Institut Riset dan Publikasi Indonesia (IRPI)
**MALCOM: Indonesian Journal of Machine Learning and Computer Science**
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1420-1430
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

# *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 Media Sosial X Menggunakan Algoritma Naive Bayes

**Diky Pramudita[1], Yuma Akbar[2], Tri Wahyudi[3]**

[1,2,3]Program Studi Teknik Informatika, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, Indonesia

E-Mail: [1]dikypramudita3@gmail.com, [2]yumekhan@gmail.com, [3]triwahyudi100390@gmail.com

*Received Jun 18th 2024; Revised Jul 28th 2024; Accepted Jul 31th 2024*
*Penulis Koresponden: Diky Pramudita*

**Abstract**

*This study shows that the public has various responses to the Indonesia Smart Card Program for College (KIP-K) which can be categorized into positive and negative sentiments. The problem studied is how the public responds to the KIP-K program expressed through social media X. This study uses a sentiment method with the Naive Bayes algorithm and the CRISP-DM approach to ensure a systematic and structured analysis process. The data collected were 1,516 tweets containing the keyword "KIP-K" through data crawling techniques using API X. The results showed that the Naive Bayes algorithm was effective with an accuracy of 84.99%, a positive sentiment precision of 83.54%, and a negative sentiment precision of 87.25%. The solution offered is the use of machine learning techniques to automatically categorize sentiment from large and unstructured text data. The benefits of this study are to provide insight to the government and stakeholders about public perceptions of the KIP-K program, which can be used as a basis for evaluating and improving the program in the future. In conclusion, the Naive Bayes algorithm can classify sentiment well using data from tweets about KIP-K, with results showing a dominance of negative sentiment. This research also contributes to the development of machine learning-based sentiment analysis methods in the field of education.*

*Keywords: CRISP-DM, Naïve Bayes, Sentiment Analysis, Smart Indonesia Card*

**Abstrak**

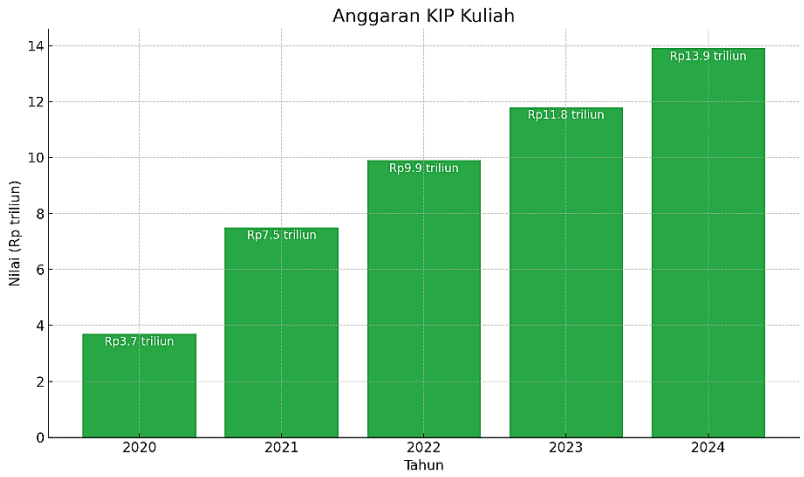
Penelitian ini menunjukkan bahwa publik memiliki berbagai tanggapan terhadap Program Kartu Indonesia Pintar Kuliah (KIP-K) yang dapat dikategorikan ke dalam sentimen positif dan negatif. Permasalahan yang diteliti adalah bagaimana tanggapan publik terhadap program KIP-K yang diungkapkan melalui media sosial X. Penelitian ini menggunakan metode analisis sentimen dengan algoritma Naive Bayes dan pendekatan CRISP-DM untuk memastikan proses analisis yang sistematis dan terstruktur. Data yang dikumpulkan sebanyak 1.516 tweet yang mengandung kata kunci "KIP-K" melalui teknik crawling data menggunakan API X. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes efektif dengan akurasi 84.99%, presisi sentimen positif 83.54%, dan presisi sentimen negatif 87.25%. Solusi yang ditawarkan adalah penggunaan teknik machine learning untuk secara otomatis mengkategorikan sentimen dari data teks yang besar dan tidak terstruktur. Manfaat dari penelitian ini adalah memberikan wawasan kepada pemerintah dan pemangku kebijakan tentang persepsi masyarakat terhadap program KIP-K, yang dapat dijadikan dasar untuk evaluasi dan perbaikan program di masa mendatang. Kesimpulannya, algoritma Naive Bayes dapat mengklasifikasikan sentimen dengan baik menggunakan data dari tweet tentang KIP-K, dengan hasil yang menunjukkan dominasi sentimen negatif. Penelitian ini juga berkontribusi dalam pengembangan metode analisis sentimen berbasis machine learning di bidang pendidikan.

Kata Kunci: Analisis Sentimen, CRISP-DM, Kartu Indonesia Pintar, Naïve Bayes

## 1. PENDAHULUAN

Pendidikan tinggi merupakan salah satu aspek penting dalam pembangunan nasional. Di Indonesia, pemerintah telah melaksanakan berbagai program untuk meningkatkan aksesibilitas dan kualitas pendidikan tinggi, termasuk Program Kartu Indonesia Pintar Perguruan Tinggi (KIP-K) [1]. Program KIP-K merupakan inisiatif pemerintah Indonesia yang bertujuan membantu siswa dari keluarga berpenghasilan rendah untuk mendapatkan pendidikan tinggi. Sistem ini memberikan berbagai bantuan kepada penerimanya, termasuk

subsidi pendidikan dan tunjangan bulanan. Selain aspek finansial, program KIP-K juga memberikan dukungan psikologis dan sosial kepada siswa dari keluarga kurang mampu. Hal ini akan meningkatkan rasa percaya diri dan motivasi mereka untuk melanjutkan pendidikan tinggi dan sukses dalam karir masa depan [2]. Meskipun program ini sudah berjalan cukup baik, namun masih ada kendala yang menghambat bagi penerima KIP-K. Salah satunya pada tahap verifikasi data ekonomi keluarga penerima manfaat seringkali menjadi tantangan yang kompleks. Kurangnya infrastruktur pendukung dan kesulitan dalam memverifikasi data secara akurat dapat menghambat penyaluran bantuan kepada mereka yang sebenarnya membutuhkannya. Dan ketidakmerataan alokasi dana yang dialokasikan untuk Program KIP-K secara merata di antara daerah-daerah yang membutuhkan. Hal ini dapat menyebabkan ketidakadilan akses terhadap program bagi masyarakat di daerah yang kurang berkembang [3].




