CENTIVE
VOL. 3, NO. 1, PP.001-010, DEC 2024

ISBN: 978-602-53004-0-0

Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2024

Analisis Sentimen Pada Media Sosial *X* Terhadap KIP Kuliah Dengan Algoritma *Random Forest Classifier*

Hanindiya Putri Almeyda[1] , Aldi Yoga Setiawan[2] , Nabilah Husen Alkaff[3], Muhamad Awiet Wiedanto Prasetyo[4]

[1,2,3,4]Sistem Informasi, Telkom University Purwokerto
Jl.D.I Panjaitan No.128, Kec. Purwokerto Selatan, Kabupaten Banyumas, Jawa Tengah 53147

[1] hanindiyaputri@student.telkomuniversity.ac.id
[2] aldiiyg@student.telkomuniversity.ac.id
[3] nabilahhusenalkaff@student.telkomuniversity.ac.id
[4] awietmwp@telkomuniversity.ac.id

Diterima pada dd-mm-yyyy, direvisi pada dd-mm-yyyy, diterima pada dd-mm-yyyy

Abstrak

Teknologi digital yang semakin berkembang telah menjadikan media sosial sebagai platform utama dalam menyampaikan opini publik, termasuk terkait isu bantuan Kartu Indonesia Pintar Kuliah (KIP Kuliah). Penelitian ini berfokus dalam menganalisis opini masyarakat terhadap KIP Kuliah di media sosial X dengan menerapkan algoritma *Random Forest Classifier*. Data dikumpulkan melalui crawling otomatis dengan kata kunci "KIPK" dan menghasilkan 2012 data. Setelah melalui proses preprocessing seperti *cleaning, tokenization,* dan *stemming,* data dilabeli menggunakan metode *Inset Lexicon* menjadi sentimen positif, negatif, dan netral. Teknik SMOTE digunakan untuk menyeimbangkan data, diikuti oleh ekstraksi fitur menggunakan TF-IDF. Model dikembangkan dengan algoritma *Random Forest Classifier* serta evaluasinya melalui *Confusion Matrix*. Pengujian yang dilakukan menunjukkan hasil akurasi sebesar 82%, f1-score 81%, serta precision dan recall masing-masing 81% dan 82%. Mayoritas sentimen terhadap KIP Kuliah adalah positif, menunjukkan persepsi publik yang baik terhadap program ini. Penelitian ini menggarisbawahi potensi analisis sentimen berbasis machine learning dalam memahami opini publik serta relevansinya untuk pengambilan keputusan di sektor pendidikan.

Kata Kunci: Analisis Sentimen, KIP Kuliah, *Random Forest Classifier*, Media Sosial *X*

Penulis Koresponden:
Aldi Yoga Setiawan
Program Studi S1 Sistem Informasi, Fakultas Rekayasa Industri, Telkom University Purwokerto
Jl. D.I Panjaitan No.128, Kec. Purwokerto Selatan, Kabupaten Banyumas, Jawa Tengah, Indonesia
Email: aldiiyg@student.telkomuniversity.ac.id

I. PENDAHULUAN

Teknologi informasi di Indonesia berkembang dengan sangat pesat ke arah yang serba digital. Akibatnya, masyarakat di seluruh dunia, termasuk masyarakat Indonesia sangat bergantung pada penggunaan teknologi digital dengan menggunakan berbagai platform media sosial misalnya *X, Instagram, Facebook, YouTube, TikTok,* dan masih banyak lagi. Media sosial *Twitter* atau *X* merujuk pada media sosial yang banyak diikuti oleh masyarakat di Indonesia dan termasuk ke dalam situs *microblogging* yang memungkinkan masyarakat sebagai penggunanya menulis serta membagikan pendapatnya terhadap berbagai topik atau isu yang ada [1]. Pesan atau opini masyarakat yang disampaikan dalam media sosial ini biasa disebut tweet atau kicauan, namun saat ini sudah diubah menjadi post atau unggahan [2]. Terdapat fitur trending di media sosial X yang akan menampilkan kata kunci masalah atau topik yang paling populer.

Dalam beberapa waktu terakhir, Kartu Indonesia Pintar Kuliah (KIP), juga dikenal sebagai KIP Kuliah, adalah salah satu isu yang paling populer di media sosial X.

