Jurnal Kajian dan Terapan Matematika
Volume 12, Edisi 2, Bulan Juli 2026, 30 – 46
http://journal.student.uny.ac.id/ojs/index.php/jktm

ANALISIS SENTIMEN MASYARAKAT TERHADAP PENERIMA KIPK BERDASARKAN PENGGUNA TWITTER (X) MENGGUNAKAN METODE SVM DAN *NAIVE BAYES*

PUBLIC SENTIMENT ANALYSIS TOWARD KIP-K BASED ON TWITTER (X) USERS USING SVM AND NAÏVE BAYES METHODS

Chintya Wijayanti*, Prodi Matematika FMIPA UNY
Dhoriva Urwatul Wutsqa, Prodi Matematika FMIPA UNY
*e-mail: cichintyaaaw@gmail.com

Abstrak
Media sosial X (Twitter) menjadi salah satu *platform* yang banyak digunakan masyarakat untuk menyampaikan opini terhadap berbagai kebijakan pemerintah, termasuk program Kartu Indonesia Pintar (KIP-K). penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap penerima KIP-K menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM), serta membandingkan performa kedua metode tersebut. Penelitian menggunakan pendekatan kuantitatif dengan metode *text mining* dan *sentiment analysis*. Data diperoleh melalui proses *crawling tweet* menggunakan kata kunci ”KIPK”, ”KIP-K”, ”KIP-Kuliah”, dan ”KIP Kuliah” pada periode 2023-2025. Tahapan *preprocessing* meliputi *case folding*, *cleaning*, normalisasi kata, *tokenizing*, *filtering (stopwords removal)*, dan *stemming*. Selanjutnya dilakukan pembobotan TF-IDF dan klasifikasi menggunakan algoritma *Naive Bayes* dan SVM. Hasil penelitian menunjukkan bahwa sentimen masyarakat terhadap penerima KIP-K didominasi oleh sentimen negatif. Metode *Naive Bayes* menghasilkan akurasi sebesar 70,3% dan 69,4% pada pembagian dta 80%:20% dan 700%:30%. Sementara itu, metode SVM menghasilkan akurasi sebesar 73,6% dan 72,3% pada pembagian data yang sama. Berdasarkan hasil tersebut, metode SVM memiliki performa lebih baik dibandingkan *Naive Bayes* dalam klasifikasi sentimen data *tweet* terkait penerima KIP-K.
Kata kunci: analisis sentimen, KIP-K, *machine learning*, *Naïve Bayes*, *Support Vector Machine*.

Abstract
Social media platform X (Twitter) has become one of the main platforms used by the public to express opinions regarding government policies, including the Kartu Indonesia Pintar Kuliah (KIP-K) program. This study aims to analyze public sentiment toward KIP-K recipients using the Naïve Bayes and Support Vector Machine (SVM) methods, as well as compare the performance of both algorithms. The study employeda quantitative approach using text mining and sentiment analysis methods. Data were collected through tweet crawling using the keywords “KIPK”, “KIP-K”, “KIP-Kuliah”, and “KIP Kuliah”. During the 2023-2025 period. The preprocessing stages included case folding, cleaning, word normalization, tokenizing, filtering (stopwords removal), and stemming. TF-IDF weighting and sentiment classification were then conducted using Naïve Bayes and SVM algorithms. The results showed that public senntimient toward KIP-K recipients was dominated by negative sentiment. The Naïve Bayes method achieved an accuracy of 70,3% and 69,4% using 80%:20% and 70%:30% data split, respectively. Meanwhile, the SVM method achieved an accuracy 73,6% and 72,3% using the same data splits. Based on these findings, the SVM method demonstrated better performance than Naïve Bayes in classifying tweet sentiment related to KIP-K recipients.
Keywords*: KIP-K, machine learning, Naïve Bayes, sentiment analysis, Support Vector Machine*

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi menyebabkan media sosial menjadi salah satu sarana utama masyarakat dalam menyampaikan opini terhadap berbagai isu publik. Media sosial memungkinkan pengguna untuk berinteraksi, berbagi informasi, serta menyampaikan pandangan secara cepat tanpa dibatasi ruang dan waktu. Salah satu *platform* yang aktif digunakan masyarakat Indonesia adalah media sosial X (Twitter). Berdasarkan data pengguna media sosial di Indonesia, jumlah pengguna Twitter terus mengalami peningkatan dari tahun ke tahun seiring dengan meningkatnya penggunaan internet dan perangkat digital. Kondisi tersebut menjadikan media sosial sebagai sumber data yang potensial dalam mengetahui persepsi masyarakat terhadap suatu kebijakan pemerintah.

Salah satu program pemerintah yang banyak dibahas di media sosial adalah Kartu Indonesia Pintar Kuliah (KIP-K). Program KIP-K merupakan bantuan pendidikan dari pemerintah yang ditujukan kepada mahasiswa dari keluarga kurang mampu untuk meningkatkan pemerataan akses pendidikan tinggi. Program ini memberikan bantuan biaya pendidikan dan biaya hidup bagi mahasiswa penerima sehingga diharapkan mampu mengurangi hambatan ekonomi dalam menempuh pendidikan tinggi. Namun, dalam pelaksanaannya, program KIP-K masih menimbulkan berbagai tanggapan dari masyarakat, baik berupa dukungan maupun kritik terkait ketidaktepatan sasaran penerima bantuan.

Berbagai opini mengenai penerima KIP-K banyak disampaikan melalui media sosial Twitter dalam bentuk komentar, kritik, maupun dukungan. Sebagian masyarakat memberikan sentimen positif karena program ini dinilai membantu mahasiswa kurang mampu untuk melanjutkan pendidikan. Akan tetapi, tidak sedikit pula masyarakat yang memberikan sentimen negatif terkait dugaan ketidaktepatan sasaran penerima bantuan, kecemburuan sosial, hingga penyalahgunaan dana bantuan. Banyaknya opini yang muncul di media sosial menyebabkan proses identifikasi sentimen secara manual menjadi tidak efekktif dan berpotensi menimbulkan subjektivitas dalam penilaian. Oleh karena itu, diperlukan suatu metode yang mampu mengelompokkan opini masyarakat secara otomatis dan objektif.

