

ANALISIS SENTIMEN ULASAN PENGGUNA ACCESS BY KAI MENGGUNAKAN METODE WORD2VEC DAN ALGORITMA SVM

Ditha Lozera Devi^{1*}, Amalia Anjani Arifiyanti², Seftin Fitri Ana Wati³

^{1,2,3}Universitas Pembangunan Nasional “Veteran” Jawa Timur; Jl. Rungkut Madya No.1, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur; Telp. (031) 870 6369/ Fax. (031) 870 6372

Received: 12 Juli 2024

Accepted: 31 Juli 2024

Published: 7 Agustus 2024

Keywords:

Klasifikasi, Analisis Sentimen, Word2vec, SVM, Access by KAI.

Correspondent Email:

*dithalozeradevi2002@gmail.com

Abstrak. Beberapa ulasan dari pengguna KAI Access menyatakan sering terjadi gangguan pada saat pemesanan tiket. Hingga akhirnya pada tanggal 10 Agustus 2023 PT Kereta Api Indonesia melakukan peluncuran aplikasi Access by KAI sebagai bentuk upgrade dari aplikasi sebelumnya. Dengan adanya ulasan yang diberikan pengguna untuk aplikasi, perlu dilakukan analisis sentimen untuk melihat bagaimana pendapat dan reaksi pengguna dalam menggunakan aplikasi Access by KAI. Data ulasan pengguna diambil dari Google Play Store dan App Store. SEMMA dipilih sebagai metode pengembangan model data mining dengan tahapan dimulai dari Sample, Explore, Modify, Model, dan Assess. Analisis sentimen dilakukan dengan menggunakan metode Word2vec (CBOW dan Skip-gram) sebagai metode ekstraksi fitur dan 4 kernel SVM yang digunakan yaitu kernel linear, kernel polynomial, kernel RBF, dan kernel sigmoid. Hasil dari delapan skenario model klasifikasi yang dilakukan dengan menggabungkan metode Word2vec dan algoritma Support Vector Machine, dihasilkan satu skenario terbaik yaitu skenario model yang menggunakan algoritma SVM kernel RBF dengan metode Skip-Gram ditambah metode oversampling SMOTE dihasilkan nilai akurasi 81% dan nilai AUC sebesar 0.81.

Abstract. Some reviews from KAI Access users stated that there were frequent interruptions when booking tickets. Until finally on August 10, 2023 PT Kereta Api Indonesia launched the Access by KAI application as an upgrade from the previous application. With the reviews given by users for the application, it is necessary to conduct sentiment analysis to see how users' opinions and reactions are in using the Access by KAI application. User review data is taken from Google Play Store and App Store. SEMMA was chosen as the model development method data mining with stages starting from Sample, Explore, Modify, Model, and Assess. Sentiment analysis is carried out using the Word2vec method (CBOW and Skip-gram) as a feature extraction method and 4 SVM kernels used, namely linear kernel, polynomial kernel, RBF kernel, and sigmoid kernel. The results of eight classification model scenarios carried out by combining the Word2vec method and the Support Vector Machine algorithm, produced one of the best scenarios, that is the model scenario using the RBF kernel SVM algorithm with the Skip-Gram method plus the SMOTE oversampling method resulting in an accuracy value of 81% and an AUC value of 0.81.

1. PENDAHULUAN

Di era perkembangan teknologi yang semakin maju dan terdepan ini, tidak dapat dipungkiri semua orang bergantung pada teknologi. Dalam berbagai aspek di kehidupan sehari-hari, semua manusia membutuhkan teknologi untuk membantu mempermudah pekerjaan mereka. Salah satunya adalah aktivitas pembelian tiket kereta api. Sekarang untuk membeli tiket kereta api tidak perlu datang ke stasiun, pembeli dapat melakukan pembelian tiket dari rumah. Hal ini dikarenakan pada tahun 2014, PT. Kereta Api Indonesia yang merupakan perusahaan Badan Usaha Milik Negara (BUMN) berinovasi dengan meluncurkan aplikasi KAI Access [1]. Tahun 2020, aplikasi ini bahkan memenangkan penghargaan “the best innovation in payment system” (bukan bank) dalam Artajasa Award. *Fitur* dalam KAI Access diantaranya yaitu, KA lokal, KA antar kota, KA bandara, melihat jadwal kereta, melihat kode dan barcode pemesanan, dan cek posisi kereta api [2].

PT. KAI menyatakan pada Juni 2023 jumlah pengguna KAI Access yang tercatat sebanyak 12,4 juta pengguna. Sedangkan untuk pengguna aktif sebanyak 6,1 juta pengguna. *Rating* ulasan pada Play Store sebesar 3.2. Beberapa ulasan dari pengguna yang menyatakan bahwa *loading* aplikasi lambat, batas waktu pemesanan tiket yang terlalu cepat, dan sering terjadi gangguan pada saat pemesanan tiket [3]. Hingga akhirnya pada tanggal 10 Agustus 2023 PT Kereta Api Indonesia melakukan peluncuran aplikasi Access by KAI sebagai bentuk *upgrade* dari aplikasi sebelumnya yaitu KAI Access. Terdapat 4 *fitur* baru yang ditawarkan pada aplikasi ini antara lain, reservasi hotel, *loyalty poin*, *trip planner*, dan *live tracking*. Dari permasalahan diatas diperlukan klasifikasi ulasan dari pengguna untuk mengetahui sentimen pengguna Access by KAI.

