Published online at: **https://journal.rekayasainformatika.com/**

JURNAL REIN
(REKAYASA INFORMATIKA)

Vol. x No. x (2024) hal 16 - 21

Analisis Sentimen Masyarakat Terhadap Indonesia Vs Uzbekistan Menggunakan Smote (Synthetic Minority Over-sampling Technique) Dan Knn (K-Nearest Neighbor)

Ade Rocky Saputra[1], Muhammad Anugrah Hakiki[2],Hafiz Irsyad[3]
[1,2,3]Prodi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Kota Palembang, Indonesia
[1]dekyraa1003@mhs.mdp.ac.id, [2]m.anugrahhakiki@mhs.mdp.ac.id, [3]hafizirsyad@mdp.ac.id

Abstract

Football is one of the most popular sports in the world, including in Indonesia. The match between Indonesia and Uzbekistan in the AFC U-23 Championship attracted widespread public attention. This study aims to analyze public sentiment towards the match using the Synthetic Minority Over-sampling Technique (SMOTE) and the K-Nearest Neighbor (KNN) algorithm. The study also examines the impact of SMOTE implementation on the performance of the sentiment classification model. The results indicate that the application of SMOTE led to a decrease in the performance of the KNN model, with a 10% reduction in accuracy, a 16% reduction in precision, a 3% reduction in recall, and an 11% reduction in F1-Score. Additionally, sentiment analysis revealed that the majority of public sentiment towards the match outcome was negative.

Kata Kunci: Football; KNN; Sentiment Analysis; SMOTE; Sentimen

1. Pendahuluan

Sepak bola adalah salah satu olahraga yang paling populer di dunia, termasuk di Indonesia. Pertandingan-pertandingan sepak bola, terutama yang melibatkan tim nasional, selalu menarik perhatian luas masyarakat. Salah satu pertandingan yang menarik perhatian baru-baru ini adalah pertandingan antara Indonesia dan Uzbekistan di Piala Asia U-23. Pertandingan ini tidak hanya penting dari segi kompetisi, tetapi juga menjadi cerminan dari dukungan dan harapan masyarakat terhadap tim nasional mereka.

Di era digital saat ini, media sosial dan *platform online* menjadi tempat utama bagi masyarakat untuk mengekspresikan pendapat dan perasaan mereka terhadap berbagai peristiwa, termasuk pertandingan sepak bola. Analisis sentimen masyarakat terhadap pertandingan ini dapat memberikan gambaran yang jelas tentang bagaimana perasaan publik, baik positif maupun negatif, terhadap performa tim nasional. Informasi ini sangat berharga bagi pelatih, pemain, dan manajemen tim untuk memahami bagaimana persepsi publik dapat mempengaruhi moral dan strategi tim di masa mendatang.

Namun, mengumpulkan dan menganalisis data dari media sosial menghadirkan tantangan tersendiri. Data yang diperoleh dari media sosial sering kali tidak seimbang (*imbalanced*), di mana jumlah sentimen positif atau negatif bisa jauh lebih banyak dibandingkan yang lain. Untuk mengatasi masalah ini, teknik seperti *Synthetic Minority Over-sampling Technique* (SMOTE) dapat digunakan untuk menyeimbangkan data. SMOTE adalah pembangkitan data minoritas sebanyak data mayoritas [1].

Setelah data diseimbangkan, langkah selanjutnya adalah mengklasifikasikan sentimen tersebut. Algoritma K-Nearest Neighbor adalah algoritma pembelajaran terawasi yang dapat digunakan untuk tugas klasifikasi dan regresi. Ini bekerja dengan menemukan K titik data terdekat ke sampel yang diberikan, dan menggunakan label kelas atau nilai dari titik data ini untuk membuat prediksi tentang label kelas atau nilai sampel [2]. Dengan menggunakan SMOTE untuk penyeimbangan data dan KNN untuk klasifikasi, terdapat kemungkinan didapatkan model yang memiliki performa yang lebih baik.