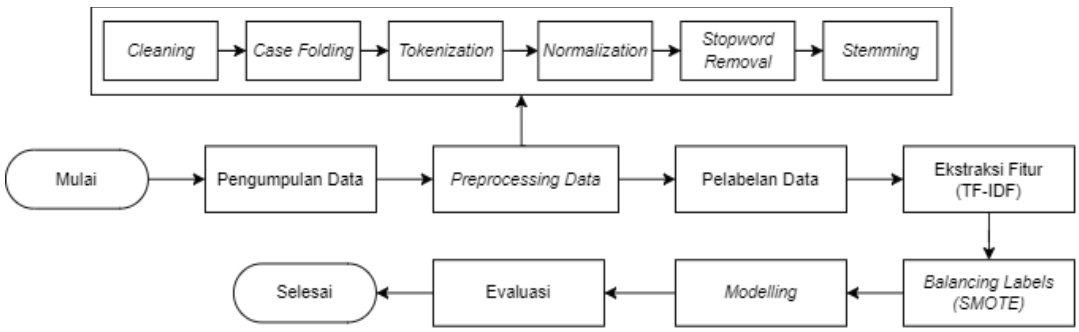
Kartu Indonesia Pintar Kuliah (KIP Kuliah) adalah program bantuan pemerintah yang telah terlaksana dari tahun 2020 sampai dengan saat ini [3]. KIP Kuliah adalah program beasiswa bantuan pemerintah yang ditujukan bagi mahasiswa tidak mampu serta berprestasi dengan tujuan program ini dapat memberikan kesempatan pendidikan untuk seluruh mahasiswa di Indonesia. Berdasarkan pencarian di media sosial X, Peneliti mengidentifikasi beberapa isu yang sering disorot oleh masyarakat yang diantaranya adalah terkait KIP Kuliah. KIP Kuliah menjadi isu terhangat dan kontroversial baru-baru ini.

Analisis sentimen adalah teknik analisis yang digunakan untuk menentukan apakah pendapat atau perasaan yang diungkapkan dalam teks termasuk dalam kategori positif, negatif, atau netral [4]. Analisis sentimen dibutuhkan saat menyaring opini atau komentar di masyarakat karena analisis sentimen dapat membantu memahami bagaimana opini publik berkembang terhadap suatu kasus yang sedang terjadi. Dari penjelasan tersebut, dapat disimpulkan bahwa analisis sentimen adalah proses menentukan pendapat atau perasaan seseorang melalui teks. Sentimen dapat dikategorikan menjadi sentimen positif, negatif, atau netral.

Dalam era teknologi saat ini, analisis sentimen dapat dilakukan dengan menggunakan bantuan dari kecerdasan buatan ataupun yang lebih dikenal dengan sebutan *Artificial Intelligent (AI)*. Salah satu cabang dari *AI* adalah *Machine Learning*. *Machine Learning* atau mesin pembelajar ialah kecerdasan buatan yang berfokus pada belajar dari data, yaitu mengembangkan sistem yang dapat belajar secara "mandiri" tanpa diprogram ulang oleh penggunanya atau manusia [5]. Penelitian ini menggunakan algoritma *Random Forest Classifier*. Dalam penelitian ini, Algoritma *Random Forest* dipilih karena memiliki banyak keunggulan dalam analisis sentimen, terutama dalam konteks data tidak terstruktur seperti teks media sosial. Sebuah studi oleh [6] menggunakan algoritma ini untuk menganalisis sentimen pengguna *TikTok*, menunjukkan efektivitasnya dalam menangani dataset yang beragam dan penuh *noise*. Hasil penelitian tersebut menegaskan kemampuan *Random Forest* dalam menghasilkan klasifikasi yang akurat bahkan pada data dengan polaritas sentimen yang bervariasi dengan nilai akurasi 80%. Selain itu, penelitian [7] membuktikan keandalan *Random Forest* dalam klasifikasi ulasan aplikasi digital seperti PeduliLindungi. Dengan penerapan teknik balancing data untuk mengatasi ketidakseimbangan, algoritma ini berhasil memberikan hasil yang konsisten dan akurat dengan akurasi 72%. Keunggulan tersebut menjadikan *Random Forest* pilihan yang sangat cocok untuk tugas analisis sentimen berbasis teks. Algoritma tersebut termasuk dalam kategori *supervised learning* yang artinya *machine learning* dilatih untuk dapat mengenali data yang telah melalui proses pelabelan sehingga data tersebut memiliki label khusus.

II. METODE PENELITIAN

Penelitian ini dilakukan melalui berbagai tahapan penelitian. Beberapa langkah metode penelitian ini tercantum melalui Gambar 1.



Gambar 1. Diagram Alir Penelitian

A. Pengumpulan Data

Tahap pengumpulan data dalam penelitian ini menggunakan teknik pengambilan data otomatis (*crawling*) dari unggahan di media sosial X. Proses *crawling* dilakukan secara otomatis berdasarkan kata kunci yang diberikan oleh pengguna untuk mengumpulkan data [8]. Kata kunci yang digunakan adalah "KIPK" dengan rentang waktu pengumpulan dari 29 April 2024 hingga 1 Mei 2024. Alasan pengambilan data pada rentang waktu tersebut adalah KIP-Kuliah menduduki posisi trending pada aplikasi X. Pemilihan kata kunci "KIPK" dan bukan "KIP Kuliah" karena pada trending X menunjukkan kata "KIPK" dan bukan "KIP Kuliah". Untuk proses ini, digunakan platform *Google Colab* yang menjalankan bahasa pemrograman Python dan memanfaatkan library tweet harvest. Dataset yang diperoleh terdiri dari 2012 baris data, yang kemudian disimpan dalam file berformat CSV (*comma separated value*) sehingga dapat diakses melalui perangkat lunak seperti Microsoft Excel.

