Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

Analisis Sentimen Publik Terhadap Program KIP-Kuliah Menggunakan Algoritma Random Forest pada Media Sosial X

Rosdiyanah Rosdiyanah, Dinda Lestarini*

Fakultas Ilmu Komputer, Sistem Informasi, Universitas Sriwijaya, Palembang, Indonesia
Email: [1]09031482326020@student.unsri.ac.id, [2,*]dinda.lestarini@gmail.com
Email Penulis Korespondensi: dinda.lestarini@gmail.com
Submitted: **16/05/2025**; Accepted: **31/05/2025**; Published**: 01/06/2025**

Abstrak−Program KIP-Kuliah atau Kartu Indonesia Pintar Kuliah (KIP-K) adalah bantuan pendanaan oleh pemerintah bagi mahasiswa yang mempunyai kesulitan ekonomi namun ingin merasakan kesempatan mengenyam pendidikan di perguruan tinggi. Program ini tidak lepas dari perbincangan masyarakat, terutama terkait isu penyalahgunaan dana oleh penerima, inkonsistensi pencairan dana, dan pemalsuan berkas pendaftaran. Problematika ini menjadikan pandangan masyarakat bahwa program KIP-K seringkali masih salah sasaran. Penelitian bertujuan untuk mengkaji sentimen atau persepsi masyarakat terhadap program KIPK-K menggunakan algoritma *Random Forest* yang dikombinasikan dengan Word2Vec sebagai teknik pembobotan kata dan *Random Oversampling* (ROS) sebagai teknik *balancing* untuk mengatasi ketidakseimbangan data. Dataset yg diperoleh berasal dari media sosial X (*Twitter*) sebanyak 4423 *tweet* dengan kata kunci "kip-k" atau "kipk" dan dengan rentan waktu selama tahun 2024. Kinerja model menunjukkan akurasi yang tinggi sebesar 96,57%, presisi, *recall*, dan *f1-score* di angka yang sama yaitu 97%. Hasil menunjukkan bahwa model mampu menganalisis sentimen dengan akurat dan mempertahankan performa yang seimbang antar kedua kelas sentimen. Berdasarkan temuan penelitian ini, algoritma *Random Forest* yang dikombinasikan dengan Word2Vec dan *Random Oversampling* (ROS) dapat menghasilkan akurasi tinggi dan mampu mengatasi ketidakseimbangan data.

Kata Kunci: Analisis Sentimen; KIPK; *Random Forest*; Word2Vec; *Random Oversampling*

Abstract−KIP-Kuliah program or The Indonesia Smart College Card (KIP-K) is a funding assistance from the government for students with economic difficulties who want to experience educational opportunities in higher education. This program can't be separated from public discussion, especially regarding the issue of misuse of funds by recipients, inconsistency in fund disbursement and falsification of registration files. These problems make the public view that the KIP-K program is often still misdirected. The research aims to examine public sentiment or perception towards the KIP-K program using Random Forest algorithm combined with Word2Vec as a word weighting technique and Random Oversampling (ROS) as a balancing technique to overcome data imbalance. The dataset obtained comes via platform X or Twitter) a total of 4423 tweets with the keywords "kip-k" or "kipk" and with a vulnerable time during 2024. The model's performance demonstrated a high accuracy of 96,57%, precision, recall, and f1-score at the same value of 97%. The results indicate that the model is effective in analyzing sentiment accurately and maintaining a balanced performance between the two sentiment classes. Based on research in this study, the Random Forest algorithm combined with Word2Vec and Random Oversampling (ROS) can produce high accuracy and can overcome data imbalance.

Keywords: Sentiment Analysis; KIPK; *Random Forest*; Word2Vec; *Random Oversampling*

1. PENDAHULUAN

Pemerintah Indonesia sebagai wadah pendukung kesempatan mengenyam pendidikan di perguruan tinggi dalam misinya memaksimalkan aksesibilitas pendidikan salah satunya dengan bantuan Program Kartu Indonesia Pintar Kuliah (KIP-K) atau KIP-Kuliah. Program ini merupakan bantuan yang diberikan pemerintah bagi mahasiswa dengan kesulitan ekonomi dan ingin meneruskan pendidikan ke jenjang universitas atau perguruan tinggi [1]. Bantuan KIP-K yang diberikan adalah berupa bantuan dana pendidikan serta bantuan biaya hidup bulanan yang diberikan persemester [2], [3]. Selain guna menyambung pendidikan, KIP-K diberikan supaya dapat mencegah potensi tingginya angka anak putus sekolah [4]. Kebermanfaatan Program KIP-K yang diberikan kepada mahasiswa berprestasi namun terhalang ekonomi besar harapan agar di masa depan dapat memutus rantai kemiskinan dengan membantu perekonomian keluarga setelah lulus dan mendapat pekerjaan, serta memberikan sumbangsih kepada Negara dalam peningkatan sumber daya manusia yang kompeten.

Portal Informasi Indonesia [5] memaparkan data total penerima serta total anggaran dana program KIP-K dari tahun 2020 sampai dengan 2025. Pada tahun 2020 total penerima berjumlah 552.706 dengan total anggaran Rp3,7 triliun, pada tahun 2021 naik sejumlah 674.187 dengan total anggaran Rp7.5 triliun. Kemudian data konsisten naik pada tahun-tahun selanjutnya seperti pada tahun 2022 penerima berjumlah 780.014 dengan anggaran Rp9,9 triliun, pada tahun 2023 penerima berjumlah 913.636 dengan total anggaran Rp11,8 triliun, pada tahun 2024 total penerima berjumlah 985.557 dengan anggaran sejumlah Rp13,9 triliun. Pada tahun 2025 pemerintah juga telah menganggarkan kenaikan berjumlah 1.040.192 penerima dengan anggaran Rp14,69 triliun. Data ini konsinten menunjukkan kenaikan tiap tahunnya baik pada jumlah penerima maupun dana anggaran program. Jumlah yang kompleks ini menggambarkan tanggung jawab pemerintah di sektor pendidikan dalam memangkas kesenjangan sosial ekonomi, namun ini juga menjadi tantangan dalam menetapkan sasaran penerima bantuan program KIP-K kepada yang benar-benar membutuhkan.

Program KIP-K dengan segala manfaat yang didapatkan tentunya akan menjadi sebuah perbincangan masyarakat terkait isu, pendapat, dukungan, saran maupun kritik terutama pada media sosial X (*Twitter*) salah satunya mengenai pelanggaran dan penyalahgunaan dana KIP-K terkait hedonisme [6]. Permasalahan terkait KIP-K ini juga

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

dikemukakan pada penelitian [7] yang mengemukakan beberapa problematika seperti inkonsistensi pencairan dana KIP-K dan pemalsuan berkas persyaratan pendaftaran. Sehingga beberapa problematika tersebut yang dibahas dalam perbincangan program KIP-K masih dianggap seringkali tidak tepat sasaran.