Adanya ulasan yang diberikan pengguna untuk aplikasi, perlu dilakukan analisis sentimen untuk melihat bagaimana pendapat dan reaksi pengguna dalam menggunakan aplikasi Access by KAI. *Sentiment analysis* dapat digunakan untuk menganalisis dokumen dengan terperinci dan merangkum masukan dari pengguna, sehingga pihak pengembang dapat memperoleh hasil analisis tersebut untuk meningkatkan kualitas aplikasi [4]. Pada penelitian ini, proses klasifikasi menggunakan

data ulasan aplikasi Access by KAI dari Google Play Store dan App Store yang bersifat *open source* yang dapat diakses semua orang. Oleh karena itu, perlu dilakukan proses *scraping* atau pengambilan data berupa ulasan dan *rating* dari pengguna. Kemudian data tersebut akan diproses untuk dianalisis agar mendapatkan nilai akurasi. Data ulasan aplikasi yang akan dikumpulkan dalam penelitian ini adalah data ulasan dari awal upgrade aplikasi menjadi Access by KAI di Google Play Store dan App Store pada tanggal 10 Agustus 2023.

Terdapat dua jenis analisis sentimen yaitu *deep learning models* dan *traditional models*. *Deep learning models* diantaranya yaitu, *Deep Learning Neural Network (DNN)*, *Learning and Teaching Support Network (LTSM)*, *Recurrent Neural Network (RNN)*, dan *Convolutional Neural Network (CNN)*. Sedangkan untuk model tradisional yaitu *Support Vector Machine*, *Naïve Bayes*, dan *Maximum Entropy Classifier* [5]. Analisis sentimen dapat menggunakan beberapa metode klasifikasi. Beberapa metode klasifikasi yang sering digunakan yaitu *Support Vector Machine (SVM)*, *Naïve Bayes*, dan *K-Nearest Neighbor (KNN)*. Pada penelitian ini akan menggunakan algoritma *Support Vector Machine (SVM)* sebagai metode klasifikasi sentimen dan metode *Word2vec* untuk ekstraksi fitur. Setelah proses pembuatan model selesai, diperlukan evaluasi performa model dengan menggunakan *confusion matrix* untuk mengukur tingkat akurasi dari model yang dipakai sebelumnya. Dengan adanya penelitian ini diharapkan dapat menghasilkan nilai akurasi dari klasifikasi sentimen dan mengetahui pengaruh metode *Word2vec* terhadap performa model klasifikasi.

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen merupakan suatu proses yang dimulai dari pengumpulan data, pengolahan data dan evaluasi data yang berupa pendapat atau opini publik terhadap produk atau topik tertentu [6]. *Opinion mining* juga dapat disamakan dengan *sentiment analysis*, keduanya berfokus pada pengelompokan pendapat ke dalam label positif dan negatif [7]. Fungsi dari analisis sentimen adalah untuk menggolongkan setiap teks pada sumber dari media sosial dalam bentuk kalimat atau dokumen yang kemudian ditentukan

kategorinya antara negatif, positif, atau netral [8].

2.2 Access by KAI

Access by KAI merupakan aplikasi yang diluncurkan oleh PT Kereta Api Indonesia sebagai hasil perbaikan dari aplikasi sebelumnya yaitu KAI Access, dengan tujuan memberikan kenyamanan pada perjalanan penumpang melalui *fitur-fitur* yang inovatif dan terbaru [9]. Perbaikan ini terjadi atas permintaan dan masukan penumpang melalui ulasan pada Google Play Store dan App Store. Setelah pembaruan, aplikasi ini memiliki tampilan antarmuka pengguna yang lebih menarik dari segi visual dan desain terlihat lebih segar dan menarik. Fitur baru yang dihadirkan dalam aplikasi ini yaitu melacak lokasi terkini, reservasi hotel, perencanaan perjalanan, dan keuntungan dari program loyalitas [10].

2.3 SEMMA

SEMMA merupakan metode yang dikembangkan oleh SAS Institute. Metodologi SEMMA terdiri dari 5 fase (*Sample, Explore, Modify, Model, dan Assess*). Dengan adanya 5 fase tersebut, dapat dipastikan eksplorasi data menjadi lebih komprehensif dan pengembangan model menjadi lebih efisien [11].

2.4 Word2Vec

Word2Vec adalah model *Neural Network* (NN) yang mengodekan informasi semantik dari setiap istilah pada kumpulan teks yang diberikan data latih tanpa label. Model ini mengevaluasi kemiripan kosinus antar vektor kata untuk memahami kemiripan semantik. Vektor yang sama dimiliki oleh kata-kata yang memiliki makna yang sama. Sedangkan vektor yang beragam dimiliki oleh kata-kata yang memiliki makna yang berbeda [12]. Word2vec terbagi menjadi dua jenis, yaitu *Skip-Gram* dan *Continous Bag of Words* (CBOW).

2.5 Support Vector Machine

SVM adalah salah satu algoritma dari *Machine Learning* yang berasal dari teori pembelajaran statistik. SVM diperkenalkan oleh Vladimir Vapnik sebagai model pembelajaran mesin berbasis kernel untuk tugas klasifikasi dan regresi. *Support Vector Classifier* menyediakan metode yang disebut 'kernel trick'. Metode tersebut digunakan untuk mengatasi masalah pada algoritma SVM.