B. *Preprocessing Data*

Tahap berikutnya setelah pengumpulan data adalah *preprocessing*, yaitu langkah awal pengolahan data untuk mendapatkan data dalam bentuk lebih bersih, terstruktur, serta siap dilakukan analisis. *Preprocessing* mencakup beberapa tahap: *cleaning*, *tokenization*, *normalization*, *stopword removal*, dan *stemming* [4]. Tahap ini penting karena data yang terkumpul belum sepenuhnya siap untuk analisis sentimen, sering kali masih terdapat noise yang mengganggu. Proses *preprocessing* membantu meningkatkan kualitas data, mengurangi noise, mengoptimalkan efisiensi, mengurangi bias, dan meningkatkan akurasi hasil analisis. Beberapa penjelasan pada tahapan *preprocessing* yaitu sebagai berikut:

1) *Cleaning*
   Cleaning merupakan proses guna membersihkan data terhadap kata dan simbol tanda baca yang tidak diperlukan pada proses klasifikasi [9]. Karakter yang akan dihapus pada tahap ini yaitu seperti URL, username atau nama pengguna, tagar atau hashtag serta tanda baca lainnya.
2) *Case Folding*
   Case folding merujuk pada proses mengganti semua huruf kapital (*upper case*) menjadi huruf kecil (*lower case*) [10].
3) *Tokenization*
   Tokenization merupakan proses penguraian teks menggunakan *token* atau potongan yang lebih kecil seperti kata [7]. Contoh tahapan *tokenization*: 'daftar program KIPK itu gampang' menjadi 'daftar', 'program', 'KIPK', 'itu', 'gampang'.
4) *Normalization*
   Normalization merujuk pada tahapan memperbaiki kata yang tidak baku serta singkatan bentuk tertentu yang dilakukan menggunakan kamus [9].
5) *Stopword Removal*
   Stopwords removal merupakan proses menghapus kata yang berfungsi sebagai penghubung kata pada kalimat. Contoh kata umum yaitu 'yang', 'dan', 'di', 'ke' yang ada dalam daftar stopwords [6].
6) *Stemming*
   Stemming adalah tahapan untuk mengembalikan kata yang mempunyai imbuhan kembali ke bentuk dasar. Tahapan *stemming* akan mengeliminasi kata-kata imbuhan atau akhiran pada teks untuk memperoleh kata dasarnya [11]. *Stemming* akan digunakan dengan bantuan *library* Sastrawi.

C. Pelabelan Data

Penelitian ini menggunakan metode kamus *Inset Lexicon* untuk analisis sentimen, di mana model dibangun dengan pendekatan berbasis lexicon. Metode ini memanfaatkan kata-kata

yang diberi bobot berdasarkan skor polaritas untuk menentukan opini publik terhadap suatu topik atau isu [12]. Data dikelompokkan ke dalam tiga kategori label yaitu positif, negatif, dan netral.

D. Ekstraksi Fitur (TF-IDF)

Penelitian ini menerapkan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dalam mengekstraksi fitur. Metode TF-IDF berfungsi memberikan bobot pada tiap kata dalam dokumen untuk mengukur signifikansi kata tersebut [13] . Tujuan langkah ini adalah untuk meningkatkan akurasi klasifikasi dengan menghitung frekuensi kata tertentu dan membandingkannya dengan proporsi kemunculannya di seluruh dokumen. Perhitungan bobot pada setiap dokumen dapat dilakukan menggunakan persamaan pada rumus dibawah ini [14].

$$w_{i,j} = tf_{i,j}\ log\ log\ \left(\frac{N}{df_i}\right) \quad (1)$$

Keterangan :