Analisis Sentimen merupakan metode yang diterapkan untuk mengetahui opini, perasaan, sikap ataupun emosi publik yang nantinya akan digunakan untuk pengklasifikasian dengan mengkategorikannya menjadi sentimen positif maupun negatif [8]. Hasil dari Analisis Sentimen dapat membantu pemerintah dalam merumuskan kebijakan guna pengambilan keputusan terkait pengembangan program KIP-K kedepannya [9]. Dalam lingkup penelitian ini memanfaatkan analisis sentimen untuk mengindentifikasi opini publik terhadap program KIP-K pada media sosial X (Twitter).

Berbagai penelitian terdahulu terkait topik yang serupa telah dilakukan. Seperti pada penelitian yang menganalisis tanggapan masyarakat terhadap program KIP-K di jejaring sosial X dengan mengaplikasikan model *Naïve Bayes* serta pendekatan CRISP-DM yang menghasilkan akurasi sebesar 84,99% dimana model cukup baik dalam mengklasifikasikan sentimen walaupun masih terdapat sentimen negatif yang mendominan [2]. Penelitian serupa juga dilakukan oleh [10] yang mengklasifikasikan opini masyarakat dengan menerapkan algoritma *Support Vector Machine* (SVM) menghasilkan akurasi 86,27% dengan temuan adanya bias model terhadap sentimen negatif yang dominan terkait penyalahgunaan batuan program KIP-K dan salah satu saran yang diberikan peneliti tersebut yaitu terkait teknik penyeimbangan dataset. Penelitian serupa lainnya juga menganalisis terkait tanggapan publik terhadap program KIP-K dimana peneliti tersebut membandingkan algoritma *Naïve Bayes* (NB)*, Support Vector Machine* (SVM) dan *Random Forest* dengan temuan menunjukan *Random Forest* lebih unggul dibading *Naïve Bayes* (NB) ataupun *Support Vector Machine* (SVM). *Random Forest* (RF) memberikan akurasi 96% dan naik menjadi 100% setelah dilakukan teknik penyeimbangan data menggunakan SMOTE [11].

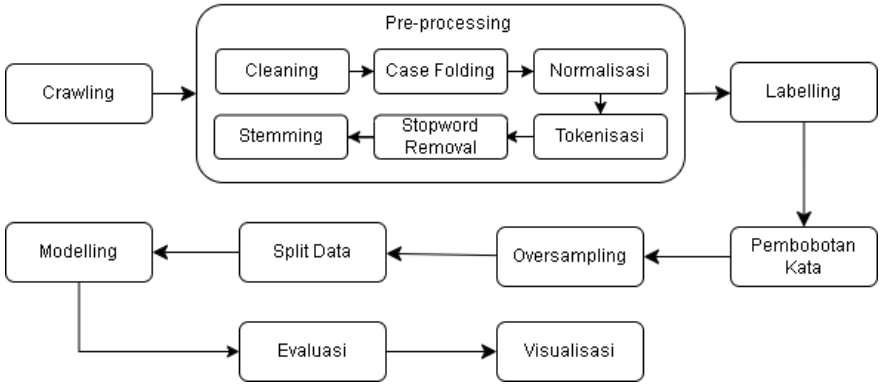
Penelitian lain yang juga membahas sentimen berdasarkan aspek EDOM (evaluasi dosen oleh mahasiswa) dimana peneliti mengkombinasikan model Word2Vec sebagai teknik ekstraksi fitur, model *Convolutional Neural Network* (CNN) serta menerapkan beberapa teknik penyeimbangan dataset atau *oversampling.* Dari perpaduan CNN, Word2Vec dan berapa Teknik *Oversampling* mendapatkan hasil yang bagus pada peningkatan tingkat akurasi model dan mengatasi ketidaksimbangan data [12]. Penelitian lain juga dilakukan oleh [13] yang membahas pengaruh Word2Vec pada klasifikasi sentimen terhadap performa *Deep Learning* LSTM yaitu salah satunya dipengaruhi oleh arsitektur CBOW menunjukkan akurasi terbaik dengan akurasi 97,12%.

Permasalahan yang telah dipaparkan serta beberapa penelitian terdahulu dapat menjadi rujukan dalam pengembangan analisis sentimen yang lebih akurat dan mencakup beragam sudut pandang. Dengan demikian, penelitian ini bertujuan untuk menganalisis sentimen atau persepsi publik terhadap program KIP-K menggunakan algoritma *Random Forest* yang dikombinasikan dengan Word2Vec dan *Random Oversampling*. *Random Forest* dipilih karena dinilai dapat menghasilkan akurasi yang lebih baik, Word2Vec digunakan karena keunggulannya dalam menghasilkan makna semantik melalui vektor yang serupa antar kata yang mirip dan *Random Oversampling* dipilih untuk mengatasi permasalahan dataset yang tidak seimbang. Sehingga perpaduan antara Random Forest, Word2Vec, dan *Random Oversampliling* diharapkan dapat meningkatkan performa model. Penelitian ini tidak hanya memperluas cakupan dari penelitian sebelumnya, tetapi juga menghadirkan kontribusi baru yang substansial dalam menafsirkan persepsi publik terkait program pendidikan terkhusus KIP-K. Selain daripada itu, juga diharapkan temuan dari penelitian ini bisa memberikan masukan penting pada pemerintah serta para pemangku kepentingan dalam memengembangkan program KIP-K di Indonesia.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menganalisis persepsi atau sentimen publik terhadap program KIP-K di Indonesia dengan tahapan penelitian tertampil pada Gambar 1.



Gambar 1. Tahap Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

Gambar 1 merepresentasikan tahapan penelitian yang dimulai dari proses pengumpulan data (*crawling*) dari media sosial X (Twitter), kemudian *pre-processing* yang dimulai dengan *cleaning*, *case folding*, normalisasi, tokenisasi, *stopword removal* serta *stemming*, lalu proses *labelling* dilakukan menggunakan *pre-trained model* dari *hugging face*, lalu proses pembobotan kata atau ekstraksi fitur menggunakan Word2Vec, selanjutnya proses *oversampling* menggunakan teknik *random oversampling*, lanjut proses distribusi dataset pelatihan dan dataset pengujian (*split*), setelah itu tahap *modeling* menggunakan *random forest*, kemudian evaluasi dan visualisasi.

2.2 Pengumpulan Data (*Crawling*)

Penelitian ini memanfaatkan data *tweet* yang dihimpun dari jejaring sosial X menggunakan metode *crawling* pada *Google Colab* yang mengimplementasikan bahasa pemrograman *Pyhton* dan *Library Tweet Harvest*. Data dikumpulkan menggunakan rentan waktu dan juga kata kunci yang berkaitan dengan program KIP-K.

