COMPARISON OF SVM, NAÏVE BAYES, AND LOGISTIC REGRESSION ALGORITHMS FOR SENTIMENT ANALYSIS OF FRAUD AND BOTS IN PURCHASING CONCERT TICKET

KOMPARASI ALGORITMA SVM, NAÏVE BAYES, DAN LOGISTIC REGRESSION UNTUK ANALISIS SENTIME PENIPUAN DAN BOT DALAM PEMBELIAN TIKET KONSER

Vania Agresia[1], Ryan Randy Suryono[2]
[1,2]Universitas Teknokrat Indonesia
Jl. ZA. Pagar Alam No. 9-11, Labuan Ratu, Kec. Kedaton, Kota Bandar Lampung, Indonesia
Email: vania_agresia@teknokrat.ac.id[1], ryan@teknokrat.ac.id[2]

Abstract *- Music concerts are highly anticipated entertainment events, but they are often subject to fraud and the use of bots in online ticket purchases, to the detriment of fans and organisers. Fans may lose confidence in the ticket system and reduce interest in the event. For organizers, it can reduce the event's reputation and finances. This research aims to analyse public sentiment regarding this issue by comparing three classification algorithms: Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. Data taken from Twitter which contains comments related to fraud and bots. The methods used include data crawling, preprocessing, sentiment labelling, and model evaluation. Preprocessing includes data cleaning, case folding, tokenising, stopwords, and stemming. Sentiment labelling is done manually or by human annotators. The results showed that SVM had the best accuracy of 91.27%, followed by Logistic Regression (90.03%) and Naïve Bayes (77.70%). Applying SMOTE to overcome class imbalance and improve the performance of negative sentiment models. This research emphasizes the importance of choosing the right algorithm and using SMOTE to improve the accuracy of sentiment analysis regarding fraud and bots in concert ticket purchases. The research results can be applied to improve bot usage detection systems and provide insight for organizers.*

Keywords *- sentiment analysis, bots, music concerts, fraud.*

Abstrak - Konser musik adalah acara hiburan yang dinantikan, namun sering kali menjadi sasaran penipuan dan penggunaan bot dalam pembelian tiket online, merugikan penggemar dan penyelenggara. Penggemar dapat kehilangan kepercayaan pada sistem tiket dan mengurangi minat pada acara. Bagi penyelenggara dapat menurunkan reputasi acara dan keuangan. Penelitian ini bertujuan menganalisis sentimen publik terkait isu tersebut dengan membandingkan tiga algoritma klasifikasi: Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression. Data diambil dari Twitter yang berisi komentar terkait penipuan dan bot. Metode yang digunakan meliputi crawling data, preprocessing, pelabelan sentimen, dan evaluasi model. Preprocessing mencakup data cleaning, case folding, tokenizing, stopword, dan stemming. Pelabelan sentimen dilakukan secara manual atau human annotators. Hasil penelitian menunjukkan SVM memiliki akurasi terbaik 91,27%, diikuti oleh Logistic Regression (90,03%) dan Naïve Bayes (77,70%). Penerapan SMOTE untuk mengatasi ketidakseimbangan kelas dann meningkatkan performa model sentimen negatif. Penelitian ini menekankan pentingnya pemilihan algoritma yang tepat dan penggunaan SMOTE untuk meningkatkan akurasi analisis sentimen terkait penipuan dan bot dalam pembelian tiket konser. Hasil penelitian dapat diaplikasikan untuk meningkatkan sistem deteksi penggunaan bot dan memberikan wawasan bagi penyelenggara.

Kata Kunci - analisis sentimen, bot, konser musik, penipuan.

I. PENDAHULUAN

Konser musik merupakan sebuah acara atau *event* yang menghadirkan musisi atau artis yang menunjukkan karyanya dihadapan para penonton. Konser musik merupakan *event* yang masuk ke kategori spesial *event* [1]. Konser-konser memiliki dampak besar bagi penggemar dan masyarakat Indonesia. Konser musik ialah satu dari kegiatan hiburan yang sering ditunggu. Konser musik dapat menjadi pengalaman mendalam bagi penggemar dan memberikan waktu untuk menyaksikan secara langsung serta menikmati rasa pada suasana konser [2]. Dalam era modernisasi perkembangan teknologi yang sangat pesat dalam pembelian tiket konser secara online. Pembelian tiket konser membuat para penggemar antusias untuk mendapatkan tiket akses ke acara musik favorit. Namun, seiring dengan popularitas acara konser sistem penjualan tiket online menjadi sasaran bagi para penipu dan bot. Penipuan dan penggunaan bot menjadi masalah serius yang dapat merugikan penggemar yang sah dan penyelenggara acara. Penggunaan bot dilakukan untuk pembelian tiket dengan jumlah besar dengan waktu yang cepat, kemudian penipuan memanfaatkan kesempatan untuk menjual kembali tiket palsu dengan harga yang tinggi. Penjualan tiket dengan harga mahal ini menyebabkan pembeli yang benar-benar ingin menonton konser tetap membeli tiket dengan harga yang begitu mahal. Seorang calo tiket pun ikut serta pada kegiatan penipuan, menjual tiket palsu atau tidak valid pada pembeli yang tidak hati-hati [3].