Salah satu pendekatan yang dapat digunakan untuk memahami opini masyarakat adalah analisis sentimen. Analisis sentimen merupakan bagian dari *text mining* yang digunakan untuk mengelompokkan opini ke dalam kategori tertentu, seperti positif, negatif, dan netral. Analisis sentimen memanfaatkan teknik *Natural Language Processing* (NLP) dan *machine learning* untuk mengolah data teks dalam jumlah besar sehingga dapat memberikan gambaran mengenai kecenderungan opini masyarakat terhadap suatu topik. Dalam penelitian analisis sentimen, tahapan *preprocessing* memiliki peran penting karena data media sosial umumnya bersifat tidak terstruktur dan banyak menggunakan bahasa informal, singkatan, maupun *slang*. Tahapan *preprocessing* seperti *case folding*, *cleaning*, normalisasi kata, *tokenizing*, *filtering (stopwords removal*), dan *stemming* dilakukan agar data dapat diproses secara lebih optimal.

Metode *machine learning* yang sering digunakan dalam analisis sentimen adalah *Naïve Bayes* dan *Support Vector Machine* (SVM). *Naïve Bayes* merupakan metode klasifikasi probabilistik yang sederhana dan efisien dalam pengolahan data teks. Meskipun menggunakan asumsi independensi antar fitur, metode ini tetap mampu menghasilkan performa klasifikasi yang cukup baik pada analisis sentimen. Sementara itu, *Support Vector Machine* (SVM) dikenal sebagai metode klasifikasi yang mampu menangani data berdimensi tinggi dan memiliki performa yang baik dalam klasifikasi teks karena dapat menemukan *hyperplane* optimal sebagai pemisah antar kelas data.

Beberapa penelitian sebelumnya menunjukkan bahwa metode *Naive Bayes* dan SVM dapat digunakan secara efektif dalam analisis sentimen media sosial. Penelitian oleh Amelia et al. (2024) menunjukkan bahwa metode SVM mampu memberikan performa klasifikasi yang baik dalam analisis sentimen terhadap program KIP-Kuliah di media sosial Twitter. Penelitian

lain oleh Ningsih et al. (2024) yang membandingkan metode SVM dan *Naive Bayes* pada analisis sentimen Twitter terkait mobil listrik menunjukkan bahwa SVM menghasilkan tingkat akurasi yang lebih tinggi dibandingkan *Naive Bayes*. Selain itu, Husen et al. (2023) juga menyatakan bahwa tahapan *preprocessing* memiliki pengaruh penting terhadap kualitas hasil klasifikasi sentimen pada data media sosial.

Berdasarkan penelitian sebelumnya, analisis sentimen menggunakan metode *Naive Bayes* dan SVM telah banyak diterapkan pada berbagai isu publik. Namun, penelitian mengenai analisis sentimen masyarakat terhadap penerima KIP-K masih relatif terbatas, khususnya yang membandingkan performa kedua algoritma tersebut pada data media sosial Twitter. Oleh karena itu, penelitian ini memiliki nilai kebaruan dalam bentuk penerapan dan perbandingan metode *Naive Bayes* dan SVM pada analisis sentimen terkait penerima KIP-K menggunakan data Twitter. Selain itu, penelitian ini diharapkan dapat memberikan gambaran mengenai kecenderungan opini masyarakat terhadap penerima KIP-K serta menjadi bahan evaluasi bagi pemerintah dalam meningkatkan kualitas pelaksanaan program bantuan pendidikan.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap penerima KIP-K berdasarkan data pengguna media sosial Twitter (X) menggunakan metode *Naive Bayes* dan *Support Vector Machine* (SVM), serta membandingkan performa kedua metode tersebut dalam klasifikasi sentimen.

METODE

Deskripsi Data

Data yang digunakan dalam penelitian ini berasal dari twitter. Data yang digunakan adalah data teks berupa *tweet* yang berisi sentimen pengguna twitter. Sentimen twitter yang digunakan adalah sentimen yang membahas mengenai penerima Kartu Indonesia Pintar Kuliah (KIP-K). Teknik pengumpulan data menggunakan metode *crawling* dengan bantuan Google Colab dan *Tweet Harvest* (Satria & Matias, 2023). Pengambilan data dilakukan menggunakan *keyword* ”KIPK”, ”KIP-K”, ”KIP-Kuliah”, dan ”KIP Kuliah” pada rentang waktu 1 Januari 2023 hingga 31 Desember 2025 dengan batas pengambilan sebanyak 10.000 *tweet*. Label data yang digunakan dalam penelitian ini adalah positif, negatif, dan netral. Jumlah data yang berhasil dikumpulkan dan digunakan dalam penelitian ini sebanyak 9866 *tweet*.

Klasifikasi dan Evaluasi Model

Penelitian ini menggunakan metode *Naive Bayes* dan Support Vector Machine (SVM) untuk melakukan klasifikasi sentimen masyarakat terhadap penerima Kartu Indonesia Pintar Kuliah (KIP-K). Sebelum proses klasifikasi dilakukan, data teks terlebih dahulu melalui proses *preprocessing* dan pembobotan TF-IDF sehingga dapat diubah menjadi representasi numerik.

1. *Naive Bayes*

Naïve Bayes merupakan metode klasifikasi probabilistik yang menggunakan pendekatan Teorema Bayes dalam menentukan probabilitas suatu data terhadap kelas tertentu (Zhang, 2004). Metode ini memiliki asumsi bahwa setiap fitur bersifat independen satu sama lain. Pada penelitian ini, metode Naïve Bayes digunakan untuk mengklasifikasikan sentimen tweet ke dalam kategori positif, negatif, dan netral.

Secara matematis, Teorema Bayes dapat dituliskan sebagai berikut:

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Dengan,

A : Data yang akan diklasifikasikan
B : Kelas atau kategori data
$P(A)$: Peluang munculnya data *A*

$P(A \mid B)$: Peluang data A didapat dari kelas B
$P(B)$: Peluang munculnya kelas B
$P(B \mid A)$: Peluang kelas B didapat dari data A

2. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan metode klasifikasi yang bekerja dengan mencari *hyperplane* optimal sebagai pemisah antar kelas data (Cortes & Vapnik, 1995). Metode ini dikenal mampu menangani data berdimensi tinggi seperti data teks sehingga banyak digunakan dalam analisis sentimen.