2.3 *Pre-processing*

Tahapan pada *pre-processing* diimplementasikan dalam penelitian ini setelah tahap *crawling* selesai. *Pre-processing* diterapkan untuk membersihkan data yang tidak relevan maupun ketidakteraturan lalu mengubahnya ke dalam format beraturan untuk siap diproses [14]. Proses yang dilakukan pada tahap *pre-processing* terkhusus di penelitian ini meliputi *cleaning*, lalu *case folding*, normalisasi, tokenisasi, *stopword removal*, serta *stemming*.

2.3.1 *Cleaning*

Cleaning merupakan proses membersihkan dataset dari elemen tidak penting dan tidak relevan. Elemen-elemen yang tidak akan digunakan adalah seperti emoji, simbol, angka, *username* (@), URL, dan *hashtag* (#) yang biasanya terdapat pada *tweet* di media sosial X [9].

2.3.2 *Case Folding*

Tahap *Case Folding* merupakan tahap merubah keseluruhan huruf besar di dataset menjadi *lowercase* atau huruf kecil dengan tujuan efisiensi dan memudahkan analisis teks karena semua huruf sudah dalam format yang sama [14].

2.3.3 Normalisasi

Normalisasi adalah tahap mentransformasi kata non baku di dataset menjadi kata yang baku atau bentuk yang standar. Tahap normalisasi ini diterapkan supaya membantu dalam pengolahan data dan memaksimalkan tingkat akurasi pada analisis text [15].

2.3.4 Tokenisasi

Tokenisasi merupakan tahapan dimana teks yang di fragmentasi dari satu kalimat menjadi kata atau pecahan yang lebih kecil [15]. Tujuan dari penerapan tokenisasi dalam pemrosesan bahasa alami adalah memudahkan dalam memahami makna dari teks pada dataset [16].

2.3.5 *Stopword Removal*

Stopword Removal ialah tahap menghapus kata-kata tidak penting atau tidak ada maknanya serta tidak mempunyai kaitan dengan isi teks pada dataset. *Stopword* adalah kata yang bersifat umum yakni berupa kata sambung seperti “di”, “ke”, “dan”, “dari”, dll. Kata-kata inilah yang tidak digunakan dan akan dihilangkan pada tahap *stopword removal* [17].

2.3.6 *Stemming*

Stemming adalah tahapan dalam menyederhanakan kata ke dalam bentul dasar atau menghapus imbuhan pada kata dalam teks [14]. Contoh tahap *stemming* biasanya seperti “menghapus”, “dihapuskan”, “penghapusan” akan disederhanakan menjadi kata dasar “hapus”. Namun perlu diwaspadai, terkadang tahap ini dapat menyerupai makna yang sama namun memiliki perbedaan arti kata [10].

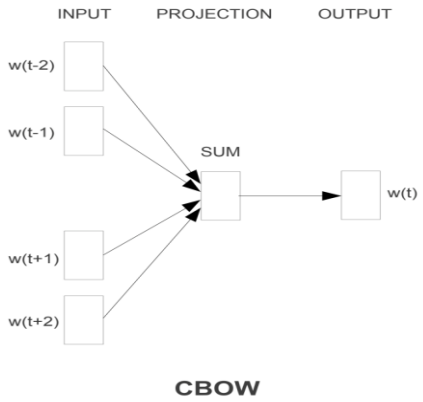
2.4 *Labelling*

Labelling merupakan tahap dimana dataset diberi label dalam beberapa kategori. Dalam penelitian ini pengklasifikasian label dilakukan dalam dua kategori yaitu sentimen negatif dan sentimen positif. *Labelling* diterapkan dengan memanfaatkan model sentimen yang sudah dibentuk yaitu *pre-trained* model dari *website Hugging Face* dimana model tersebut yaitu SMSA atau *bert base Indonesian 1.5G sentiment analysis* [12]. Model ini digunakan karena kemampuannya dalam memprediksi sentimen positif dan negatif. Setelah model berhasil memprediksi sentimen, dilakukan pengecekan manual terhadap hasil yang telah diprediksi dengan tujuan memastikan keakuratan dan kualitas data. Sehingga pada proses *labelling* ini tidak hanya berpaku pada hasil otomatis dari model saja tetapi juga melibatkan verifikasi atau pengecakan manual.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

2.5 Pembobotan Kata

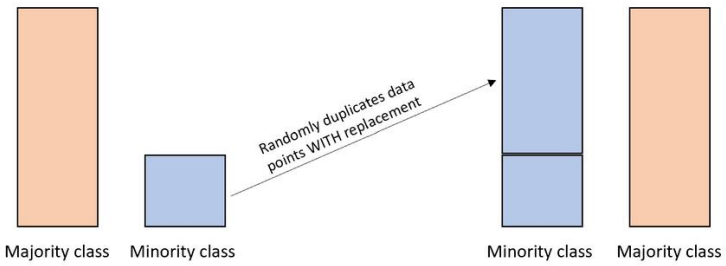
Teknik pembobotan kata pada penelitian ini yang digunakan dalam ekstraksi fitur adalah Word2Vec (*Word to Vector*). Word2Vec berfungsi sebagai pendekatan word embedding yang merubah kata menjadi vector berbentuk angka dalam dataset yang diproses oleh algoritma pembelajaran mesin [12]. Word2Vec dipilih karena kemampuannya dalam menangkap makna semantik dan sintaksis pada kata di dalam kumpulan teks. Arsitektur yang akan digunakan adalah *Continuous Bag of Words* (CBOW). CBOW dikonsepkan dalam memprediksi target kata dengan mengadopsi kata-kata terdekat sebagai konteks [13]. CBOW memiliki tiga lapisan berupa input layer, projection layer, dan output layer dimana CBOW ini terdapat pada Gambar 2.



Gambar 2. Arsitektur CBOW

2.6 *Random Oversampling*

Random Oversampling (ROS) merupakan teknik atau cara yang diimplementasikan guna memecahkan masalah ketidakseimbangan data antar kelas sentimen. *Random Oversampling* bekerja menduplikat sampel yang berasal dari kelas minoritas dengan cara acak hingga data seimbang dengan kelas mayoritas [18]. Teknik ROS ini diterapkan setelah tahap pembobotan kata atau ekstraksi fitur supaya data yang diproses adalah berbentuk representasi hasil vektor dari Word2Vec [19]. Penelitian oleh [20] menggambarkan ilustrasi dari teknik *Random Oversampling* seperti disajikan pada Gambar 3.



Gambar 3. Ilustrasi *Random Oversampling*

2.7 Split Data

Split data atau pembagian data dilakukan setelah teknik penyeimbangan dataset rampung dilakukan. Dataset dikelompokkan menjadi dua kategori yakni data pelatihan serta data pengujian dengan bobot 80:20. 80% dataset selanjutnya diimplementasikan menjadi data latih dan 20% digunakan untuk data uji. Pembagian data dilakukan dengan tujuan untuk mengalokasikan data yang akan dipakai dalam pelatihan model menggunakan data latih dan data yang digunakan untuk evaluasi performa model menggunakan data uji.

