

## **COMPARISON OF NAÏVE BAYES CLASSIFIER, SUPPORT VECTOR MACHINE, RANDOM FOREST ALGORITHMS FOR PUBLIC SENTIMENT ANALYSIS OF KIP-K PROGRAM ON TWITTER**

Humaidi Ali<sup>\*1</sup>, Nirwana Hendrastuty<sup>2</sup>

<sup>1,2</sup>Information System, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Indonesia

Email: <sup>1</sup>[humaidi\\_ali@teknokrat.ac.id](mailto:humaidi_ali@teknokrat.ac.id), <sup>2</sup>[nirwanahendrastuty@teknokrat.ac.id](mailto:nirwanahendrastuty@teknokrat.ac.id)

(Article received: July 15, 2024; Revision: August 14, 2024; published: December 29, 2024)

### **Abstract**

The Kartu Indonesia Pintar Kuliah (KIP-K) program has become a hot topic of conversation on social media Twitter, with various public sentiments regarding its implementation. The program is regulated through Minister of Education and Culture Regulation (Permendikbud) No. 10/2020, which notes an increase in the number of recipients from 552,706 in 2020 to 985,577 in 2024. However, controversy has arisen due to the alleged misuse of KIP-K funds by some influencers to support lavish lifestyles. This study aims to compare the performance of Naive Bayes, Support Vector Machine, and Random Forest algorithms in classifying public sentiment towards the KIP-K program. The research dataset was obtained from Twitter with a total of 6,842 tweets collected using crawling techniques in the time span of 2023 to 2024. The dataset was then processed through the preprocessing stage to produce clean data. The three algorithms were tested to evaluate the accuracy of each model in predicting public sentiment. The test results show that the Random Forest algorithm has the best performance with 100% accuracy, followed by Support Vector Machine with 99% accuracy, and Naive Bayes with 91% accuracy after optimization (SMOTE). Based on these findings, Random Forest proved to be the most effective algorithm in classifying sentiments related to the KIP-K program. It is hoped that the results of this research can help the management of the KIP-K program to be more targeted by providing a better understanding of public perception.

**Keywords:** sentiment analysis, KIP-K, naive bayes, random forest, SMOTE, SVM, twitter.

## **KOMPARASI ALGORITMA NAÏVE BAYES CLASSIFIER, SUPPORT VECTOR MACHINE, RANDOM FOREST UNTUK ANALISIS SENTIMEN PUBLIK PROGRAM KIP-K DI TWITTER**

### **Abstrak**

Program Kartu Indonesia Pintar Kuliah (KIP-K) telah menjadi perbincangan hangat di media sosial Twitter, dengan berbagai sentimen publik mengenai pelaksanaannya. Program ini diatur melalui Peraturan Menteri Pendidikan dan Kebudayaan (Permendikbud) Nomor 10 Tahun 2020, yang mencatat peningkatan jumlah penerima dari 552.706 pada tahun 2020 menjadi 985.577 pada tahun 2024. Meskipun demikian, kontroversi muncul akibat dugaan penyalahgunaan dana KIP-K oleh beberapa *influencer* untuk mendukung gaya hidup mewah. Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Naive Bayes*, *Support Vector Machine*, dan *Random Forest* dalam mengklasifikasikan sentimen publik terhadap program KIP-K. Dataset penelitian diperoleh dari Twitter dengan total 6.842 tweet yang dikumpulkan menggunakan teknik *crawling* dalam rentang waktu 2023 hingga 2024. Dataset tersebut kemudian diproses melalui tahapan *preprocessing* untuk menghasilkan data yang bersih. Pengujian ketiga algoritma dilakukan untuk mengevaluasi tingkat akurasi masing-masing model dalam memprediksi sentimen publik. Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memiliki performa terbaik dengan akurasi sebesar 100%, diikuti oleh *Support Vector Machine* dengan akurasi 99%, dan *Naive Bayes* dengan akurasi 91% setelah dilakukan optimasi (SMOTE). Berdasarkan temuan ini, *Random Forest* terbukti sebagai algoritma paling efektif dalam mengklasifikasikan sentimen terkait program KIP-K. Diharapkan hasil penelitian ini dapat membantu pengelolaan program KIP-K agar lebih tepat sasaran dengan memberikan pemahaman yang lebih baik mengenai persepsi publik.

**Kata kunci:** analisis sentimen, KIP-K, naive bayes, random forest, SMOTE, SVM, ,twitter.

## 1. PENDAHULUAN

Pendidikan adalah salah satu alat utama untuk meningkatkan kualitas sumber daya manusia (SDM), meskipun pendidikan merupakan investasi jangka panjang yang manfaatnya tidak dapat dirasakan secara langsung. Proses pendidikan membutuhkan waktu lama dan biaya signifikan[1]. Setiap negara berupaya menyediakan pendidikan berkualitas guna meningkatkan kesejahteraan masyarakat. Untuk mencapai tujuan ini, pemerintah menyelenggarakan berbagai program, termasuk beasiswa. Beasiswa adalah bentuk dukungan keuangan bagi siswa atau mahasiswa untuk mendukung pendidikan mereka. Diharapkan, beasiswa ini dapat meningkatkan kualitas SDM serta memperluas akses pendidikan bagi kelompok masyarakat yang rentan atau kurang mampu[2].

Salah satu bentuk beasiswa di Indonesia adalah Kartu Indonesia Pintar Kuliah (KIP Kuliah) yang diatur dalam Peraturan Menteri Pendidikan dan Kebudayaan (PERMENDIKBUD) No. 10 Tahun 2020. KIP Kuliah adalah bantuan biaya pendidikan yang bertujuan untuk memberikan kesempatan bagi siswa dari keluarga miskin atau rentan agar tetap dapat mengenyam pendidikan tinggi[3]. Program ini diharapkan dapat berkontribusi pada peningkatan kualitas SDM di Indonesia sehingga mampu bersaing secara global dan mendukung pertumbuhan ekonomi nasional[4].

Dalam konteks ekonomi Indonesia, masih banyak siswa yang berasal dari keluarga kurang mampu namun berprestasi secara akademik. Pemerintah menyediakan beasiswa untuk kelompok ini dengan harapan bahwa pendidikan mereka akan bermanfaat bagi masyarakat luas di masa depan[5]. Data penerima bantuan KIP Kuliah selama periode 2020-2024 menunjukkan peningkatan yang konsisten setiap tahunnya. Pada tahun 2020, sebanyak 552.706 mahasiswa menerima bantuan ini, dan jumlahnya bertambah menjadi 674.187 pada tahun 2021. Meskipun pandemi Covid-19 melanda, jumlah penerima terus meningkat hingga mencapai 780.014 mahasiswa pada tahun 2022 dan 913.636 penerima pada tahun 2023. Pada tahun 2024, angka penerima KIP Kuliah mendekati 1 juta, mencapai 985.577 mahasiswa[6].

Di era digital, media sosial berfungsi sebagai platform utama bagi masyarakat untuk mengungkapkan pendapat tentang berbagai isu, termasuk kebijakan pemerintah. Salah satu platform tersebut adalah media sosial *Twitter*, yang memiliki 206 juta pengguna aktif di seluruh dunia, termasuk 18,4 juta pengguna di Indonesia. Hal ini menjadikannya salah satu platform terbesar untuk diskusi mengenai program KIP Kuliah[7]. Seperti sebuah pemberitaan yang sempat viral di media sosial *Twitter*, beberapa waktu lalu, mengenai penyalahgunaan KIP Kuliah oleh sejumlah *influencer* yang menggunakan dana tersebut untuk berfoya-foya.

Hal ini memicu kekecewaan di kalangan netizen, karena beberapa penerima KIP terlihat menggunakan uang beasiswanya untuk berpesta, bahkan hingga ke luar negeri. Berita ini berawal dari kemarahan netizen terhadap perilaku beberapa oknum yang memamerkan gaya hidup mewah di media sosial, yang kemudian menyebar luas dan menjadi viral di platform tersebut[8].