**Gambar 1.** Anggaran KIP Kuliah Sejak Tahun 2020-2024

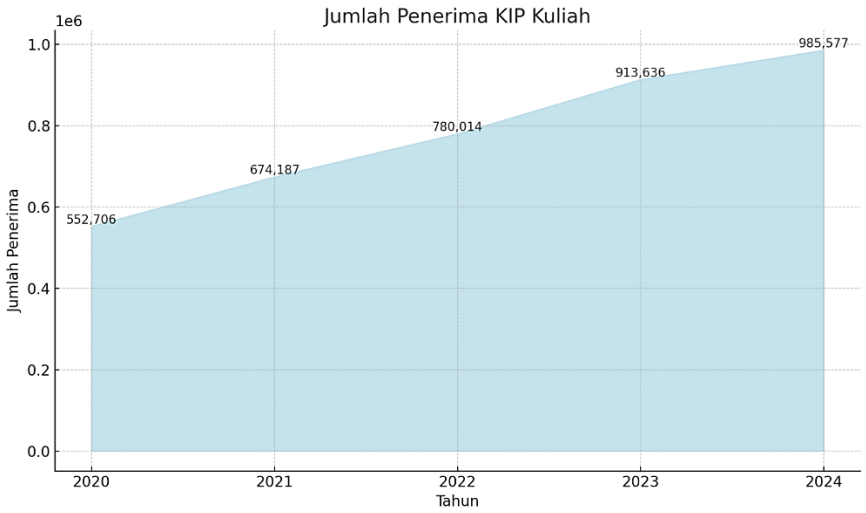
Berdasarkan informasi Pusat Layanan Pembiayaan Pendidikan (Puslapdik) Kemendikbudristek, Terlihat pada gambar 1 bahwa anggaran untuk KIP Kuliah selalu naik setiap tahunnya. Hal ini bisa dilihat dari anggaran yang masih di angka Rp3,7 triliun pada tahun 2020, kemudian naik Rp3,8 triliun menjadi Rp7,5 triliun pada tahun 2021. Momen pandemi Covid-19 tidak membuat anggaran KIP Kuliah menurun, Justru pada tahun 2022 anggarannya kembali naik Rp2,4 triliun menjadi Rp9,9 triliun pada tahun 2022. Tahun 2023 terjadi lagi peningkatan sebesar Rp1,9 triliun. Dalam rancangan anggaran tahun 2024, dananya naik lagi Rp2,1 triliun menjadi Rp13,9 triliun. Ini berarti dalam 5 tahun berturut-turut terdapat total Rp46,8 triliun uang digelontorkan untuk KIP Kuliah, dengan total peningkatan sebesar Rp10,2 triliun selama 2020-2024[4]. Adanya sumber dana yang memadai ini mencerminkan komitmen bersama untuk meningkatkan akses pendidikan tinggi dan mengurangi kesenjangan sosial ekonomi di Indonesia melalui Program KIP-K. Namun, meskipun tersedianya anggaran, masih terdapat tantangan dalam pengelolaan dan distribusi dana secara efektif, yang menjadi bagian dari permasalahan yang ada [5].

Terdapat pula data mengenai jumlah penerima bantuan KIP Kuliah selama 2020-2024. Terlihat bahwa lompatan naik penerima bantuan ini terus terjadi selama 5 tahun terakhir. Dari yang semula sejumlah 552.706 mahasiswa pada tahun 2020, terjadi kenaikan menjadi 674.187 mahasiswa di tahun 2021. Kembali, pandemi Covid-19 tetap membuat penerima bantuan tidak mengalami penurunan. Tahun 2022 terdapat 780.014 penerima, dan tahun 2023 terdapat 913.636 penerima KIP Kuliah pada tahun 2023. Tahun 2024 ini, jumlahnya hampir mencapai 1 juta, dengan 985.577 penerima [6], hal tersebut terdata pada gambar 2.

Berdasarkan pencarian di media sosial *X*, Peneliti mengidentifikasi beberapa masalah yang sering disorot oleh masyarakat terkait isu bagi penerima bantuan program KIP-K. Banyak sekali komentar dari masyarakat yang positif dan negatif, terkait program ini yang sudah berlangsung cukup lama. Analisis sentimen merupakan suatu pendekatan yang digunakan untuk menggali opini dan respons masyarakat terhadap suatu objek, topik, atau peristiwa tertentu [7]. Dalam konteks penelitian ini, fokusnya adalah untuk mengeksplorasi opini dan tanggapan masyarakat terhadap Program Kartu Indonesia Pintar Kuliah (KIP-K) melalui media sosial *X*. Dengan memanfaatkan metode *Naive Bayes*, penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam tentang bagaimana persepsi masyarakat terhadap program tersebut.

Pada penelitian terdahulu, memfokuskan pada prediksi penerimaan Kartu Indonesia Pintar (KIP) untuk memastikan bantuan tepat sasaran. Permasalahan yang diangkat adalah adanya ketidaksesuaian penerimaan KIP, di mana siswa/i yang seharusnya tidak memenuhi syarat malah menerima bantuan, sedangkan siswa/i yang membutuhkan tidak mendapatkan. Penelitian ini menggunakan algoritma *K-Nearest Neighbor* (*KNN*) dan *Naïve Bayes* untuk memprediksi penerimaan KIP, dengan hasil menunjukkan bahwa

algoritma *KNN* memiliki performa lebih baik dengan akurasi rata-rata 96,9%, dibandingkan *Naïve Bayes* yang memperoleh akurasi 86,87% [8].




**Gambar 2.** Jumlah Penerima KIP Kuliah

Kemudian untuk peneliti terdahulu lainnya, membahas prediksi calon penerima mahasiswa KIP. Dengan algoritma *Decision Tree*, menggunakan metode klasifikasi yang merepresentasi struktur pohon dimana setiap node merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas [9]. Adapun atribut yang digunakan yaitu program keahlian, nilai rata-rata, status KK, penghasilan, dan kartu bantuan. Sehingga mempermudah bagi instansi pendidikan yang ingin melakukan seleksi penerima bantuan KIP dengan cepat. Berbeda dengan penelitian ini, tidak hanya memprediksi penerimaan KIP, tetapi juga menganalisis sentimen publik terhadap Program Kartu Indonesia Pintar Kuliah (KIP-K) melalui media sosial *X* [10]. Sehingga dapat memahami persepsi masyarakat terhadap program KIP-K. Dengan menggunakan algoritma *Naïve Bayes* dan pendekatan *CRISP-DM*. Meskipun memiliki fokus yang berbeda, kedua penelitian ini sama-sama memberikan kontribusi dalam evaluasi dan peningkatan efektivitas program KIP di Indonesia, baik dalam aspek prediksi penerimaan maupun pemahaman sentimen publik[11].