2.8 *Modeling*

Pada tahap pemodelan, data akan diklasifikasi menggunakan algoritma *Random Forest* yang merupakan metode *ensemble learning* atau gabungan dari bebeapa model yang mengembangkan beberapa pohon keputusan (*decision trees*) bersumber dari sampel acak yang diambil dari data latih. Tiap pohon akan menghasilkan prediksi lalu keputusan final diputuskan berlandaskan *voting mayority* atau suara mayoritas dari semua pohon. model ini mempunyai kehandalan dalam memberikan tingkat akurasi yang tinggi dan tahan dari overfitting [21]. Menurut penelitian oleh [22] tahap pembentukan pohon keputusan diawali dengan memperhitungkan angka *Gini Impurity*, *Average Gini Impurity* serta *Information Gain* yang terdapat pada persamaan 1,2, dan 3.

$$Gini\ Impurity\ = 1 - \sum_{i=1}^{n}(Pi)^2 \quad (1)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

$$Average\ Gini\ Impurity = \frac{n}{i}\ x\ Gini\ Impurity \quad (2)$$

$$Information\ Gain = Gini\ Impurity - Average\ Gini\ Impurity \quad (3)$$

Persamaan diatas memilki keterangan informasi berupa n sebagai jumlah *term*, i sebagai jumlah dokumen dan *Pi* sebagai probabilitas kemunculan *term*.

2.9 Evaluasi

Evaluasi hasil performa model *Random forest* yang dikombinasikan dengan Word2Vec dan *Random Oversampling* di penelitian ini adalah menerapkan *Confusion Matrix* serta *Classification Report. Confusion Matrix* menggambarkan banyaknya data yang terklasifikasi oleh model dengan benar ataupun tidak. *Confusion Matrix* terdapat empat kompenen penting yaitu TP (*True Positive),* (TN) *True Negative,* (FP) *False Positive, dan* (FN) *False Negative.* Berdasarkan keempat matriks ini dapat diperhitungkan beberapa metrik-metrik evaluasi yaitu akurasi, presisi, recall, dan *f1-score* yang biasanya ditamplikan dalam bentuk *Classsification Report* [12]. Penelitian [11] merepresentasikan matrik-matrik tersebut ke persamaan 4,5,6, dan 7.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (4)$$

$$Precision = \frac{TP}{TF+FP} \quad (5)$$

$$Recall = \frac{TF}{TF+FN} \quad (6)$$

$$F1\text{-}Score = 2x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (7)$$

Persamaan 4,5,6, dan 7 memiliki keterangan informasi berupa TP sebagai *True Positive,* TN sebagai *True Negative*, FP sebagai *False Positive* dan FN sebagai *False Negative*.

2.10 Visualisasi

Visualisasi pada penelitian ini adalah menggunakan grafik batang untuk distribusi sentimen, *heatmap confusion matrix*, grafik *classification matrix*, dan *wordcloud. Wordcloud* adalah gambaran berupa kumpulan teks atau kata-kata dengan frekuensi kemunculan tinggi maka akan semakin besar ukuran teksnya dengan penempatan yang acak pada tampilan *wordcloud* [14].

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data (*Crawling*)

Pengumpulan data dijalankan dengan mengaplikasikan teknik *crawling* pada jejaring sosial X (Twitter). *Crawling* dijalankan dengan bahasa pemrograman *Python* dan mengimplementasikan *Library Tweet Harvest* di *Google Colab*. Data *tweet* berhasil dikumpulkan sebanyak 4423 tweet dengan menggunakan kata kunci “kipk” atau “kip-k” dan data juga menggunakan rentan waktu dari Januari sampai dengan Desember 2024 atau selama tahun 2024. Kemudian hasil data *crawling* disimpan dalam format csv dan xlsx yang selanjutnya akan diolah pada analisis sentimen. Contoh dataset hasil *crawling* terdapat pada Tabel 1.

Tabel 1. Hasil Pengumpulan Data

Username	*Tweet*
AthY_ReNaita	@sbmptnfess Halah jaman sekarang juga banyak orang kaya dapet kipk karena memalsukan data. Saingan anak2 kurang mampu buat dapat kipk bukan hanya dari kalangan bawah tapi juga anakÂ² orang mampu yang nggak tau diri.
ArafahNailatul	Terimakasih karena berkat beasiswa ini saya dapat melanjutkan pendidikan saya di PTN. Saya mendapatkan fasilitas yang sama dengan mahasiswa Reguler lainnya tanpa adanya perbedaan mahasiswa kipk. Semoga jangkauan info Kipk lebih merata. #TerimakasihKIPKuliah

3.2 Hasil *Pre-processing*

Tahap *Pre-processing* dijalankan setelah proses crawling selesai. Pre-processing yang dilakukan pada penelitian ini adalah meliputi *Cleaning, Case Folding*, Normalisasi, Tokenisasi, *Stopword Removal*, dan *Stemming*. *Cleaning* merupakan proses menghilangkan emoji, simbol, angka, *username* / *mention* (@), URL dan *hashtag (#). Case Folding* ialah langkah mentransformasi keseluruhan huruf besar pada teks dan menjadikannya huruf kecil. Normalisasi merupakan tahap mengubah kalimat tidak baku menjadi kalimat baku dengan mengadopsi kamus kata baku bersumber dari *Kaggle*. Tokenisasi merupakan proses memfragmentasi suatu kalimat menjadi kata-kata atau pecahan kata. Proses *Cleaning, Case Folding*, Normalisasi, dan Tokenisasi dijalankan dengan menggunakan pendekatan *Regex (Regular Expressions).* Kemudian *Stopword Removal* merupakan tahap menghilangkan kata umum yang tidak bermakna.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

Stemming merupakan tahap menyederhanakan kata atau menghapus imbuhan. *Stopword Removal* dan *Stemming* dijalankan dengan *Library Sastrawi.* Temuan dari semua tahapan *Pre-processing* tersebut terdapat pada Tabel 2.