Beberapa penelitian terkait masalah dan solusi serupa telah dilakukan. Salah satunya adalah studi yang menganalisis sentimen terhadap Program KIP-K di media sosial *Twitter*. Sentimen tersebut diklasifikasikan menjadi sentimen positif dan negatif. Dengan menggunakan algoritma *Naïve Bayes* dan pendekatan CRISP-DM, penelitian ini mencapai tingkat akurasi sebesar 84,99%, dengan presisi sentimen positif sebesar 83,54% dan presisi sentimen negatif sebesar 87,25% [6].

Penelitian lain melakukan analisis sentimen masyarakat terhadap penerimaan Beasiswa KIP Kuliah menggunakan metode *Support Vector Machine* (SVM). Hasil penelitian menunjukkan bahwa metode ini mencapai akurasi sebesar 81,82%, dengan presisi 80%, *recall* 82%, dan *F1-Score* 78%. Hal ini menunjukkan bahwa metode tersebut cukup efektif dalam penelitian terkait penerimaan Beasiswa KIP Kuliah[9].

Penelitian sebelumnya mengimplementasikan algoritma *Naïve Bayes* dalam proses klasifikasi untuk penentuan penerima beasiswa KIP Kuliah. Algoritma *Naïve Bayes* digunakan untuk mengklasifikasikan data calon penerima beasiswa berdasarkan beberapa kriteria seleksi, yakni hasil seleksi berkas, ujian online, dan wawancara. Menggunakan dataset berjumlah 100 data calon penerima, penelitian ini mencapai tingkat akurasi 98,89% berdasarkan perhitungan *confusion matrix*[10].

Penelitian lain menggunakan analisis sentimen untuk mengklasifikasikan persepsi publik terhadap program Merdeka Belajar Kampus Merdeka melalui komentar di media sosial, dengan algoritma *Support Vector Machine* (SVM) dan teknik fitur TF-IDF. Hasil klasifikasi menunjukkan 287 komentar netral, 242 komentar positif, dan 91 komentar negatif. Model SVM dengan *kernel linear* memiliki akurasi 82,25%, presisi 79,12%, dan *recall* 77,70%. Untuk sentimen negatif, model topik mencapai akurasi 80,79%, presisi 78,76%, dan *recall* 66,46% menggunakan *10-fold cross validation*[11].

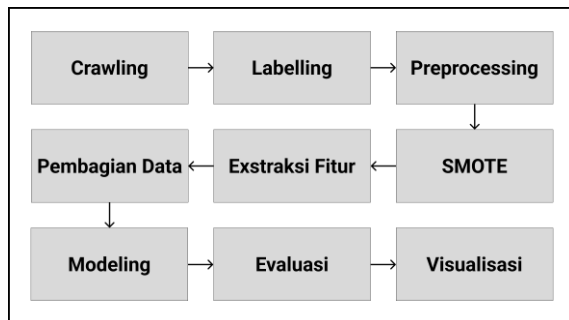
Adapun penelitian yang menganalisis sentimen pengguna terhadap aplikasi Ruangguru dengan membandingkan tiga model klasifikasi *Naive Bayes*, *Random Forest*, dan *Support Vector Machine*. Hasil menunjukkan bahwa model *Random Forest* memiliki performa terbaik dengan akurasi 97,16% dan skor AUC 0,996 menggunakan validasi silang. Model *Support Vector Machine* mencapai akurasi 96,01% dengan AUC 0,543, sedangkan *Naive Bayes* mencapai akurasi 94,16% dengan AUC 0,999[12].

Selanjutnya, peineilitian lain meimbandingkan algoritma *Random Forest*, *Support Vector Machine* (SVM), dan *Naïvei Bayeis* dalam analisis seintimein teirhadap stigma keiseihatan meintal. Hasilnya, model SVM meincapai akurasi seibeisar 86,11%, seimeintara *Random Forest* meimiliki akurasi 82,55%, dan *Naïvei Bayeis* meincatat akurasi seibeisar 78,19% [13].

Beirdasarkan masalah yang telah diuraikan seibeilumnya seirta peineilitian-peineilitian teirkait, peineilitian ini beirtujuan untuk meimbandingkan algoritma *Naïvei Bayeis*, *Support Vector Machine* (SVM), dan *Random Forest* dalam analisis seintimein teirhadap program KIP Kuliah. Tidak seperti penelitian sebelumnya yang lebih fokus pada akurasi satu algoritma seperti *Naïve Bayes* atau SVM dalam analisis sentimen program KIP Kuliah. Peineilitian ini diharapkan dapat meimbeirikan wawasan meingeinai pola seintimein masyarakat yang beirmanfaat untuk peirbaikan dan peingeilolaan program KIP Kuliah di masa deipan.

**2. METODE PENELITIAN**

Pada Metode penelitian dalam studi ini dirancang untuk memberikan pemahaman mendalam terkait sentimen terhadap program KIP-K yang diambil dari media sosial *Twitter*, dengan membandingkan kinerja algoritma *Random Forest*, SVM, dan *Naive Bayes*. Pada Gambar 1, dijelaskan tahapan-tahapan yang meliputi pengumpulan data (*crawling*), pelabelan, praproses data, penerapan SMOTE, ekstraksi fitur, pembagian data, pemodelan, evaluasi, dan visualisasi. Untuk gambaran lengkap mengenai tahapan metode penelitian ini, dapat dilihat pada Gambar 1.

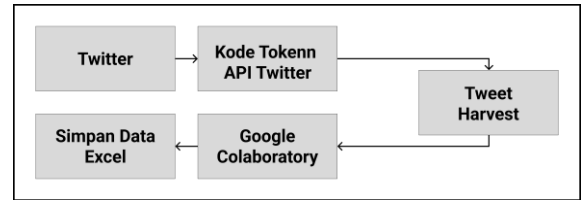


Gambar 1. Metode Penelitian

**2.1. Pengumpulan Data**

Dalam penelitian ini, data dikumpulkan melalui metode *crawling* dengan menggunakan *library Harvest*[14]. Metode *crawling* data adalah teknik otomatis untuk mengumpulkan data dari *website* menggunakan program komputer, memungkinkan pengambilan data secara efisien dan terorganisir[15]. Penelitian ini menggunakan data *tweet* dalam bahasa Indonesia mengenai opini publik tentang program KIP Kuliah di *Twitter*[16]. Data diperoleh dari *twitter* menggunakan *Python* di *Google Colaboratory*,

*Keyword* yang digunakan dalam pengumpulan data ini meliputi "KIP Kuliah" dan "beasiswa KIP." Jumlah *tweet* yang berhasil dikumpulkan sebanyak 6.842 *tweet*, dengan data yang mencakup rentang waktu antara tahun 2023 hingga 2024. Tahapan proses *crawling* data KIP Kuliah dapat dilihat pada Gambar 2.



Gambar 2. Proses *crawling*

**2.2. Pelabelan Data**

Pelabelan Data adalah proses penandaan atau pengkategorian data mentah dengan label atau tag tertentu untuk tujuan analisis atau pemodelan. Hal ini akan menghasilkan skor sentimen yang digunakan untuk membagi komentar menjadi sentimen positif dan negatif. Klasifikasi ini bertujuan untuk menentukan nilai sentimen dari komentar yang dianalisis[17]. dapat dilihat pada persamaan 1.

$$S_{sentiment} = \begin{cases} positive & \text{if } S_{positive} > S_{negative} \\ negative & \text{if } S_{positive} < S_{negative} \end{cases} \quad (1)$$

Dimana:

$S_{sentiment}$  : Kalimat sentiment pada dataset

$S_{positive}$  : Kalimat sentimen positif pada dataset

$S_{negative}$  : Kalimat sentimen negatif pada dataset.