Persamaan *hyperplane* pada metode SVM dapat dituliskan sebagai berikut:

$$\vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}} + b = 0 \tag{2}$$

Dengan,
$\vec{\boldsymbol{w}}$ = Bobot *vector* berukuran $(1 \times n)$
$\vec{\boldsymbol{x}}_{\boldsymbol{i}}$ = *Vector* data berukuran $(1 \times n)$
b = Bias

Pada penelitian ini digunakan kernel linear yang dinyatakan dengan fungsi berikut:

$$K(\vec{\boldsymbol{x}}, \vec{\boldsymbol{x}_{\boldsymbol{i}}}) = sum(\vec{\boldsymbol{x}} \,.\, \vec{\boldsymbol{x}_{\boldsymbol{i}}}) \tag{3}$$

Dengan,
$\vec{\boldsymbol{x}}$ = vektor input berukuran $(1 \times n)$
$\vec{\boldsymbol{x}_{\boldsymbol{i}}}$ = vektor input yang memiliki nilai pada setiap fitur berukuran $(1 \times n)$

3. Evaluasi Model

Evaluasi performa model klasifikasi dilakukan menggunakan *confusion matrix* berdasarkan nilai *accuracy, precision, recall*, dan *F1-score* (Tharwat, 2018).

Nilai *accuracy* digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan data dan dihitung menggunakan persamaan berikut:

$$Akurasi = \frac{TP+TNe+TNt}{Total} \tag{4}$$

Nilai precision digunakan untuk mengukur ketepatan hasil prediksi positif oleh model dan dihitung menggunakan persamaan berikut:

$$Presisi = \frac{TP}{TP+FP} \tag{5}$$

Nilai *recall* digunakan untuk mengukur kemampuan model dalam menemukan data positif dan dihitung menggunakan persamaan berikut:

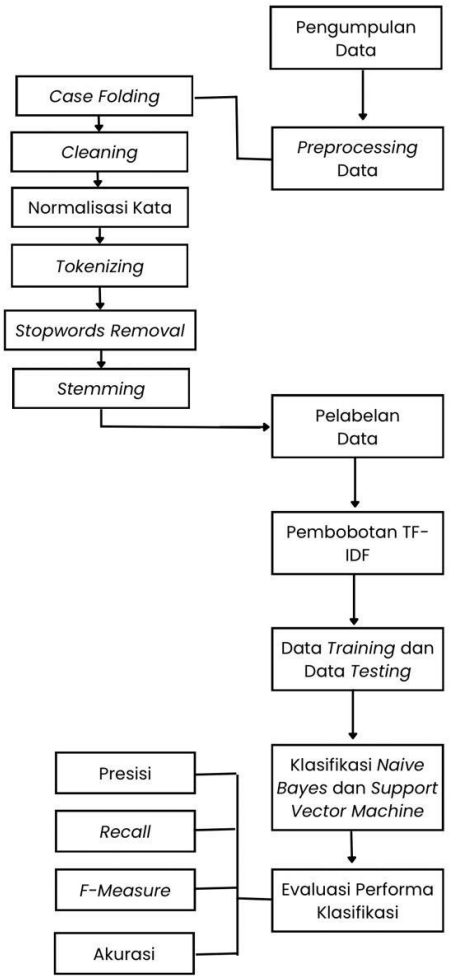
$$Recall = \frac{TP}{TP+FNe+FNt} \tag{6}$$

Nilai F1-score digunakan untuk mengukur keseimbangan antara precision dan recall dan dihitung menggunakan persamaan berikut:

$$F1 - Score = 2 \times \frac{presisi \times recall}{presisi+recall} \tag{7}$$

Prosedur Analisis Data

Pada Gambar 1 terdapat beberapa proses yang dilakukan dalam penelitian ini setelah dilakukan pengumpulan dan pelabelan data, yaitu *preprocessing* data, pelabelan data, pembobotan TF-IDF, pembagian data *training* dan *testing*, klasifikasi menggunakan metode *Naive Bayes* dan *Support Vector Machine* (SVM), serta evaluasi performa model menggunakan *confusion matrix*. Hasil evaluasi kemudian dibandingkan berdasarkan nilai akurasi, presisi, *recall*, dan *f-measure* pada masing-masing metode klasifikasi.



Gambar 1. Tahapan Analisis Data

1. Data penelitian dikumpulkan dari media sosial X (Twitter) dengan menggunakan metode *crawling* berdasarkan kata kunci yang telah ditentukan.
2. Data yang telah diperoleh kemudian dipersiapkan melalui tahap *preprocessing* sebelum dilakukan proses selanjutnya. Tahapan ini meliputi *case folding*, *cleansing*, normalisasi kata, *tokenizing*, *stopwords removal*, dan *stemming*.
3. Data yang telah melalui tahap *preprocessing* selanjutnya diberikan label sentimen berupa positif, negatif, atau netral. Proses pelabelan dilakukan sebagai acuan (data aktual) dalam proses klasifikasi.
4. Data penelitian kemudian dibagi menjadi dua bagian, yaitu data latih (*training*) dan data uji (*testing*). Pembagian data dilakukan secara acak dengan perbandingan tertentu.
5. Membangun model klasifikasi menggunakan algoritma *Naive Bayes* dan *Support Vector Machine* (SVM) berdasarkan data latih yang telah disiapkan.

6. Melakukan proses klasifikasi terhadap data uji menggunakan model yang telah dibangun untuk mengetahui hasil prediksi sentimen.
7. Melakukan evaluasi kinerja model klasifikasi menggunakan *confusion matrix* dengan menghitung nilai akurasi, presisi, *recall*, *F-1 Score*. Hasil evaluasi digunakan untuk membandingkan performa metode *Naïve Bayes* dan SVM.

HASIL DAN PEMBAHASAN

Hasil

Algoritma klasifikasi yang digunakan sebagai pendekatan analisis sentimen pada penelitian ini yaitu *Naive Bayes* dan *Support Vector Machine* (SVM) dengan menerapkan pembbobotan TF-IDF pada dataset *tweet* terkait penerima Kartu Indonesia Pintar Kuliah (KIP-K). Penelitian ini menggunakan dua skenario pembagian data, yaitu 80%:20% dan 70%:30%, untuk mengetahui performa masing-masing metode dalam klasifikasi sentimen. Oleh karena itu, terdapat empat model klasifikasi yang dibandingkan untuk memperoleh metode terbaik dalam mengklasifikasikan sentimen masyarakat terhadap penerima KIP-K. Proses pembentukan model dilakukan melalui beberapa tahapan, yaitu *preprocessing* data, pelabelan sentimen, pembobotan TF-IDF, pembagian data *training* dan *testing*, klasifikasi menggunakan metode *Naive Bayes* dan SVM, serta evaluasi model menggunakan *confusion matrix* berdasarkan nilai akurasi, presisi, *recall*, dan *f-1 score*.