Tabel 2. Hasil *Pre-processing*

Tahapan	*Pre-processing*
Full Text	@sbmptnfess Halah jaman sekarang juga banyak orang kaya dapet kipk karena memalsukan data. Saingan anak2 kurang mampu buat dapat kipk bukan hanya dari kalangan bawah tapi juga anakÂ² orang mampu yang nggak tau diri.
Cleaning	Halah jaman sekarang juga banyak orang kaya dapet kipk karena memalsukan data Saingan anak kurang mampu buat dapat kipk bukan hanya dari kalangan bawah tapi juga anak orang mampu yang nggak tau diri
Case Folding	halah jaman sekarang juga banyak orang kaya dapet kipk karena memalsukan data saingan anak kurang mampu buat dapat kipk bukan hanya dari kalangan bawah tapi juga anak orang mampu yang nggak tau diri
Normalisasi	halah jaman sekarang juga banyak orang kaya dapat kipk karena memalsukan data saingan anak kurang mampu buat dapat kipk bukan hanya dari kalangan bawah tapi juga anak orang mampu yang tidak tau diri
Tokenisasi	['halah', 'jaman', 'sekarang', 'juga', 'banyak', 'orang', 'kaya', 'dapat', 'kipk', 'karena', 'memalsukan', 'data', 'saingan', 'anak', 'kurang', 'mampu', 'buat', 'dapat', 'kipk', 'bukan', 'hanya', 'dari', 'kalangan', 'bawah', 'tapi', 'juga', 'anak', 'orang', 'mampu', 'yang', 'tidak', 'tau', 'diri']
Stopword Removal	['halah', 'jaman', 'sekarang', 'banyak', 'orang', 'kaya', 'kipk', 'memalsukan', 'data', 'saingan', 'anak', 'kurang', 'mampu', 'buat', 'kipk', 'bukan', 'kalangan', 'bawah', 'anak', 'orang', 'mampu', 'tau', 'diri']
Stemming	halah jaman sekarang banyak orang kaya kipk palsu data saing anak kurang mampu buat kipk bukan kalang bawah anak orang mampu tau diri

3.3 Hasil *Labelling*

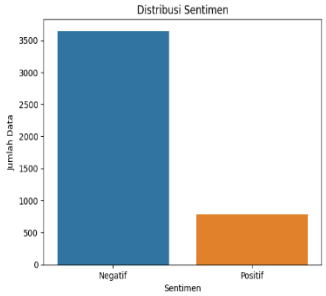
Pada tahap *Labeling* mengimplementasikan model yang telah dibentuk yakni *pre-trained* model berasal dari *Hugging Face Website* bernama model SMSA atau *bert base Indonesian* 1.5G *sentiment analysis* yang akan mengklasifikasikan sentimen secara otomatis ke dalam kategori positif dan negatif. Setelah model berhasil diterapkan, maka akan dilakukan pengecekan manual terhadap hasil prediksi model untuk memastikan keakuratan dan kesesuaian konteks.

Setelah berhasil menerapkan model prediksi, peneliti melakukan pengecekan secara satu persatu dan ternyata masih terdapat beberapa kalimat yang salah prediksi atau tidak sesuai makna, sehingga harus dikoreksi ke label yang seharusnya. Hal ini menunjukkan bahwa model prediksi pelabelan ini tidak 100% akurat dan membutuhkan tahap pengecakan setelahnya yang memakan banyak waktu. Setelah proses pengecekan dan koreksi selesai, peneliti dapat memastikan bahwa hasil pelabelan sudah benar dan akurat. Dengan demikian, tahap pelabelan ini tidak hanya berpaku pada hasil prediksi otomatis saja tetapi juga melibatkan verifikasi atau pengecekan manual untuk menghasilkan dataset yang lebih sesuai konteks dan lebih akurat. Hasil *Labelling* terdapat pada Tabel 3.

Tabel 3. Hasil *Labelling*

Tweet	Label
halah jaman sekarang juga banyak orang kaya dapet kipk karena memalsukan data saingan anak kurang mampu buat dapat kipk bukan hanya dari kalangan bawah tapi juga anak orang mampu yang nggak tau diri	Negatif
terimakasih karena berkat beasiswa ini saya dapat melanjutkan pendidikan saya di ptn saya mendapatkan fasilitas yang sama dengan mahasiswa reguler lainnya tanpa adanya perbedaan mahasiswa kipk semoga jangkauan info kipk lebih merata	Positif

Distribusi hasil *Labelling* digambarkan dalam grafik batang berbentuk *bar chart* untuk mengetahui seberapa banyak kalimat yang dikategorikan ke dalam label positif dan juga negatif seperti pada Gambar 4 yang merepresentasikan distribusi sentimen.



Gambar 4. Distribusi Sentimen

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

Gambar 4 menunjukkan hasil distribusi sentimen analisis program KIP-K yang dikategorikan ke dalam sentimen negatif dan sentimen positif. Distribusi menunjukan sentimen negatif lebih dominan berjumlah 3643 data dan sentimen positif berjumlah 780 data.

3.4. Hasil Pembobotan Kata

Pembobotan kata dalam penelitian ini adalah mengimplementasikan Word2Vec. Word2Vec diterapkan untuk tujuan mengubah teks menjadi vektor berbasis numerik dengan dimensi 100. Pada tiap kata dalam teks dilakukan pelatihan menggunakan konteks jendela (*window=5*) dan *min_count*=1 serta *workers* = 4. Rata-rata vektor dari keseluruhan kata pada teks akan menghasilkan merepresentasikan teks. Proses Word2Vec diterapkan pada seluruh data tanpa memisahkan dataset ke data pelatihan dan dan data pengujian terlebih dahulu yang bermaksud agar model berhasil menciptakan representasi kata yang menyeluruh dengan konteks semua dataset. Contoh representasi vektor dari kata-kata penting pada teks atau corpus diperlihatkan pada Gambar 5.

	dim_0	dim_1	dim_2	dim_3	dim_4	dim_5	dim_6	dim_7	dim_8	dim_9
kipk	-0.374902	0.731466	0.437253	-0.055619	0.227647	-1.597956	0.491978	2.260435	-0.744021	-0.295200
mahasiswa	-0.346822	0.643539	0.363131	-0.087032	0.257182	-1.395975	0.436335	1.992848	-0.634406	-0.253829
beasiswa	-0.327344	0.605122	0.355108	-0.075846	0.212535	-1.337276	0.411714	1.874727	-0.584789	-0.256867

Gambar 5. Representasi Vektor Word2Vec

Gambar 5 menampilkan contoh kata-kata penting seperti “kipk”, “mahasiswa”, dan “beasiswa dengan angka vektor pada tiap dimensinya. Pada contoh ditampikan angka vektor dari 10/100 dimensi yang dibentuk.

3.5 Hasil *Random Oversampling*

Berdasarkan gambar 4 distribusi sentimen yang menunjukan dominan sentimen negatif dibanding sentimen positif mengindikasikan ketidakseimbangan data antar kelasnya. Untuk mengatasi ketidakseimbangan data maka diterapkanlah teknik *Oversampling* menggunakan *Random Oversampling* (ROS). Teknik ini dijalankan dengan cara menduplikasi data pada kelas minoritas (positif) dan menjadikannya seimbang dengan jumlah pada kelas mayoritas (negatif) yang bertujuan untuk meningkatkan kinerja model klasifikasi. Representasi data yang telah diproses ditampilkan pada Tabel 4.