**2.3. Preprocessing**

*Preprocessing* adalah tahap di mana data yang diperoleh melalui *crawling* sebanyak 5.638 data akan diproses lebih lanjut. Teknik yang digunakan dalam pra-proses data meliputi *Cleansing*, *Case Folding*, *Tokenizing*, *stopword*, *Filtering*, dan *Stemming*[18]. Dalam penelitian ini ada beberapa tahapan pada *preprocessing* sebagai berikut.

**2.3.1. Cleansing**

*Cleansing* adalah proses pembersihan data teks dari komponen yang tidak diperlukan atau dianggap sebagai noise. Komponen yang akan dihilangkan antara lain karakter HTML, simbol *emoticon*, *hashtag* (#), *username account* (@username), *retweet* (RT), *link URL*, dan alamat *website*[18].

**2.3.2. Case Folding**

*Case Folding* adalah tahap pra-pemrosesan yang mengubah semua karakter dalam sebuah string menjadi format huruf tertentu, seperti mengonversinya ke huruf kecil atau huruf besar[19].

### 2.3.3. Tokenizing

*Tokenization* atau *tokenizing* adalah proses memecah suatu kalimat menjadi kata-kata individu. Ini melibatkan pemisahan seluruh karakter dari suatu kumpulan dokumen menjadi bagian-bagian kecil berupa kata. Bagian-bagian kecil ini disebut sebagai token[18].

### 2.3.4. Stopwords

*Stopword* adalah proses menghapus kata-kata yang dianggap tidak relevan atau terlalu umum dalam teks atau dokumen[20]. *Stopword* biasanya berupa kata-kata umum yang sering muncul dalam banyak kalimat dalam jumlah besar, seperti kata penghubung, contohnya: "adalah", "ke", "di", "dari", "dan"[21].

### 2.3.5. Stemming

*Stemming* adalah proses mengubah kata-kata yang sudah melalui tahap *filtering* menjadi bentuk dasarnya dengan menghapus imbuhan pada teks. Contohnya, kata "mempertahankan" diubah menjadi "tahan", "mengakui" menjadi "akui", dan "melindungi" menjadi "lindung"[22].

## 2.4. Pembagian Data

Pembagian data dalam analisis sentimen dengan rasio 70:30 adalah tahap pembagian data, yakni proses membagi dataset menjadi data latih dan data uji. Pada penelitian ini, 70% dari data dialokasikan sebagai data latih dan 30% sebagai data uji. Data latih digunakan untuk membangun model berdasarkan opini dari setiap kategori, yaitu positif dan negatif. Sedangkan data uji berfungsi untuk menguji model yang sudah dibangun agar diperoleh model yang paling optimal.

## 2.5. Ekstraksi Fitur

Fitur ekstraksi merupakan langkah penting sebelum klasifikasi, dengan tujuan menghasilkan fitur melalui pembobotan kata. Pembobotan ini dihitung untuk meningkatkan akurasi dalam klasifikasi data. TF-IDF adalah metode yang digunakan untuk memberi bobot pada kata sebagai strategi klasifikasi dokumen. Proses pembobotan menggunakan TF-IDF melibatkan penghitungan TF (*Term Frequency*), yang menunjukkan frekuensi kemunculan suatu kata dalam teks, dan IDF (*Inverse Document Frequency*), yang menghitung probabilitas kemunculan kata tersebut di seluruh teks[23]. Berikut ini adalah perhitungan pembobotan dengan metode TF-IDF. dapat dilihat pada persamaan 2.

$$TF * IDF(d, t) = TF(d, t) * \log \frac{N}{df(t)} \quad (2)$$

Dimana:

*TF* : Merupakan nilai dari *term frequency*

*IDF(d, t)* : Merupakan nilai dari *inverse document frequency*

*(d, t)* : Merupakan banyaknya term *t* pada dokumen *d*

*N* : Jumlah dari semua kumpulan dokumen

*df(t)* : Merupakan jumlah dokumen yang memiliki *term t*.

## 2.6. SMOTE

*Synthetic Minority Over-sampling Technique* (SMOTE), adalah salah satu teknik over-sampling yang efektif dan cocok digunakan untuk menangani dataset yang tidak seimbang, karena dapat mengatasi masalah *overfitting* selama proses *oversampling* pada kelas *minoritas*. Teknik *Synthetic Minority Over-sampling Technique* (SMOTE) bekerja dengan menambah jumlah data pada kelas *minoritas* sehingga menjadi seimbang dengan kelas *mayoritas*. [24] *Synthetic Minority Over-sampling Technique* (SMOTE) dipilih untuk mengatasi ketidakseimbangan data karena teknik ini mempertahankan pola distribusi data dengan lebih baik daripada metode over-sampling tradisional yang hanya menduplikasi data. SMOTE memungkinkan model untuk mempelajari karakteristik *minoritas* yang lebih bervariasi, meningkatkan akurasi prediksi dan generalisasi model..

## 2.7. Pembangunan Model

Pada tahap pemodelan, dilakukan klasifikasi data menggunakan tiga algoritma, *Naive Bayes*, *Support Vector Machine* dan *Random Forest*. Model ini dikembangkan menggunakan bahasa pemrograman *Python* dengan memanfaatkan *library sklearn*. Berikut penjelasan lebih lanjut terkait ketiga algoritma tersebut:

### 2.7.1. Naive Bayes Classifier (NBC)

*Naive Bayes* adalah salah satu algoritma yang digunakan dalam proses penambangan teks[25]. Metode ini didasarkan pada teorema Bayes dan mampu melakukan klasifikasi dengan kinerja yang mirip dengan *decision tree* dan *neural network*. *Naive Bayes* terbukti efektif dengan tingkat akurasi dan kecepatan tinggi ketika diterapkan pada *database* dengan dataset besar[26]. Persamaan yang mendasari metode pada *teorema Bayes* dapat dilihat pada persamaan 3.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (3)$$

Dimana:

*X* :Data dengan Kelas yang belum diketahui.

*H* :*Hipotesis* data merupakan suatu kelas spesifik.

*P(H|X)* :*Probabilitas hipotesis H* berdasarkan kondisi *X* (*posteriori probabilitas*).

$P(H)$  : Probabilitas hipotesis H (prior probabilitas).

$P(X|H)$  : Probabilitas X berdasarkan kondisi pada hipotesis H.

$P(X)$  : Probabilitas X..

### 2.7.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma Machine Learning yang menggunakan fungsi *hyperplane* pada data untuk membentuk area-area bagi setiap kelas. *Hyperplane* merupakan fungsi yang berfungsi sebagai pemisah antara kelas-kelas tersebut. Dalam memprediksi kelas suatu data, SVM akan memberikan label berdasarkan area kelas di mana data tersebut berada. SVM umumnya digunakan pada dataset besar yang diambil dari situs online dan menjadi populer karena penggunaannya dalam klasifikasi teks[27]. Prinsip SVM adalah membangun *hyperplane* dengan margin seimbang antara kelas-kelas, sehingga tidak condong ke salah satu kelas. SVM menggunakan tiga jenis *kernel radial*, *linier*, dan *sigmoid*. Dalam penelitian ini, digunakan *kernel linier*, yang direpresentasikan pada persamaan 4 dan 5.

$$(w \cdot x_i + b) \leq 1, y_i = -1 \quad (4)$$

$$(w \cdot x_i + b) \leq 1, y_i = 1 \quad (5)$$

Dimana:

$x_i$  : Data ke-i

$w \cdot x_i$  : Nilai bobot untuk kelas data ke-i

$b$  : Nilai bias

$y_i$  : Kelas data ke-i.

