Perbandingan Performa Algoritma *Naïve Bayes, SVM* dan *Random Forest*: Studi Kasus Analisis Sentimen Pengguna Sosial Media *X*

Putri Cahyani[1)], **Lufty Abdillah**[2)]

[1,2)] Sistem Infornasi, Fakultas Ilmu Komputer dan Desain, Universitas Kalbis
Jalan Pulomas Selatan Kav. 22, Jakarta 13210
Email: putri.cahyani112@gmail.com
Email: lufty.abdillah@kalbis.ac.id

Abstract: *Sentiment analysis was explored to understand social media users' opinions towards the Indonesian Capital City (IKN) through the X platform with machine learning and lexicon-based algorithms. This research uses three algorithms: Naïve Bayes, Support Vector Machine (SVM), and Random Forest. The aim of this research is to test and compare the performance of the three algorithms to determine the best in classifying sentiment data from the X platform. The data consists of 10,000 tweets collected using the crawling method with the Python Harvest Library and Node.js, using keywords related to IKN. Based on the algorithm performance test, it was concluded that SVM had the highest performance compared to Naïve Bayes and Random Forest, producing an accuracy of 87%, precision 87%, recall 87%, and f-1 score 87%. This research uses the CRISP-DM Data Mining framework to ensure a structured and systematic approach to the analysis process.*
Keywords: *sentiment analysis, X, naïve bayes, support vector machine, random forest, IKN.*

Abstrak: *Analisis sentimen dieksplorasi untuk memahami opini pengguna media sosial terhadap Ibu Kota Nusantara (IKN) melalui platform X dengan algoritma machine learning dan berbasis lexicon-based. Penelitian ini menggunakan tiga algoritma: Naïve Bayes, Support Vector Machine (SVM), dan Random Forest. Tujuan penelitian ini adalah menguji dan membandingkan performa ketiga algoritma tersebut untuk menentukan yang terbaik dalam pengklasifikasian data sentimen dari platform X. Data terdiri dari 10.000 tweet yang dikumpulkan menggunakan metode crawling dengan Library Python Harvest dan Node.js, menggunakan kata kunci terkait IKN. Berdasarkan uji performa algoritma, disimpulkan bahwa SVM memiliki performa tertinggi dibandingkan Naïve Bayes dan Random Forest, dengan menghasilkan accuracy 87%, precision 87%, recall 87%, dan f-1 score 87%. Penelitian ini menggunakan framework Data Mining CRISP-DM untuk memastikan pendekatan terstruktur dan sistematis dalam proses analisis.*
Kata kunci: *analisis sentimen, X, naïve bayes, support vector Machine, random forest, IKN.*

I. PENDAHULUAN

Opini merupakan suatu istilah konsep yang luas, mencangkup aspek-aspek seperti sentimen, evaluasi, penilaian atau sikap serta informasi mengenai suatu pendapat individu atau kelompok. Sentimen dapat dianalisis dengan memahami, menganalisis dan mengekstraksi suatu pendapat tekstual, sehingga dapat mengetahui opini yang diungkapkan tersirat dalam sentimen positif, negatif atau netral [1].

Sentimen analisis adalah suatu bagian dari penelitian dengan sistem komputasi untuk menganalisis pendapat, sentimen dan emosi yang diungkapkan secara tekstual. Sentimen analisis juga dikenal sebagai *opinion mining* ataupun cabang *text mining*, yang berfokus pada penambangan data dari sebuah komentar dalam teks, ekstraksi atribut, serta melakukan analisis dari data yang telah dikumpulkan menggunakan teknik *Machine Learning* [2]. Sentimen dapat dilihat melalui berbagai jenis *platform* media sosial yang faktanya saat ini media sosial dapat digunakan oleh semua kalangan melalui perangkat *smartphone* yang dilengkapi beragam fitur canggih dan komponen pendukung seperti koneksi internet, salah satunya yaitu *platform X* [3].

X merupakan jejaring sosial dengan memberi kebebasan bagi penggunanya untuk dapat melakukan interaksi secara terbuka kepada lembaga maupun masyarakat melalui pesan singkat yang dikenal dengan kicauan atau *tweet.* Sentimen yang dihasilkan melalui *tweet* akan melewati proses analisis dengan tujuan untuk melakukan suatu pengklasifikasian teks [4].

Text mining merujuk pada proses eksplorasi dan analisis data teks dalam

berskala besar dan juga sebagai alat untuk memanfaatkan *unstructured data* dengan melakukan analisis untuk memperoleh suatu informasi baru lalu mengidentifikasi suatu pola atau model signifikan dan hubungan yang tersembunyi pada data [5]. *Text mining* dapat diterapkan untuk menganalisis media sosial, mengingat bahwa media sosial dapat menghasilkan sejumlah data yang tidak terstruktur, terutama bentuk data teks [6]. *X* merupakan salah satu sumber data untuk penambangan teks tidak hanya dari nama akun saja melainkan melalui kata kunci, hashtag serta tanggal status

Dari banyaknya opini atau sentimen di *X* salah satunya yaitu konteks mengenai Ibu Kota Nusantara (IKN) yang memicu beragam tanggapan mendukung dan menentang. Dari pemilihan letak hingga pengesahan Undang-Undang Nomor 3 Tahun 2022 yang mengacu pada Ibu Kota Nusantara (IKN). Perancangan Undang Undang IKN diketahui juga minim dalam melibatkan partisipasi warga Indonesia hal tersebut dinilai Undang-Undang Ibu Kota Nusantara (IKN) minim akan keterlibatan masyarakat Indonesia sehingga tidak sejalan dengan prinsip demokrasi yang menjadi dasar ideologi negara [7].