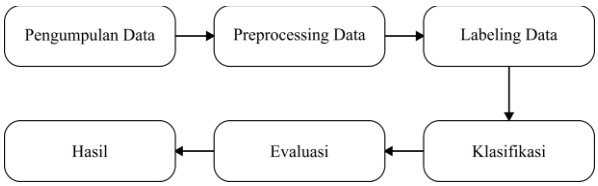
Dengan demikian, penelitian ini bertujuan untuk melakukan analisis bagaimana penerapan SMOTE dapat berpengaruh pada performa model. Penelitian ini juga bertujuan untuk menganalisis sentimen masyarakat terhadap pertandingan antara Indonesia dan Uzbekistan pada Piala Asia U-23 menggunakan metode SMOTE dan KNN.

Hasil analisis ini diharapkan dapat memberikan wawasan mendalam tentang perasaan dan opini publik, serta menyajikan informasi berharga bagi para pemangku kepentingan di dunia sepak bola Indonesia, termasuk pelatih, pemain, manajemen tim, dan federasi sepak bola, untuk meningkatkan performa dan strategi tim di masa depan.

Received: 19-06-2024 | Accepted: 29-06-2024 | Published Online: 25-07-2024

2. Metode Penelitian

Pada bagian ini menjelaskan tahapan yang dilakukan untuk menganalisis komentar-komentar yang dilakukan dalam 6 langkah, yaitu pengumpulan data, preprocessing, labeling data, klasifikasi, dan evaluasi seperti yang terlihat pada gambar 1 dan akan dijelaskan lebih lanjut pada sub bab berikutnya.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Data dikumpulkan dari video YouTube yang menampilkan sorotan pertandingan Indonesia melawan Uzbekistan dalam AFC U23 Asian Cup Qatar 2024. Komentar-komentar dikumpulkan menggunakan Netlytic sebagai alat bantu pengumpulan data. Jumlah data yang terkumpul mencapai 1250 komentar. Data yang didapatkan berupa *id, author, description, guid, to, likecount, link, pubdate, replycount, title*, dan *authorChannelUrl*.

2.2 Preprocessing Data

Preprocessing Data adalah teknik yang digunakan untuk mengekstrak informasi dari data mentah. Teknik ini berfungsi untuk menghilangkan noise dalam data mentah, sehingga data tersebut siap untuk diproses lebih lanjut. Dalam proses ini, data mentah diubah menjadi kumpulan data yang terstruktur, dengan memperhatikan gaya penulisan dan akurasi penulisan [1]. Bagian ini memiliki beberapa tahap seperti data *cleaning, case folding, normalization, stopword removal, tokenization, dan stemming.*

Tahap data cleaning adalah proses membersihkan data dari data *noise* dan tidak konsisten [3]. Data cleaning menghapus kolom-kolom yang tidak diinginkan dari dataset dengan tujuan mempersempit fokus dan meningkatkan kualitas data, tahap ini juga menghapus data-data duplikat. Langkah selanjutnya adalah *case folding* , *case folding* adalah proses mengubah seluruh huruf menjadi huruf kecil [4].*Case folding* akan mengubah huruf pada kolom deskripsi menjadi huruf kecil, dilanjutkan dengan *normalization*. *Normalization* adalah proses teks yang dilakukan untuk memperbaiki kata-kata yang salah eja atau kata yang disingkat [5]. *Normalization* dilakukan untuk mengganti kata-kata tidak sesuai dengan Bahasa Indonesia dengan ejaan yang benar. Kemudian menghilangkan kata-kata yang umum digunakan dan tidak mempunyai Informasi yang berharga pada suatu konteks atau biasa disebut *stopword removal* [6] ,seperti "yang", "di", "dan", "ke", dan lain-lain. Kemudian, dilakukan *tokenization*. *Tokenization* adalah tugas memisahkan deretan kata di dalam kalimat,paragraf atau halaman menjadi token atau potongan kata tunggal atau *trimmed word*[7], untuk membagi kalimat menjadi kata-kata, dan terakhir, dilakukan *stemming*. *Stemming* adalah metode untuk mencari kata dasar dari sebuah kata[8].*Stemming* disini dilakukan untuk menghapus imbuhan-imbuhan dari kata-kata, seperti "-an", "-nya", dan "me-", dengan tujuan mencari kata dasar dari setiap kata pada kalimat serta menghilangkan simbol-simbol yang tidak diinginkan

2.3. *Labeling Data*

Pada tahap ini, proses dilakukan secara manual menggunakan Google Spreadsheet dengan menambahkan fitur "sentimen" pada setiap data dan memberikan label sesuai dengan sentimennya, baik itu positif, negatif, atau netral. Langkah ini memiliki peran penting dalam pelatihan model nantinya karena membantu dalam mengklasifikasikan data berdasarkan sentimennya.