Kernel bekerja dengan cara mengubah data dari dimensi tertentu (2D) ke dalam dimensi tinggi (3D) sehingga data non-linier dapat dipisahkan secara linier [13]. Terdapat 4 jenis kernel dalam SVM diantaranya *linear kernel, polynomial kernel, RBF (Radial Basic Function) kernel, dan sigmoid kernel*.

2.6 Confusion Matrix

Evaluasi model yang digunakan untuk klasifikasi algoritma *supervised learning* adalah *Confusion Matrix*. *Confusion Matrix* digambarkan dengan baris dan kolom, baris sendiri berisi jumlah kejadian sebenarnya (aktual) dan kolom berisi jumlah kejadian hasil prediksi model [14]. Tabel 1 berikut merupakan hasil tabel *confusion matrix* 2 kelas yaitu positif dan negatif.

Tabel 1. Confusion matrix

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

TP (*True Positif*) yang berarti hasil prediksi data positif sudah akurat. FP (*False Positif*) yang berarti data terprediksi positif tetapi data sebenarnya adalah negatif. FN (*False Negatif*) yang berarti data terprediksi negatif tetapi data sebenarnya adalah positif. TN (*True Negatif*) yang berarti hasil prediksi data negatif sudah akurat. Dari data tersebut, maka dapat dilakukan perhitungan nilai akurasi, presisi, *recall*, dan *F1-score*.

2.7 AUC

AUC (*Area Under Curve*) adalah metrik yang digunakan untuk evaluasi kinerja model klasifikasi. Metrik ini dapat menilai seberapa baik model dapat membedakan dua kelas, yaitu kelas positif dan kelas negatif. Nilai AUC berkisar antara 0 sampai dengan 1. Semakin tinggi nilai AUC maka semakin akurat kinerja model dalam memprediksi data [15]. Tabel 2 berikut merupakan tingkatan dari nilai AUC.

Tabel 2. Tingkatan nilai AUC

Nilai AUC	Level Klasifikasi
0,91 – 1,00	Sangat baik
0,81 – 0,90	Baik
0,71 – 0,80	Cukup
0,60 – 0,70	Kurang
< 0,60	Gagal

3. METODE PENELITIAN

Pada tahap ini SEMMA dipilih sebagai metode pengembangan model data mining. SEMMA merupakan singkatan dari tahapannya sendiri yaitu *Sample, Explore, Modify, Model, dan Assess*. Berikut penjelasan setiap tahapan dari SEMMA:

3.1. Sample

Pada tahap ini dilakukan *scraping* atau pengumpulan data ulasan pengguna Access by KAI pada Google Play Store menggunakan API *google-play-scraper*, sedangkan pada App Store menggunakan API *app_store_scraper*. Dengan menggunakan kedua API tersebut, data dapat diambil dengan memasukkan ID dari aplikasi. Pada Google Play Store id aplikasi Access by KAI yaitu 'com.kai.kaiticketing', sedangkan pada App Store menggunakan angka yaitu '901804734'.

3.2. Explore

Explore merupakan tahapan untuk menggali informasi data yang akan digunakan. Tahap ini juga biasa disebut dengan EDA (*Exploratory Data Analysis*) yang bertujuan untuk memahami struktur dan karakteristik data. Visualisasi datanya menggunakan grafik dan plot seperti histogram dan *violin plot*. Manfaat dari tahap ini yaitu dapat melihat data yang kosong dan *outlier* pada data yang dapat memengaruhi analisis dan model.

3.3. Modify

Pada tahap ini terdiri beberapa proses untuk mengolah data dari kolom teks yang dipilih. Proses ini akan membantu data agar dapat dilakukan klasifikasi di tahap selanjutnya. Berikut adalah proses-proses yang dilakukan pada tahap *modify* dalam penelitian ini:

3.3.1. Pelabelan Data

Pembagian label ditentukan dengan cara dibagi menjadi 2 kelas, yaitu label positif dan label negatif. Apabila Rating ulasan lebih dari 3, maka data akan diberi label positif. Sedangkan, jika Rating ulasan sama dengan atau kurang dari 3 maka diberi label negatif [16].

3.3.2. Preprocessing Data

Pada tahap *pre-processing* terdapat beberapa tahapan yang harus dilakukan untuk membersihkan data, sehingga data dapat diolah dengan baik. Tahapan dalam *pre-processing*

data meliputi: *Cleansing data* yaitu menghapus emoticon, karakter, simbol, angka, maupun tanda baca; *Case folding* yaitu mengubah huruf kapital menjadi huruf kecil; *Stopword removal* yaitu menghapus kata yang tidak bermakna seperti 'yang', 'pada', 'dan', 'dari', 'untuk', 'itu', 'dengan', dsb.; *Tokenizing* yaitu membuat teks menjadi kumpulan kata; dan *Stemming* yaitu menghilangkan kata imbuhan seperti 'ber-', 'me-', 'di-', '-kan', '-an', dsb.