1. *Preprocessing* data

Tahap pertama yang dilakukan dalam *preprocessing* data adalah *case folding*. *Case folding* bertujuan untuk mengubah semua huruf menjadi seragam hurufnya yaitu huruf kecil (*lowercase*). Tabel 1 merupakan contoh dari proses *case folding*.

Tabel 1. *Case Folding*

No	*Tweet*	Hasil *Case Folding*
5.	kuliah nder. coba daftar kipk. kmaren aku gratis bhkn dpt uang semesteran dri kipk. ambil ptn yg deket domisili jd ga perlu byk keluar uang (aku kmaren hrs ngekos jd ga bisa nabung). why aku suruh kuliah? emg seberguna itu kah? SANGAT BERGUNA	kuliah nder. coba daftar kipk. kmaren aku gratis bhkn dpt uang semesteran dri kipk. ambil ptn yg deket domisili jd ga perlu byk keluar uang (aku kmaren hrs ngekos jd ga bisa nabung). why aku suruh kuliah? emg seberguna itu kah? sangat berguna
33.	@sbmptnfess Aku keterima KIPK tapi masih bingung soal pencairan dana sksksk	@sbmptnfess aku keterima kipk tapi masih bingung soal pencairan dana sksksk
122.	@softcookiezs @starc0zy @schfess Katanya kipk byk yg ga tepat sasaran, masih byk yg mampu tetap dpat jir	@softcookiezs @starc0zy @schfess katanya kipk byk yg ga tepat sasaran, masih byk yg mampu tetap dpat jir

Tahap yang kedua adalah *cleaning*. *Cleaning* merupakan proses membersihkan dokumen dari kata atau karakter yang tidak diperlukan untuk mengurangi *noise*. Karakter yang

dihapus meliputi karakter HTML, emoji, *hashtag* (#), *username* (*@username*), URL, email, angka, dan tanda baca yang tidak diperlukan. Tabel 2 merupakan contoh dari proses *cleaning*.

Tabel 2. *Cleaning*

No	*Tweet*	Hasil *Cleaning*
5.	kuliah nder. coba daftar kipk. kmaren aku gratis bhkn dpt uang semesteran dri kipk. ambil ptn yg deket domisili jd ga perlu byk keluar uang (aku kmaren hrs ngekos jd ga bisa nabung). why aku suruh kuliah? emg seberguna itu kah? SANGAT BERGUNA	kuliah nder coba daftar kipk kmaren aku gratis bhkn dpt uang semesteran dri kipk ambil ptn yg deket domisili jd ga perlu byk keluar uang aku kmaren hrs ngekos jd ga bisa nabung why aku suruh kuliah emg seberguna itu kah sangat berguna
33.	@sbmptnfess Aku keterima KIPK tapi masih bingung soal pencairan dana sksksk	aku keterima kipk tapi masih bingung soal pencairan dana sksksk
122.	@softcookiezs @starc0zy @schfess Katanya kipk byk yg ga tepat sasaran, masih byk yg mampu tetap dpat jir	katanya kipk byk yg ga tepat sasaran masih byk yg mampu tetap dpat jir

Tahap yang ketiga adalah normalisasi kata. Normalisasi kata yaitu proses menyeragamkan kata yang memiliki makna sama tetapi ditulis dalam bentuk berbeda. Tahap ini dilakukan untuk memperbaiki kata tidak baku, bahasa gaul (*slang*), dan kata singkatan menjadi bentuk baku agar data lebih konsisten dan hasil analisis lebih akurat. Tabel 3 merupakan contoh dari proses normalisasi kata.

Tabel 3. Normalisasi Kata

No	*Tweet*	Hasil Normalisasi Kata
5.	kuliah nder coba daftar kipk kmaren aku gratis bhkn dpt uang semesteran dri kipk ambil ptn yg deket domisili jd ga perlu byk keluar uang aku kmaren hrs ngekos jd ga bisa nabung why aku suruh kuliah emg seberguna itu kah sangat berguna	kuliah coba daftar kipk kemarin aku gratis bahkan dapat uang semesteran dari kipk ambil perguruan tinggi negeri yang dekat domisili jadi tidak perlu banyak keluar uang aku kemarin harus ngekos jadi tidak bisa nabung mengapa aku menyuruh kuliah memang seberguna itu kah sangat berguna
33.	@sbmptnfess Aku keterima KIPK tapi masih bingung soal pencairan dana sksksk	aku keterima kipk tapi masih bingung soal pencairan dana

No	*Tweet*	Hasil Normalisasi Kata
122.	@softcookiezs @starc0zy @schfess Katanya kipk byk yg ga tepat sasaran, masih byk yg mampu tetap dpat jir	katanya kipk banyak yang tidak tepat sasaran masih banyak yang mampu tetap dapat

Tahap yang keempat yaitu *tokenizing*. *Tokenizing* merupakan proses untuk memecah teks menjadi unit-unit yang lebih kecil, seperti kata atau token, sehingga memudahkan proses analisis dan pengolahan teks lebih lanjut (Liu, 2012). Proses ini dapat mempermudah perhitungan polaritas untuk melakukan *labelling*. Tabel 4 merupakan contoh dari proses *tokenizing*.