2.4 Klasifikasi

Proses ini menggunakan total 793 data yang telah diproses sebelumnya, terdiri dari 596 data dengan sentimen negatif dan 197 data dengan sentimen positif. Data-data ini kemudian dibagi menjadi data training sebanyak 85% dan data testing sebanyak 15%.

Pada proses ini, dilakukan dua skenario berbeda. Skenario pertama tidak melibatkan penyeimbangan data menggunakan SMOTE pada tahap awalnya. Sedangkan skenario kedua melibatkan penyeimbangan data menggunakan SMOTE. Kedua skenario ini bertujuan untuk membandingkan tingkat performa model yang menggunakan SMOTE dengan model yang tidak menggunakan SMOTE.

Pada skenario pertama, proses klasifikasi dimulai tanpa mengaplikasikan oversampling. Data yang sudah diproses sebelumnya digunakan sebagaimana adanya untuk melatih model, dan langkah selanjutnya adalah mencari nilai optimal untuk parameter k dengan menguji rentang nilai k dari 1 hingga 10. Nilai k terbaik kemudian diterapkan pada model KNN.

Pada skenario kedua, langkah pertama adalah mengaplikasikan oversampling pada data menggunakan metode SMOTE. Langkah ini bertujuan untuk meningkatkan representasi kelas sentimen minoritas, mengoreksi ketidakseimbangan dalam dataset, dan meningkatkan performa model dalam mengidentifikasi kelas yang kurang banyak terwakili. Setelah penyeimbangan data, dilakukan pencarian nilai optimal untuk parameter k dengan menguji rentang nilai k dari 1 hingga 10. Nilai k terbaik kemudian diterapkan pada model KNN.

Algoritma KNN dimulai dengan memilih sebuah titik atau observasi dari data training yang telah diberi label. Selanjutnya, KNN mencari K tetangga terdekat dari titik yang dipilih, di mana K merupakan nilai yang telah ditentukan sebelumnya. Proses ini melibatkan perhitungan jarak antara titik yang dipilih dengan K

tetangga terdekat menggunakan matrik jarak seperti Euclidean distance atau Manhattan distance. Label kelas yang paling sering muncul dari K tetangga terdekat akan dijadikan sebagai label kelas untuk titik yang dipilih.

Model KNN dilatih dengan menggunakan data training untuk menentukan nilai optimal K dan membangun model klasifikasi. Setelah model dibangun, data testing digunakan untuk menguji akurasi model. Dengan demikian, model yang telah dibuat dapat digunakan untuk mengklasifikasikan data baru yang belum pernah dilihat sebelumnya[9].

2.5 Evaluasi

Confusion matrix adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [10] . *Confusion matrix* digunakan untuk mengevaluasi kinerja algoritma klasifikasi dalam analisis sentimen. *Confusion Matrix* merupakan alat serta metode yang digunakan untuk mengevaluasi visual dalam konsep *Machine Learning*. Kolom pada *Confusion Matrix* menggambarkan hasil kelas prediksi, serta baris menggambarkan hasil kelas yang sebenarnya [11]. *Confusion matrix* menjadi sumber informasi apakah model yang digunakan memiliki performa yang baik atau tidak, yang dapat dilihat dari angka-angka di dalamnya. *Matrix confusion* terdiri dari empat bagian: TP (*True Positive)*; TN (*True Negative*); FP (*False Positive)*; dan FN (*False Negative*). Angka pada variabel TP (True Positif) dan variabel TN (True Negative) merepresentasikan total prediksi benar oleh model, sementara angka pada variabel FP (False Positive) dan variabel FN (False Negative) merepresentasikan total prediksi salah. Performa model dihitung dengan menghitung nilai accuracy, precision, recall, dan F1-Score, menggunakan rumus yang dapat dilihat pada persamaan (1), (2), (3), dan (4).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precission = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2\ x\ Precission\ x\ Recall}{Precission + Recall} \quad (4)$$

TP adalah jumlah data kelas positif yang diprediksi dengan benar sebagai kelas positif .

FN adalah jumlah data kelas positif yang diprediksi secara salah sebagai kelas negatif .

TN adalah jumlah data kelas negatif yang diprediksi dengan benar sebagai kelas negatif.