$\boldsymbol{W_{i,j}}$: bobot kata i pada dokumen j

$\boldsymbol{tf_{i,j}}$: banyaknya kemunculan (frekuensi) kata i pada dokumen j

$\boldsymbol{N}$: jumlah seluruh dokumen yang ada

$\boldsymbol{df_i}$: jumlah dokumen yang mendapati kata i

E. Balancing Labels

Penelitian ini menggunakan teknik penyeimbangan label atau *balancing labels* yang bertujuan untuk menyeimbangkan label atau data. Balancing labels dalam penelitian ini dilakukan untuk mengatasi ketidakseimbangan jumlah data antar label setelah proses pelabelan. Contoh sintetis label minoritas dibuat dengan teknik SMOTE, yang merupakan singkatan dari teknik *over-sampling minoritas sintetis*. Ini meningkatkan jumlah data yang berkaitan dengan label minoritas. Algoritme SMOTE menggunakan teknik oversampling untuk menyeimbangkan kembali set pelatihan awal untuk memperbaiki ketidakseimbangan distribusi kelas atau label dalam data [10]. Metode utama SMOTE adalah menghasilkan contoh sintetis daripada hanya meniru contoh dari kelas minoritas. Karena ketidakseimbangan label yang diamati, SMOTE digunakan dalam penelitian ini karena membutuhkan metode untuk mengimbangi label [11]. Hal ini dilakukan untuk memastikan distribusi data menjadi lebih seimbang. Untuk memastikan bahwa model yang dibangun dapat melakukan klasifikasi dengan lebih akurat dan tidak bias terhadap label tertentu, langkah ini sangat penting. Selain itu, penelitian oleh [10] membuktikan keandalan metode SMOTE dalam proses analisis sentimen terhadap ulasan tentang wisata Baturaden dimana metode SMOTE digunakan untuk menyeimbangkan data yang semula memiliki distribusi tidak merata antara sentimen positif, netral, dan negatif. Setelah diterapkan, distribusi kelas menjadi seimbang, yang memungkinkan model klasifikasi seperti *Random Forest*, SVM, KNN dan *Naive Bayes* untuk memberikan akurasi yang lebih baik dan hasil prediksi yang lebih stabil.

F. Modelling

Pada tahap *modelling*, penelitian ini menerapkan algoritma *Random Forest Classifier* untuk menganalisis sentimen dari data yang telah diproses sebelumnya. *Random Forest Classifier* adalah teknik pembelajaran mesin dalam kategori pembelajaran yang diawasi, di mana data label digunakan untuk mempelajari model. Proses ini dimulai setelah tahap *preprocessing*, pelabelan, dan ekspos fitur dengan menggunakan teknik *Frekuensi Terma-Inverse Dokumen Frekuensi* (TF-IDF). *Random Forest Classifier* merupakan algoritma gabungan dari setiap teknik *decision tree* yang ada, kemudian digabungkan dan dikombinasikan menjadi sebuah model [15]. Metode

klasifikasi *Random Forest Classifier* dapat diibaratkan sebagai kumpulan model klasifikasi berbasis pohon keputusan [6].

Decision Tree dimulai dengan menghitung nilai *entropy*. Nilai *entropy* berfungsi sebagai referensi untuk menentukan tingkat normatif pada node atribut dan nilai gain informasi. Selain itu, ia juga berfungsi sebagai penentu tingkat informatif pada node atribut dan nilai gain informasi. Untuk menghitung nilai *entropy*, gunakan rumus berikut [7] :

$$Entropy\ (Y) = -\sum ip(c|Y)log2p(c|Y) \quad (2)$$

Keterangan:

Y= Himpunan Kasus

$p(c|Y)$ = Proporsi nilai Y terhadap kelas c

G. Evaluasi Menggunakan *Confusion Matrix*

Penelitian ini menggunakan *Confusion Matrix* sebagai parameter pada tahapan evaluasi model. *Confusion Matrix* merujuk pada metode yang dipergunakan dalam proses mengukur hasil sebuah model dengan pengukuran tingkat *accuracy*, *precission*, *recall* dan *f1-score*. Rumus dari setiap pengukuran dapat diamati pada formula dibawah ini [16].

$$Accuracy\ = \frac{True\ positive + True\ negative}{\sum\ \ data} \times 100\% \quad (3)$$

$$Precission\ = \frac{True\ positive}{True\ positive + False\ positive} \quad (4)$$

$$Recall\ = \frac{True\ positive}{True\ positive + False\ negative} \quad (5)$$

$$F1 - score\ = 2 \times \frac{Precission \times Recall}{Precission + Recall} \quad (6)$$

Tabel 1 berikut menunjukkan bentuk tabel *Confusion Matrix*.

Tabel 1. Confusion Matrix

Confusion Matrix		**Nilai Sebenarnya**	
		positive (0)	***negative (1)***
Prediksi	***positive (0)***	*TP*	*FP*
	negative (1)	*FN*	*TN*

Keterangan :

TP (*True Positive*): nilai sebenarnya positif dan diperkirakan benar sebagaimana positif

TN (*True Negative*): nilai sebenarnya negatif dan diperkirakan benar sebagaimana negatif

FP (*False Positive*): nilai sebenarnya negatif namun salah diperkirakan sebagaimana positif

FN (*False Negative*): nilai sebenarnya positif namun salah diperkirakan sebagaimana negatif

III. HASIL DAN PEMBAHASAN

Hasil penelitian ini menggunakan metode *Random Forest Classifier* untuk mengevaluasi persepsi data *tweet* di media sosial *X* tentang KIP Kuliah. Untuk membuat hasil analisis lebih akurat, beberapa tahap *preprocessing* dilakukan terlebih dahulu pada data yang digunakan. Bagian dari analisis ini adalah membagi data ke dalam kategori sentimen positif, negatif, dan netral.