Tabel 4. *Tokenizing*

No	*Tweet*	Hasil *Tokenizing*
5.	kuliah coba daftar kipk kemarin aku gratis bahkan dapat uang semesteran dari kipk ambil perguruan tinggi negeri yang dekat domisili jadi tidak perlu banyak keluar uang aku kemarin harus ngekos jadi tidak bisa nabung mengapa aku menyuruh kuliah memang seberguna itu kah sangat berguna	‘kuliah’, ‘coba’, ‘daftar’, ‘kipk’, ‘kemarin’, ‘aku’, ‘gratis’, ‘bahkan’, ‘dapat’, ‘uang’, ‘semesteran’, ‘dari’, ‘kipk’, ‘ambil’, ‘perguruan’, ‘tinggi’, ‘negeri’, ‘yang’, ‘dekat’, ‘domisili’, ‘jadi’, ‘tidak’, ‘perlu’, ‘banyak’, ‘keluar’, ‘uang’, ‘aku’, ‘kemarin’, ‘harus’, ‘ngekos’, ‘jadi’, ‘tidak’, ‘bisa’, ‘nabung’, ‘mengapa’, ‘aku’, ‘menyuruh’, ‘kuliah’, ‘memang’, ‘seberguna’, ‘itu’, ‘kah’, ‘sangat’, ‘berguna’
33.	aku keterima kipk tapi masih bingung soal pencairan dana	‘aku’, ‘keterima’, ‘kipk’, ‘tapi’, ‘masih’, ‘bingung’, ‘soal’, ‘pencairan’, ‘dana’
122.	katanya kipk banyak yang tidak tepat sasaran masih banyak yang mampu tetap dapat	‘katanya’, ‘kipk’, ‘banyak’, ‘yang’, ‘tidak’, ‘tepat’, ‘sasaran’, ‘masih’, ‘banyak’, ‘yang’, ‘mampu’, ‘tetap’, ‘dapat’

Tahap yang kelima adalah *filtering (stopwords removal)*. *Filtering (stopwords removal*) merupakan tahap menghapus kata-kata yang sering muncul tetapi tidak memiliki makna penting dalam analisis sentimen. Kata yang sering muncul tetapi tidak memberikan pengaruh terhadap proses klasifikasi dihapus, seperti ”dan”, ”di”, ”ke”, ”yang”, dan ”atau” (Tala, 2003). Tabel 5 merupakan contoh dari proses *filtering*.

Tabel 5. *Filtering*

No	*Tweet*	Hasil *Filtering*
5.	‘kuliah’, ‘coba’, ‘daftar’, ‘kipk’, ‘kemarin’, **‘aku’**, ‘gratis’, **‘bahkan’**,	‘kuliah’, ‘coba’, ‘daftar’, ‘kipk’, ‘kemarin’, ‘gratis’, ‘dapat’, ‘uang’,

No	*Tweet*	Hasil *Filtering*
	‘dapat’, ‘uang’, ‘semesteran’, **‘dari’**, ‘kipk’, ‘ambil’, ‘perguruan’, ‘tinggi’, ‘negeri’, **‘yang’**, ‘dekat’, ‘domisili’, **‘jadi’**, **‘tidak’**, ‘perlu’, ‘banyak’, ‘keluar’, ‘uang’, **‘aku’**, ‘kemarin’, **‘harus’**, ‘ngekos’, **‘jadi’**, ‘tidak’, ‘bisa’, ‘nabung’, **‘mengapa’**, **‘aku’**, ‘menyuruh’, ‘kuliah’, **‘memang’**, ‘seberguna’, ‘itu’, ‘kah’, ‘sangat’, ‘berguna’	‘semesteran’, ‘kipk’, ‘ambil’, ‘perguruan’, ‘tinggi’, ‘negeri’, ‘dekat’, ‘domisili’, ‘perlu’, ‘banyak’, ‘keluar’, ‘uang’, ‘kemarin’, ‘ngekos’, ‘tidak’, ‘bisa’, ‘nabung’, ‘menyuruh’, ‘kuliah’, ‘seberguna’, ‘sangat’, ‘berguna’
33.	**‘aku’**, ‘keterima’, ‘kipk’, **‘tapi’**, **‘masih’**, ‘bingung’, **‘soal’**, ‘pencairan’, ‘dana’	‘keterima’, ‘kipk’, ‘bingung’, ‘pencairan’, ‘dana’
122.	**‘katanya’**, ‘kipk’, **‘banyak’**, **‘yang’**, ‘tidak’, ‘tepat’, ‘sasaran’, **‘masih’**, **‘banyak’**, **‘yang’**, ‘mampu’, ‘tetap’, ‘dapat’	‘kipk’, ‘tidak’, ‘tepat’, ‘sasaran’, ‘mampu’, ‘dapat’

Tahap yang keenam adalah *Stemming*. Stemming merupakan proses mengubah kata berimbuhan menjadi kata dasar dengan menghapus imbuhan pada awal maupun akhir kata. Tabel 6 merupakan contoh dari proses *stemming*.

Tabel 6. *Stemming*

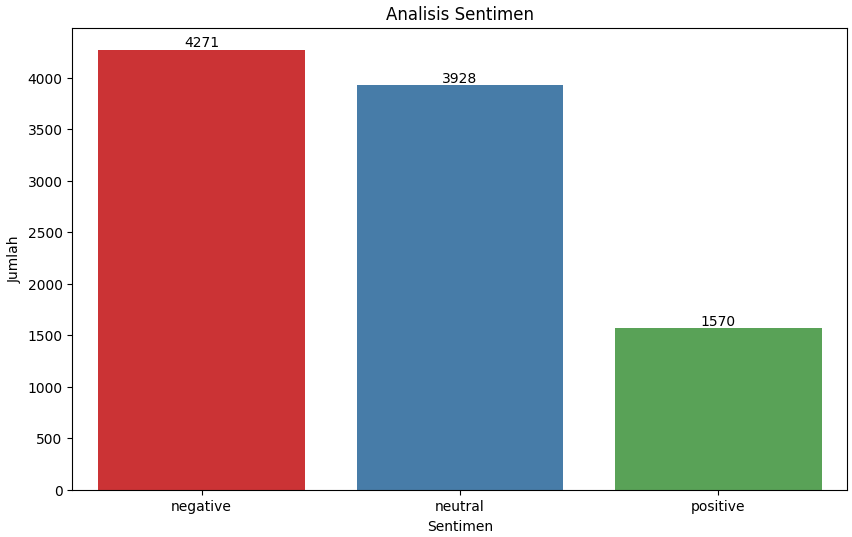
No	*Tweet*	Hasil *Stemming*
5.	‘kuliah’, ‘coba’, ‘daftar’, ‘kipk’, ‘kemarin’, ‘gratis’, ‘dapat’, ‘uang’, **‘semesteran’**, ‘kipk’, ‘ambil’, ‘perguruan’, ‘tinggi’, ‘negeri’, ‘dekat’, ‘domisili’, ‘perlu’, ‘banyak’, ‘keluar’, ‘uang’, ‘kemarin’, **‘ngekos’**, ‘tidak’, ‘bisa’, ‘nabung’, ‘menyuruh’, ‘kuliah’, **‘seberguna’**, ‘sangat’, **‘berguna’**	‘kuliah’, ‘coba’, ‘daftar’, ‘kipk’, ‘kemarin’, ‘gratis’, ‘dapat’, ‘uang’, ‘semester’, ‘kipk’, ‘ambil’, ‘perguruan’, ‘tinggi’, ‘negeri’, ‘dekat’, ‘domisili’, ‘perlu’, ‘banyak’, ‘keluar’, ‘uang’, ‘kemarin’, ‘kos’, ‘tidak’, ‘bisa’, ‘nabung’, ‘suruh, ‘kuliah’, ‘guna, ‘sangat’, ‘guna
33.	**‘keterima’**, ‘kipk’, ‘bingung’, **‘pencairan’**, ‘dana’	‘terima’, ‘kipk’, ‘bingung’, ‘cair, ‘dana’
122.	‘kipk’, ‘tidak’, ‘tepat’, ‘sasaran’, ‘mampu’, ‘dapat’	‘kipk’, ‘tidak’, ‘tepat’, ‘sasaran’, ‘mampu’, ‘dapat’