Tabel 4. Hasil Random Oversampling

Sentimen	Sebelum ROS	Sesudah ROS
Positif	780	3643
Negatif	3643	3643

3.6 Pemodelan *Random Forest*

Pembentukan model pada penelitian ini adalah menggunakan algoritma *Random Forest Classifier.* Sebelum menerapkan algoritma, dilakukan terlebih dahulu pembagian data dari hasil penerapan teknik *Oversampling* sebelumnya. Data dibagi menjadi data latih dan data uji menggunakan metode *train-test split* dengan perbandingan 80:20. Split data ini diterapkan supaya data yang dipakai pada pelatihan model berbeda dengan data yang dipakai pada saat pengujian. Pada tahap *modelling,* data latih yang berjumlah 80% diimplementasikan pada pelatihan model, sisanya yaitu 20% dataset uji akan diterapkan pada pengujian model guna mendapatkan hasil performa dari klasifikasi model. Penerapan algoritma *Random Forest Classifier* disajikan pada Gambar 6.

```
[ ]  from sklearn.ensemble import RandomForestClassifier

     rf = RandomForestClassifier(random_state=42)
     rf.fit(X_train, y_train)
```

Gambar 6. Algoritma *Random Forest*

Gambar 6 merepresentasikan algoritma *Random Forest* yang mengimplementasikan *library scikit-learn. Random Forest Classifier* membentuk objek *classifier* (rf) dengan parameter *random_state*=42. Parameter ini diterapkan supaya hasil *training* pada saat dijalankan akan tetap konsisten. X_train sebagai fitur atau input dan y_train sebagai label atau target dalam melatih model *Random Forest.*

3.7 Evaluasi Model

Berdasarkan hasil evaluasi performa model menggunakan 20% data uji yang telah melalui proses *balancing* atau *oversampling* sebanyak 728 sentimen positif dan 730 sentimen negatif dengan total 1458 sampel, merepresetasikan

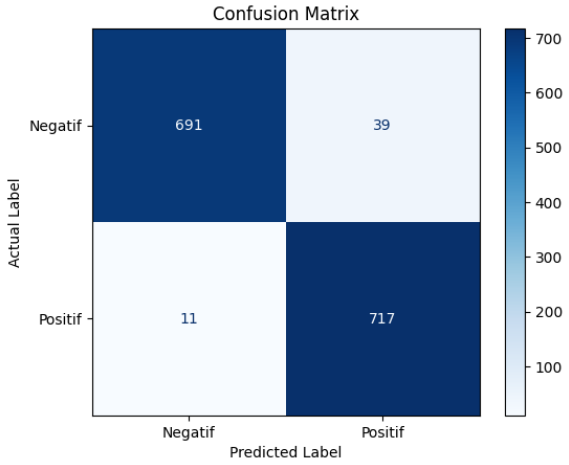
Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

bahwa model *Random Forest* yang dikombinasikan dengan Word2Vec dan *Random Oversampling* menghasilkan performa dengan sangat baik. Model memberikan hasil akurasi performa sebesar 96.84%, dan angka presisi, *recall*, serta *f1-score* berdasarkan makro sebesar 97%. Lebih lanjut analisis hasil performa model Random Forest di penelitian ini akan menerapkan *Confusion Matrix* dan *Classification Report.*

Tabel 5. *Confusion Matrix*

Class	*Predicted Negative*	*Predicted Positive*
Actual Negative	691	39
Actual Positive	11	717

Hasil analisis pada Tabel 5 menunjukkan *Confusion Matrix* yang mencakup komponen (TN) *True Negative*, (FP) *False Positive*, (FN) *False Negative*, dan (TP) *True Positive*. Dari 730 sampel negatif, sebanyak 691 (TN) menunjukkan jumlah opini negatif yang berhasil diprediksi dengan benar dan sebanyak 39 (FP) menunjukkan opini negatif yang keliru diprediksi sebagai opini positif. Kemudian dari 728 sampel positif, sebanyak 717 (TP) menunjukkan opini positif yang berhasil diprediksi dengan benar dan sebanyak 11 (FN) menunjukkan opini positif yang salah diprediksi menjadi negatif oleh model. Hasil tersebut membuktikan bahwa performa klasifikasi positif dan negatif adalah sangat baik dikarenakan jumlah *True Positive* (717) dan True Negative (691) berjumlah cukup tinggi sedangkan untuk jumlah *False Positive* (39) dan *False Negative* (11) relatif di angka yang masih rendah.



Gambar 7. *Heatmap Confusion Matrix*

Gambar 7 merupakan representasi *Confusion Matrix* yang berbentuk *Heatmap. Heatmap* yang menunjukan warna gelap adalah hasil yang akurat sedangkan yang berwarna terang adalah hasil prediksi yang salah [23]. Berlandaskan *heatmap* ini menampilkan bahwa model mempunyai performa akursi antar kelas positif dan negatif yang hampir seimbang.

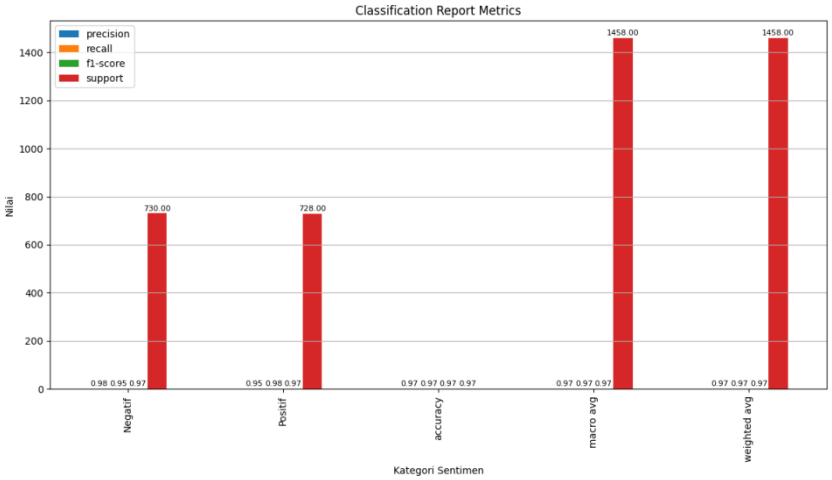
Hasil metrik akurasi, presisi, *recall* dan *f1-score* sebelumnya, disajikan ke dalam bentuk *Classification Report* pada Tabel 6.