Twitter merupakan salah satu platform media sosial yang memiliki menu untuk pemakainya memberi serta meminta informasi dalam bentuk teks, gambar, serta video. Tweet terdapat seperti opini atau pendapat pada sebuah kejadian yang sedang berlangsung atau sudah terjadi [4]. Dalam media sosial Twitter berbagai macam komentar dicurahkan terhadap kejadian penipuan dan penggunaan bot saat pembelian tiket konser. Banyak komentar negatif dan positif mengenai penipuan dan pengunaan bot. dari banyaknya komentar dari pemakai Twitter, lalu komentar ini bisa diamati memakai cara analisis sentimen. Analisis sentimen ialah bagian dari pengklasifikasian teks yang berkaitan dengan pengolahan bahasa alami, linguistik komputasional, serta penambangan teks. Tujuan dari analisis ini ialah untuk memahami opini, penilaian, dan emosi tentan suatu topik [5]. Dalam hal ini, analisis sentimen menjadi metode paling efektif untuk mengevaluasi opini publik, seperti ulasan suatu peristiwa, layanan, atau produk tertentu.

Penelitian ini bertujuan untuk melakukan komparasi antara ketiga algoritma yaitu SVM, Naïve Bayes, dan Logistic Regression, untuk meningkatkan akurasi dalam analisis sentimen terkait penipuan dan pengunaan bot dalam pembelian tiket konser. Dengan melakukan komparasi dari berbagai metode dapat ditemukan metode klasifikasi yang paling efektif [6]. Setiap algortima memiliki kekuatan dan kelemahannya masing-masing dalam mengklasifikasikan sentimen positif, negatif, atau netral [7]. Hasil penelitian ini diharapkan bisa membantu penyelenggara memantau keluhan pengguna terkait penipuan dan penggunaan bot dalam pembelian tiket dan dapat memperbaiki sistem penjualan agar tidak dapat kecurangan saat pembelian tiket.

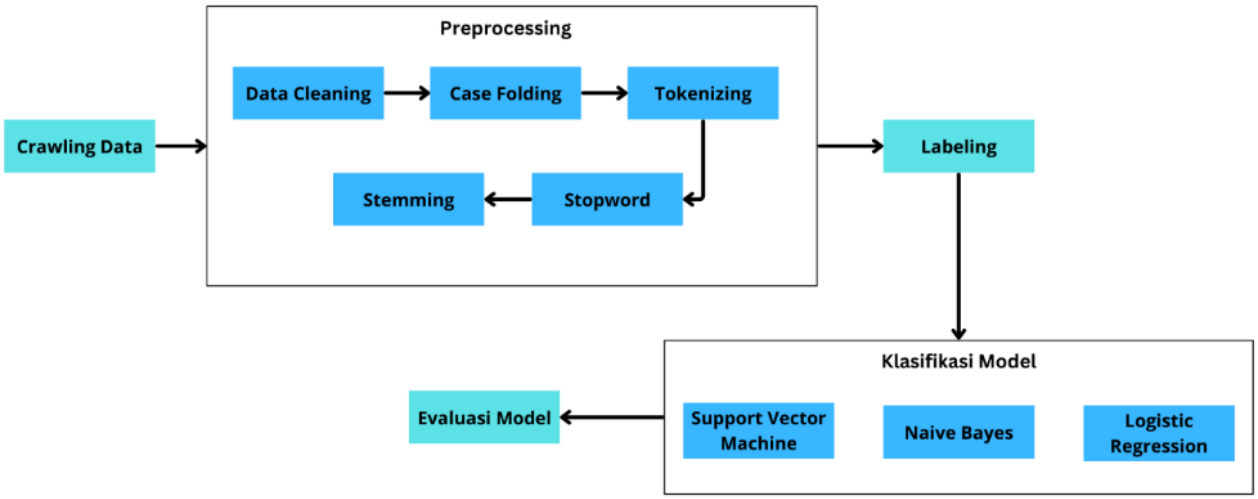
II. SIGNIFIKASI STUDI

A. Penelitian Terdahulu

Sebelumnya sudah dilaksanakan penelitian analisis sentimen memakai naïve bayes pada komentar netizen terhadap pembubaran konser NCT 127. Hasil penelitian ini mendapatkan akurasi sebanyak 82,01%, precision sebanyak 75,21%, recall sebanyak 95,50%, serta f1-score sebanyak 84,15%. Lalu penelitian berikutnya sentimen pada analisis tingkah laku fans coldplay memakai naïve bayes mendapat jumlah sebanyak 80,25%. Selanjutnya penelitian dari komparasi algoritma K-NN, naïve bayes, serta SVM untuk prediksi kelulusan mahasiswa tingkat akhir yang mendapatkan akurasi algoritma K-NN 87,8%, naïve bayes 82,6%, SVM 73,6. Yang di mana pada penelitian ini membandingkan tiga algoritma dengan hasil bahwa algoritma K-NN lebih tinggi (baik) dibandingkan dengan naïve bayes serta SVM [8]. Selanjutnya penelitian analisis senntimen evaluasi pembelajaran tatap muka dengan metode logistic regression mendapatkan jumlah sebanyak 78,57%, precision 76,92%, serta f1-score 80% [9]. Kemudian penelitian berikutnya dengan judul komparasi algoritma SVM serta naïve bayes untuk analisis sentimen dalam metaverse mendapatkan akurasi 90,3% dann algoritma naïve bayes mendapatkan akurasi 84,23%. Artinya pada penelitian analisis sentimen ini pada algortima SVM lebih tingi (baik) dari pemakaian algoritma naïve bayes [10]. Dari beberapa penelitian tersebut menunjukkan SVM unggul dalam mengklasifikasikan data tinggi, namun mengalami kesulitan saat dataset yang tidak seimbang, sehingga memerlukan teknik SMOTE yang menjadi kelemahan pada penelitian tersebut. Kemudian metode Naïve Bayes juga digunakan karena kesederhanaannya, namun penelitian sebelumnya dataset yang didapatkan lebih sedikit dan juga kurang efektif karena mengalami data yang tidak seimbang. Penemuan ini dapat digunakan penyelenggara dan penyedia platform penjualan tiket untuk memantau keluhan pengemar secara efektif. Dengan analisis sentimen mereka dapat memperbaiki sistem penjualan untuk mendeteksi dan mencegah kecurangan yang berguna untuk memastikan penjualan yang lebih adil.