3.3.3. Ekstraksi Fitur (Word2Vec)

Word2vec terbagi menjadi dua jenis, yaitu *Skip-Gram* dan *Continuous Bag of Words (CBOW)*. CBOW digunakan untuk memprediksi kata yang kosong, sedangkan *Skip-Gram* digunakan untuk memprediksi kata di sekitar target. Tabel 3 berikut adalah contoh prediksi dari CBOW dan Skip-Gram dengan seberapa jauh prediksi kata ($n_window = 2$).

Tabel 3. Perbandingan CBOW and Skip-gram.

CBOW	Skip-Gram
halo nama <u>saya</u> adalah ditha	<u>halo nama</u> saya <u>adalah ditha</u>
(halo, saya)	(saya, halo)
(nama, saya)	(saya, nama)
(adalah, saya)	(saya, adalah)
(ditha, saya)	(saya, ditha)

3.3.4. Data Splitting

Data ulasan hasil proses-proses sebelumnya akan dibagi menjadi 2 yaitu, data *train* (latih) dan data *test* (uji). Pada penelitian ini menggunakan metode *Hold Out* dengan rasio pembagian datanya yaitu 80% data latih dan 20% data uji [17].

3.3.5. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) merupakan metode *oversampling* yang akan digunakan untuk menyeimbangkan kelas yang berbeda dalam *dataset*. Metode ini bekerja dengan cara menambah jumlah kelas minoritas dan menyamakan jumlahnya dengan kelas mayoritas.

3.4. Model

Pemodelan dibagi menjadi 8 skenario klasifikasi, sebagai berikut. Skenario pertama, menggunakan CBOW dan algoritma SVM kernel linear. Skenario kedua, menggunakan

Skip-gram dan algoritma SVM kernel linear. Skenario ketiga, menggunakan CBOV dan algoritma SVM kernel polynomial. Skenario keempat, menggunakan Skip-gram dan algoritma SVM kernel polynomial. Skenario kelima, menggunakan CBOV dan algoritma SVM kernel RBF. Skenario keenam, menggunakan Skip-gram dan algoritma SVM kernel RBF. Skenario ketujuh, menggunakan CBOV dan algoritma SVM kernel sigmoid. Skenario kedelapan, menggunakan Skip-gram dan algoritma SVM kernel sigmoid.

3.5. Assess

Pada tahap ini dilakukan pengukuran kinerja dari algoritma SVM yang telah digunakan untuk klasifikasi. *Confusion matrix* digunakan sebagai metode evaluasi kinerja dengan menilai tingkat akurasi, *recall*, presisi, dan *F1-score* dari model. Setelah mendapatkan hasil dari pengukuran tersebut, dapat ditarik kesimpulan dengan membandingkan kedelapan skenario model yang digunakan. Lalu, dilakukan evaluasi kembali dengan membandingkan nilai AUC (*Area Under Curve*).

4. HASIL DAN PEMBAHASAN

4.1. Sample

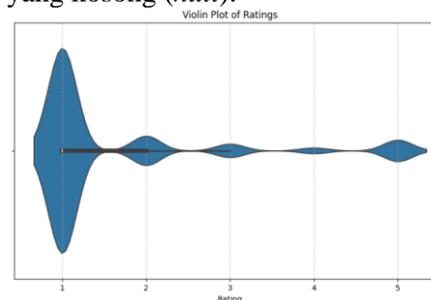
Data ulasan yang digunakan diambil dari platform Google Play Store dan App Store mulai tanggal 10 Agustus 2023 sampai 5 Februari 2024. Data diambil dengan melakukan *filter* ke satu wilayah dan bahasa dengan jumlah 16.452 data ulasan. Atribut yang digunakan untuk pemodelan adalah atribut review dan rating. Tabel 4 berikut hasil *scrapping* data.

Tabel 4. Hasil *scrapping* data

Review	Rating
ga ada gunanya sama sekali aplikasi ini.sm sekali gak layak terlebih sekelas KAI	1
Udah melek jam 00:00 tapi knp jadwal tidak tersedia,, setelah tersedia tiba” udah penuh,, ini ada calo cybernya ya?	1
MASIH PAKAI KALENDER 2023.TOLONG DI UPDATE	2
Tll, setiap kali tutup aplikasi selalu logout akun. Tll	5
Lebih enak yg dulu. Metode pembayaran malah cuma pake ovo doang skrng,-	3

4.2. Explore

Eksplorasi data diperlukan setelah selesai melakukan *scrapping* data. Data ulasan hasil proses *scrapping* tersebut dilakukan analisis untuk mengetahui tipe data, jumlah baris dan kolom, nilai *mean*, median, nilai terbesar, nilai terkecil, kuartil, dan jumlah data *null*. Tipe data atribut review adalah *object*, sedangkan tipe data atribut rating adalah *integer*. Pada atribut review terdapat 16452 entri data dengan 16044 data unik. Kata yang sering muncul adalah “ok” dengan frekuensi kemunculannya 54 kali. Sedangkan pada atribut rating terdapat 165452 entri data. Rata-rata rating / *mean* adalah 1,58 dengan std (standar deviasi) 1,18. Nilai rating minimum adalah 1 dan maksimumnya adalah 5. Nilai kuartil 1 (25%) adalah 1, nilai kuartil 2 (50%) adalah 1, dan nilai kuartil 3 (75%) adalah 2. Pada atribut review dan rating tidak terdapat data yang kosong (*null*).