2. Pelabelan data

Data hasil *preprocessing* selanjutnya diberikan label sentimen berupa positif, negatif, dan netral. Proses pelabelan dilakukan secara manual berdasarkan opini yang terkandung pada

setiap *tweet* terkait penerima Kartu Indonesia Pintar Kuliah (KIP-K). Sentimen positif diberikan pada *tweet* yang mengandung dukungan atau apresiasi terhadap program KIP-K, sentimen negatif diberikan pada *tweet* yang mengandung kritik atau keluhan, sedangkan sentimen netral diberikan pada *tweet* yang mengandung kritik atau keluhan, sedangkan sentimen netral diberikan pada *tweet* yang bersifat informatif dan tidak menunjukkan kecenderungan opini tertentu. Hasil pelabelan data pada penelitian ini ditunjukkan pada Gambar 2.

Gambar 2. Visualisasi Distribusi Sentimen Hasil Pelabelan



3. Pembobotan TF-IDF

Data hasil *preprocessing* selanjutnya melalui tahap pembobotan kata menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Metode TF-IDF digunakan untuk mengubah data teks menjadi representasi numerik berdasarkan tingkat kemunculan suatu kata pada dokumen (*term frequency*) dan tingkat kepentingan kata pada seluruh dokumen (*inverse document frequency*). Metode TF-IDF digunakan karena mampu memberikan bobot yang lebih besar kata-kata penting yang sering muncul dalam suatu dokumen, tetapi jarang muncul pada dokumen lainnya. Dengan demikian, kata yang memiliki pengaruh besar terhadap penentuan sentimen akan memiliki nilai bobot yang lebih tinggi. Perhitungan TF-IDF dilakukan menggunalan persamaan berikut.

$$(TF\ IDF)(t,d,D) = TF(t,d) \times IDF(t,D) \tag{8}$$

dengan nilai *Inverse Document Frequency* (IDF) dihitung menggunakan persamaan berikut.

$$IDF(t,D) = \log \frac{N}{df_t} \tag{9}$$

Keterangan:
IDF (*t, D*) = Jumlah kata (*t*) yang tersebar dari kumpulan dokumen (*D*)
df_t = Jumlah dokumen (*d*) yang mengandung kata (*t*)
N = Jumlah seluruh dokumen (*d*)

Hasil pembobotan TF-IDF selanjutnya digunakan sebagai *input* dalam proses klasifikasi sentimen menggunakan metode *Naive Bayes* dan *Support Vector Machine* (SVM).

4. Pembagian Data Latih dan Data Uji

Data latih (*training data*) merupakan data yang digunakan untuk membangun model klasifikasi sehingga dapat digunakan dalam proses prediksi data baru. Sementara itu, data uji (*testing data*) digunakan untuk mengevaluasi performa model klasifikasi yang telah dibangun menggunakan data latih (Mitchell, 1997). Oleh karena itu, data latih dan data uji memiliki keterkaitan dalam proses klasifikasi sentimen, di mana model yang dibentuk dari data latih akan diuji menggunakan data uji untuk memperoleh hasil prediksi.

Pada penelitian ini digunakan dua skenario pembagian data, yaitu 80%:20% dan 70%:30%. Penggunaan dua skenario pembagian data bertujuan untuk mengetahui pengaruh proporsi data terhadap performa model klasifikasi sentimen. Rincian pembagian data latih dan data uji ditunjukkan pada Tabel 13 dan Tabel 14.

Tabel 7. Data Latih 80% dan Data Uji 20%

Jenis Data	Persentase	Jumlah
Data Latih	80%	7915
Data Uji	20%	1954
Total *Tweet*	100%	9769

Berdasarkan Tabel 7, diketahui bahwa pada skenario pembagian data 80%:20%, sebanyak 7915 *tweet* digunakan sebagai data latih dan 1954 *tweet* digunakan sebagai data uji.

Tabel 8. Data Latih 70% dan Data Uji 30%

Jenis Data	Persentase	Jumlah
Data Latih	70%	6838
Data Uji	30%	2931
Total *Tweet*	100%	9769

Berdasarkan Tabel 8, diketahui bahwa pada skenario pembagian data 70%:30%, sebanyak 6838 *tweet* digunakan sebagai data latih dan 2931 *tweet* digunakan sebagai data uji. Pembagian data tersebut selanjutnya digunakan dalam proses klasifikasi sentimen menggunakan metode *Naive Bayes* dan *Support Vector Machine* (SVM).

5. Klasifikasi menggunakan metode *Naive Bayes*

Proses klasifikasi sentimen menggunakan metode *Naive Bayes* dilakukan menggunakan dua skenario pembagian data, yaitu 80%:20% dan 70%:30%. Pengujian dilakukan terhadap data *tweet* yang telah melalui tahap *preprocessing*, pelabelan sentimen, pembobotan TF-IDF, serta pembagian data latih dan data uji. Evaluasi metode *Naive Bayes* dilakukan meenggunakan *confusion matrix* untuk mengetahui performma model dalam mengklasifikasikan sentimen data *tweet* terkait penerima KIP-K.

Hasil *confusion matrix* pada pembagian data latih 80% dan data uji 20% ditunjukkan pada Tabel 9.