B. Metode Penelitian

Metode yang diterapkan menggunakan tiga algortima Support Vector Machine (SVM), Naïve Bayes, serta Logistic Regression. Pada penelitian ini ditunjukkan pada gambar 1 metode yang dilaksanakan dengan tahapan *crawling data* maupun pengumpulan data dari platform twitter terkait penipuan serta penggunaan bot dalam pembelian tiket konser, preprocessing data untuk pembersihan data yang sudah dikumpulkan, labeling untuk memberi label pada setiap data atau entitas, pembobotan setiap kata, dan evaluasi model dengan menguji metode SVM, Naïve Bayes, serta Logistic Regression untuk analisis sentimen.



Gambar 1. Tahapan Penelitian

1. Crawling Data

Crawling ialah sebuah metode yang dipakai untuk mendapatkan informasi yang terdapat pada web [11]. Pada proses *crawling data* adalah tahap awal untuk melakukan penelitian ini. Pengumpulan data atau *crawling data* diperoleh dari media sosial twitter mengunakan keyword yang berkaitan dengan topik penelitian. Melakukan proses *crawling data* mennggunakan bantuan *auth token*agar dapat mengakses dan mengambil data dari platform twitter [12].

2. Preprocessing

Preprocessing data adalah mempersiapkan data mentah untuk melakukan persiapan data yang akan digunakan untuk analisis lebih lanjut dan dilakukan pemodelan [13]. Pada tahap ini akan dilakukan 5 tahapan, yaitu :

A. Data Cleaning

Membersihkan teks adalah menjadi tahap awal pada analisis teks yang bertujuan untuk membersihkan kata yang tidak berkaitan seperti hashtag, tanda kutip, emoticon, karakter atau simbol.

TABEL I.
DATA CLEANING

Tweet	***Data Cleaning***
[17s] sedih bgt caratland Indo. mau war tiket sendiri kalah sama bot. mau beli yg di calo banyak yang scam. aku yg kemaren war sendiri sumpah sih sakit hati bgt dan doa orang tersakiti tuh dikabulkan semoga akan ada karmanya deh yg pake bot dan scam tiket	sedih bgt caratland Indo. mau war tiket sendiri kalah sama bot. mau beli yg di calo banyak yang scam. aku yg kemaren war sendiri sumpah sih sakit hati bgt dan doa orang tersakiti tuh dikabulkan semoga akan ada karmanya deh yg pake bot dan scam tiket

B. Case Folding

Case folding merupakan proses yang mengubah data tweet menjadi huruf kecil atau lowercase [14]. Tujuan dilakukan case folding untuk memastikan konsistensi dalam penggolaan data.

TABEL II.
CASE FOLDING

Tweet	*Case Folding*
sedih bgt caratland Indo. mau war tiket sendiri kalah sama bot. mau beli yg di calo banyak yang scam. aku yg kemaren war sendiri sumpah sih sakit hati bgt dan doa orang tersakiti tuh dikabulkan semoga akan ada karmanya deh yg pake bot dan scam tiket	sedih banget caratland indo mau perang tiket sendiri kalah sama bot mau beli yang di makelar banyak yang scam aku yang kemarin perang sendiri sumpah sih sakit hati banget dan doa orang tersakiti itu dikabulkan semoga akan ada karmanya deh yang pakai bot dan scam tiket

C. Tokenizing

Tokenizing adalah proses untuk pemecahann kalimat menjadi unit-unit lebih kecil, yang disebut token, seperti kata, frasa atau karakter [15].

Tabel III.
TOKENIZING

Tweet	*Tokenizing*
sedih banget caratland indo mau perang tiket sendiri kalah sama bot mau beli yang di makelar banyak yang scam aku yang kemarin perang sendiri sumpah sih sakit hati banget dan doa orang tersakiti itu dikabulkan semoga akan ada karmanya deh yang pakai bot dan scam tiket	['sedih', 'banget', 'caratland', 'indo', 'mau', 'perang', 'tiket', 'sendiri', 'kalah', 'sama', 'bot', 'mau', 'beli', 'yang', 'di', 'makelar', 'banyak', 'yang', 'scam', 'aku', 'yang', 'kemarin', 'perang', 'sendiri', 'sumpah', 'sih', 'sakit', 'hati', 'banget', 'dan', 'doa', 'orang', 'tersakiti', 'itu', 'dikabulkan', 'semoga', 'akan', 'ada', 'karmanya', 'deh', 'yang', 'pakai', 'bot', 'dan', 'scam', 'tiket']

D. Stopword

Stopword adalah tahap menghilangkan kata-kata yang diangap tidak sesuai atau diangap kurang mempunyai arti penting yang dapat mempengaruhi dalam analisis sentimen [16]. dengan menghapus stopword model analisis sentimen akan lebih akurat.