A. Pengumpulan Data

Pada tahap pengumpulan data penelitian ini, kata kunci "KIPK" digunakan untuk melakukan *crawling* data di media sosial *X* dari 29 April 2024 hingga 1 Mei 2024. *Crawling* data dilakukan di *Google Colab* menggunakan bahasa pemrograman *Python* dan *library tweet harvest*. Dataset dengan 2012 baris total dikumpulkan dan kemudian disimpan dalam file *csv* dengan nilai yang dipisahkan dengan koma. Tabel 2 berikut menunjukkan contoh dataset contoh.

Tabel 2. Contoh Sample Dataset

No	Tweet
1.	@messierous @undipmenfess ya bener dong? kalau hanya miskin tapi gak ada potensi akademik yang baik apakah cukup buat menerima kipk?
2.	@riansazyn kaa benerr aku ada jg org yg di kampus penerima kipk kaya org2 di atas semmenjaak ku tau jd sebel bgt tiap liat dia pedahal temenku yg niat ga mampu pinter ga dpeet bahkan sampe cabut kuliah:))
	.....
2011.	@diarydidu emg seterbuka itu kah dokumennya penerima kipk?
2012.	Bayangin lo udah miskin apply kipk terus cari kos yang murah (biasanya agak jauh) eh ada yang nyinyir anak kipk ga boleh bawa motor oasu

B. *Pre-processing*

Pada *preprocessing* dilakukan beberapa tahapan seperti *cleaning*, *case folding*, *tokenization*, normalisasi kata, *stopword removal*, dan *stemming*. Tahapan-tahapan ini bertujuan mengoptimalkan data agar siap untuk dianalisis. *Preprocessing* dibantu menggunakan *library sastrawi* karena sumber data berbahasa Indonesia. Dalam *stopword* yang dilakukan untuk menghapus kata-kata penghubung dalam kalimat menggunakan *file stopword* dalam format *txt* yang berisi daftar kata penghubung berbahasa Indonesia. Kemudian *stemming* melakukan pengubahan terhadap kata yang berimbuhan menjadi bentuk dasarnya menggunakan *library sastrawi*. Contoh hasil dari *preprocessing* data dapat dilihat pada tabel berikut.

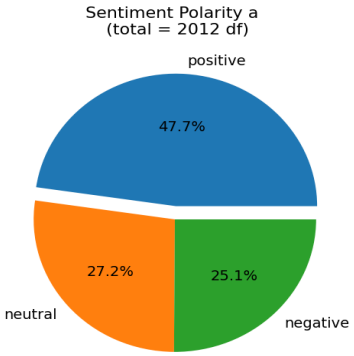
Tabel 3. Tabel Hasil *Preprocessing*

Preprocessing	**Hasil**
Data Tweets	@undipmenfess Kalo mahasiswanya emang berprestasi banyak kok beasiswa lain selain KIPK yang persyaratannya ga harus MISKIN
Cleaning	Kalo mahasiswanya emang berprestasi banyak kok beasiswa lain selain KIPK yang persyaratannya ga harus MISKIN
Case Folding	kalo mahasiswanya emang berprestasi banyak kok beasiswa lain selain kipk yang persyaratannya ga harus miskin
Tokenization	['kalo', 'mahasiswanya', 'emang', 'berprestasi', 'banyak', 'kok', 'beasiswa', 'lain', 'selain', 'kipk', 'yang', 'persyaratannya', 'ga', 'harus', 'miskin']

Normalisasi	['kalau', 'mahasiswanya', 'emang', 'berprestasi', 'banyak', 'kok', 'beasiswa', 'lain', 'selain', 'kipk', 'yang', 'persyaratannya', 'ga', 'harus', 'miskin']
Stopword Removal	['mahasiswanya', 'berprestasi', 'beasiswa', 'kipk', 'persyaratannya', 'ga', 'miskin']
Stemming	mahasiswa prestasi beasiswa kipk syarat ga miskin

C. Pelabelan

Proses Pelabelan menggunakan data dari *Senticnet* untuk dapat memproses skor polaritas sentimen positif, negatif, dan netral. Setiap sentimen memiliki skor persentase masing-masing. Skor sentimen positif lebih atau sama dengan 1, sentimen negatif kurang dari atau sama dengan 0, dan ketika skornya diluar dari skor positif atau negatif maka skor tersebut dinilai netral. Berikut hasil dan detail pelabelan.



Gambar 2. Diagram Hasil Pelabelan

Berdasarkan Gambar 2 di atas, dari total 2012 data terlihat bahwa polaritas 47.7% sentimen masuk ke dalam label positif, 25.1% sentimen memiliki label negatif dan 27.2% sentimen memiliki label netral. Hasil pelabelan ini termasuk ke dalam hasil data tidak seimbang atau *imbalance data*. Pada langkah berikutnya, data yang tidak seimbang ini akan diseimbang dengan metode penyeimbangan data, juga dikenal sebagai teknik *balancing* data.