### 2.7.3. Random Forest

Random Forest adalah salah satu metode yang digunakan untuk klasifikasi dan regresi. Metode ini termasuk dalam ensemble pembelajaran, di mana beberapa pohon keputusan (*decision tree*) digunakan sebagai *base classifier* yang dikembangkan dan digabungkan. Beberapa aspek penting dari *Random Forest* meliputi penggunaan *bootstrap sampling* untuk membangun setiap pohon prediksi. Setiap pohon keputusan menggunakan prediktor secara acak untuk melakukan prediksi, dan hasil akhir *Random Forest* diperoleh dengan mengombinasikan prediksi dari semua pohon. Untuk klasifikasi, kombinasi ini dilakukan melalui pemungutan suara terbanyak (*majority vote*), sedangkan untuk regresi menggunakan rata-rata[12]. dapat menggunakan persamaan 6.

$$f(x) = \text{Average}(f_1(x), f_2(x), \dots, f_n(x)) \quad (6)$$

Dimana:

$f(x)$  : Hasil prediksi

$f_{1-n}(x)$  : Hasil prediksi dari setiap pohon keputusan ke-n

$(x)$  : Inputan data.

## 2.8. Evaluasi Model

Pada tahap evaluasi ini, digunakan *Confusion Matrix* untuk mengevaluasi kinerja klasifikasi algoritma terhadap jumlah data berlabel yang diprediksi[28]. Evaluasi model dilakukan untuk mengukur kinerja model klasifikasi yang telah dikembangkan dan menentukan tingkat akurasi. Dalam penelitian ini, evaluasi menggunakan metode *confusion matrix*. Data uji yang dikumpulkan dikelompokkan ke dalam empat kategori: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Akurasi diperoleh dengan menghitung rasio antara jumlah total nilai *True* dan keseluruhan data[26]. *Confusion Matrix* digunakan karena memberikan detail kesalahan prediksi, seperti *false positives* dan *false negatives*, sehingga lebih informatif daripada akurasi saja. Metode ini juga memungkinkan perhitungan *precision*, *recall*, dan *F1-score*, yang penting untuk data tidak seimbang, sehingga evaluasi model menjadi lebih komprehensif dan mendukung pemilihan model yang optimal. setiap elemen dalam *Confusion Matrix* yang direpresentasikan pada persamaan berikut.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (7)$$

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$f1 - \text{score} = \frac{2 \times \text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad (10)$$

Dimana:

*True Positive* (TP) : Jumlah data yang berhasil diidentifikasi dengan benar sebagai positif oleh model.

*False Positive* (FP) : Jumlah data yang model salah mengklasifikasikan sebagai positif.

*True Negative* (TN) : Jumlah data yang berhasil diidentifikasi dengan benar sebagai negative oleh model

*False Negative* (FN) : Jumlah data yang model salah mengklasifikasikan sebagai *negative*

## 3. HASIL DAN PEMBAHASAN

### 3.1. Pengumpulan Data

Pengumpulan data dilakukan dengan metode crawling pada media sosial *Twitter*. *Keyword* yang digunakan dalam pengumpulan data ini meliputi "KIP Kuliah" dan "beasiswa KIP." Jumlah tweet yang berhasil dikumpulkan sebanyak 6.842 *tweet*, dengan data yang mencakup rentang waktu antara tahun 2023 hingga 2024. Data tersebut kemudian disimpan dalam

format *Excel* sebelum diolah lebih lanjut untuk analisis sentimen. Proses *crawling* dilakukan menggunakan *Google Colab* dengan bahasa pemrograman *Python*, menggunakan *library pandas* sebagai dasar perintah untuk pengambilan data. Dataset yang digunakan dalam pemodelan dapat dilihat pada Tabel 1.

Tabel 1. Sampel Hasil Pengumpulan Data

Username	Tweet
astoriasm	kuliah gratis full program KIP yang dirancang pak mulyono bjirr cuman ganti nama dri Bidikmisi yg dirancang era SBY into KIP @centildikit Kuliah juga ada KIP-K. Walaupun kadang salah sasaran tpi ya menurutku berguna banget buat yang membutuhkan. Tergantung gaya hidupnya kalau hedon ya wassalam. Banyak temen2ku yg hedon bgitu
urmineck	

### 3.2. Labelling Data

Dalam penelitian ini, dilakukan perhitungan skor *polaritas* berdasarkan jumlah kata yang teridentifikasi melalui kamus leksikon, sehingga diperoleh sentimen dengan label positif dan negatif. Langkah selanjutnya adalah memeriksa hasil pelabelan sentimen berdasarkan skor *polaritas* yang didapatkan. Proses pelabelan sentimen ini dikategorikan ke dalam dua kelas, yaitu sentimen positif dan negatif. Hasil pelabelan data program KIP-Kuliah pada penelitian ini dapat dilihat pada tabel 2.

Tabel 2. Hasil Labelling

Tweet	Label
kuliah gratis full program KIP yang dirancang pak mulyono bjirr cuman ganti nama dri Bidikmisi yg dirancang era SBY into KIP @centildikit Kuliah juga ada KIP-K. Walaupun kadang salah sasaran tpi ya menurutku berguna banget buat yang membutuhkan. Tergantung gaya hidupnya kalau hedon ya wassalam. Banyak temen2ku yg hedon bgitu	Positif
	Neigatif

### 3.3. Preprocessing

*Preproceissing* adalah langkah penting untuk meimbeersihkan dan meinyiapkan data teiks seibeilum analisis dilakukan. Tujuannya adalah untuk meingkatkan kualitas data, meinghapus eileimein yang tidak reileivan, dan meimformat teiks agar leibih eifisiein diproseis oleih algoritma. Seilain itu, *preproceissing* beirtujuan untuk meingubah data meinjadi format yang leibih mudah dan eifeiktif untuk peingolahan seilanjutnya. Beirikut adalah beibeirapa langkah umum dalam *preproceissing* data untuk analisis seintimein:

#### 3.3.1. Cleansing

Tahap pertama dari proses *preprocessing* adalah cleaning dimana data harus dibersihkan dari hal-hal yang dianggap tidak penting. Pada tahap ini bertujuan untuk menghilangkan karakter atau simbol seperti

tautan URL, nama pengguna atau mention (@username), hashtag (#), retweet, dan emotikon dari teks, sehingga analisis didasarkan pada data yang akurat. Hasil *cleansing* dapat dilihat pada tabel 3.

Tabel 3. Hasil Proses *Cleiansing*

<i>Cleiansing</i>
kuliah gratis full program KIP yang dirancang pak mulyono bjirr cuman ganti nama dri Bidikmisi yg dirancang era SBY into KIP Kuliah juga ada KIPK Walaupun kadang salah sasaran tpi ya menurutku berguna banget buat yang gaya hidupnya kalau hedon ya wassalam Banyak temenku yg hedon bgitu

#### 3.3.2. Case Folding

*Case folding* adalah proses mengonversi seluruh teks dalam dokumen menjadi huruf kecil untuk mengatasi inkonsistensi dalam penggunaan huruf kapital. Proses ini mempertimbangkan sensitivitas huruf besar dan kecil. Hasil dari *case folding* ini ditampilkan pada tabel 4.

Tabel 4. Hasil Proses *Case Folding*

<i>Case Folding</i>
kuliah gratis full program kip yang dirancang pak mulyono bjirr cuman ganti nama dri bidikmisi yg dirancang era sby into kip kuliah juga ada kirk walaupun kadang salah sasaran tpi ya menurutku berguna banget buat yang gaya hidupnya kalau hedon ya wassalam banyak temenku yg hedon bgitu

#### 3.3.3. Tokenizing

*Tokenizing* adalah proses membagi kalimat atau teks menjadi kata-kata terpisah. Dengan melakukan *tokenizing*, kita dapat lebih mudah menghitung frekuensi kemunculan setiap kata dalam teks tersebut. Contoh hasil *tokenizing* dapat dilihat pada tabel 5.