Tabel 6. *Classification Report*

Sentiment	*Precision*	*Recall*	*F1-Score*	*Support*
Negative	0.98	0.95	0.97	730
Positive	0.95	0.98	0.97	728
Accuracy			0.97	1458
Macro Avg	0.97	0.97	0.97	1458
Weighted Avg	0.97	0.97	0.97	1458
Accuracy	0.9657064471879286			

Tabel 6 menunjukkan *Classification Report* dari model *Random Forest* yang diterapkan dalam klasifikasi sentimen terhadap program KIP-K yang memberikan hasil performa dengan sangat baik. Pada sentimen negatif, model memberikan angka presisi sebesar 0,98 yang berarti 98% prediksi klasifikasi negatif adalah memang benar negatif. Nilai *recall* dengan angka 0,95 yang berarti model berhasil mengidentifikasi 95% dari seluruh data negatif. Kemudian *f1-score* menunjukkan angka 0,97 yang artinya model mempunyai kemampuan yang seimbang dalam mengklasifikasi dan memprediksi sentimen negatif dengan tepat. Pada sentimen positif, model memberikan angka presisi mencapai 0,95 dimana 95% prediksi positif dipastikan benar. Angka *recall* sebesar 0,98 yang artinya model 98% berhasil mengenali opini positif. *F1-score* dengan angka 0,97 menunjukkan bahwa model sama baiknya dengan model sentimen negatif dalam kemampuan mengklasifikasikan opini positif. Selain dalam bentuk tabel distribusi, pada Gambar 8 juga menampilkan *Classification Report Metrics* ke bentuk grafik batang.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

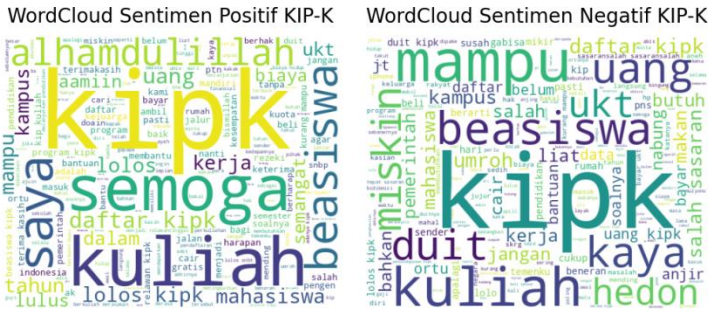


Gambar 8. *Clasification Report Metrics*

Berdasarkan dari hasil semua metrik yang diperoleh pada tahap evaluasi ini, model memberikan kinerja atau performa dengan sangat baik dalam mengklasifikasikan kedua kelas sentimen. Akurasi keseluruhan mencapai angka 0,97 atau 97% membuktikan bahwa model mempunyai kehandalan dalam memprediksi opini positif dan juga negatif. Nilai tersebut menggambarkan keakuratan prediksi pada keseluruhan dengan kesalahan yang sangat minim. Kemudian nilai *macro average* dan *weighted average* menampilkan hasil performa yang cukup imbang dan tanpa tendensi di salah satu kelas. Secara keseluruhan, model membuktikan performa yang stabil dalam mengidentifikasi kedua sentimen.

3.8 Visualisasi *Wordcloud*

Wordcloud digunakan dalam penelitian ini untuk menampilkan dan menganalisis data teks dengan frekuensi kemunculan tinggi. *Wordcloud* ditampilkan dalam 2 gambar yaitu *wordcloud* sentimen positif dan *wordcloud* sentimen negative seperti pada Gambar 9.



Gambar 9. Visualisasi *Wordcloud*

Gambar 9 merepresentasikan *wordcloud* sentimen negatif dan sentimen positif. Hasil menunjukkan terdapat perbedaan antara kedua *Wordcloud* tersebut. *Wordcloud* sentimen positif menyajikan kata-kata dengan frekuensi kemunculan tinggi seperti “kipk”, “beasiswa”, “kuliah”, “semoga”, “Alhamdulillah”. Kata-kata ini menunjukkan opini positif yang mengandung makna harapan serta rasa syukur terhadap adanya program atau beasiswa kuliah kipk. Sedangkan pada *wordcloud* sentimen negatif menampilkan kata-kata dengan frekuensi kemunculan tinggi seperti “kipk”, “beasiswa”, “kuliah”, “uang” atau ”duit”, “miskin”, “kaya”, “mampu”, “hedon”, “salah sasaran”. Kata kata tersebut mengarah ke opini negatif dimana beasiswa kuliah kipk ini terdapat indikasi kesenjangan ekonomi dan ketidakpuasan terhadap alokasi program. Sehingga perbedaan ini menunjukkan bahwa walaupun ada yang mendukung program KIP-K, namun masih banyak keresahan yang perlu diatasi terutama oleh pemerintah atau pemangku kepentingan untuk membenahi mekanisme serta meningkatkan efektivitas program KIP-K di Indonesia.

4. KESIMPULAN

Penelitian ini menerapkan algoritma *Random Forest* dengan kombinasi Word2Vec dan *Random Oversampling* (ROS). Model di evaluasi dengan bagian data 80:20 dengan data pelatihan sejumlah 80% serta data pengujian sejumlah 20%. Hasil distribusi data dengan total 4423 data yang digunakan menunjukkan sejumlah 3643 data sentimen negatif dan 780 data berupa sentimen positif menunjukkan ketidakseimbangan data yang menjadi tantangan dalam penelitian ini, oleh karena itu diterapkanlah teknik *Random Oversampling* (ROS). Penerapan model dilakukan pada penelitian ini

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

memberikan hasil performa klasifikasi dengan sangat baik. Akurasi yang dicapai yaitu sebesar 96,57%, selain itu presisi, *recall,* dan *f1-score* rata-rata menghasilkan angka sebesar 0,97 atau 97%. Angka ini membuktikan bahwa model memiliki kemampuan dalam membedakan dan mengenali sentimen negatif serta sentimen positif dengan akurat dan seimbang. Algoritma *Random Forest* yang dikombinasikan dengan Word2Vec dan *Random Oversampling* terbukti mampu memberikan akurasi tinggi dan mampu mengatasi ketidakseimbangan data. Hasil dari eksplorasi data sentimen menunjukkan bahwa persepsi publik terhadap program KIP-K pada media sosial X (*Twitter*) didominasi oleh sentimen negatif sebanyak 82% dari keseluruhan data yang didapatkan. Terbukti melalui visualisasi *wordcloud* sentimen negatif yang mengandung kata-kata kritikan berkaitan dengan proses seleksi, alokasi, dan transparansi program. Temuan ini diharapkan mampu membentuk salah satu bahan dalam membantu pemerintah atau pemangku jabatan untuk memahami persepsi publik, mengevaluasi, dan meningkatkan efektivitas program KIP-K. Selain itu penelitian ini juga menghadirkan kontribusi baru yang substansial dalam menafsirkan persepsi publik terkait program KIP-K. Lebih lanjut, penelitian ini merekomendasikan penggunaan model lain untuk tahap labeling, serta eksplorasi penggunaan algoritma *deep learning* yang lebih kompleks seperti *Long Short Term Memory* (LSTM) dan *Bidirectional Encoder Representations from Transformers* (BERT). Penelitian mendatang juga dapat membandingkan beberapa algortima sekaligus untuk mengetahui tingkat performa terbaik. Selain itu teknik *balancing* atau penyeimbangan data seperti *hybrid sampling* dan *undersampling*, serta pengumpulan data dari berbagai *platform* media juga direkomendasikan guna memaksimalkan dan mendapatkan perbandingan performa model.