Penelitian ini bertujuan untuk mengevaluasi sentimen publik terhadap Program KIP-K di media sosial *X*,memberikan pandangan mendalam tentang aspek-aspek yang mempengaruhi persepsi masyarakat, memahami tren sentimen dari waktu ke waktu, dan memberikan dasar informasi untuk pengambilan keputusan strategis. Urgensinya terletak pada kemampuan memberikan wawasan mendalam terhadap Program KIP-K. Evaluasi ini mencakup pengujian data menggunakan metrik seperti akurasi, presisi, *recall*, dan *f1-score*. Penelitian ini juga melihat sejauh mana kinerja klasifikasi algoritma *Naive Bayes* dengan menggunakan pendekatan *CRISP-DM*, dengan fokus pada akurasi, presisi, dan *recall* [12].

Rujukan pada penelitian terdahulu menunjukkan bahwa peningkatan akurasi dapat dicapai dengan menambahkan jumlah data penelitian. Selain itu, penelitian dapat diperkaya dengan penambahan parameter pada pengujian algoritma, seperti parameter tuning pada *Naive Bayes*[13]. Dengan pendekatan ini, penelitian selanjutnya diharapkan dapat lebih mendalam, menghasilkan hasil yang lebih akurat, dan lebih relevan dalam konteks analisis sentimen program KIP-K. Harapannya, penelitian ini akan berperan sebagai indikator berharga dalam menilai sejauh mana analisis sentimen yang diimplementasikan pada program KIP-K dapat menjadi dasar evaluasi yang lebih baik untuk penelitian berikutnya, serta memberikan rekomendasi yang bermanfaat untuk peningkatan program di masa mendatang.

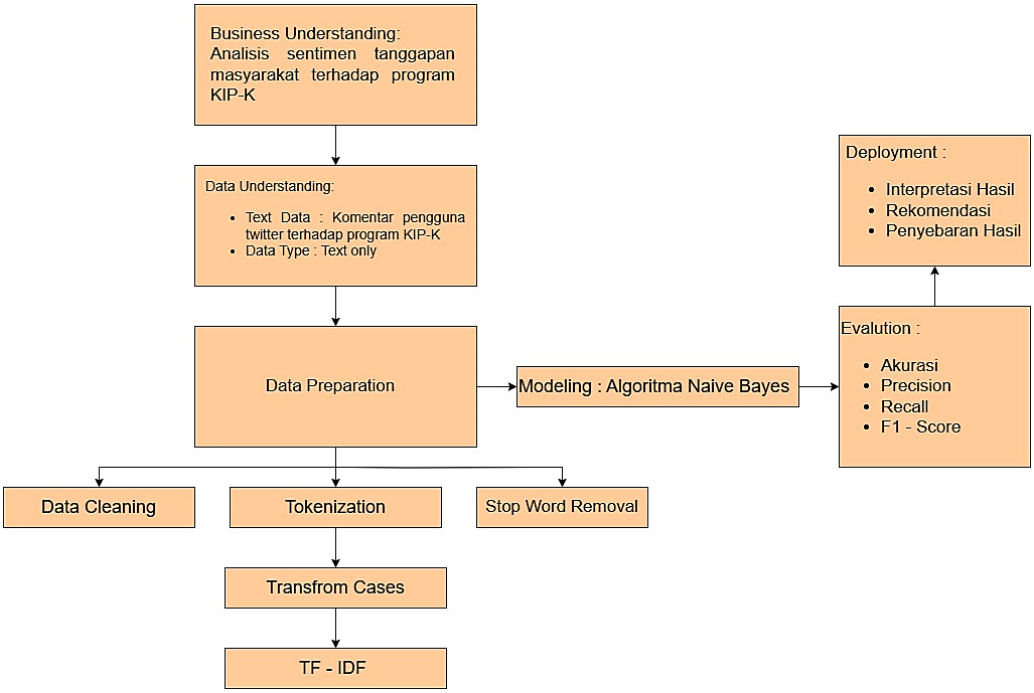
## 2. BAHAN DAN METODE

Dalam penelitian ini, penerapan metodologi dilakukan dengan menggunakan pendekatan *Cross-Industry Standard Process for Data Mining* (CRISP-DM) untuk mengelola dan menganalisis data yang diperoleh dari *X* terkait sentimen publik terhadap Program Kartu Indonesia Pintar Kuliah (KIP-K). Penelitian ini menggunakan metode kualitatif dengan dua sumber data utama yaitu data primer dan data sekunder. Data primer diperoleh melalui observasi langsung, yaitu proses *crawling* data dari media sosial *X* menggunakan *API X*. Data sekunder dikumpulkan dengan membaca dan memahami literatur dari penelitian-penelitian lain yang relevan, yang tersedia di perpustakaan maupun sumber-sumber digital lainnya [14].

Data primer dalam penelitian ini dikumpulkan melalui proses *crawling* data dari *X*. *Crawling* data adalah metode pengumpulan data otomatis dari media sosial, yang dalam hal ini berfokus pada opini dan

sentimen pengguna *X* terkait Program Kartu Indonesia Pintar Kuliah (KIP-K). Pengumpulan data dilakukan dengan memanfaatkan *X* API untuk mengekstraksi *tweet* yang mengandung kata kunci terkait KIP-K. Proses ini melibatkan beberapa langkah, termasuk autentikasi *API*, pengambilan data, dan penyimpanan data mentah dalam format yang dapat dianalisis lebih lanjut. Data yang diambil mencakup teks *tweet*, tanggal posting, dan informasi pengguna yang relevan. Observasi ini memastikan bahwa data yang dikumpulkan adalah yang paling relevan dan terbaru.

Data sekunder dalam penelitian ini diperoleh dari studi literatur. Peneliti melakukan tinjauan terhadap berbagai penelitian sebelumnya yang terkait dengan analisis sentimen, algoritma *Naive Bayes*, dan penerapan analisis sentimen pada data media sosial *X*. Literatur yang digunakan berasal dari buku-buku akademik, jurnal ilmiah, dan sumber-sumber terpercaya lainnya yang dapat diakses dari perpustakaan. Studi literatur ini bertujuan untuk memberikan landasan teoritis yang kuat bagi penelitian serta membantu dalam memahami metodologi yang telah diterapkan dalam penelitian serupa. Selain itu, literatur ini juga membantu dalam mengidentifikasi celah penelitian yang dapat diisi oleh penelitian ini dan memperkuat argumen serta hasil yang diperoleh.