Tabel 4. Detail Hasil Pelabelan

Tweets setelah stemming	***Polarity_score***	**Label**
daftar kipk sadar kampus swasta masuk kampus anak kaya iya tertawa	3.954	positif
kampus juta semester gunain uang kipk nya kuliah iya beli tiket konser deh		
guna kipk nonton konser harga jt beli tiket	-1.751	negatif
gatau juang gadapet kipk juang mati an cepat lulus biar ga bebanin ortunya		
ukt	0.291	netral

Hasil pelabelan sentimen yang didapat yaitu sejumlah 960 sentimen positif, 505 sentimen negatif, 547 sentimen netral dari jumlah 2012 dataset. Pada Tabel 3 di atas adalah contoh dari data *tweets* setelah proses *stemming*, *polarity score* untuk setiap *tweets* serta label untuk setiap *tweets* yang tertera. Hasil *polarity score* menunjukkan beberapa sample pelabelan sentimen positif, negatif, dan netral. Untuk memvisualkan kata-kata dari semua kategori (positif, negatif, dan netral), dapat menggunakan visualisasi melaui *Word Cloud*.

D. *TF-IDF*

Tahap *TF-IDF* mengubah data teks atau kalimat yang telah diproses pada tahap *stemming* ke dalam bentuk matriks numerik. Dengan memberi bobot pada kata-kata tertentu berdasarkan frekuensi kemunculannya dalam dataset secara keseluruhan, matriks ini membantu memilih kata-kata yang paling relevan dan penting.

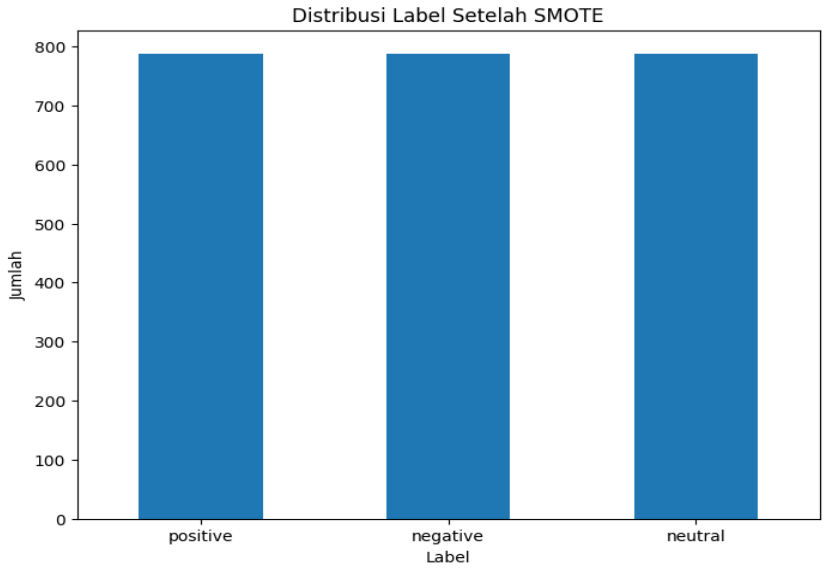
	aaaa	aaaaaa	aamiin	aamiinn	abai	abak	abang	abiezz	abis	about	...	yundip	yuran	yutub	yyaa	yyyy	zaa	zalim	zaman	zero	zuzurly
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2007	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2008	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2009	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2010	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2011	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 3. Hasil *TF-IDF*

Pada Gambar diatas, data yang telah diproses oleh *TF-IDF* sebanyak 2012 kata unik. Setiap kata unik diproses menjadi bentuk numerik ke dalam matriks.

E. *Balancing* Label

Balancing Label memiliki tujuan untuk menangani masalah ketimpangan pada setelah dilakukan pelabelan. Berdasarkan hasil pelabelan terjadi ketimpangan antara positif, negatif, dan netral. Dalam menjalankan *balancing* pada label, diperlukan teknik *SMOTE* yang dapat meningkatkan tingkat dataset dari label minoritas sehingga membuat pelabelan menjadi lebih seimbang. Label minoritas dengan menggunakan *SMOTE* kemudian memilih label terdekat menghasilkan meningkatnya data label minoritas melalui data-data baru. Pada proses ini, data dilakukan *split* menjadi dua kategori: data uji dan data latih. Data latih dibagi 80% kemudian data uji dibagi 20%. Setelah dilakukan proses *SMOTE*, dataset menjadi seimbang di setiap labelnya masing masing memiliki dataset dengan jumlah 788.