TABEL IV.
STOPWORD

Tweet	*Stopword*
['sedih', 'banget', 'caratland', 'indo', 'mau', 'perang', 'tiket', 'sendiri', 'kalah', 'sama', 'bot', 'mau', 'beli', 'yang', 'di', 'makelar', 'banyak', 'yang', 'scam', 'aku', 'yang', 'kemarin', 'perang', 'sendiri', 'sumpah', 'sih', 'sakit', 'hati', 'banget', 'dan', 'doa', 'orang', 'tersakiti', 'itu', 'dikabulkan', 'semoga', 'akan', 'ada', 'karmanya', 'deh', 'yang', 'pakai', 'bot', 'dan', 'scam', 'tiket']	['sedih', 'banget', 'caratland', 'indo', 'mau', 'perang', 'tiket', 'sendiri', 'kalah', 'sama', 'bot', 'mau', 'beli', 'makelar', 'banyak', 'scam', 'aku', 'kemarin', 'perang', 'sendiri', 'sumpah', 'sih', 'sakit', 'hati', 'banget', 'doa', 'tersakiti', 'dikabulkan', 'semoga', 'karmanya', 'deh', 'pakai', 'bot', 'scam', 'tiket']

E. Stemming

Stemming adalah proses yang bertujuan untuk menubah kata menjadi bentuk dasar, seperti menghilangkan imbuhan. Tujuan lainnya yaitu menurani variasi kata yang memiliki arti sama.

TABEL V.
STEMMING

Tweet	*Stemming*
['sedih', 'banget', 'caratland', 'indo', 'mau', 'perang', 'tiket', 'sendiri', 'kalah', 'sama', 'bot', 'mau', 'beli', 'yang', 'di', 'makelar', 'banyak', 'yang', 'scam', 'aku', 'yang', 'kemarin', 'perang', 'sendiri', 'sumpah', 'sih', 'sakit', 'hati', 'banget', 'dan', 'doa', 'orang', 'tersakiti', 'itu', 'dikabulkan', 'semoga', 'akan', 'ada', 'karmanya', 'deh', 'yang', 'pakai', 'bot', 'dan', 'scam', 'tiket']	['sedih', 'banget', 'caratland', 'indo', 'mau', 'perang', 'tiket', 'sendiri', 'kalah', 'sama', 'bot', 'mau', 'beli', 'makelar', 'banyak', 'scam', 'aku', 'kemarin', 'perang', 'sendiri', 'sumpah', 'sih', 'sakit', 'hati', 'banget', 'doa', 'sakit', 'kabul', 'moga', 'karma', 'deh', 'pakai', 'bot', 'scam', 'tiket']

3. Labeling

Labeling atau pelabelan dalam analisis ini dilakukan secara manual atau *human annotators* di mana data yang telah dikumpulkann diberi label berdasarkan sentimen, yaitu positif, negatif, dan netral [17].

4. Klasifikasi Model

Klasifikasi model merupakan proses analisis data yang bertujuan untuk menevaluasi tingkat akurasi dari berbagai algortima. Kemudian hasil dari algortima yang sudah diterapkan akan dibandingkan. Pada penelitian ini menerapkan klasifikasi model *Support Vector Machine* (SVM), Naïve Bayes, dan Logistic Regression.

A. Support Vector Machine

Support Vector Machine termasuk algortima machine learning yang dapat dipakai pada kasus klasifikasi dan regresi. SVM adalah algoritma pembelajaran mesin yang kuat dalam klasifikasi. SVM adalah satu dari teknik yang seringkali dipakai untuk mendeteksi polaritas data tekstual [18]. Berikut rumusnya :

$$f(x) = sign(w^T + b) \qquad (1)$$

B. Naïve Bayes

Naïve bayes ialah satu dari model klasifikasi yang paling banyak dipakai pada tahap text mining. Naïve bayes disesuaikan dengan Teorema Bayes yang menggambarkan probabilitas suatu kelas berdasarkan data yang diberikan [19].

$$P(X|Y) = \frac{P(Y|X) \times P(X)}{P(Y)} \qquad (2)$$

C. Logistic Regression

Logistic Regression merupakan teknik statistik yang diterapkan untuk menganalisis kaitan antara variabel independen serta variabel dependen yang bersifat biner atau hanya memiliki dua kategori [20].

5. SMOTE

SMOTE *(Synthetic Minority Oversampling Technique)* adalah teknik untuk mengolah ketidakseimbangan kelas dalam kumpulan data. Yang mana jumlah data dalam satu kelompok lebih banyak daripada dengan kelompok lainnya. Ketidakseimbangan bisa menyebabkan bentuk lebih condong memprediksi kelas mayoritas. Penerapan SMOTE juga efektif dalam menangkap pola penting dari kelas minoritas dan mendapatkan hasil yang lebih akurat.

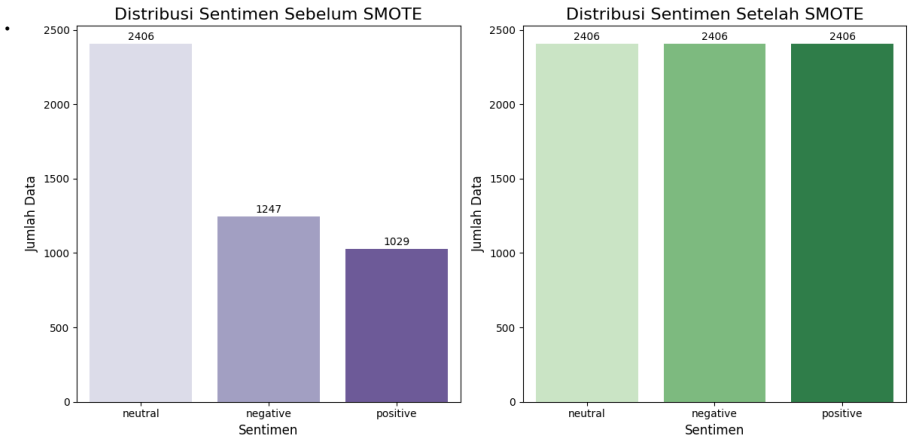
6. Evaluasi Model

Evaluasi model merupakan proses penting dalam analisis sentimen untuk menilai akurasi dan efektivitas hasil prediksi sentimen yang dihasilkan oleh model yang digunakan. Proses ini menghasilkan berbagai metrik berupa akurasi, presisi, recall, serta F1 score [21].