**Gambar 3.** *Metode CRISP-DM*

### 2.1. *Business Understanding*

Tahap pertama dalam metodologi *CRISP-DM* adalah memahami konteks bisnis dan tujuan analisis [15]. Dalam konteks penelitian ini, tujuan utama adalah untuk menganalisis sentimen publik terhadap program Kartu Indonesia Pintar (KIP) Kuliah. Program ini merupakan inisiatif pemerintah Indonesia untuk membantu biaya pendidikan mahasiswa yang kurang mampu. Dengan menganalisis sentimen publik yang diekspresikan melalui *tweet*, peneliti dapat memperoleh gambaran umum mengenai persepsi masyarakat terhadap program ini, apakah cenderung positif, atau negatif.

### 2.2. *Data Understanding*

Data yang digunakan dalam penelitian ini berasal dari platform *X*, yang dikumpulkan menggunakan teknik *crawling* dengan memanfaatkan *API* dari *X*. Proses pengumpulan data dilakukan dengan menyaring *tweet* yang mengandung kata kunci terkait Kartu Indonesia Pintar Kuliah (KIP-K). Dalam penelitian ini, sebanyak 1.516 *tweet* telah berhasil dikumpulkan melalui proses *crawling* tersebut. Setiap *tweet* yang dikumpulkan telah dipastikan relevansinya dengan topik penelitian, sehingga data yang diperoleh dapat memberikan informasi yang akurat dan sesuai dengan kebutuhan analisis[16]. Setelah data terkumpul, langkah selanjutnya adalah mempelajari struktur, kualitas, dan distribusi data tersebut.Analisis awal dilakukan untuk memahami karakteristik data, seperti panjang teks, variasi bahasa yang digunakan, dan tingkat keterlibatan pengguna. Selain itu, dilakukan juga pengidentifikasian *tweet* yang relevan dengan topik penelitian untuk memastikan bahwa hanya data yang relevan dan bermakna yang akan digunakan dalam analisis sentimen. Tahap ini sangat penting untuk menjamin bahwa data yang digunakan memiliki kualitas yang baik dan dapat mendukung hasil penelitian yang valid.

### 2.3. *Data Preparation*

Pada tahapan ini data yang dikumpulkan dari media sosial *X* kemudian diproses lebih lanjut untuk diberikan sentimen. Setelah data yang telah didapatkan diberikan sentimen, data teresbut akan masuk tahapan data *preprocessing*. Beberapa tahap *preprocessing* yang dilakukan adalah Pembersihan data merupakan menghapus duplikasi, mention, iklan serta *tweet* yang tidak relevan dengan penelitian [17]. *Transform Case*, merupakan proses mengubah semua huruf pada teks menjadi huruf kecil semua atau sebaliknya. Tahap pelebelan data adalah proses di mana data mentah, seperti *tweet*, dikategorikan atau diberi label berdasarkan atribut tertentu. Dalam konteks analisis sentimen, label ini biasanya menunjukkan sentimen dari teks tersebut, apakah positif atau negatif. *Tokenizing*, dilakukan pemecahan kalimat menjadi kata. *Stopwords*, pada tahap ini dilakukan penghapusan kata-kata yang dianggap tidak sesuai atau sering muncul seperti,"dan","di". *Stemming* adalah proses mengubah kata yang berimbuhan menjadi kata dasar. Serta *Term Weighting (TF-IDF)* merupakan proses untuk memberikan pembobotan kata dengan mencari nilai *Term Frequency (TF)*, kemudian mencari nilai *Document Frequency (DF)*, lalu mencari nilai *Invers Document Frequency (IDF)* setelah itu baru menghitung bobot [18].

### 2.4. *Modeling*

Pada tahap ini, dilakukan pemodelan terhadap data *tweet* yang telah melalui proses pembersihan dan *preprocessing*. Data *tweet* tersebut dibagi menjadi data *training* dan data *testing* untuk memastikan bahwa model dapat belajar dari sebagian data dan diuji pada data yang belum pernah dilihat sebelumnya. Proses ini terbagi menjadi dua langkah utama, pembagian data (*split data*) dan validasi silang (*cross-validation*). Dalam pembagian data, peneliti membagi data *tweet* tersebut dengan perbandingan 80:20 antara data *training* dan data *testing*. Selanjutnya, pada tahap *cross-validation*, data dibagi secara acak menjadi 10 bagian *(10-fold cross-validation)*. Teknik ini membantu memastikan bahwa model tidak mengalami *overfitting* dan memberikan evaluasi performa yang lebih komprehensif. Setiap bagian secara bergantian digunakan sebagai data *testing* sementara bagian lainnya digunakan sebagai data *training*, sehingga setiap data point diuji secara menyeluruh. Untuk mengatasi ketidakseimbangan kelas, digunakan teknik *SMOTE*. Pada tahap pemodelan, peneliti menggunakan algoritma *Naive Bayes* untuk mengidentifikasi sentimen positif dan negatif dari setiap *tweet*. Algoritma ini dipilih karena kemampuannya dalam menangani teks dan menghasilkan hasil yang baik dalam klasifikasi teks[19]. Formula dari algoritma *Naive Bayes* ditunjukkan pada persamaan 1.

$$P(X|H) = P(H|X)\,P(H) / P(X) \tag{1}$$

### 2.5. *Evalution*

Model yang dibangun dievaluasi untuk menilai kinerjanya dalam mengklasifikasikan sentimen *tweet*. Metrik yang digunakan untuk evaluasi termasuk akurasi, *precision, recall,* dan *F1-score*. Hasil evaluasi membantu menentukan apakah model telah memenuhi tujuan penelitian dan memberikan insight yang berguna [20]. Persamaan 2-6 diguanakan untuk menghitung nilai evaluasi.

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2}$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP++FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP++FN}} \tag{4}$$

$$\text{Spesifikasi} = \frac{\text{TN}}{\text{TN++FP}} \tag{5}$$

$$\text{F1 Score} = \frac{\text{2 x Recall x Presisi}}{\text{Recall + Presisi}} \tag{6}$$

### 2.6. *Deployment*

Tahapan deployment adalah tahapan terakhir dari penelitian ini, yaitu membuat visualisasi data pada penelitian ini. Tujuan melakukan visualisasi data adalah untuk menampilkan beberapa kata yang sering muncul berdasarkan sentimennya. Hal ini tentu akan memudahkan saat melakukan analisis terhadap opini yang disampaikan oleh masyarakt melalaui media sosial *X* tanpa perlu membaca semua komentar pada media sosial *X* satu-persatu.