Gambar 4. Hasil *Balancing* Label menggunakan *SMOTE*

F. *Modelling*

Pada *Modelling* dilakukan inisiasi penggunaan model *Random Forest Classifier* pada data yang telah dilakukan proses *preprocessing*, pelabelan, *TF-IDF*, dan *balancing*. Hasil dari tahap ini diantaranya *cross-validation* sebesar (0.85, 0.82, 0.80, 0.80, 0.83), kemudian skor rata-rata sebesar 0.82, dan standar deviasi sebesar 0.018. *Cross validation* digunakan untuk menilai performa algoritma

Random Forest dengan membagi data menjadi beberapa lipatan (*folds*). *K-folds* pada proses *modelling* menggunakan 5 *k-folds*.

```
Skor cross-validation: [0.84566596 0.82029598 0.8012685  0.79915433 0.83474576]
Skor rata-rata: 0.8202261081449098
Standar deviasi: 0.018228669854317592
```

Gambar 5. Skor Hasil *Modelling*

Gambar diatas menunjukkan hasil pengujian *cross-validation* dimana nilai dari akurasi 5 *k-folds* model *random forest* yaitu 0.85, 0.82, 0.80, 0.80, dan 0.83. Kemudian nilai rata-rata dari keseluruhan *cross-validation* yaitu sebesar 0.82 dan nilai standar deviasi antar *k-folds* sebesar 0.018.

G. Evaluasi

Pada evaluasi menggunakan *confusion matrix* untuk menghitung berapa banyak prediksi benar dan salah tentang sentimen positif, negatif, dan netral yang dapat dilihat dari akurasi serta prediksi nilai. Hasil dari tahap evaluasi ini didapat akurasi sebesar 0.82, *F1-Score* sebesar 0.81, *Precision* sebesar 0.81, dan *Recall* sebesar 0.82.

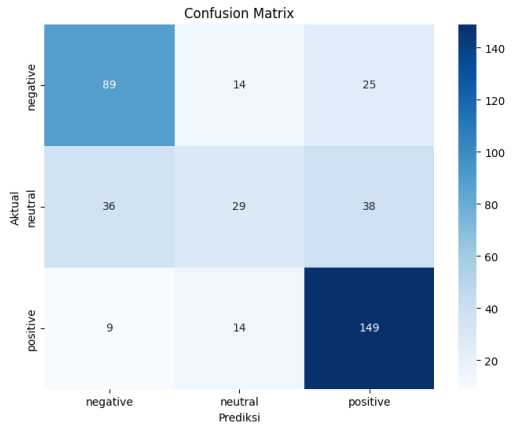
```
Skor Akurasi : [0.8329809725158562, 0.813953488372093, 0.7949260042283298, 0.8076109936575053, 0.8305084745762712]
Rata-rata Akurasi : 0.8159959866700109

Skor F1-score : [0.8302165875934667, 0.8120467363163281, 0.793450937874204, 0.8040341639960698, 0.8268354648429349]
Rata-rata F1-score : 0.8133167781246007

Skor Precision: [0.8391080897752647, 0.8128834028844939, 0.7964750751268977, 0.8061155543813501, 0.8344276187829951]
Rata-rata Precision: 0.8178019481902004

Skor Recall: [0.8329809725158562, 0.813953488372093, 0.7949260042283298, 0.8076109936575053, 0.8305084745762712]
Rata-rata Recall: 0.8159959866700109
```

Gambar 6. Skor Akurasi, *F1-Score*, *Precision* dan *Recall*



Gambar 7. *Confusion Matrix*

Hasil *Confusion Matrix* ini menunjukkan 149 label positif dapat diprediksi dengan benar, 29 label netral diprediksi secara benar, dan 89 label negatif diprediksi dengan benar.

H. Visualisasi *Word Cloud*

Visualisasi data *Word Cloud* menampilkan kata-kata populer atau sering muncul pada *tweet*. *Word cloud* menampilkan data dalam bentuk kumpulan kata dengan ukuran yang mencerminkan frekuensi atau pentingnya kata agar lebih mudah untuk menemukan topik atau tema yang populer.

Gambar 8. *Word Cloud* pada Sentimen Positif

Pada Gambar 8 di atas, "Kipk, ga, iya, kuliah, banget, kaya, beasiswa" adalah beberapa kata yang paling banyak dibahas dalam sentimen positif visualisasi data *Word Cloud*.

Gambar 9. *Word Cloud* pada Sentimen Negatif

Kata-kata "Kipk, ga, spill, miskin, banget, kuliah, teman" adalah beberapa kata yang paling banyak dibicarakan dalam sentimen negatif visualisasi data *Word Cloud*, seperti yang terlihat pada Gambar 9 di atas.

Gambar 10. *Word Cloud* Sentimen Netral

"Kipk, banget, uang, undip, kaya, kuliah" merupakan kumpulan kata yang paling banyak dibicarakan dalam sentimen netral visualisasi data *Word Cloud*, seperti yang terlihat pada Gambar 10 di atas.

IV. KESIMPULAN

Hasil studi menunjukkan analisis sentimen terhadap KIP Kuliah di media sosial *X* menggunakan algoritma *Random Forest Classifier* berhasil dilakukan dengan baik. Proses pengumpulan data dan *preprocessing* yang tepat berkontribusi pada akurasi model yang mencapai 82% serta nilai *f1-score* sebesar 81%, *precision* sebesar 81%, dan *recall* sebesar 82%. Pelabelan sentimen menunjukkan bahwa mayoritas opini publik cenderung positif terhadap program KIP Kuliah. Penelitian ini menegaskan pentingnya penggunaan teknologi analisis data dalam memahami dinamika opini masyarakat di era digital, serta memberikan dasar bagi penelitian lebih dalam di bidang analisis sentimen dan implementasi kecerdasan buatan di sektor pendidikan.

