S., Ransi, N., Surimi, L., Tenriawaru, A. & Saidi, L. O. (2022). Analisis Sentimen Masyarakat

Terhadap Toko Online Aplikasi Shopee Menggunakan Metode *Multinomial Naïve Bayes*. *Jurnal Matematika Komputasi Dan Statistika*, *2*(2), 68–75. https://doi.org/10.33772/jmks.v2i2.9

Saputro, E., & Rosiyadi, D. (2022). Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes. *Bianglala Informatika*, *10*(1), 42–47. https://doi.org/10.31294/bi.v10i1.11739

Siregar, R. R. A., Sinaga, F. A., & Arianto, R. (2017). Aplikasi Penentuan Dosen Penguji Skripsi Menggunakan Metode TF-IDF dan Vector Space Model. *Computatio: Journal of Computer Science and Information Systems*, *1*(2), 171. https://doi.org/10.24912/computatio.v1i2.1014

Syukron, A., & Subekti, A. (2018). Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika*, *5*(2), 175–185. https://doi.org/10.31311/ji.v5i2.4158

Tarju, T., & Wahidi, R. (2017). Pengaruh Metode Latihan Terhadap Peningkatan Passing Dalam Permainan Sepak Bola. *JUARA: Jurnal Olahraga*, *2*(2), 66. https://doi.org/10.33222/juara.v2i2.35

Verawati, I., & Audit, B. S. (2022). Algoritma Naïve Bayes Classifier Untuk Analisis Sentiment Pengguna Twitter Terhadap Provider By.u. *Jurnal Media Informatika Budidarma*, *6*(3), 1411. https://doi.org/10.30865/mib.v6i3.4132

Wibawa. (2018). Metode-metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134.