## 3. HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil dan pembahasan performa algoritma *Naive Bayes* dengan *metode CRISP-DM* yang digunakan dalam penelitian ini.Setelah data terkumpul, langkah selanjutnya adalah mempelajari

struktur, kualitas, dan distribusi data tersebut. Analisis awal dilakukan untuk memahami karakteristik data, seperti panjang teks, variasi bahasa yang digunakan, dan tingkat keterlibatan pengguna. Selain itu, dilakukan juga pengidentifikasian *tweet* yang relevan dengan topik penelitian untuk memastikan bahwa hanya data yang relevan dan bermakna yang akan digunakan dalam analisis sentimen. Tahap ini sangat penting untuk menjamin bahwa data yang digunakan memiliki kualitas yang baik dan dapat mendukung hasil penelitian yang valid.

### 3.1 Tahap *Pre-processing*

Sebelum digunakan untuk data klasifikasi, terlebih dahulu diproses pada tahap *pre-processing*. Tahap penting dalam proses data mining di mana data mentah disiapkan untuk analisis. Pada tahap ini, data dibersihkan dari duplikasi dan kesalahan, kemudian diubah menjadi format yang lebih mudah dianalisis. Pada tahap ini, data yang diperoleh dari media sosial *X* dibersihkan dari duplikasi dan kesalahan untuk memastikan keakuratan dan relevansi analisis sentimen terhadap Program Kartu Indonesia Pintar Kuliah. Proses pembersihan data dimulai dengan menghapus *tweet* yang duplikat. Selain itu, *tweet* yang tidak relevan atau tidak diperlukan juga dihapus, termasuk *tweet* yang mengandung spam, iklan, atau konten yang tidak berkaitan dengan topik penelitian. Hasil akhir dari pembersihan ini adalah 1488 data *tweet* siap digunakan.

**Table 1.** Pembersihan Data

| No | Sebelum | Sesudah |
|---|---|---|
| 1 | @Fitria Orang Miskin Tetep Punya Berpeluang Kuliah Selama Dapet Kipkuliah Orang Berkelas Menengah Yang Kasian Daftar Kipkuliah Dan Tidak Sesuai Syarat Ekonomi Harus Mencari Beasiswa Sambil Kerja | Orang Miskin Tetep Punya Berpeluang Kuliah Selama Dapet Kipkuliah Orang Berkelas Menengah Yang Kasian Daftar Kipkuliah Dan Tidak Sesuai Syarat Ekonomi Harus Mencari Beasiswa Sambil Kerja |
| 2 | @Awang_uye Menjadi Satu Fakta Karenanya Ada Kasus Mahasiswa Dapat Beasiswa Kipkuliah Padahal Bapaknya Menengah Keatas Dan Bahkan Bisa Ganti Mobil Baru | Menjadi Satu Fakta Karenanya Ada Kasus Mahasiswa Dapat Beasiswa Kipkuliah Padahal Bapaknya Menengah Keatas Dan Bahkan Bisa Ganti Mobil Baru |

Tahap selanjutnya yaitu *Transfrom Cases*, merupakan salah satu langkah penting dalam *preprocessing* data *tweet* untuk penelitian analisis sentimen terhadap Program Kartu Indonesia Pintar Kuliah. Proses ini bertujuan untuk mengubah semua teks dalam dataset menjadi huruf kecil, hasil dari proses ini ditunjukkan pada tabel 2.

**Table 2.** *Transfrom Cases*

| No | Sebelum | Sesudah |
|---|---|---|
| 1 | Orang Miskin Tetep Punya Berpeluang Kuliah Selama Dapet Kipkuliah Orang Berkelas Menengah Yang Kasian Daftar Kipkuliah Dan Tidak Sesuai Syarat Ekonomi Harus Mencari Beasiswa Sambil Kerja | orang miskin tetep punya berpeluang kuliah selama dapet kipkuliah orang berkelas menengah yang kasian daftar kipkuliah tidak sesuai syarat ekonomi harus mencari beasiswa sambil kerja |
| 2 | Menjadi Satu Fakta Karenanya Ada Kasus Mahasiswa Dapat Beasiswa Kipkuliah Padahal Bapaknya Menengah Keatas Dan Bahkan Bisa Ganti Mobil Baru | menjadi satu fakta karenanya ada kasus mahasiswa dapat beasiswa kipkuliah padahal bapaknya menengah keatas dan bahkan bisa ganti mobil baru |

Setelah data dibersihkan data yang siap digunakan untuk analisis sebanyak 1488 data, langkah berikutnya adalah memberikan label pada setiap *tweet* berdasarkan sentimen yang diekspresikan. Pelabelan ini dilakukan secara manual yaitu sebanyak 400 data, yang nantinya akan digunakan sebagai data latih. Proses pelabelan dilakukan dengan operator *union*, untuk pelabelan data dibagi menjadi dua yaitu positif dan negatif, hasil dari proses ini ditunjukkan pada tabel 3.

**Table 3.** Tahap Pelebelan

| No | Label | Teks |
|---|---|---|
| 1 | Negatif | orang miskin tetep punya berpeluang kuliah selama dapet kipkuliah orang berkelas menengah yang kasian daftar kipkuliah tidak sesuai syarat ekonomi harus mencari beasiswa sambil kerja |
| 2 | Negatif | menjadi satu fakta karenanya ada kasus mahasiswa dapat beasiswa kipkuliah padahal bapaknya menengah keatas dan bahkan bisa ganti mobil baru |

Proses berikutnya adalah *tokenizing*, yaitu proses memecah teks menjadi unit-unit yang lebih kecil, yang disebut token. Dalam konteks pemrosesan bahasa alami *(Natural Language Processing/NLP)*, token

biasanya berupa kata, frasa, atau karakter individual. Proses ini merupakan langkah penting dalam *preprocessing* data teks karena memungkinkan analisis lebih mendalam terhadap struktur dan makna dari teks tersebut, hasil dari proses ini ditunjukkan pada tabel 4.

**Table 4.** Tahap *Tokenizing*

| No | Sebelum | Sesudah |
|---|---|---|
| 1 | orang miskin tetep punya berpeluang kuliah selama dapet kipkuliah orang berkelas menengah yang kasian daftar kipkuliah tidak sesuai syarat ekonomi harus mencari beasiswa sambil kerja | “orang”, “miskin”, “tetep”, “punya”, “berpeluang”, “kuliah”, “selama”, “dapet”, “kipkuliah”, “orang”, “berkelas”, “menengah”, “yang”, “kasian”, “daftar”, “kipkuliah”, “tidak”, “sesuai”, “syarat”, “ekonomi”, “harus”, “mencari”, “beasiswa”, “sambil”, “kerja” |
| 2 | menjadi satu fakta karenanya ada kasus mahasiswa dapat beasiswa kipkuliah padahal bapaknya menengah keatas dan bahkan bisa ganti mobil baru | “menjadi”, “satu”, “fakta”, “karenanya”, “ada”, “kasus”, “mahasiswa”, “dapat”, “beasiswa”, “kipkuliah”, “padahal”, “bapaknya”, “menengah”, “keatas”, “dan”, “bahkan”, “bisa”, “ganti”, “mobil”, “baru” |

Selanjutnya proses *Stemming*, adalah pemrosesan bahasa alami (*NLP*) yang bertujuan untuk mengurangi kata-kata ke bentuk dasarnya atau akar katanya. Dalam konteks analisis sentimen terhadap Program Kartu Indonesia Pintar Kuliah di media sosial *X*, *stemming* digunakan untuk menyederhanakan variasi kata menjadi bentuk dasar yang sama, sehingga memudahkan analisis teks, hasil dari proses ini ditunjukkan pada tabel 5.