III. HASIL DAN PEMBAHASAN

C. Dataset

Dari data set dalam penelitian ini memperoleh total 4684 tweet yang berkaitan dengan topik yang dianalisis. Data yang digunakan adalah data yang diperoleh dari media sosial Twitter. Dari data yang sudah diperoleh pada proses pelabelan menunjukan jumlah sentimen positif sebanyak 1029, kemudian sentimen negatif sebanyak 1247, serta sentimen netral berjumlah 2406. Ketidakseimbangan antara tiga kelas tersebut diatasi melalui optimasi metode SMOTE untuk menyeimbangkan jumlah data antara positif, negatif, dan netral. Keseimbangan ini membantu model dalam mempelajari karakteristik dari setiap kelas secara lebih seimbang. Penerapan sebelum SMOTE dan setelah SMOTE bisa dilihat dalam gambar 2.



Gambar. 2 Perbandigan SMOTE

D. Tahap Pengujian

Dalam tahap pengujian menngevaluasi tiga algortima klasifikasi yaitu, SVM, Naïve Bayes, dan Logistic Regession. Dari tiga algoritma yang sudah diuji menunjukkan bahwa SVM memberikan jumlah paling tinggi daripada dengan Naïve Bayes serta Logitic Regression. Proses ini memandingkan pembagian data sebanyak 80% untuk training serta 20% untuk testing. SVM memberikan akurasi 91,27%, Naïve Bayes dengan jumlah 77,70%, dan Logistic Regression memberikan akurasi 90,03%. Berdasarkan model algoritma yang diterapkan, model ini menggunakan data yang telah dioptimalkan dengan metode SMOTE, serta membandingkan dengan data yang belum dioptimalkan. Hasil dari model algoritma bisa dilihat dalam tabel 6 dan 7. Karakteristik SVM memiliki keunggulan dibanding algoritma lain pada penelitian ini. Penggunaan kernel pada SVM memberikan fleksibilitas untuk menangani pola non-linear yang sering ditemukan pada data sentimen. Hal ini membuat SVM lebih unggul dalam medeteksi polaritas sentimen dengan akurasi yang tinggi. Dibandingkan dengan Naïve Bayes dan Logistic Regression lebih sederhana dan efisien, tetapi terbatas dalam akurasi dengan data yang lebih kompleks.

TABEL VI.
MODELLING SEBELUM SMOTE

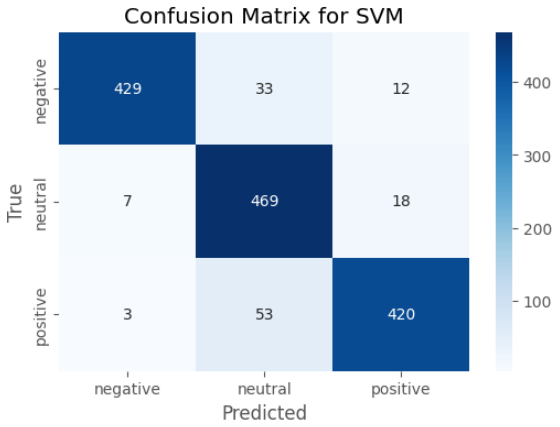
Modelling	***Sentimen***	***Precision***	***Recall***	***F1-Score***
SVM	Positif	0.86	0.81	0.84
	Negatif	0.94	0.81	0.87
	Netral	0.88	0.96	0.92
Naïve Bayes	Positif	0.86	0.81	0.84
	Negatif	0.94	0.81	0.87
	Netral	0.88	0.96	0.92
Logistic Regression	Positif	0.90	0.80	0.84
	Negatif	0.96	0.81	0.88
	Netral	0.86	0.98	0.92

Setelah perbandingan SMOTE algoritma SVM menunjukkan peningkatan dalam skor precision, recall, serta F1-Score. Peningkatan pada sentimen positif precision menjadi 0.93, recall 0.88, F1-Score 0.91. Sentimen negatif, precision 0.98, recall 0.91, F1-Score 0.94. Dan sentimen netral precision 0.85, recall 0.95, dan F1-Score 0.89. Untuk algoritma Naive Bayes mengalami sedikit penurunan nilai. Kemudian algoritma Logistic Regresion menunjukkan peningkatan nilai precision, recall, serta F1-Score, namun dalam sentimen netral menunjukkan penurunan.