Gambar 1. Violin plot atribut rating

Pada gambar 1 diatas, dapat dilihat bahwa *outlier* data ditunjukkan pada bagian yang lebih pipih yaitu pada nilai rating 4. Hal ini berarti rating 4 memiliki jumlah yang paling sedikit dibandingkan rating lainnya. Sedangkan, *violin* yang lebih tebal dan panjang menunjukkan area dimana sebagian besar rating berada. Hal tersebut ditunjukkan pada rating 1 yang berarti memiliki jumlah paling banyak dibandingkan rating lainnya.

4.3. Modify

4.3.1. Pelabelan Data

Pelabelan data dilakukan dengan membagi data menjadi 2 kelas sentimen, yaitu kelas positif dan negatif. Label ditentukan dengan melihat nilai Rating ulasan yang telah diberikan pengguna. Rating 1-3 diberi label negatif dan rating 4-5 diberi label positif. Tabel 5 berikut merupakan hasil pelabelan data.

Tabel 5. Hasil pelabelan data

Review	Rating	Sentiment
ga ada gunanya sama sekali aplikasi ini.sm sekali gak layak terlebih sekelas KAI	1	negatif
Udah melek jam 00:00 tapi knp jadwal tidak tersedia,, setelah tersedia tiba” udah penuh,, ini ada calo cybernya ya?	1	negatif
MASIH PAKAI KALENDER 2023.TOLONG DI UPDATE	2	negatif
Tll, setiap kali tutup aplikasi selalu logout akun. Tll	5	positif
Lebih enak yg dulu. Metode pembayaran malah cuma pake ovo doang skrng,-	3	positif

4.3.2. Preprocessing Data

Pada tahap ini dilakukan serangkaian tahapan untuk membersihkan dan menyesuaikan data agar kualitas data lebih baik. Adapun serangkaian tahapan dalam preprocessing data yaitu *cleansing*, *case folding*, *stopword removal*, *tokenizing*, dan *stemming*. Tabel 6 di bawah ini merupakan hasil *preprocessing* data.

Tabel 6. Hasil preprocessing data

Review	Data hasil preprocessing
ga ada gunanya sama sekali aplikasi ini.sm sekali gak layak terlebih sekelas KAI	['ga', 'guna', 'aplikasi', 'sm', 'gak', 'layak', 'kelas', 'kai']
Udah melek jam 00:00 tapi knp jadwal tidak tersedia,, setelah tersedia tiba” udah penuh,, ini ada calo cybernya ya?	['udah', 'melek', 'jam', 'knp', 'jadwal', 'sedia', 'sedia', 'udah', 'penuh', 'calo', 'cybernya', 'ya']
MASIH PAKAI KALENDER 2023.TOLONG DI UPDATE	['pakai', 'kalender', 'tolong', 'update']
Tll, setiap kali tutup aplikasi selalu logout akun. Tll	['tll', 'kali', 'tutup', 'aplikasi', 'logout', 'akun', 'tll']

Lebih enak yg dulu. ['enak', 'yg', 'Metode pembayaran 'metode', 'bayar', 'malah cuma pake ovo 'pake', 'ovo', 'doang skrng,-' 'doang', 'skrng']

4.3.3. Ekstraksi Fitur

Ekstraksi fitur akan dilakukan dengan 2 jenis teknik dalam metode Word2Vec yaitu *Continous Bag of Words* (CBOW) dan *Skip-Gram*. Sebelum melakukan ekstraksi fitur menggunakan Word2vec, perlu dipastikan tipe data yang akan diekstraksi adalah *list*. Hal ini dikarenakan metode Word2Vec mempelajari konteks kata berdasarkan urutan kata dalam kalimat. Dalam kode programnya, parameter *sg=0* merupakan kode untuk CBOW, sedangkan parameter *sg=1* merupakan kode untuk *Skip-gram*.

4.3.4. Data Splitting

Metode *hold out* digunakan sebagai teknik *splitting* data pada penelitian ini. Data ulasan akan dibagi menjadi 2 jenis dengan rasio pembagian datanya yaitu 80% data latih dan 20% data uji.

4.3.5. SMOTE

SMOTE digunakan untuk menyeimbangkan data ulasan pada penelitian ini. Hal ini dikarenakan jumlah data negatif lebih banyak ketimbang data positif, sehingga data positif jumlahnya ditambah hingga sama dengan jumlah data negatif. Hasilnya pada tabel 7 dibawah ini.

Tabel 7. Jumlah data sebelum dan sesudah SMOTE

Output
Jumlah data positif sebelum SMOTE: 1150
Jumlah data negatif sebelum SMOTE: 10695
Jumlah data positif setelah SMOTE: 10695
Jumlah data negatif setelah SMOTE: 10695

4.4. Model

Pada tahap model ini dilakukan klasifikasi ulasan dengan menggunakan algoritma *Support Vector Machine* (SVM). Skenario pemodelan menggunakan 4 kernel SVM yaitu *kernel linear*, *kernel polynomial*, *kernel RBF*, *kernel sigmoid*. Kode program untuk inialisasi kernel SVM adalah `svm_model = SVC(kernel='linear', random_state=42)`, parameter kernel dapat diubah sesuai nama kernel yang digunakan.