**Table 5.** Tahap *Stemming*

| No | Sebelum | Sesudah |
|---|---|---|
| 1 | orang miskin tetep punya berpeluang kuliah selama dapet kipkuliah orang berkelas menengah yang kasian daftar kipkuliah tidak sesuai syarat ekonomi harus mencari beasiswa sambil kerja | orang miskin tetep punya peluang kuliah selama dapet kipkuliah orang kelas menengah yang kasian daftar kipkuliah tidak sesuai syarat ekonomi harus cari beasiswa sambil kerja |
| 2 | menjadi satu fakta karenanya ada kasus mahasiswa dapat beasiswa kipkuliah padahal bapaknya menengah keatas dan bahkan bisa ganti mobil baru | menjadi satu fakta karena ada kasus mahasiswa dapat beasiswa kipkuliah padahal bapak menengah atas dan bahkan bisa ganti mobil baru |

Tahap selanjutnya *Stopwords* adalah kata-kata yang sering diabaikan atau dihapus selama pemrosesan teks karena tidak memberikan banyak informasi tentang konten atau makna utama teks. Menghapus *stopwords* membantu mengurangi jumlah fitur dalam dataset teks, sehingga membuat analisis lebih efisien dan fokus pada kata-kata yang lebih relevan dan bermakna, hasil dari proses ini ditunjukkan pada tabel 6.

**Table 6.** Tahap *Stopwords*

| No | Sebelum | Sesudah |
|---|---|---|
| 1 | orang miskin tetep punya peluang kuliah selama dapet kipkuliah orang kelas menengah yang kasian daftar kipkuliah tidak sesuai syarat ekonomi harus cari beasiswa sambil kerja | orang miskin punya peluang kuliah orang menengah kasian tidak sesuai syarat ekonomi harus cari beasiswa sambil kerja |
| 2 | menjadi satu fakta karena ada kasus mahasiswa dapat beasiswa kipkuliah padahal bapak menengah atas dan bahkan bisa ganti mobil baru | kasus mahasiswa beasiswa kipkuliah bisa ganti mobil baru |

Tahap selanjutnya yaitu, *Term Frequency-Inverse Document Frequency* (*TF-IDF*) adalah teknik yang digunakan dalam penambangan teks dan pemrosesan bahasa alami untuk mengevaluasi seberapa penting suatu kata dalam sebuah dokumen relatif terhadap kumpulan dokumen (*corpus*), hasil dari proses ini ditunjukkan pada tabel 7.

**Table 7.** Tahap *TF-IDF*

| No | program | beasiswa | kipkuliah | sangat | membantu |
|---|---|---|---|---|---|
| 1 | 0,875 | 0,785 | 0,899 | 0,523 | 0,190 |
| 2 | 0,576 | 0,422 | 0,721 | 0,280 | 0,126 |

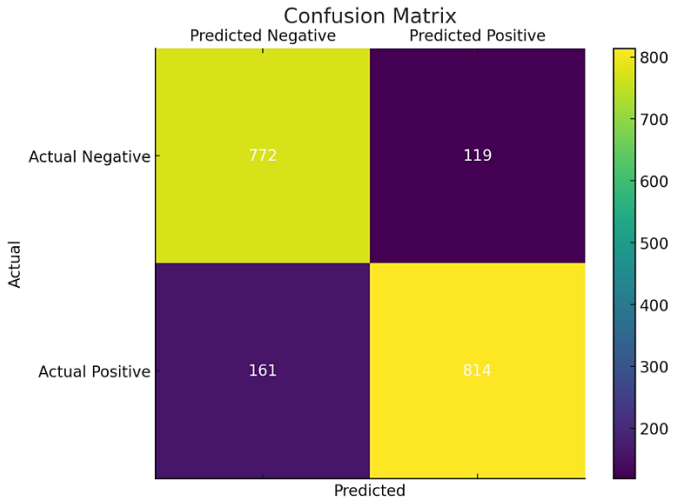
### 3.2 Pembuatan Model

Pada tahap pemodelan, proses yang dilakukan menghasilkan sebuah model klasifikasi yang mampu mengklasifikasikan sentimen *tweet* terkait program Kartu Indonesia Pintar (KIP) Kuliah menggunakan algoritma *Naive Bayes* [18]. Berikut beberapa tahapan yang dilakukan:

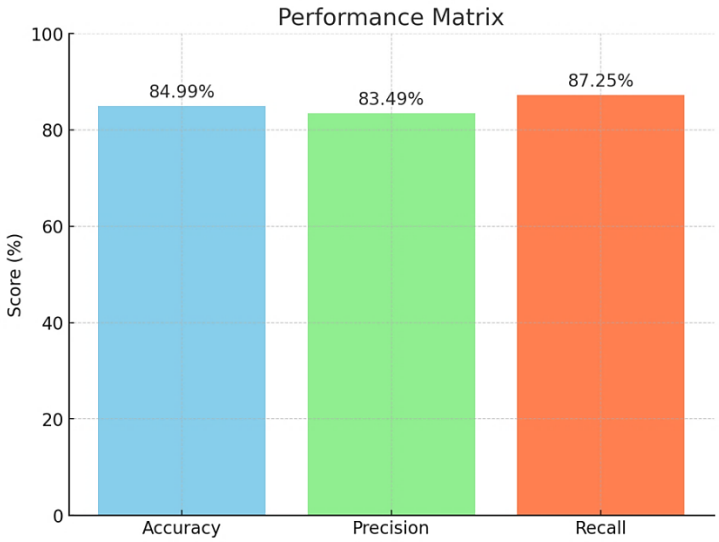
1. Tahap pertama membuat model klasifikasi *Naive Bayes* dan data latih yang nanti akan digunakan dalam proses sentimen analisis.
2. Tahap selanjutnya menggabungkan data latih *(training data)* dan data uji *(testing data)*, Untuk memastikan data yang siap digunakan pada pelatihan ulang model atau analisis lanjutan. Langkah-langkah penggabungan data ini dilakukan untuk menyatukan kedua set data dengan operator *union* yang sebelumnya sudah dipisahkan, sehingga bisa mendapatkan gambaran yang lebih menyeluruh tentang performa model dan distribusi data.