TABEL VII.
MODELLING SETELAH SMOTE

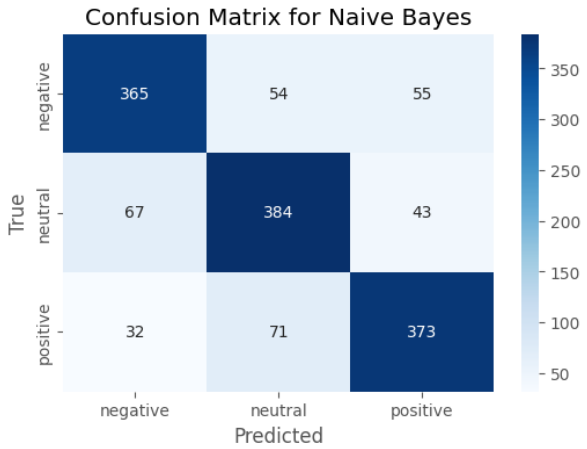
Modelling	*Sentimen*	*Precision*	*Recall*	*F1-Score*
SVM	Positif	0.93	0.88	0.91
	Negatif	0.98	0.91	0.94
	Netral	0.85	0.95	0.89
Naïve Bayes	Positif	0.79	0.78	0.79
	Negatif	0.79	0.77	0.78
	Netral	0.75	0.78	0.77
Logistic Regression	Positif	0.91	0.88	0.90
	Negatif	0.98	0.88	0.92
	Netral	0.83	0.95	0.88

Dalam menentukan model algoritma yang paling optimal, peneliti melakukan ekperimen dengan membandingkan nilai-nilai pada confusion matrix dari tiga algoritma yang digunakan. Hasil dari perbandingan setiap algoritma bisa dilihat dalam gambar 3.



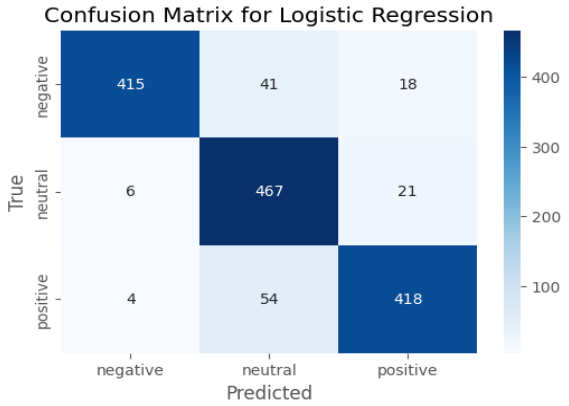
Gambar 3. Confusion Matrix SVM

Confusion matrix untuk model klasifikasi SVM menunjukkan bahwa dari total data yang diuji, model berhasil mengklasifikasi 429 teks sebagai kategori negatif dengan benar. Namun terdapat 33 teks yang termasuk ke dalam kategori positif dan 12 teks termasuk ke dalam kategori netral yang salah diklasifikasikan menjadi negatif (false positives). Untuk kategori positif, model berhasil mengidentifikasi 469 teks dengan benar. Namun terdapat 7 teks yang termasuk ke dalam kategori negatif dan 18 teks termasuk ke dalam kategori netral yang salah diklasifikasikan sebagai (false positives). Pada kategori netral, model berhasil mengklasifikasikan 420 teks dengan benar.namun terdapat 3 teks yang termasuk dalam kategori negatif dan 53 teks yang seharusnya termasuk ke dalam kategori positif yang salah diklasifikasikan sebagai netral (false positives). Hasil dari confusion matrix ini menyatakan bahwa model lebih efektif pada mengidentifikasi teks positif serta negatif dengan akurasi yang cukup baik.



Gambar 4. Cofusion Matrix Naïve Bayes

Confusion matrix untuk model klasifikasi Naive Bayes menyatakan bahwa dari total data yang diuji, model berhasil mengklasifikasikan 365 teks sebagai negatif dengan benar. Namun 54 teks yang sebenarnya termasuk ke dalam ketegori netral dan 55 teks termasuk ke dalam kategori positif yang salah diklasifikasikan sebagai negatif. Untuk kategori netral, model berhasil mengidentifikasi 384 teks dengan benar. Namun 67 teks yang sebenarnya termasuk dalam kategori negatif dan 43 teks yang sebenarnya termasuk ke dalam kategori positif yang salah diklasifikasikan sebagai netral. Pada kategori positif, model berhasil mengklasifikasikan 373 teks dengan benar. Namun terdapat 32 teks yang seharusnya termasuk dalam kategori negatif dan 71 teks yang seharusnya masuk ke dalam kategori netral yang slah diklasifikasikan sebagai positif. Hasil dari confusion matrix ini menyatakan bahwa model lebih efektif pada mengidentifikasi teks netral dan positif dengan akurasi yang cukup baik.

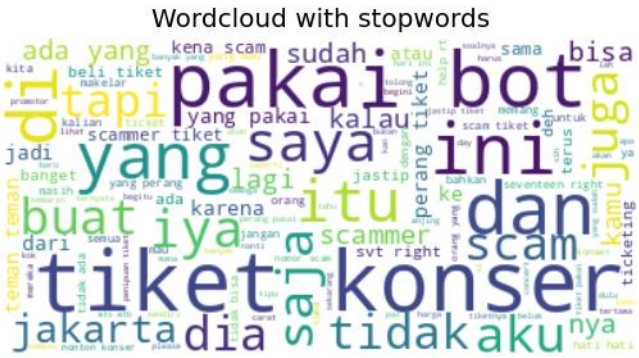


Gambar 5. Confusion Matrix Logistic Regression

Confusion matrix untuk model Logistic Regression menunjukkan bahwa dari total data yang diuji, model berhasil mengklasifikasikan 415 teks sebagai negatif dengan benar. Namun 41 teks yang sebenarnya masuk dalam kategori netral dan 18 teks yang termasuk dalam kategori positif yang salah diklasifikan sebagai negatif. Untuk kategori netral, model berhasil mengindentifikasi 467 teks dengan benar. Namun, terdapat teks yang seharusnya termasuk dalam kategori negatif dan 21 teks seharusnya termasuk dalam kategori positif yang salah diklasifikasikan sebagai netral. Pada kategori positif, model berhasil mengklasifikasi 418 teks dengan benar. Namun, terdapat 4 teks yang seharusnya dalam kategori negatif dan 54 teks yang seharusnya termasuk dalam kategori netral yang salah diklasifikasikan sebagai positif. Hasil dari confusion matrix ini menyatakan bahwa model lebih efektif pada mengidentifikasi teks netral dan positif.