Tabel 5. Hasil Proses *Tokenizing*

<i>Tokenizing</i>
['kuliah', 'gratis', 'full', 'program', 'kip', 'yang', 'dirancang', 'pak', 'mulyono', 'bjirr', 'cuman', 'ganti', 'nama', 'dri', 'bidikmisi', 'yg', 'dirancang', 'eira', 'sby', 'into', 'kip'] ['kuliah', 'juga', 'ada', 'kirk', 'walaupun', 'kadang', 'salah', 'sasaran', 'tpi', 'ya', 'meinerutku', 'beirguna', 'bangeit', 'buat', 'yang', 'gaya', 'hidupnya', 'kalau', 'heidon', 'ya', 'wassalam', 'banyak', 'teimeinku', 'yg', 'heidon', 'bgitu']

#### 3.3.4. StopWord

Tahapan *stopword* bertujuan untuk menghilangkan kata-kata umum yang sering muncul dalam jumlah besar dan tidak memiliki makna signifikan. Proses ini dilakukan dengan menggunakan algoritma *stoplist* untuk menghapus kata-kata yang kurang relevan, atau dengan *wordlist* untuk menyimpan kata-kata yang memiliki arti penting. Hasil dari proses *stopword* dapat dilihat pada Tabel 6.

Tabel 6. Hasil Proses *StopWord*

<i>StopWord</i>
['kuliah', 'gratis', 'full', 'program', 'kip', 'yang', 'dirancang', 'mulyono', 'bjirr', 'cuman', 'ganti', 'nama', 'dri', 'bidikmisi', 'yg', 'dirancang', 'eira', 'sby', 'kip']

[kuliah', 'kipk', 'kadang', 'salah', 'sasaran', 'tpi', 'ya', 'meinurutku', 'beirguna', 'bangeit', 'yang', 'gaya', 'hidupnya', 'heidon', 'ya', 'wassalam', 'teimeinku', 'yg', 'heidon', 'bgitu']

### 3.3.5. Stemming

Pada proses *stemming*, setiap kata diubah menjadi bentuk dasarnya dengan menghilangkan imbuhan. Proses *stemming* ini dilakukan menggunakan *library Sastrawi*. Hasil dari proses *stemming* dapat dilihat pada Tabel 7.

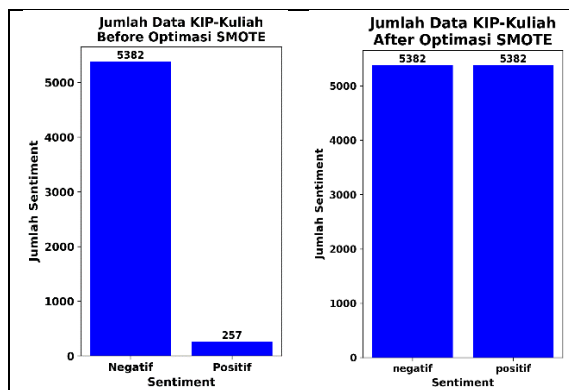
Tabel 7. Hasil Proses *Steimming*

<i>Stemming</i>
kuliah gratis full program kip yang rancang mulyono bjirr cuman ganti nama dri bidikmisi yg rancang eira sby kip kuliah kipk kadang salah sasar tpi ya turut guna bangeit yang gaya hidup heidon ya wassalam teimeinku yg heidon bgitu

### 3.4. Perbandingan Akurasi

Dalam penelitian ini, data yang dihasilkan setelah melalui tahap *preprocessing* pada analisis sentimen terdiri dari 257 data positif dan 5.382 data negatif, dari jumlah 6.842 data *tweet*. Ketidakseimbangan antara data sentimen positif dan negatif ini menyebabkan model algoritma, seperti *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*, cenderung belajar dari kata-kata negatif, yang berdampak pada tingginya akurasi dan performa klasifikasi untuk kelas negatif. Oleh karena itu, penilaian model algoritma tidak hanya berfokus pada nilai akurasi, tetapi juga mempertimbangkan metrik dari *classification report*, seperti *precision*, *recall*, dan *F1-score*, untuk menentukan model yang paling efektif.

Untuk mengatasi masalah ketidakseimbangan data, penelitian ini mengimplementasikan optimasi *Synthetic Minority Over-sampling Technique* (SMOTE), dengan tujuan agar model algoritma tidak hanya terfokus pada satu jenis sentimen. Dengan penerapan SMOTE, jumlah data *minoritas* akan disamakan dengan jumlah data *majoritas*. Hasil optimasi SMOTE dapat dilihat pada Gambar 3 berikut.



Gambar 4. Hasil *Beforei Afteri* SMOTEi

Berdasarkan Gambar 4, terdapat perbandingan antara data sebelum dan sesudah penerapan optimasi *Synthetic Minority Over-sampling Technique*

(SMOTE). Dengan diterapkannya optimasi SMOTE, jumlah data untuk setiap kategori sentimen menjadi sama, sehingga dalam penelitian ini terdapat 5.382 data untuk sentimen positif dan negatif. Optimasi SMOTE memastikan bahwa model algoritma dapat mempelajari sentimen secara seimbang tanpa adanya ketidakseimbangan antara data *minoritas* dan *majoritas*.

Setelah data seimbang, langkah berikutnya adalah melakukan pelatihan dan pengujian agar model algoritma dapat belajar dari data baru yang telah dioptimasi. Dalam penelitian ini, setiap model algoritma, yaitu *Naïve Bayes*, SVM, dan *Random Forest*, menggunakan 70% data untuk pelatihan dan 30% untuk pengujian. Data yang digunakan mencakup data sebelum dan sesudah optimasi SMOTE, sehingga hasil dari setiap algoritma dapat dibandingkan untuk menentukan model dengan performa terbaik. Rincian lebih lanjut mengenai hasil dari model algoritma dapat dilihat pada Tabel 8 dan 9 berikut ini.

Tabel 8. Hasil Akurasi *Beiforei* SMOTEi

<i>Model</i>	<i>Class</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
NBC	Neigatif	95%	95%	100%	98%
	Positif		00%	00%	00%
SVM	Neigatif	96%	95%	100%	98%
	Positif		100%	05%	10%
RF	Neigatif	96%	96%	100%	98%
	Positif		100%	09%	16%

Dari hasil pembagian data sebesar 70% untuk pelatihan dan 30% untuk pengujian sebelum penerapan optimasi SMOTE, ketiga algoritma yaitu *Naïve Bayes*, SVM, dan *Random Forest* menunjukkan akurasi yang cukup tinggi. *Naïve Bayes* mencapai akurasi 95%, sementara SVM dan *Random Forest*, masing-masing mencapai akurasi 96%. Namun, dalam penelitian ini, akurasi saja tidak cukup untuk menilai kemampuan model dalam membedakan sentimen positif dan negatif. Oleh karena itu, kinerja algoritma dilihat lebih lanjut melalui metrik *precision*, *recall*, dan *F1-Score*.

Tabel 9. Hasil Akurasi *After* SMOTE

<i>Model</i>	<i>Class</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
NBC	Negatif	91%	99%	84%	91%
	Positif		86%	99%	92%
SVM	Negatif	99%	100%	98%	99%
	Positif		98%	100%	99%
RF	Negatif	100%	100%	100%	100%
	Positif		100%	100%	100%

Setelah mengoptimasi model dengan *Synthetic Minority Over-sampling Technique* (SMOTE), terdapat variasi peningkatan dan penurunan pada nilai akurasi ketiga algoritma. Pada algoritma *Naïve Bayes*, akurasi menurun dari 95% menjadi 91%. Untuk sentimen negatif, *precision* meningkat dari 95% menjadi 99%, sementara *recall* turun dari 100% menjadi 84%, dan *F1-Score* menurun dari 98% menjadi 91%. Penurunan dan peningkatan ini juga terlihat pada algoritma SVM, di mana akurasinya

meningkat dari 96% menjadi 99%. Peningkatan ini diikuti oleh perbaikan pada *precision*, yang naik dari 95% menjadi 100%, *recall* dari 100% menjadi 98%, dan *F1-Score* dari 98% menjadi 99%. Pada algoritma *Random Forest*, akurasi meningkat dari 96% menjadi 100%, dengan *precision* naik dari 96% menjadi 100%, *recall* tetap di 100%, dan *F1-Score* juga meningkat dari 98% menjadi 100%.