FP adalah jumlah data kelas negatif yang diprediksi secara salah sebagai kelas positif.

3. Hasil dan Pembahasan

3.1 Hasil

3.1.1 Preprocessing Data

Preprocessing yang dilakukan pada data khususnya pada fitur “description” menghasilkan data sebagai berikut:

Tabel 1. Data hasil preprocessing data

No	Sebelum	Sesudah
1	Tenang aja.. ini ciri2 negara masuk Piala Dunia.	tenang aja ciri2 negara masuk piala dunia
2	Highlight 20 menit gol ke 2 kagak ditayangin lawak倮藁	highlight 20 menit gol 2 kagak ditayangin lawak
3	Kalian Garuda muda hebat. Ayo kejar cita" Kalian. 鉠ь潳鉠ь潳鉠	kalian garuda muda hebat ayo kejar cita kalian
4	Uda gak heran kalah , kenapa ,,? Karna timnas indonesia kalau uda masuk semi final apa lagi pas final , mendadak jadi amnesia semua pemainnya , lupa skillnya 倮藁倮藁倮藁倮藁	udah gak heran kalah karna timnas indonesia kalau udah masuk semi final apa pas final dadak jadi amnesia semua main lupa skillnya

3.1.2 Labeling Data

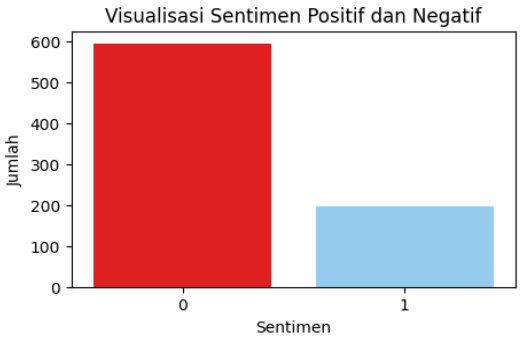
Proses labeling sentimen pada data yang juga sudah melalui proses *preprocessing* adalah sebagai berikut:

Tabel 2. Data yang sudah dilabeli

id	description	sentimen
1	tenang aja ciri2 negara masuk piala dunia	positif
2	highlight 20 menit gol 2 kagak ditayangin lawak	negatif
3	kalian garuda muda hebat ayo kejar cita kalian	positif
4	udah gak heran kalah karna timnas indonesia kalau udah masuk semi final apa pas final dadak jadi amnesia semua main lupa skillnya	negatif

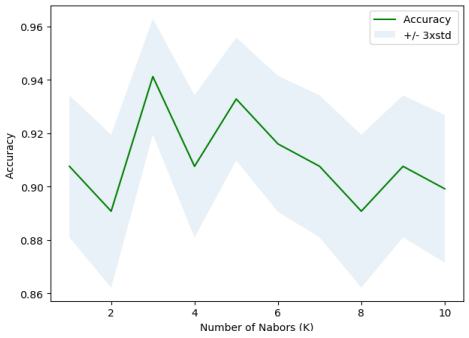
3.1.3 Klasifikasi dan Evaluasi

Pada skenario pertama, yang dilakukan tanpa menggunakan SMOTE, terlihat bahwa perbandingan jumlah data positif dan negatif cukup jauh berbeda.



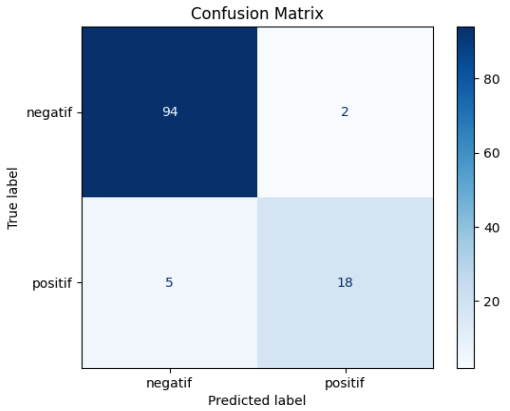
Gambar 2. Perbandingan data tanpa menggunakan SMOTE

Kemudian, pencarian nilai k yang paling optimal dengan cara menguji nilai k dari 1 sampai 10 menggunakan data tersebut.