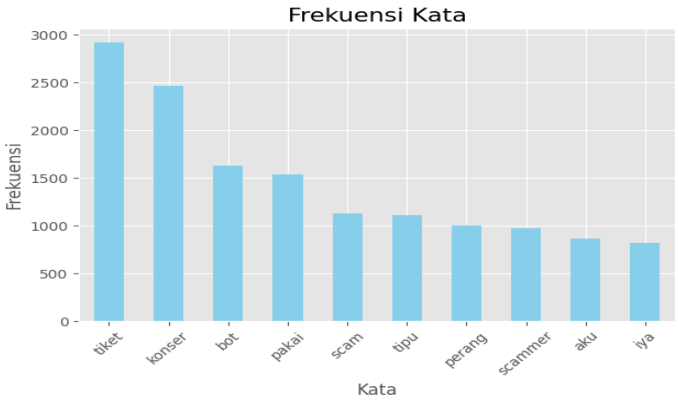
E. Visualisasi Data

Wordcloud merupakan sebuah visualisasi yang menggambarkan kumpulan kata yang paling sering muncul [22]. Visualisasi ini bertujuan untuk mempermudah identifikasi kata-kata yang paling sering digunakan. Hasil visualisasi wordcloud ditunjukkan pada Gambar. Kata-kata yang paling sering muncul aka ditampilkan dengan ukuran yang lebih besar, seperti “tiket konser”, “bot”, dan “scam” yang menunjukkan tingginya pembahasan terhadap penggunaan bot dan penipuan pada pembelian tiket konser.



Gambar 6. Visualisasi Wordcloud

Dalam analisis frekuensi kata, visualisasi wordcloud digunakan untuk menampilkan sepuluh kata yang paling sering muncul pada dataset ini ialah “tiket”, “konser”, “bot”, serta “tipu”. Hasil memunjukkan pada fokus utama pada penipuan tiket konser dan penggunan bot dalam pembelian tiket konser. Hasil ini memberikan gambaran yang jelas terkait pembahasan pada analisis ini.



Gambar 7. Frekuensi Kata

IV. KESIMPULAN

Sesuai dengan hasil penelitian pada 4685 data ditemukan bahwa memiliki sentimen positif 1029 data, sentimen negatif 1247 data, sedangkan sentimen netral 2406 data. Penelitian ini dilaksanakan komparasi algoritma SVM, Naïve Bayes, serta Logistic Regression. Penelitian ini menyatakan bahwa algoritma Support Vector Machine (SVM) memberikan performa terbaik dalam analisis sentimen terkait penipuan dan penggunaan bot dalam pembelian tiket konser, dengan akurasi yang mencapai 91,27%. SVM terbukti efektif dalam mengklasifikasikan sentimen positif, negatif, serta netral. Metode SMOTE yang diterapkan pada penilitian ini berhasil menyeimbangkan jumlah data antar kelas, yang memberikan kontribusi dalam peningkatan performa model, terutama pada kategori sentimen negatif. SVM paling diuntungkan dengan pengunaan metode SMOTE, karena dapat memanfaatkan margin yang lebih baik untuk memisahkan sentimen negatif lebih akurat. Sementara Naïve Bayes dan Logistic Regression menunjukkan peningkatan namun tidak sebesar SVM, cenderung lebih sederhana. Hal ini disebabkan oleh kesederhanaan kedua algoritma yang kurang fleksibel dalam menangani ketidakseimbangan data karena tidak menggunakan metode SMOTE. Penelitian ini menunjukkan pentingnya pemilihan algoritma yang tepat dan penerapan teknik SMOTE dalam analisis sentimen. Untuk penelitian selanjutnya, disarakan untuk mendapatkan dataset lebih banyak dan penggunaan algoritma yang lebih kompleks.

REFERENSI

[1] M. H. Raihardi and A. Parlindungan, “Jurnal Ekonomi Revolusioner PENGARUH HARGA DAN PROMOSI TERHADAP KEPUTUSAN PEMBELIAN TIKET DALAM SEBUAH EVENT (STUDI KASUS PADA EVENT KONSER PEMAIN RASA),” vol. 7, no. 6, pp. 296–302, 2024.

[2] D. F. Zahra and C. Carkiman, “Pengalaman Pelanggan Membeli Tiket Konser Coldplay: Menambang Ulasan Online Berdasarkan Pemodelan Topik Dan Analisis Sentimen,” *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 8, no. 2, p. 243, 2024, doi: 10.52362/jisamar.v8i2.1426.

[3] A. S. R. S. Andra Gustian Yamin, Muhammad Alief Surur, “Analisa Pelanggaran Etika dalam Industri Hiburan (Studi Kasus: Penipuan Calo Tiket Konser Coldplay),” vol. 10, no. 14, pp. 126–134, 2024.