4.5. Assess

Pada tahap assess dilakukan penilaian kinerja dari algoritma SVM dengan menggunakan *confusion matrix*. Penilaian kinerja berupa tingkat akurasi, presisi, *recall*, dan *F1-score* yang dapat mengukur seberapa baik model dalam memprediksi atau mengklasifikasikan data.

4.5.1. SVM Linear & CBOW

Hasil pengujian SVM Linear menggunakan metode ekstraksi fitur CBOW dan metode *oversampling* SMOTE menghasilkan nilai akurasi 76%. Sedangkan hasil pengujian SVM Linear menggunakan metode ekstraksi fitur CBOW tanpa SMOTE menghasilkan nilai akurasi 89%. Hasil perbandingan *confusion matrix* pada tabel 8 berikut.

Tabel 8. Perbandingan SVM Linear & CBOW SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	226	635	2011	90
Tanpa SMOTE	0	0	2646	316

4.5.2. SVM Linear & Skip-gram

Hasil pengujian SVM Linear menggunakan metode ekstraksi fitur *Skip-Gram* dan metode *oversampling* SMOTE menghasilkan nilai akurasi 78%. Sedangkan hasil pengujian SVM Linear menggunakan metode ekstraksi fitur *Skip-Gram* tanpa SMOTE menghasilkan nilai akurasi 89%. Hasil perbandingan *confusion matrix* pada tabel 9 berikut.

Tabel 9. Perbandingan SVM Linear & Skip-gram SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	231	554	2092	85
Tanpa SMOTE	0	0	2646	316

4.5.3. SVM Polynomial & CBOW

Hasil pengujian SVM Polynomial menggunakan metode ekstraksi fitur CBOW dan metode *oversampling* SMOTE menghasilkan nilai akurasi 64%. Sedangkan hasil pengujian SVM Polynomial menggunakan metode ekstraksi fitur CBOW tanpa SMOTE menghasilkan nilai akurasi 89%. Hasil perbandingan *confusion matrix* pada tabel 10 berikut.

Tabel 10. Perbandingan SVM Polynomial & CBOW SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	250	1000	1646	66
Tanpa SMOTE	0	0	2646	316

4.5.4. SVM Polynomial & Skip-gram

Hasil pengujian SVM Polynomial menggunakan metode ekstraksi fitur *Skip-Gram* dan metode *oversampling* SMOTE menghasilkan nilai akurasi 71%. Sedangkan hasil pengujian SVM Polynomial menggunakan metode ekstraksi fitur *Skip-Gram* tanpa SMOTE menghasilkan nilai akurasi 89%. Hasil perbandingan *confusion matrix* pada tabel 11 berikut.

Tabel 11. Perbandingan SVM Polynomial & Skip-gram with SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	240	771	1875	76
Tanpa SMOTE	0	0	2646	316

4.5.5. SVM RBF & CBOW

Hasil pengujian SVM RBF menggunakan metode ekstraksi fitur CBOW dan metode *oversampling* SMOTE menghasilkan nilai akurasi 78%. Sedangkan hasil pengujian SVM RBF menggunakan metode ekstraksi fitur CBOW tanpa SMOTE menghasilkan nilai akurasi 89%. Hasil perbandingan *confusion matrix* pada tabel 12 berikut.

Tabel 12. Perbandingan SVM RBF & CBOW SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	211	541	2105	105
Tanpa SMOTE	0	0	2646	316

4.5.6. SVM RBF & Skip-gram

Hasil pengujian SVM RBF menggunakan metode ekstraksi fitur *Skip-Gram* dan metode *oversampling* SMOTE menghasilkan nilai akurasi 81%. Sedangkan hasil pengujian SVM RBF menggunakan metode ekstraksi fitur *Skip-Gram* tanpa SMOTE menghasilkan nilai akurasi 89%. Hasil perbandingan *confusion matrix* pada tabel 13 berikut.

Tabel 13. Perbandingan SVM RBF & Skip-gram SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	217	459	2187	99

Tanpa SMOTE	0	0	2646	316
-------------	---	---	------	-----

4.5.7. SVM Sigmoid & CBOW

Hasil pengujian SVM Sigmoid menggunakan metode ekstraksi fitur CBOW dan metode *oversampling* SMOTE menghasilkan nilai akurasi 52%. Sedangkan hasil pengujian SVM Sigmoid menggunakan metode ekstraksi fitur CBOW tanpa SMOTE menghasilkan nilai akurasi 84%. Hasil perbandingan *confusion matrix* pada tabel 14 berikut.

Tabel 14. Perbandingan SVM Sigmoid & CBOW SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	158	1258	1388	158
Tanpa SMOTE	82	237	2409	234

4.5.8. SVM Sigmoid & Skip-gram

Hasil pengujian SVM Sigmoid menggunakan metode ekstraksi fitur *Skip-Gram* dan metode *oversampling* SMOTE menghasilkan nilai akurasi 54%. Sedangkan hasil pengujian SVM Sigmoid menggunakan metode ekstraksi fitur *Skip-Gram* tanpa SMOTE menghasilkan nilai akurasi 84%. Hasil perbandingan *confusion matrix* pada tabel 15 berikut.