Gambar 3. Hasil pencarian nilai k terbaik skenario pertama

Hasilnya menunjukkan bahwa k = 3 memiliki akurasi tertinggi hingga mencapai 0.94, menunjukkan kinerja yang baik dalam mengklasifikasikan data. Kemudian, nilai confusion matrix dari k = 3 dapat dilihat pada gambar 5.



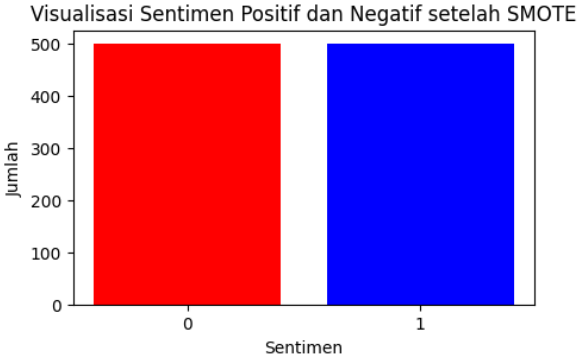
Gambar 4. Confusion matrix skenario pertama dengan k = 3

Berdasarkan *confusion matrix* di atas, dapat dihitung nilai *accuracy, precision, recall,* dan *f1-score*.

Tabel 3. Tabel evaluasi skenario pertama

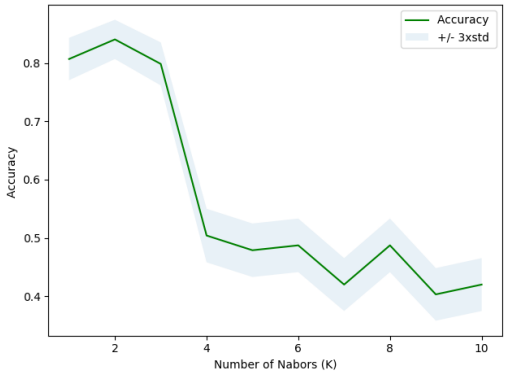
	Negatif	Positif	Rata-rata
Precision	0.95	0.90	0.92
Recall	0.98	0.78	0.88
F1-Score	0.96	0.84	0.90
Accuracy		0.94	

Selanjutnya, pada skenario kedua, pengembangan model KNN dimulai dengan melakukan oversampling pada data menggunakan SMOTE agar meningkatkan representasi minoritas kelas sentimen, memperbaiki keseimbangan dataset, dan meningkatkan kinerja model dalam mengenali kelas yang kurang banyak diwakili.



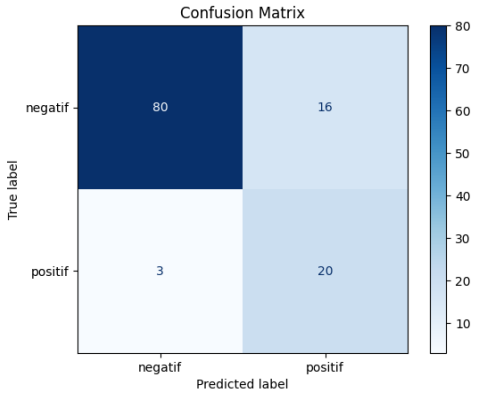
Gambar 5. Data hasil oversampling menggunakan SMOTE

Kemudian, nilai k dari 1 sampai 10 dilakukan pengujian kembali untuk mencari k yang paling optimal.



Gambar 6. Hasil pencarian nilai k terbaik skenario kedua

Didapatkan k = 2 yang memiliki akurasi tertinggi hingga mencapai 0.84. Nilai confusion matrix-nya dapat dilihat pada gambar 5.



Gambar 7. Confusion matrix skenario pertama dengan k = 2

Didapatkan nilai accuracy, precision, recall, dan f1-score berdasarkan confusion matrix di atas.

Tabel 4. Tabel evaluasi skenario kedua

	Negatif	Positif	Rata-rata
Precision	0.96	0.56	0.76
Recall	0.83	0.87	0.85
F1-Score	0.89	0.68	0.79
Accuracy		0.84	

3.2 Pembahasan

Berdasarkan hasil klasifikasi dan evaluasi yang dilakukan dengan dua skenario, maka didapatkan perbandingan performa antara kedua model.