[4] N. Q. Rizkina and F. N. Hasan, “Analisis Sentimen Komentar Netizen Terhadap Pembubaran Konser NCT 127 Menggunakan Metode Naive Bayes,” *J. Inf. Syst. Res.*, vol. 4, no. 4, pp. 1136–1144, 2023, doi: 10.47065/josh.v4i4.3803.

[5] F. F. Mailo and L. Lazuardi, “Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia,” *J. Inf. Syst. Public Heal.*, vol. 4, no. 1, pp. 28–36, 2019.

[6] L. Abdillah Fudholi, N. Rahaningsih, and R. Danar Dana, “Sentimen Analisis Perilaku Penggemar Coldplay Di Media Sosial Twitter Menggunakan Metode Naive Bayes,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 4150–4159, 2024, doi: 10.36040/jati.v8i3.9827.

[7] H. Purnomo, “Jurnal Pepadun Analisis Sentimen Opini Masyarakat Terhadap Penggunaan ChatGPT di Bidang Pendidikan Berbasis Twitter Jurnal Pepadun,” vol. 5, no. 3, pp. 275–285, 2024.

[8] A. Putri *et al.*, “Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 20–26, 2023, doi: 10.57152/malcom.v3i1.610.

[9] S. A. Assaidi and F. Amin, “Analisis Sentimen Evaluasi Pembelajaran Tatap Muka 100 Persen pada

Pengguna Twitter menggunakan Metode Logistic Regression," *J. Pendidik. Tambusai*, vol. 6, no. 2, pp. 13217–13227, 2022.

[10] D. N. Novianti, D. F. Shiddieq, F. F. Roji, and W. Susilawati, "Comparison of Support Vector Machine and Naïve Bayes Algorithms for Sentiment Analysis of the Metaverse," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. April, pp. 231–239, 2024.

[11] R. Maria, R. U. Umayah, S. Mahardinny, D. N. Kalana, and D. D. Saputra, "Analisis Sentimen Persepsi Masyarakat Terhadap Penggunaan Aplikasi My Pertamina Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes Classifier," *J. Komput. Antart.*, vol. 1, no. 1, pp. 1–10, 2023, [Online]. Available: https://ejournal.mediaantartika.id/index.php/jka%0Ahttps://ejournal.mediaantartika.id/index.php/jka/article/view/1%0Ahttps://ejournal.mediaantartika.id/index.php/jka/article/download/1/1

[12] A. Wibowo, Firman Noor Hasan, Rika Nurhayati, and Arief Wibowo, "Analisis Sentimen Opini Masyarakat Terhadap Keefektifan Pembelajaran Daring Selama Pandemi COVID-19 Menggunakan Naïve Bayes Classifier," *J. Asiimetrik J. Ilm. Rekayasa Inov.*, vol. 4, pp. 239–248, 2022, doi: 10.35814/asiimetrik.v4i1.3577.

[13] S. Chohan, A. Nugroho, A. M. B. Aji, and W. Gata, "Analisis Sentimen Pengguna Aplikasi Duolingo Menggunakan Metode Naïve Bayes dan Synthetic Minority Over Sampling Technique," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 2, pp. 139–144, 2020, doi: 10.31294/p.v22i2.8251.

[14] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Comparison of Naïve Bayes and Support Vector Machine Methods in Twitter Sentiment Analysis," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020.

[15] M. A. Yaqin and M. Faid, "Perbandingan Algoritma Klasifikasi untuk Prediksi Sentimen Emoji dalam Teks," vol. 01, no. 01, pp. 18–25, 2024.

[16] D. N. Herisnan and M. Elwinda, "Analisis sentimen terhadap resesi ekonomi global di indonesia menggunakan hybrid linear regression – naive bayes sentiment analysis of global economic recession in indonesia using hybrid linear regression – naive bayes," vol. 7, pp. 1495–1501, 2024.

[17] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 785–795, 2022, doi: 10.35957/jatisi.v9i2.1835.

[18] A. N. Syafia, M. F. Hidayattullah, and W. Suteddy, "Studi Komparasi Algoritma SVM Dan Random Forest Pada Analisis Sentimen Komentar Youtube BTS," *J. Inform. J. Pengemb. IT*, vol. 8, no. 3, pp. 207–212, 2023, doi: 10.30591/jpit.v8i3.5064.

[19] A. Kusuma and A. Nugroho, "Analisa Sentimen Pada Twitter Terhadap Kenaikan Tarif Dasar Listrik Dengan Metode Naïve Bayes," *J. Ilm. Teknol. Inf. Asia*, vol. 15, no. 2, p. 137, 2021, doi: 10.32815/jitika.v15i2.557.

[20] A. Pratama, A. C. Nurcahyo, and L. Firgia, "Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes," *Pros. CORISINDO 2023*, pp. 116–121, 2023, [Online]. Available: https://stmikpontianak.org/ojs/index.php/corisindo/article/view/30%0Ahttps://stmikpontianak.org/ojs/index.php/corisindo/article/download/30/22

[21] L. Rangga, A. Tarigan, T. Informatika, R. Forest, O. Fitur, and F. Selection, "OPTIMALISASI FITUR DENGAN FORWARD SELECTION PADA ESTIMASI TINGKAT PENYAKIT PARU-PARU MENGGUNAKAN ALGORITMA," vol. 8, no. 5, pp. 10341–10348, 2024.

[22] P. Agusia, M. Uli, A. Manurung, V. Calista, and V. C. Mawardi, "Pemanfaatan Word Cloud Pada Analisis Sentimen Dalam Menggali Persepsi Publik," pp. 25–30, 2024.