Untuk meineintukan algoritma teirbaik, peineiliti meilakukan peirbandingan kineirja beirdasarkan *confusion matrix* dari keitiga modeil teirseibut. Keitiga *confusion matrix* meinggambarkan peirforma modeil *Naive Bayeis Classifieir*, *SVM*, dan *Random Foreist* seibeilum peineirapan teiknik *Syntheitic Minority Oveir-sampling Teichniquei* (SMOTEi). Hasil peirbandingan ini dapat dilihat pada tabel 10 dan 11 seibagai beirikut.

Tabel 10. *Confusion Matrix Beiforei* SMOTEi

Model	Prediction Class	Actual Class	
		Pred. Positif	Pred. Negatif
NBC	Neigatif	0	80
	Positif	0	1612
SVM	Neigatif	0	76
	Positif	4	1612
RF	Neigatif	7	73
	Positif	0	1612

*Naive Bayeis Classifier* menunjukkan performa terburuk, tidak berhasil mendeteksi data positif (*True Positive* = 0, *False Negative* = 80), meskipun dapat memprediksi kelas negatif dengan benar (*True Negative* = 1612).

*SVM* menghasilkan 1612 prediksi benar untuk kelas negatif (*True Negative* = 1612), namun memiliki kinerja yang buruk dalam mendeteksi kelas positif. Dari 80 data positif, hanya 4 yang berhasil terdeteksi dengan benar (*True Positive* = 4), sementara 76 data positif salah diprediksi sebagai negatif (*False Negative* = 76).

*Random Forest* menunjukkan performa yang sedikit lebih baik dibandingkan *SVM*. Model ini juga memprediksi 1612 data kelas negatif dengan benar (*True Negative* = 1612) dan berhasil mendeteksi 7 dari 80 data positif (*True Positive* = 7), meskipun masih terdapat 73 data positif yang salah diprediksi sebagai negatif (*False Negative* = 73).

Meskipun *Random Forest* lebih unggul dalam mendeteksi kelas positif dibandingkan *SVM* dan *Naive Bayeis*, masih diperlukan peningkatan, sehingga optimasi seperti penerapan *Syntheitic Minority Oveir-sampling Teichniquei* (SMOTEi), sangat diperlukan untuk mengatasi ketidakseimbangan data dan meningkatkan deteksi kelas positif.

Tabel 11. *Confusion Matrix Afteir* SMOTEi

Model	Prediction Class	Actual Class	
		Pred. Positif	Pred. Negatif
NBC	Neigatif	262	16
	Positif	1575	1377
SVM	Neigatif	1591	0
	Positif	28	1611
RF	Neigatif	1591	0

Seiteilah Setelah penerapan teknik *Synthetic Minority Over-sampling* (SMOTE) mengalami peningkatan signifikan dalam mendeteksi kelas positif. *Naive Bayeis*, juga menunjukkan peningkatan, berhasil mendeteksi 1575 *instance* positif dengan benar dan hanya membuat 16 kesalahan (*False Negative*), meskipun masih memiliki 262 kesalahan dalam memprediksi kelas negatif (*False Positive*).

*SVM* mampu memprediksi 1611 *instance* kelas negatif dengan benar (*True Negative*), hanya membuat 28 kesalahan (*False Positive*), dan mendeteksi 1591 *instance* positif dengan benar (*True Positive*) tanpa kesalahan (*False Negative*), menunjukkan peningkatan besar dalam deteksi data positif.

Sementara itu, *Random Forest*, mencapai performa sempurna, memprediksi 1639 *instance* negatif dan 1591 *instance* positif dengan benar, tanpa kesalahan pada kedua kelas (*False Positive* dan *False Negative* = 0). Secara keseluruhan, SMOTE berhasil meningkatkan kinerja ketiga model, terutama dalam mendeteksi kelas positif.

Sebelum Sebelum penerapan *Synthetic Minority Over-sampling Technique* (SMOTE), ketiga model *Naive Bayeis Classifier*, *Support Vector Machine* (*SVM*), dan *Random Forest* lebih efektif dalam mendeteksi kelas negatif namun lemah pada kelas positif, dengan *Naive Bayeis* memiliki performa terburuk dan *Random Forest* terbaik. Setelah penerapan SMOTE, kemampuan ketiga model dalam mendeteksi kelas positif meningkat tajam. *Random Forest* mencapai akurasi sempurna, *SVM* hampir tanpa kesalahan, sementara *Naive Bayeis* masih tertinggal terutama pada kelas negatif. SMOTE terbukti efektif meningkatkan performa pada data yang tidak seimbang, khususnya dalam mendeteksi kelas positif.

### 3.5. Visualisasi

#### WordCloud Analisis Sentiment Kartu Indonesia Pintar



Gambar 5. Wordcloud KIP Kuliah



Penelitian ini menggunakan *wordcloud* untuk menganalisis data teks secara efektif. *Wordcloud* menampilkan kata-kata yang sering muncul dengan ukuran font yang mencerminkan frekuensinya semakin sering kata muncul, semakin besar ukurannya visualisasi *wordcloud* terkait topik KIP dapat dilihat sebagai berikut.

Gambar 5 menunjukkan gabungan dari sentimen positif dan negatif mengenai program KIP Kuliah. Kata-kata seperti "terima", "kipk", "kuliah", "uang", dan "daftar" kembali mendominasi. Kata-kata ini mewakili berbagai perasaan, baik positif maupun negatif, yang dirasakan oleh para mahasiswa mengenai program ini.

**WordCloud Positif Analisis Sentiment Kartu Indonesia Pintar**



Gambar 6. *Wordcloud* Positif

Gambar 6 menunjukkan visualisasi *wordcloud* dengan sentimen positif, kata-kata yang sering muncul seperti "kuliah", "uang", "terima", "beasiswa", "hidup", dan "keuarga" menunjukkan bahwa KIP sangat membantu mahasiswa melanjutkan pendidikan tanpa membebani keluarga.

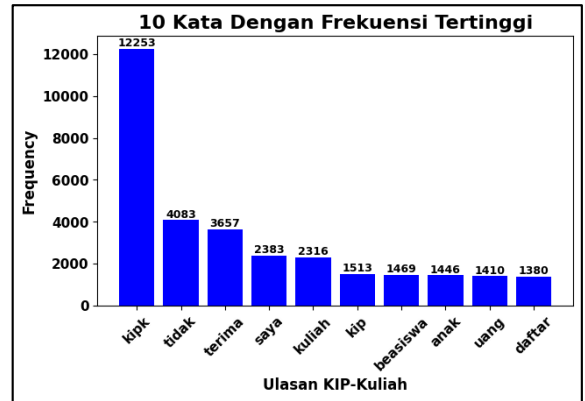
**WordCloud Negatif Analisis Sentiment Kartu Indonesia Pintar Kuliah**



Gambar 7. *Wordcloud* Negatif

Berdasarkan gambar 7 visualisasi *wordcloud* sentimen negatif, kata-kata yang sering muncul seperti "kip", "uang", "tunggak", "daftar", "biaya", dan "utang" menunjukkan kekecewaan terkait keterlambatan pembayaran beasiswa KIP.

Penggunaan frekuensi kata bertujuan untuk memahami seberapa sering kata-kata tertentu muncul dalam analisis sentimen. Penjelasan lebih rinci mengenai kata-kata dengan frekuensi tertinggi dapat dilihat pada gambar 8 di bawah ini:



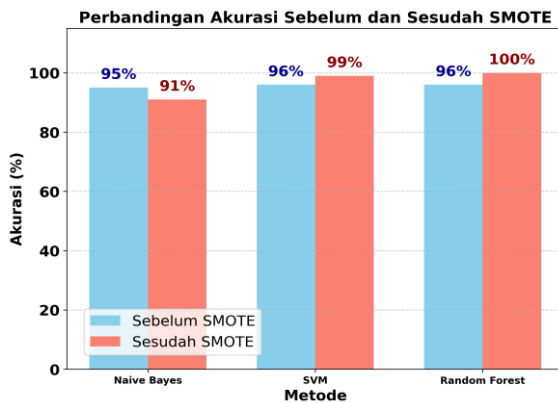
Gambar 8. Frekuensi Kata

Berdasarkan gambar 8, terlihat 10 kata dengan frekuensi tertinggi, di mana frekuensi kata yang sering muncul adalah "kipk," "tidak," "terima," "saya," dan "kuliah." Sehingga dapat disimpulkan bahwa pembahasan terkait program KIP-Kuliah banyak berfokus pada penerimaan program, perasaan penerima, dan istilah-istilah yang berhubungan dengan program pendidikan tinggi. Kata "kipk" dan "tidak" menempati posisi tertinggi, yang mungkin menunjukkan ulasan yang banyak membahas penerimaan KIP atau ketidakpuasan terhadap prosesnya.