Tabel 9. *Confusion Matrix* Data Latih 80% dan Data Uji 20%

Aktual	Prediksi			Total
	Negatif	Netral	Positif	
Negatif	729	87	38	854
Netral	221	532	33	786
Positif	156	46	112	314
Total	1106	665	183	1954

Berdasarkan Tabel 9, diperoleh nilai evaluasi model yang terdiri atas *accuracy*, *precision*, *recall*, dan *f-1 score*. Hasil evaluasi metode *Naïve Bayes* pada pembagian data 80%:20% ditunjukkan pada Tabel 10.

Tabel 10. Hasil Evaluasi *Naive Bayes* Data Latih 80% dan Data Uji 20%

Evaluasi	Nilai
Accuracy	70,3%
Precision	70,2%
Recall	70,2%
F-1 Score	70,2%

Berdasarkan Tabel 10, metode *Naive Bayes* menghasilkan nilai akurasi sebesar 70,3%. Nilai tersebut menunjukkan bahwa sebanyak 70,3% data *tweet* berhasil diklasifikasikan dengan benar oleh model. Selain itu, diperoleh nilai *precision*, *recall,* dan *f-1 score* sebesar 70,2%, yang menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengidentifikasi dan mengklasifikasikan sentimen data *tweet*.

Hasil *confusion matrix* pada pembagian data latih 70% dan data uji 30% ditunjukkan pada Tabel 11.

Tabel 11. *Confusion Matrix* Data Latih 70% dan Data Uji 30%

Aktual	Prediksi			Total
	Negatif	Netral	Positif	
Negatif	1085	158	38	1281
Netral	336	792	51	1179
Positif	240	75	156	471
Total	1661	1025	245	2931

Berdasarkan Tabel 11, diperoleh nilai evaluasi model yang terdiri atas *accuracy*, *precision*, *recall*, dan *f-1 score*. Hasil evaluasi metode *Naïve Bayes* pada pembagian data 70%:30% ditunjukkan pada Tabel 12.

Tabel 12. Hasil Evaluasi *Naive Bayes* Data Latih 80% dan Data Uji 20%

Evaluasi	Nilai
Accuracy	69,4%
Precision	69,4%
Recall	69,4%
F-1 Score	69,4%

Berdasarkan Tabel 12, metode *Naive Bayes* menghasilkan nilai akurasi sebesar 69,4%.Nilai tersebut menunjukkan bahwa sebanyak 69,4% data *tweet* berhasil diklasifikasikan dengan benar oleh model. Selain itu, diperoleh nilai *precision*, *recall,* dan *f-1 score* sebesar 69,4%, yang menunjukkan bahwa performa model pada pembagian data 70%:30% sedikit lebih rendah dibandingkan pembagian data 80%:20%.

6. Klasifikasi menggunakan *Support Vector Machine* (SVM)

Proses klasifikasi sentimen menggunakan metode SVM dilakukan menggunakan dua skenario pembagian data, yaitu 80%:20% dan 70%:30%. Pengujian dilakukan terhadap data *tweet* yang telah melalui tahap *preprocessing*, pelabelan sentimen, pembobotan TF-IDF, serta pembagian data latih dan data uji. Evaluasi metode SVM dilakukan meenggunakan *confusion matrix* untuk mengetahui performma model dalam mengklasifikasikan sentimen data *tweet* terkait penerima KIP-K.

Hasil *confusion matrix* pada pembagian data latih 80% dan data uji 20% ditunjukkan pada Tabel 13.

Tabel 13. *Confusion Matrix* SVM Data Latih 80% dan Data Uji 20%

Aktual	Prediksi			Total
	Negatif	Netral	Positif	
Negatif	671	140	43	854
Netral	137	618	31	786
Positif	110	55	149	314
Total	918	813	223	1954

Berdasarkan Tabel 13, diperoleh nilai evaluasi model yang terdiri atas *accuracy*, *precision*, *recall*, dan *f-1 score*. Hasil evaluasi metode SVM pada pembagian data 80%:20% ditunjukkan pada Tabel 14.

Tabel 14. Hasil Evaluasi SVM Data Latih 80% dan Data Uji 20%

Evaluasi	Nilai
Accuracy	73,6%
Precision	73,6%
Recall	73,6%
F-1 Score	73,6%

Berdasarkan Tabel 14, metode SVM menghasilkan nilai akurasi sebesar 73,6%. Nilai tersebut menunjukkan bahwa sebanyak 73,6% data *tweet* berhasil diklasifikasikan dengan benar oleh model. Selain itu, diperoleh nilai *precision*, *recall,* dan *f-1 score* sebesar 73,6%, yang menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengidentifikasi dan mengklasifikasikan sentimen data *tweet*.

Hasil *confusion matrix* pada pembagian data latih 70% dan data uji 30% ditunjukkan pada Tabel 15.

Tabel 15. *Confusion Matrix* SVM Data Latih 70% dan Data Uji 30%

Aktual	Prediksi			Total
	Negatif	Netral	Positif	
Negatif	986	239	56	1281
Netral	208	917	54	1179
Positif	176	79	216	471
Total	1370	1235	326	2931

Berdasarkan Tabel 15, diperoleh nilai evaluasi model yang terdiri atas *accuracy*, *precision*, *recall*, dan *f-1 score*. Hasil evaluasi metode SVM pada pembagian data 70%:30% ditunjukkan pada Tabel 16.

Tabel 16. Hasil Evaluasi SVM Data Latih 70% dan Data Uji 30%

Evaluasi	Nilai
Accuracy	72,3%
Precision	72,3%
Recall	72,3%
F-1 Score	72,3%

Berdasarkan Tabel 16, metode SVM menghasilkan nilai akurasi sebesar 72,3%.Nilai tersebut menunjukkan bahwa sebanyak 72,3% data *tweet* berhasil diklasifikasikan dengan benar oleh model. Selain itu, diperoleh nilai *precision*, *recall,* dan *f-1 score* sebesar 72,3%, yang menunjukkan bahwa performa model pada pembagian data 70%:30% sedikit lebih rendah dibandingkan pembagian data 80%:20%.