Tabel 5. Perbandingan accuracy, precision, recall, dan f1-score

	NON-SMOTE	SMOTE
Accuracy	94%	84%
Precision	92%	76%
Recall	88%	85%
F1-Score	90%	79%

Tabel 5 menunjukkan bahwa model mengalami penurunan performa setelah penerapan SMOTE. Penurunan tersebut tercermin dari berkurangnya accuracy sebesar 10%, precision sebesar 16%, recall sebesar 3%, dan F1-Score sebesar 11%. Ini menunjukkan bahwa dalam kasus ini, penyeimbangan data menggunakan SMOTE justru menurunkan performa model.

Didapatkan juga visualisasi antara sentimen positif dan negatif masyarakat terhadap pertandingan Indonesia vs Uzbekistan pada Piala Asia U-23 menggunakan *library wordcloud* pada video youtube yang berjudul "HIGHLIGHT! Indonesia (0) vs (2) Uzbekistan | AFC U23 ASIAN CUP QATAR 2024".

Gambar 8. Confusion matrix hasil KNN dengan k = 2

Gambar 9. Confusion matrix hasil KNN dengan k = 2

Secara umum, mayoritas komentar negatif dari masyarakat berfokus pada keputusan yang dibuat oleh wasit yang dianggap salah. Sementara itu, komentar positif lebih banyak berisi dukungan dan harapan untuk timnas Indonesia di masa mendatang.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, didapatkan kesimpulan bahwa penerapan SMOTE pada algoritma K-Nearest Neighbor (K-NN) untuk klasifikasi sentimen terhadap pertandingan Indonesia vs Uzbekistan pada Piala Asia U-23 menyebabkan penurunan performa model. Terjadi penurunan dalam accuracy sebesar 10%, precision sebesar 16%, recall sebesar 3%, dan F1-Score sebesar 11%. Dari analisis ini juga terungkap bahwa sentimen masyarakat terhadap hasil pertandingan cenderung negatif.

References

[1] A. N. Kasanah, Muladi, dan U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN", *RESTI*, Vol. 3 No. 2(2019) 196 - 201, Augustus 2019. http://jurnal.iaii.or.id/index.php/RESTI/article/view/945

[2] S. D. Prasetyo, S S Hilabi, dan F Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN", *KomtekInfo*, Vol. 10 No. 1 Hal: 1-7, Maret. 2023. https://jkomtekinfo.org/ojs/index.php/komtekinfo/article/view/330

[3] H, Sulastri , A, I, Gufroni, "PENERAPAN DATA MINING DALAM PENGELOMPOKAN PENDERITA THALASSAEMIA" , *Jurnal Nasional Teknologi Dan Sistem Informasi*, VOL. 03 NO. 02 (2017) 299-305. https://teknosi.fti.unand.ac.id/index.php/teknosi/article/view/335

[4] F. S. Jumeilah , "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian", *RESTI*, Vol. 1 No. 1 (2017) 19 – 25. http://www.jurnal.iaii.or.id/index.php/RESTI/article/view/11.

[5] J. J. A. Limbong ,"Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor",*UKSW*, 2022 https://repository.uksw.edu/handle/123456789/23206

[6] A, Guswandri, R, P, Cahyono, "PENERAPAN SENTIMEN ANALIS MENGGUNAKAN METODE NAÏVE BAYES DAN SVM",*Jurnal Ilmu Data* ,Volume 2 (12), 2022 http://ilmudata.org/index.php/ilmudata/article/view/295

[7] A, Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier ", *SNATI*, 2014

https://journal.uii.ac.id/Snati/article/view/3284

[8] A, A, Magriyanti, "ANALISIS PENGEMBANGAN ALGORITMA PORTER STEMMING DALAM BAHASA INDONESIA ", *OSF*, 2018 https://osf.io/preprints/inarxiv/7ge4v/

[9] I. H. Kusuma dan N Cahyono, "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor", *JPIT*, Vol.8, No.3 Hal: 1-6, September. 2023. http://ejournal.poltekharber.ac.id/index.php/informatika/article/view/5734

[10] D Normawati, S A Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter" , *J-SAKTI*, Volume 5 Nomor 2, September 2021, pp. 697-711 http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

[11] C. A. Pamungkas dan W. W. Widiyanto, "KLASIFIKASI INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022 DENGAN SUPPORT VECTOR MACHINE", *SINOV*, Vol 2 No. 3 , pp. 1 - 7, November. 2022. https://journal.sinov.id/index.php/juisik/article/view/407