**3.6. Visualisasi Perbandingan Model**

Penerapan *Synthetic Minority Over-sampling Technique (SMOTE)* memberikan hasil yang bervariasi pada setiap metode klasifikasi. Pada metode *Naive Bayes*, akurasi sebelum penerapan SMOTE mencapai 95%, namun mengalami penurunan menjadi 91% setelah SMOTE diterapkan. Penurunan akurasi sebesar 4% ini kemungkinan disebabkan oleh karakteristik algoritma *Naive Bayes* yang mengandalkan distribusi probabilitas, sehingga data sintesis dari SMOTE dapat mengganggu kinerja algoritma ini. Sebaliknya, pada metode *Support Vector Machine (SVM)*, SMOTE meningkatkan akurasi dari 96% menjadi 99%, menunjukkan bahwa SVM dapat memanfaatkan data tambahan untuk kelas minoritas dan menghasilkan model yang lebih seimbang. Pada metode *Random Forest*, SMOTE meningkatkan akurasi dari 96% menjadi 100%, mengindikasikan bahwa algoritma ini sangat diuntungkan oleh data yang lebih seimbang dan mampu melakukan klasifikasi dengan lebih akurat.

tanpa bias. Secara keseluruhan, SMOTE terbukti efektif dalam meningkatkan akurasi pada metode SVM dan *Random Forest*, tetapi tidak memberikan manfaat yang serupa pada metode *Naïve Bayes* karena asumsi independensi antar fitur yang terganggu oleh data *sintetis*. Perbandingan akurasi model dapat dilihat pada Gambar 9.



Gambar 9. Perbandingan Akurasi

#### 4. DISKUSI

Hasil penelitian ini membandingkan tiga model algoritma *Naïve Bayes*, SVM, dan *Random Forest*, menunjukkan bahwa algoritma *Naïve Bayes* mencapai akurasi 91%, algoritma SVM mencapai akurasi 99% dan *Random Forest* unggul dalam mengklasifikasikan opini publik di media sosial, dengan akurasi mencapai 100%. Temuan ini konsisten dengan beberapa studi sebelumnya, tetapi hasilnya lebih signifikan, mengingat banyak penelitian yang menggunakan algoritma serupa mencatatkan akurasi lebih rendah.

Seperti penelitian yang menganalisis sentimen terhadap Program KIP-K di media sosial *Twitter*, menggunakan algoritma *Naïve Bayes* dengan pendekatan *Cross-Industry Standard Process for Data Mining (CRISP-DM)* hanya mencapai akurasi sebesar 84,99% [6]. Sedangkan penelitian lain yang menggunakan algoritma *Support Vector Machine (SVM)* untuk analisis sentimen terhadap penerimaan Beasiswa Kartu Indonesia Pintar Kuliah hanya mencatatkan akurasi sebesar 81,82% [9]. Selanjutnya terdapat penelitian lainnya tentang aplikasi Ruangguru, di mana *Random Forest* berhasil mencapai akurasi terbaik sebesar 97,16% [11], melampaui performa *Support Vector Machine (SVM)* dan *Naïve Bayes*.

Namun, Penelitian ini memiliki keterbatasan pada ketidakseimbangan data yang dapat mempengaruhi hasil klasifikasi sentimen, terutama untuk kategori yang jarang muncul. Meskipun SMOTE membantu mengatasi ketidakseimbangan ini, penerapannya harus hati-hati untuk menghindari bias atau *overfitting*. Hasilnya penelitian ini dapat membantu pemerintah memahami opini publik terkait program seperti KIP-K, memungkinkan

strategi komunikasi yang lebih responsif, efektif, dan transparan.

#### 5. KESIMPULAN

Berdasarkan hasil penelitian, analisis sentimen terhadap opini masyarakat mengenai program Kartu Indonesia Pintar Kuliah (KIP-K) di platform media sosial *Twitter* telah berhasil dilakukan. Penelitian ini membandingkan tiga algoritma, yaitu *Naïve Bayes*, *Support Vector Machine (SVM)*, dan *Random Forest*, untuk mengevaluasi performa model klasifikasi. Hasil evaluasi model dari ketiga algoritma dilakukan dengan pembagian data 70:30, 70% data training dan 30% data testing. Dari data yang digunakan, terdiri dari 5.639 data, menghasilkan 5.382 data dengan sentimen negatif dan 257 data dengan sentimen positif. Ketidakseimbangan data ini menjadi tantangan utama dalam penelitian ini, untuk mengatasi masalah ini, teknik *Synthetic Minority Over-sampling Technique (SMOTE)* diterapkan agar jumlah data dikelas *minoritas* seimbang dengan kelas *mayoritas*.

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* memiliki performa terbaik dibandingkan dengan *Naïve Bayes* dan SVM. *Naïve Bayes* mencapai akurasi awal sebesar 95%, tetapi menurun menjadi 91% setelah dioptimalkan menggunakan *Synthetic Minority Over-sampling Technique (SMOTE)*. Sebaliknya, SVM mengalami peningkatan akurasi dari 96% menjadi 99% setelah optimasi. *Random Forest* memberikan hasil terbaik, dengan akurasi awal 96% yang meningkat menjadi 100% setelah optimasi menggunakan SMOTE. Hasil ini menegaskan efektivitas SMOTE dalam mengatasi ketidakseimbangan data, terutama dalam meningkatkan kinerja model untuk sentimen positif.

Penelitian ini merekomendasikan eksplorasi algoritma pembelajaran mendalam seperti *Long Short-Term Memory (LSTM)* dan *Bidirectional Encoder Representations from Transformers (BERT)* serta teknik optimasi seperti *undersampling* atau *hybrid sampling* untuk penelitian lanjutan. Pengumpulan data dari berbagai platform media sosial dan penggunaan teknik ensemble seperti *voting* dan *stacking* juga disarankan untuk meningkatkan performa model.

Secara praktis, model analisis sentimen yang dikembangkan dapat diimplementasikan sebagai sistem pemantauan respons publik real-time terhadap program KIP-K, memungkinkan respons cepat dari pemerintah. Penelitian ini berkontribusi pada analisis sentimen kebijakan publik dan menunjukkan efektivitas SMOTE dalam meningkatkan akurasi model, mendukung pengembangan sistem pemantauan kebijakan yang responsif dan berbasis data.