Pembahasan

Hasil penelitian menunjukkan bahwa opini masyarakat terhadap penerima KIP-Kuliah di media sosial X (Twitter) memiliki kecenderungan sentimen popsitif, negatif, dan netral. Sentimen Sentimen positif umumnya menunjukkan dukungan masyarakat terhadap program KIP-Kuliah dalam membantu mahasiswa dari keluarga kurang mampu untuk melanjutkan pendidikan tinggi. Sementara itu, sentimen negatif lebih banyak berkaitan dengan isu ketidaktepatan sasaran penerima bantuan, kecemburuan sosial, dan stigma terhadap penerima KIP-Kuliah. Adapun sentimen netral cenderung berupa informasi atau pernyataan yang tidak menunjukkan kecenderungan opini tertentu.

Pada penelitian ini digunakan metode Naïve Bayes dan Support Vector Machine (SVM) dengan pembobotan TF-IDF untuk mengklasifikasikan sentimen masyarakat terhadap penerima KIP-Kuliah. Pengujian dilakukan menggunakan dua skenario pembagian data, yaitu 80%:20% dan 70%:30%. Hasil pengujian menunjukkan bahwa metode SVM menghasilkan performa yang lebih baik dibandingkan Naïve Bayes pada kedua skenario pembagian data.

Tabel 17. Perbandingan Nilai Akurasi Model

Pembagian Data Latih dan Data Uji	**Akurasi Model**	
	Naïve Bayes	**SVM**
80%:20%	70,3%	73,6%
70%:30%	69,4%	72,3%

Berdasarkan Tabel 17, nilai akurasi tertinggi diperoleh metode SVM pada pembagian data 80%:20%, yaitu sebesar 73,6%, sedangkan metode *Naive Bayes* memperoleh akurasi tertinggi sebesar 70,3%. Hasil tersebut menunjukkan bahwa metode SVM lebih baik dalam mengklasifikasikan sentimen data *tweet* dibandingkan metode Naïve Bayes. Selain itu, penggunaan data latih yang lebih besar pada pembagian data 80%:20% menghasilkan performa model yang lebih baik dibandingkan pembagian data 70%:30%.

Hasil penelitian ini sejalan dengan penelitian Putu et al. (2024) yang menunjukkan bahwa metode SVM memiliki performa lebih baik dibandingkan Naïve Bayes dalam analisis sentimen program KIP-Kuliah. Penelitian tersebut memperoleh tingkat akurasi sebesar 92% pada metode SVM dan 79% pada metode Naïve Bayes. Selain itu, penelitian Jannan dan Negara (2023) juga menunjukkan bahwa metode SVM efektif digunakan dalam klasifikasi sentimen positif, negatif, dan netral pada data media sosial Twitter menggunakan pembobotan TF-IDF.

Perbedaan nilai akurasi pada setiap penelitian dapat dipengaruhi oleh jumlah data, distribusi kelas sentimen, tahapan *preprocessing*, metode pembobotan kata, serta pembagian data *training* dan *testing*. Selain itu, karakteristik bahasa pada media sosial yang tidak terstruktur, seperti penggunaan singkatan, bahasa tidak baku, dan campuran bahasa sehari-hari, juga dapat memengaruhi performa model klasifikasi.

Secara keseluruhan, hasil penelitian menunjukkan bahwa metode Support Vector Machine (SVM) lebih efektif dibandingkan Naïve Bayes dalam mengklasifikasikan sentimen masyarakat terhadap penerima KIP-Kuliah pada media sosial X (Twitter).

SIMPULAN

Hasil analisis sentimen menunjukkan bahwa opini masyarakat terhadap penerima KIP-Kuliah di media sosial X (Twitter) didominasi oleh sentimen negatif sebesar 43,7%, diikuti sentimen netral 40,2% dan positif 16,1%. Berdasarkan hasil klasifikasi, metode Support Vector Machine (SVM) memiliki performa lebih baik dibandingkan Naïve Bayes, dengan akurasi tertinggi sebesar 73,6% pada pembagian data 80%:20%, sedangkan Naïve Bayes memperoleh akurasi 70,3%. Pada pembagian data 70%:30%, SVM memperoleh akurasi 72,3% dan Naïve Bayes 69,4%. Dengan demikian, metode SVM dengan pembagian data 80%:20% menjadi model terbaik dalam mengklasifikasikan sentimen masyarakat terhadap penerima KIP-Kuliah.

UCAPAN TERIMA KASIH

Terima kasih kepada koordinator Prodi Matematika dan seluruh Dosen Prodi Matematika yang telah memberikan ilmu dan bimbingan hingga terselesainya artikel ini.

DAFTAR PUSTAKA

Amelia, I., Sugiyono, S., Sarimole, F. M., & Tundo, T. (2024). Analisis sentimen tanggapan pengguna media sosial X terhadap program beasiswa KIP-Kuliah dengan menggunakan algoritma Support Vector Machine (SVM). *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, 5(3), 2994–3003. https://doi.org/10.35870/jimik.v5i3.990

Arsi, P., & Waluyo, R. (2021). Analisis sentimen wacana pemindahan ibu kota Indonesia menggunakan algoritma Support Vector Machine (SVM). *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(1), 147–156. https://doi.org/10.25126/jtiik.202183944

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 121–167.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan algoritma Naive Bayes untuk analisis sentimen review data Twitter BMKG nasional. *Jurnal Tekno Kompak*, 15(1), 131–145.

Hemalatha, I., & Govardhan, A. (2012). Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science*, 1(2).

Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi berita online dengan menggunakan pembobotan TF-IDF dan cosine similarity. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(1), 306–312.

Jannan, M. F. Z., Dwi, Y., & Negara, P. (2024). Analisis sentimen masyarakat terhadap penerima beasiswa Kartu Indonesia Pintar Kuliah dengan metode Support Vector Machine. *Jurnal Informatika Terapan dan Sistem Informasi*, 4(2), 26–30. https://doi.org/10.32938/jitu.v4i2.7598

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and Its Applications*, 13(3). https://doi.org/10.15849/IJASCA.211128.11

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael, CA: Morgan & Claypool Publishers.

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies*, 9(4), 81.

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Ningsih, W., Alfianda, B., & Wulandari, D. (2024). Comparison of Naive Bayes and SVM algorithms in Twitter sentiment analysis on electric car use in Indonesia. *Jurnal RESTI*, 8(2), 556–562.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tala, F. Z. (2003). *A study of stemming effects on information retrieval in Bahasa Indonesia* (Master's thesis). University of Amsterdam.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Wankhade, M., Chandra, A., Rao, S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55, 5731–5780.

Zhang, H. (2004). The optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, 562–567.