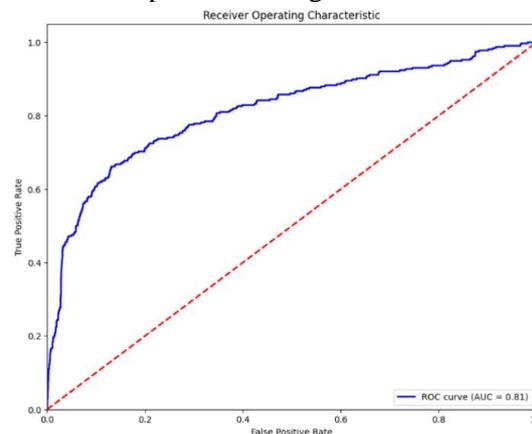
Tabel 15. Perbandingan SVM Sigmoid & Skip-gram SMOTE dan tanpa SMOTE

	TP	FP	TN	FN
SMOTE	156	1209	1437	160
Tanpa SMOTE	85	238	2408	231

4.6. Analisis Performa model

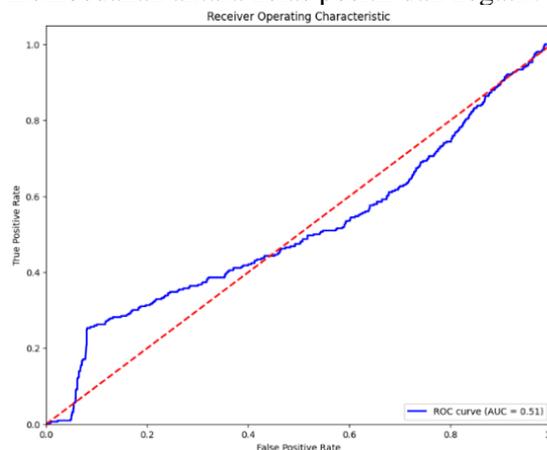
Berdasarkan kedelapan scenario yang telah diujikan, dihasilkan 3 skenario terbaik yaitu klasifikasi menggunakan algoritma SVM kernel RBF dengan metode *Skip-Gram* ditambah metode *oversampling* SMOTE dihasilkan nilai akurasi 81%, klasifikasi menggunakan algoritma SVM kernel sigmoid dengan metode CBOW dihasilkan nilai akurasi 84%, dan klasifikasi menggunakan algoritma SVM kernel sigmoid dengan metode *Skip-Gram* dihasilkan nilai akurasi 84%. Skenario tersebut akan diuji kembali untuk mendapatkan satu scenario terbaik dengan membandingkan nilai nilai AUC (*Area Under the Curve*). Gambar 2 berikut merupakan ROC Curve dari SVM kernel RBF

dengan metode *Skip-Gram* ditambah metode *oversampling* SMOTE yang menghasilkan nilai AUC sebesar 0.81 yang berarti model memiliki kemampuan yang baik dalam membedakan antara kelas positif dan negatif.



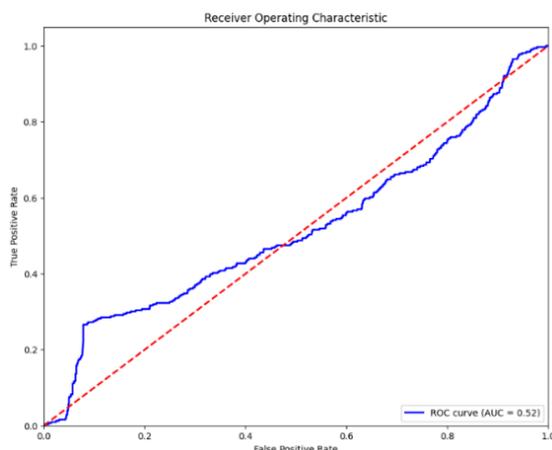
Gambar 2. ROC Curve SVM RBF & Skip-gram with SMOTE

Gambar 3 berikut merupakan ROC Curve dari SVM kernel Sigmoid dengan metode CBOW yang menghasilkan nilai AUC sebesar 0.51 yang berarti hasil prediksi model dapat dikatakan menebak secara acak. Hal ini menunjukkan bahwa model tersebut tidak dapat membedakan antara kelas positif dan negatif.



Gambar 3. ROC Curve SVM Sigmoid & CBOW

Gambar 4 berikut merupakan ROC Curve dari SVM kernel Sigmoid dengan metode *Skip-Gram* yang menghasilkan nilai AUC sebesar 0.52 yang berarti hasil prediksi model dapat dikatakan menebak secara acak. Hal ini menunjukkan bahwa model tersebut tidak dapat membedakan antara kelas positif dan negatif.



Gambar 4. ROC Curve SVM Sigmoid & Skip-gram

5. KESIMPULAN

- a. Hasil dari skenario model yang telah dilakukan menunjukkan bahwa terdapat 3 skenario yang memiliki nilai akurasi terbaik yaitu klasifikasi menggunakan algoritma SVM kernel RBF dengan metode *Skip-Gram* ditambah metode *oversampling* SMOTE dihasilkan nilai akurasi 81%, klasifikasi menggunakan algoritma SVM kernel sigmoid dengan metode CBOW dihasilkan nilai akurasi 84%, dan klasifikasi menggunakan algoritma SVM kernel sigmoid dengan metode *Skip-Gram* dihasilkan nilai akurasi 84%.
- b. Kemudian untuk memilih skenario terbaik dengan melihat nilai AUC (*Area Under the Curve*) dari ketiga skenario tersebut. Algoritma SVM kernel RBF dengan metode *Skip-Gram* ditambah metode *oversampling* SMOTE yang menghasilkan nilai AUC sebesar 0.81 sedangkan kedua skenario lainnya hanya menghasilkan nilai AUC tidak lebih besar dari 0.52. Sehingga skenario model menggunakan algoritma SVM kernel RBF dengan metode *Skip-Gram* ditambah metode *oversampling* SMOTE dipilih sebagai skenario terbaik.