#### DAFTAR PUSTAKA

[1] D. Amrizal, D. F. Nasution, and A. Amran,

- “Efektivitas Pelayanan Program Kartu Indonesia Pintar (KIP) dalam Rangka Peningkatan Kualitas Pendidikan di SMA Negeri 1 Tanjung Balai,” *Jurnal SOMASI (Sosial Humaniora Komunikasi)*, vol. 1, no. 1, pp. 11–20, 2020, doi: 10.53695/js.v1i1.27.
- [2] M. Romdoni, “Combination of Ahp and Maut Method To Determine Scholarship Recipients in Higher Education ( Case Study : Universitas Teknokrat Indonesia ) Kombinasi Metode Ahp Dan Maut Untuk Menentukan Penerima Beasiswa Pada Perguruan Tinggi ( Studi Kasus : Universitas ,” vol. 5, no. 4, pp. 1125–1135, 2024, doi: 10.52436/1.jutif.2024.5.4.1940.
- [3] Gagan Suganda, Marsani Asfi, Ridho Taufiq Subagio, and Ricky Perdana Kusuma, “Penentuan Penerima Bantuan Beasiswa Kartu Indonesia Pintar (Kip) Kuliah Menggunakan Naïve Bayes Classifier,” *JSiI (Jurnal Sistem Informasi)*, vol. 9, no. 2, pp. 193–199, 2022, doi: 10.30656/jsii.v9i2.4376.
- [4] Sabrina Fitri Jasmine, “Pengaruh Beasiswa KIP-K Terhadap Prestasi Belajar Mahasiswa Manajemen Pendidikan Angkatan 2021 Universitas Negeri Surabaya,” *Jurnal Pendidikan, Bahasa dan Budaya*, vol. 2, no. 2, pp. 61–70, 2023, doi: 10.55606/jpbb.v2i2.1437.
- [5] R. Winata and R. Khairunnisa, “Pengaruh Penggunaan Dana KIP-K Terhadap Pemenuhan Kebutuhan Mahasiswa Dalam Prespektif Ekonomi Islam (Studi Pada Mahasiswa Penerima KIP- K UIN Raden Intan Lampung),” *Business and Entrepreneurship Journal (BEJ)*, vol. 4, no. 1, pp. 8–14, 2023, doi: 10.57084/bej.v4i1.1037.
- [6] J. Homepage, 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, doi: 10.57152/malcom.v4i4.1565
- [7] 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 ),” vol. 5, no. 3, pp. 2994–3003, 2024, doi: 10.35870/jimik.v5i3.990
- [8] A. Suryasuciramdhan, M. Meliasari, B. Meilani Ifada, and M. A. Faidh, “Analisis Framing Dalam Kasus Penyalahgunaan Kip Kuliah Mahasiswi Universitas Diponegoro Di Media Sosial X Dan Media Online Kompas.Tv,” *Journal of Creative Student Research (JCSR)*, vol. 2, no. 3, pp. 40–51, 2024, [Online]. Available: <https://doi.org/10.55606/jcsrpolitama.v2i3.3824>
- [9] M. F. Z. Jannan, Y. Dwi, and P. Negara, “Analisis Sentimen Masyarakat Terhadap Penerima Beasiswa Kartu Indonesia Pintar Kuliah Dengan Metode Support Vector Machine ( Analysis of Community Sentiment Toward Indonesia Smart Card Scholarship Recipients with Support Vector Machine Method ),” pp. 26–30, 2024, doi: 10.32938/jitu.v4i2.7598.
- [10] C. Fitri, N. Halizah, and M. Kartikasari, “Implementasi Algoritma Naïve Bayes Dalam Penentuan Pemberian Beasiswa Kip Kuliah (Studi Kasus Stiki Malang) Implementation of Naïve Bayes Algorithm in Determining Kip-Kuliah Scholarship (Case Study Stiki Malang),” vol. 12, no. 2, pp. 1–6, 2024, [Online]. Available: <http://repository.stiki.ac.id/id/eprint/2220>.
- [11] D. A. Ardan, Mukhsar, G. N. A. Wibawa, B. Abapihi, D. C. Arisona, and A. Tenriawaru, “Analisis Sentimen Persepsi Publik Tentang Program Merdeka Belajar Kampus Merdeka di X Menggunakan Support Vector Machine,” *Science, and Technology (J-HEST)*, vol. 6, no. 2, pp. 151–161, 2024, doi: 10.36339/j-hest.v6i2.37.
- [12] E. Fitri, “Sentiment Analysis of the Ruanguru Application Using Naive Bayes, Random Forest and Support Vector Machine Algorithms,” *Jurnal Transformatika*, vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.
- [13] P. Elisa and A. Rahman Isnain, “Comparison of Random Forest, Support Vector Machine and Naive Bayes Algorithms To Analyze Sentiment Towards Mental Health Stigma,” *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 1, pp. 321–329, 2024, doi: 10.52436/1.jutif.2024.5.1.1817
- [14] D. S. Ningsih and R. R. Suryono, “Comparison of Naïve Bayes and Information Gain Algorithms in Cyberbullying Sentiment Analysis on Twitter Perbandingan Algoritma Naïve Bayes Dan Information Gain,” vol. 5, no. 4, pp. 1085–1091, 2024, doi: 10.52436/1.jutif.2024.5.4.1908
- [15] I. Syahrohim, S. D. Saputra, R. W. Saputra, V. H. Pranatawijaya, and R. Priskila, “Perbandingan Analisis Sentimen Setelah Pilpres 2024 Di Twitter Menggunakan Algoritma Machine Learning,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol.

- 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4249.
- [16] R. S. Arischo and D. Damayanti, "Analisis Sentimen Pinjaman Online di Twitter dengan Metode Naive Bayes Classifier dan SVM," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, p. 1120, 2024, doi: 10.30865/mib.v8i2.7406.
- [17] M. J. Palepa, N. Pratiwi, and R. Q. Rohmansa, "Analisis Sentimen Masyarakat Tentang Pengaruh Politik Identitas Pada Pemilu 2024 Terhadap Toleransi Beragama Menggunakan Metode K - Nearest Neighbor," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 1, pp. 389–401, 2024, doi: 10.29100/jupi.v9i1.4957.
- [18] 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.
- [19] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naive Bayes dan KNN," *Jurnal KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.
- [20] D. Pratmanto and F. F. D. Imaniawan, "Analisis Sentimen Terhadap Aplikasi Canva Menggunakan Algoritma Naive Bayes Dan K-Nearest Neighbors," *Computer Science (CO-SCIENCE)*, vol. 3, no. 2, pp. 110–117, 2023, doi: 10.31294/coscience.v3i2.1917.
- [21] A. Hendra and F. Fitriyani, "Analisis Sentimen Review Halodoc Menggunakan Naive Bayes Classifier," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 6, no. 2, pp. 78–89, 2021, doi: 10.14421/jiska.2021.6.2.78-89.
- [22] I. Alamsyah and R. T. Shita, "Penerapan Metode Multinomial Naive Bayes Untuk Implementation of the Multinomial Naive Bayes Method To Analyze Sentiment of User Reviews of Wahyoo," vol. 2, no. September, pp. 436–444, 2023, Available: <https://senafiti.budiluhur.ac.id/index.php/senafiti/article/view/750>
- [23] M. H. Wicaksono, M. D. Purbolaksono, and S. Al Faraby, "Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada Review Female Daily," *eProceedings of Engineering*, vol. 10, no. 3, pp. 3591–3600, 2023, [Online]. Available: <https://openlibrary.telkomuniversity.ac.id/home/catalog/id/185919.html>
- [24] M. I. Putri and I. Kharisudin, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Analisis Sentimen Data Review Pengguna Aplikasi Marketplace Tokopedia," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 5, pp. 759–766, 2022, [Online]. Available: <https://journal.unnes.ac.id/sju/index.php/prisma/>
- [25] L. A. Hayurian and N. Hendrastuty, "Comparison of Naive Bayes Algorithm and Support Vector Machine in Sentiment Analysis of Boycott Israeli Products on Twitter," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 3, pp. 731–738, 2024, [Online]. Available: <https://doi.org/10.52436/1.jutif.2024.5.3.1813>
- [26] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *Jurnal Informatika*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.
- [27] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Comparison of Naive Bayes and Support Vector Machine Methods in Twitter Sentiment Analysis," *Smatika Jurnal*, vol. 10, no. 02, pp. 71–76, 2020, doi: 10.32664/smatika.v10i02.455
- [28] T. J. Firdaus *et al.*, "SENTIMENT ANALYSIS OF THE SAMBARA APPLICATION USING THE SUPPORT," vol. 5, no. 4, pp. 1183–1192, 2024, doi: 10.52436/1.jutif.2024.5.4.2673.