UCAPAN TERIMAKASIH

Penulis menyampaikan rasa terima kasih kepada semua yang telah mendukung studi ini. Penulis turut mengucapkan terima kasih kepada Telkom University atas bantuan dan fasilitas yang diberikan selama proses penelitian. Ucapan terima kasih juga ditujukan kepada Bapak Muhamad Awiet Wiedanto Prasetyo, S.Kom., M.MSI atas arahan dan masukan yang sangat berharga dalam penyusunan penelitian ini. Penulis memiliki harapan untuk penelitian ini agar dapat membantu kemajuan ilmu pengetahuan, terutama bidang pendidikan dan analisis data.

DAFTAR PUSTAKA

[1] M. Imam Syafii, “Sentimen analisis Pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier (NBC),” 2022.

[2] M. Fikri, T. Haq, A. Hendratno, M. N. Arfa, A. F. Amanullah, and E. Sholihatin, “Analysis Of Changes In The Foreign Language Lexicon Used By The Millennial Generation In The X Application In Indonesia’ 1,” vol. 2, no. 1, 2024, [Online]. Available: https://humasjournal.my.id/index.php/HJ/index

[3] I. Arfyanti, M. Fahmi, and P. Adytia, “Penerapan Algoritma Decision Tree Untuk Penentuan Pola Penerima Beasiswa KIP Kuliah,” *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2275.

[4] Y. Julianto, D. H. Setiabudi, and S. Rostianingsih, “Analisis Sentimen Ulasan Restoran Menggunakan Metode Support Vector Machine,” 2022.

[5] A. A. Soebroto, “Buku Ajar AI, Machine Learning & Deep Learning,” 2019. [Online]. Available: https://www.researchgate.net/publication/348003841

[6] W. N. Mufidati Nur Edma, E. N. Andini, and I. Widodo, “Analisis Sentimen Pada Pengguna Tiktok Menggunakan Metode Random Forest (Studi Kasus: Jessica-Mirna),” *Journal Of Social Science Research*, vol. 4, pp. 14477–14489, 2022.

[7] M. Reza, U. Pulungan, D. E. Ratnawati, and B. Rahayudi, “Analisis Sentimen Ulasan Aplikasi PeduliLindungi dengan Metode Random Forest,” 2022. [Online]. Available: http://j-ptiik.ub.ac.id

[8] K. Zuhri, N. Adha, and O. Saputri, “Analisis Sentimen Masyarakat Terhadap Pilpres 2019 Berdasarkan Opini Dari Twitter Menggunakan Metode Naive Bayes Classifier,” 2020. [Online]. Available: https://journal-computing.org/index.php/journal-cisa/index

[9] Muttaqin *et al.*, *Pengenalan Data Mining*, 2023rd ed. Medan: Yayasan Kita Menulis, 2023.

[10] M. Afrad, C. Febrianto, S. Wijayanto, and Y. Fathoni, “Edu Komputika Journal Sentiment Analysis of Visitor Reviews on Baturaden Tourist Attraction Using Machine Learning Methods,” *Edu Komputika*, vol. 11, no. 1, 2024, [Online]. Available: https://journal.unnes.ac.id/journals/edukom/

[11] C. Magnolia, A. Nurhopipah, D. Bagus, and A. Kusuma, “Edu Komputika Journal Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter,” 2022. [Online]. Available: http://journal.unnes.ac.id/sju/index.php/edukom

[12] M. D. Al Fahreza, Ardytha Luthfiarta, Muhammad Rafid, Michael Indrawan, and Adhitya Nugraha, “Analisis Sentimen: Pengaruh Jam Kerja Terhadap Kesehatan Mental Generasi Z,” *Journal Of Applied Computer Science And Technology (JACOST)*, vol. 5, Feb. 2024.

[13] A. Wandani, “Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN, Random Forest, dan Naive Bayes,” 2021.

[14] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, “Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN,” *Jurnal KomtekInfo*, pp. 1–7, Jan. 2023, doi: 10.35134/komtekinfo.v10i1.330.

[15] R. Wahyudi *et al.*, “Analisis Sentimen pada review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine,” *JURNAL INFORMATIKA*, vol. 8, no. 2, 2021, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ji

[16] M. K. Yunita Rani, “Perbandingan Algoritma SVM Dan Naïve Bayes Pada Analisis Sentimen Kebijakan Penghapusan Kewajiban Skripsi,” *Indonesian Journal of Computer Science*, no. Penelitian ini membandingkan hasil evaluasi algoritma Support Vector Machine (SVM) dengan Naïve Bayes menggunakan 80% data latih dan 20% data uji., Oct. 2023.