DAFTAR PUSTAKA

- [1] S. Hendrik, and D. Novita. "Analisis Kepuasan Pengguna Aplikasi KAI Access Sebagai Media Pemesanan Tiket Kereta Api Menggunakan Metode EUCS." *Jurnal Teknologi Sistem Informasi*, vol. 2, pp. 162-175, 2021.
- [2] U. W. Yuda, M. B. Pratama, M. Rhamadani, and T. Sutabri, "Implementasi Teknologi Informasi Dalam Sistem Pemesanan Tiket Melalui Aplikasi KAI Access pada PT. Kereta Api Indonesia", *ijm*, vol. 1, no. 5, Nov. 2023.
- [3] N. M. Arochma, E. G. Purnaningsih, N. K. Anggreani, and A. Wulansari, "Evaluasi Kepuasan Pengguna Aplikasi KAI Access dengan Pendekatan End-User Computing Satisfaction", *JSIT*, vol. 3, no. 2, pp. 186-197, Jun. 2023.
- [4] R. Kusnadi, Y. Yusuf, A. Andriantony, R. Ardian Yaputra, and M. Caintan, "ANALISIS SENTIMEN TERHADAP GAME GENSHIN IMPACT MENGGUNAKAN BERT", *rabit*, vol. 6, no. 2, pp. 122-129, Jul. 2021.
- [5] M. Ikhsan, "Analisis Sentimen Terhadap Kenaikan Harga Bahan Bakar Minyak Menggunakan Long Short-Term Memory", *IJCSR*, vol. 2, no. 1, pp. 31-41, Feb. 2023.
- [6] Y. Akbar dan T. Sugiharto. "Analisis Sentimen Pengguna Twitter di Indonesia Terhadap ChatGPT Menggunakan Algoritma C4. 5 dan Naïve Bayes", *Jurnal Sains dan Teknologi*, vol. 5, no. 1, pp. 115-122, 2023.
- [7] S. Samsir, A. Ambiyar, U. Verawardina, F. Edi, dan R. Watrianthos. "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes", *Jurnal Media Informatika Budidarma*, vol. 5, no. 1, pp. 157-163, 2021.
- [8] S. M. Salsabila, A. A. Murtopo, dan N. Fadhilah. "Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier", *Jurnal Minfo Polgan*, vol. 11, no. 2, pp. 30-35, 2022.
- [9] D. P. Waluyaningtyas, dan D. H. Laksana. "Pengaruh Perceived Ease Of Use, Perceived Usefulness, Security, Dan Trust Terhadap Intention To Use Aplikasi Access By Kai (Survei pada Mahasiswa di Daerah Istimewa Yogyakarta)", *Innovative: Journal Of Social Science Research*, vol. 3, no. 6, pp. 9970-9981, 2023.
- [10] N. A. Pratama dan H. M. Thamrin. "Peran hubungan masyarakat PT Kereta Api Indonesia Daop 9 Jember dalam meningkatkan kualitas pelayanan kepada Masyarakat", *Journal of Advances in Accounting, Economics, and Management*, vol. 1, no. 1, pp. 1-9, 2023.
- [11] N. Hotz. "What is SEMMA?. Data Science Process Alliance", 12 Agustus 2023. Diakses pada 23 Februari 2024, dari laman <https://www.datascience-pm.com/semma/>
- [12] S. Sivakumar, L. S. Videla, T. R. Kumar, J. Nagaraj, S. Itnal, dan D. Hariitha. "Review on word2vec word embedding neural net", *In 2020 international conference on smart*

- electronics and communication (ICOSEC)*, pp. 282-290, 2020.
- [13] M. Muchtar dan R. A. Muchtar. "PERBANDINGAN METODE KNN DAN SVM DALAM KLASIFIKASI KEMATANGAN BUAH MANGGA BERDASARKAN CITRA HSV DAN FITUR STATISTIK", *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, 2024.
- [14] S. Suryani dan M. Mustakim. "Estimasi Keberhasilan Siswa dalam Pemodelan Data Berbasis Learning Menggunakan Algoritma Support Vector Machine", *Bulletin of Informatics and Data Science*, vol. 1, no. 2, pp. 81-88, 2022.
- [15] S. R. Darmawan, M. Fatchan, dan M. Maulana. "Prediction of 2024 Presidential Election using K-NN with Metric Approaches Chebyshev and Euclidean Based on Twitter Data Investigation", *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 2, pp. 475-485, (2024).
- [16] I. Karo Karo, J. Karo Karo, Y. Yuniarto, H. Hariyanto, M. Falah, and M. Ginting, "Analisis Sentimen Ulasan Aplikasi Info BMKG di Google Play Menggunakan TF-IDF dan Support Vector Machine", *josh*, vol. 4, no. 4, pp. 1423-1430, Jul. 2023.
- [17] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 83-87, 2020.