REFERENCES

[1] E. K. Martins and N. T. Toletina, "Evaluasi Pelaksanaan Kebijakan Program KIP-K Di Indonesia," *Prof. J. Komun. dan Adm. Publik*, vol. 11, no. 1, pp. 331–340, 2024, doi: 10.37676/professional.v11i1.6166.

[2] D. Pramudita, Y. Akbar, and T. Wahyudi, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Sentiment Analysis of the Indonesian Smart College Card Program on Social Media X Using the Naive Bayes Algorithm Analisis Sentimen Terhadap Program Kartu Indonesia Pintar Kuliah Pada Med," *Malcom*, vol. 4, no. October, pp. 1420–1430, 2024.

[3] "Waktunya Daftar KIP Kuliah 2025," *Portal Informasi Indonesia*, 2025. https://indonesia.go.id/kategori/pendidikan/8895/waktunya-daftar-kip-kuliah-2025?lang=1 (accessed Mar. 03, 2025).

[4] L. Z. Damayanti, W. D. Yuniarti, and S. Nur'aini, "Sistem Pendukung Keputusan Penerima Bantuan Kartu Indonesia Pintar dengan Metode Weighted Product," *J. Transform.*, vol. 20, no. 2, pp. 92–104, 2023, doi: 10.26623/transformatika.v20i2.5877.

[5] "Yuk, Mari Daftar Beasiswa KIP Kuliah 2025," *Portal Informasi Indonesia*, 2025. https://indonesia.go.id/kategori/pendidikan/8997/yuk-mari-daftar-beasiswa-kip-kuliah-2025?lang=1 (accessed Mar. 03, 2025).

[6] M. H. Arfian *et al.*, "Analisis Sentimen Pada Media Sosial Menggunakan Metode Support Vector Machine," *J. Ilmu Tek. dan Komput.*, vol. 09, 2025, doi: 10.22441/jitkom.v9i1.001.

[7] A. D. Larasati;, D. Dinda;, N. A. Aidah;, R. Gustiputri;, and S. N. R. Isyak, "Analisis Kebijakan Program Beasiswa Kartu Indonesia Pintar-Kuliah (Kip-K) Di Universitas Diponegoro," *J. Ilmu Adm. dan Stud. Kebijak.*, vol. 5, no. 1, pp. 1–22, 2022, doi: 10.48093/jiask.v5i1.91.

[8] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *J. Inform.*, vol. 15, no. 1, p. 45, 2021, doi: 10.26555/jifo.v15i1.a20111.

[9] H. C. Husada and A. S. Paramita, "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *Teknika*, vol. 10, no. 1, pp. 18–26, 2021, doi: 10.34148/teknika.v10i1.311.

[10] I. Amelia and F. M. Sarimole, "Analisis Sentimen Tanggapan Pengguna Media Sosial X Terhadap Program Beasiswa KIP-Kuliah dengan Menggunakan Algoritma Support Vector Machine (SVM)," *J. Indones. Manaj. Inform. dan Komun.*, vol. 5, no. 3, 2024, doi: doi: 10.35870/jimik.v5i3.990.

[11] H. Ali and N. Hendrastuty, "Comparison of Naïve Bayes Classifier, Support Vector Machine, Random Forest Algorithms for Public Sentiment Analysis of KIP-K Program on Twitter," *J. Tek. Inform.*, vol. 5, no. 6, pp. 1701–1712, 2024.

[12] M. Irfani and S. Khomsah, "Analisis Sentimen Berbasis Aspek pada EDOM Pembelajaran Menggunakan Metode CNN dan Word2vec," *justin J. Sist. dan Teknol. Inf.*, vol. 12, no. 3, 2024, doi: 10.26418/justin.v12i3.75610.

[13] D. I. Af'idah, D. Dairoh, S. F. Handayani, and R. W. Pratiwi, "Pengaruh Parameter Word2Vec terhadap Performa Deep Learning pada Klasifikasi Sentimen," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, 2021, doi: 10.30591/jpit.v6i3.3016.

[14] R. Aryanti, T. Misriati, and A. Sagiyanto, "Analisis Sentimen Aplikasi Primaku Menggunakan Algoritma Random Forest dan SMOTE untuk Mengatasi Ketidakseimbangan Data," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 218–227, 2023, doi: 10.47065/josyc.v5i1.4562.

[15] R. R. Salam, M. F. Jamil, Y. Ibrahim, R. Rahmaddeni, S. Soni, and H. Herianto, "Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 27–35, 2023, doi: 10.57152/malcom.v3i1.590.

[16] R. A. Sulthana, A. K. Jaithunbi, H. Harikrishnan, and V. Varadarajan, "Sentiment Analysis on Movie Reviews Dataset Using Support Vector Machines and Ensemble Learning," *Int. J. Inf. Technol. Web Eng.*, vol. 17, no. 1, 2022, doi: 10.4018/IJITWE.311428.

[17] A. Hendra and F. Fitriyani, "Analisis Sentimen Review Halodoc Menggunakan Nai ̈ve Bayes Classifier," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 2, pp. 78–89, 2021, doi: 10.14421/jiska.2021.6.2.78-89.

[18] A. Elhan, M. K. D. Hardhienata, H. Yeni, S. Wijaya Hartono, And J. Adisantoso, "Analisis Sentimen Pengguna Twitter terhadap Vaksinasi COVID-19 di Indonesia menggunakan Algoritme Random Forest dan BERT Sentiment Analysis of Twitter Users on COVID-19 Vaccines in Indonesia using Random Forest and BERT Algorithms," *J. Ilmu Komput. Agri-informatika*, vol. 9, no. 2, pp. 199–211, 2022.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 264−274
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7366

[19] U. Ependi and N. A. Ahmad, "A Novel Hybrid Classification on Urban Opinion Using ROS-RF: A Machine Learning Approach," *J. Penelit. Pendidik. IPA*, vol. 10, no. 8, pp. 5816–5824, 2024, doi: 10.29303/jppipa.v10i8.8042.
[20] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.
[21] T. D. Putra and D. Oktafiani, "Klasifikasi Sentimen Postingan Sosial Media Menggunakan Machine Learning Random Forest dan Naïve Bayes," vol. 5, pp. 2338–2347, 2025.
[22] T. F. Basar, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen Pengguna Twitter terhadap Pembayaran Cashless menggunakan Shopeepay dengan Algoritma Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, pp. 1426–1433, 2022.
[23] A. S. Muliana, D. Lestarini, and S. P. Raflesia, "Analysis of Public Sentiment on Election Results using Naive Bayes in Social Media X," *Sist. J. Sist. Inf.*, vol. 13, no. 6, pp. 2467–2478, 2024.