### 3.3 Evaluasi Model

Untuk menilai performa model secara lebih rinci penelitian ini, menggunakan *confusion matrix* yang memberikan gambaran mengenai jumlah prediksi benar dan salah yang dibuat oleh model. *Confusion matrix* ini memungkinkan peneliti untuk menghitung metrik evaluasi lainnya seperti presisi, *recall*, dan *F1-score*. Dan untuk meningkatkan performa model, peneliti menerapkan teknik *SMOTE (Synthetic Minority Over-sampling Technique)* berguna untuk menangani masalah ketidakseimbangan kelas dalam data. Untuk menunjukkan dan membuktikan hasil prediksi dari model *Naive Bayes*. Hasil evaluasi *confusion matrix* ditunjukkan seperti gambar 4 dan gambar 5.




**Gambar 4.** *Confusion Matrix*



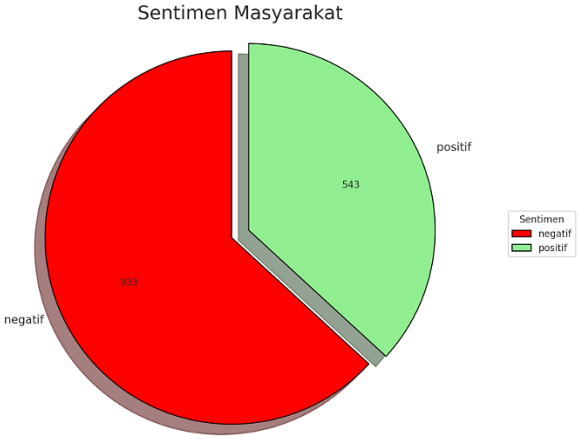

**Gambar 5.** *Performance Matrix*

Bahwa *precision* menunjukkan tingkat ketepatan data yang diprediksi positif terhadap banyaknya data yang benar diprediksi positif. Dari hasil penelitian ini, presisi untuk data bersentimen positif adalah 83.54%, sedangkan untuk data bersentimen negatif memiliki presisi sebesar 87.25%. Hal ini menunjukkan bahwa dari semua prediksi positif yang dibuat model, 83.54% adalah benar-benar positif, dan dari semua prediksi negatif yang dibuat model, 87.25% adalah benar negatif. *Recall* untuk sentimen negatif *(Specificity)* adalah 87.25%,

sehingga dapat disimpulkan bahwa model dapat menemukan kembali informasi atau data yang benar-benar negatif dengan sangat baik. Sedangkan pada sentimen positif, *recall* adalah 83.54%. Ini berarti bahwa dari semua data positif yang ada, model berhasil memprediksi 83.54% dengan benar. Nilai akurasi yang dihasilkan menggunakan model *Naive Bayes* adalah 84.99%, sehingga dapat disimpulkan bahwa algoritma *Naive Bayes* dapat mengklasifikasikan sentimen dengan baik menggunakan data dari tweet tentang Program Kartu Indonesia Pintar Kuliah.

### 3.4 Tahap Visualisasi

Tahapan ini merupakan tahap akhir dalam pembuatan laporan hasil kegiatan data mining. Pada tahap ini, laporan akhir disusun untuk menyampaikan pengetahuan yang diperoleh atau pengenalan pola dalam proses data mining, dalam hal ini, sentimen masyarakat terhadap Program Kartu Indonesia Pintar Kuliah. Hasil analisis menunjukkan bahwa mayoritas publik memiliki pandangan negatif terhadap program tersebut, dengan 933 *tweet* bersentimen negatif dibandingkan dengan 544 *tweet* bersentimen positif. Meskipun demikian, terdapat beberapa pandangan positif yang menunjukkan apresiasi terhadap program KIP Kuliah.




**Gambar 6.** Distribusi Sentimen


**Gambar 7.** *World Cloud* KIPK

Seperti gamnbar diatas *Word cloud* adalah teknik visualisasi teks di mana kata-kata yang lebih sering muncul dalam teks ditampilkan dengan ukuran yang lebih besar atau lebih menonjol. Kata-kata yang jarang muncul ditampilkan dengan ukuran yang lebih kecil. *Word cloud* membantu dalam mengidentifikasi kata-kata kunci dan tema yang mendominasi dalam kumpulan teks dengan cepat dan intuitif. Dengan *Word Cloud* dapat menampilkan kata-kata yang paling sering muncul dalam *tweet* terkait KIP Kuliah. Kata-kata yang lebih besar dalam *word cloud* menunjukkan frekuensi kemunculan yang lebih tinggi. Visualisasi ini membantu mengidentifikasi topik atau isu yang sering dibahas dalam *tweet* [18].

## 4. KESIMPULAN

Berdasarkan hasil penelitian yang bertujuan untuk menganalisis sentimen terhadap Program Kartu Indonesia Pintar Kuliah (KIP-K) menggunakan algoritma *Naive Bayes* pada data *tweet* yang dikumpulkan pada media sosial *X*. Beberapa kesimpulan dapat ditarik, Bahwa algoritma *Naive Bayes* terbukti efektif dalam menganalisis sentimen terkait program KIP-K, dengan akurasi mencapai 84.99%. Dengan presisi sentimen positif sebesar 83.54%, dan presisi sentimen negatif sebesar 87.25%. Hal ini menunjukkan bahwa algoritma

ini mampu mengidentifikasi sentimen positif dan negatif dengan cukup akurat. Hasil analisis sentimen menunjukkan bahwa sentimen negatif lebih dominan dibandingkan sentimen positif terhadap program KIP-K. Dominasi sentimen negatif ini mungkin mencerminkan permasalahan yang diidentifikasi dalam penelitian ini, yaitu proses verifikasi data ekonomi keluarga penerima manfaat yang tidak transparan, distribusi dana yang tidak merata, dan kurangnya sosialisasi serta informasi mengenai program KIP-K. Temuan ini menegaskan bahwa ada kebutuhan mendesak untuk meningkatkan transparansi, keadilan dalam distribusi dana, dan sosialisasi program untuk memastikan bahwa manfaat program KIP-K dapat dirasakan secara merata oleh masyarakat yang benar-benar membutuhkan. Berdasarkan temuan penelitian ini, beberapa saran dapat diajukan untuk meningkatkan efektivitas dan penerimaan program KIP-K di masyarakat. Dengan meningkatan transparansi dalam proses verifikasi data penerima manfaat untuk mengurangi ketidakpuasan dan meningkatkan kepercayaan masyarakat terhadap program KIP-K. Disarankan agar penelitian selanjutnya menggunakan dataset yang lebih besar dan lebih bervariasi untuk meningkatkan keakuratan dan relevansi hasil penelitian. Peneliti juga sebaiknya mempertimbangkan penggunaan algoritma lain dalam analisis sentimen. Untuk membandingkan performa dan memperoleh gambaran yang lebih komprehensif mengenai persepsi masyarakat terhadap program KIP-K. Dari segi teknis, disarankan untuk mengembangkan sebuah dashboard *real-time* yang memvisualisasikan hasil analisis sentimen terhadap program KIP-K. Dashboard ini akan memungkinkan pemangku kepentingan untuk memantau sentimen publik secara terus menerus dan responsif terhadap isu-isu yang muncul.

**REFERENSI**


[1] L. E. Wahyudi *et al.*, "Mengukur Kualitas Pendidikan di Indonesia," vol. 1, no. 1, pp. 18–22, 2022.DOI : https://jurnal.maarifnumalang.id.

[2] Kemdikbud, *Kajian Program Indonesia Pintar (Pip): Strategi Penjangkauan Anak Tidak Sekolah (Ats) Untuk Mengikuti Pendidikan Melalui Program Indonesia Pintar (Pip)*. 2018. DOI: https://doi.org/10.52626/jg.v5i3.193.

[3] T. Krisdiyanto, "Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 7, no. 1, p. 32, 2021, doi: 10.24014/coreit.v7i1.12945.

[4] J. Juli, S. N. Tanjungbalai, D. Amrizal, D. F. Nasution, and A. Imran, "Efektivitas Pelaksanaan Program Kartu Indonesia Pintar ( KIP ) Dalam Rangka Peningkatan Kualitas Pendidikan Di," vol. 1, no. 1, pp. 9–15, 2020.DOI: https://doi.org/10.53695/js.v1i1.27.

[5] D. B. Siswanto, D. Normawati, F. T. Industri, U. Ahmad, D. Yogyakarta, and K. Bantul, "Sistem Klasifikasi Monitoring dan Evaluasi Kelayakan Penerima Beasiswa UAD Menggunakan Algoritma Naïve Bayes Indonesia terkait Sumber Daya Manusia," vol. 9, pp. 2–6, 2023. DOI: https://doi.org/10.33020/saintekom.v13i2.428.

[6] P. I. P. Di, K. Tuan, K. Tahun, D. Astuti, D. Febriyanti, and M. Q. Kariem, "EVALUASI KEBIJAKAN PROGRAM KARTU INDONESIA PINTAR," vol. 4, no. 3, pp. 249–256, 2023. DOI: https://doi.org/10.55314/tsg.v4i3.435 Hal. 249-256

[7] Y. Akbar and T. Sugiharto, "Analisis Sentimen Pengguna Twitter di Indonesia Terhadap ChatGPT Menggunakan Algoritma C4.5 dan Naïve Bayes (Yuma Akbar 1*, Tri Sugiharto 2 ) Analisis Sentimen Pengguna Twitter di Indonesia Terhadap ChatGPT Menggunakan Algoritma C4.5 dan Naïve Bayes," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 115–122, 2023. DOI : https://doi.org/10.55338/saintek.v4i3.1368.

[8] P. Kip, J. Menggunakan, and K. C. Validation, "JURNAL RESTI Komparasi Algoritma K-Nearest Neighbor dan Naïve Bayes Dalam," vol. 5, pp. 1–9, 2022. DOI : http://jurnal.iaii.or.id

[9] Z. Saputra, D. Sartika, and M. H. Irfani, "Prediksi Calon Mahasiswa Penerima KIP Pada Universitas Indo Global Mandiri menggunakan Algoritma Decision Tree," vol. 4, no. 3, pp. 231–240, 2024. DOI : https://djournals.com/resolusipengambilan

[10] M. R. Romadhon, M. Faisal, and M. Imamudin, "Improving The Performance of the K-Nearest Neighbor Algorithm in the Selection of KIP Scholarship Recipients," *J. Ris. Inform.*, vol. 5, no. 4, pp. 465–470, 2023, DOI: 10.34288/jri.v5i4.575.

[11] M. Harahap, Y. Lubis, and Z. Situmorang, "Analisis Pemasaran Bisnis dengan Data Science : Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering," vol. 1, no. 1, pp. 76–88, 2022. DOI : https://doi.org/10.47709/dsi.v1i2.1348

[12] N. S. Marga, "Sentimen Analisis Tentang Kebijakan Pemerintah Terhadap Kasus Corona Menggunakan Metode Naive Bayes," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 2, no. 4, pp. 453–463, 2022, DOI: 10.33365/jatika.v2i4.1602.

[13] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 27–29, 2020, DOI: 10.30864/jsi.v15i1.332.

[14] G. S. Mengga and M. Ronal, "Analisis Tingkat Pemahaman Masyarakat Tana Toraja Terhadap

Investasi Keuangan," vol. 3, pp. 2438–2449, 2023.DOI: https://j-innovative.org/index.php/Innovative Analisis

[15] E. Laia and M. Yamin, "Penerapan Algoritma Naïve Bayes dalam Menganalisis Sentimen pada Review Pengguna E-Commerce," *Media Online)*, vol. 4, no. 1, pp. 305–316, 2023, DOI: 10.30865/klik.v4i1.1186.

[16] F. Matheos Sarimole and K. Kudrat, "Analisis Sentimen Terhadap Aplikasi Satu Sehat Pada Twitter Menggunakan Algoritma Naive Bayes Dan Support Vector Machine," *J. Sains dan Teknol.*, vol. 5, no. 3, pp. 783–790, 2024, DOI: 10.55338/saintek.v5i3.2702.

[17] A. Imron, "Analisis Sentimen Terhadap Tempat Wisata di Kabupaten Rembang Menggunakan Metode Naive Bayes Classifier," *Tek. Inform.*, pp. 10–13, 2019,DOI: https://dspace.uii.ac.id/handle/123456789/14268

[18] M. A. A. A. Solichin, "Analisis Sentimen MotoGP Mandalika Pada Twitter Menggunakan Metode Naïve Bayes," *J. Ticom Technol. Inf. Commun.*, vol. 11, no. Vol 11 No 1 (2022): Jurnal Ticom: Technology of Information and Communication, pp. 20–25, 2022, DOI: https://jurnal-ticom.jakarta.aptikom.or.id/index.php/Ticom/article/view/66/55.

[19] B. Seref, "Bayes dan Complement Naive Bayes Classifier pada Hadoop Framework", vol. 12, no. 2, pp. 24-25, DOI : 10.1109/ic-ETITE47903.2020.201.

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