Dalam analisis sentimen terdapat beberapa metode algoritma pengklasifikasian, diantaranya yaitu *Naïve Bayes, SVM* dan *Random Forest.*
Naïve Bayes bekerja sebagai penambangan teks digunakan dalam analisis sentimen dengan probabilitas sederhana yang bergantung pada Teorema Bayes dengan memiliki asumsi yang tinggi atas ketidaketergantungan [8]. Penelitian sebelumnya yang dilakukan oleh Ade Tiara S, Nur Anjeni L, dan Puput Alpria dengan topik analisis sentiment public pada *twitter* terhadap boikot produk Israel menggunakan metode *Naïve Bayes* dengan 303 *tweets* menghasilkan akurasi dengan nilai *accuracy* 95%, precision 96%, *recall* 95% dan *F1-Score* 95% [9].

Support Vector Machine (SVM) bekerja dengan mencari *hyperplane* yang optimal dengan memisahkan jarak antara masing-masing titik yang disebut kelas [10]. Penelitian sebelumnya yang dilakukan oleh Primandani Arsi dan Retno Waluyo dengan topik analisis sentimen wacana pemindahan Ibu Kota Indonesia menggunakan algoritma *Support Vector Machine* (SVM) yang menghasilkan nilai *accuracy* 96,68%, *precision* 95.82%, *recall* 94.04% dan AUC = 0,979 dengan 1.236 *tweets* (404 positif dan 832 negatif) [11].

Random Forest, sesuai dengan namanya merupakan serangkaian pohon (tree) yang menyatu membentuk hutan (*forest*) atau metode algoritma berdasarkan pohon keputusan dan menerapkan *bagging* [12]. Penelitian sebelumnya yang dilakukan oleh Thifal Fadiyah B, Dian Eka R, dan Issa Arwani dengan topik analisis sentimen pengguna *twitter* terhadap pembayaran *cashless* menggunakan *Shopeepay* dengan algoritma *Random Forest* menghasilkan nilai *accuracy* 95%*, recall* 94%*, Precision* 95% dan *F1-Score* 95% [13]. Penelitian lainnya tahun 2024 adalah analisis sentimen penutupan tiktokshop di Indonesia dengan algoritma naïve bayes. Penelitian ini menggunakan komninasi naïve bayes, TF-IDF dan blobtext untuk pemrosesan teks. Akurasi yang dihasilkan adalah 86.60% [14]. Lalu penelitian analisis sentimen pemilihan presiden berdasarkan opini dari beberapa media social. Metode yang digunakan adalah LSTM (*Long Short Term Memory Network*) dan beberapa tahap pemrosesan teks seperti tokenisasi, lematisasi, dan lainnya 15].

Penelitian ini bertujuan untuk mengklasifikasikan sentimen dalam konteks Ibu Kota Nusantara (IKN) menggunakan algoritma *Naïve Bayes, SVM* dan *Random Forest* untuk tiga kelas sentimen: positif, negatif, dan netral. Penelitian juga dilakukan seleksi untuk mencari algoritma terbaik dengan cara membandingkan tingkat performa algoritma menggunakan *dataset X* berdasarkan metrik *accuracy, recall, precision* dan *f1-Score* dengan dataset *X*. Pendekatan *text mining* yang digunakan mencakup *machine learning* dan *lexicon based* untuk analisis sentimen. Teknik *k-fold cross validation* dan SMOTE digunakan untuk meningkatkan performa model, dengan evaluasi menggunakan *Confusion Matrix*. Penelitian ini mengadopsi *framework* CRISP-DM untuk pendekatan analisis data yang terstruktur dan sistematis.

II. METODE PENELITIAN

Pada bagian ini menjelaskan mengenai alur maupun metode yang digunakan dalam penelitian.

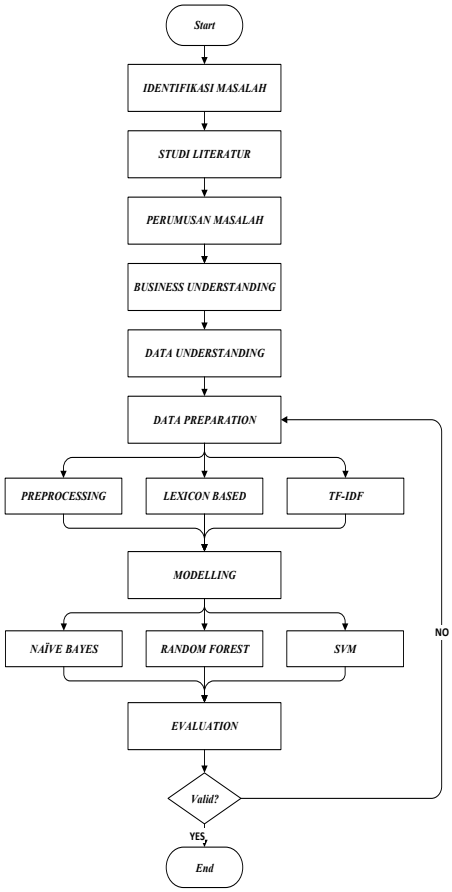
A. Metode Pengumpulan data

Pengumpulan data dari media sosial *X* melalui proses *crawling data* yang hasilnya disimpan dalam format CSV menggunakan python yang dijalankan dengan *Library Python* yakni Harvest dan Node.js melalui *code editor google collaboration.* Jenis data yang dikumpulkan termasuk data primer yang didapatkan secara langsung pada sumber media sosial *X* yang berisi opini pengguna yang diambil dari 2 Februari 2024 hingga 26 April 2024 sejumlah kurang lebih 10.000 data setelah IKN disahkan.

B. Metode *Framework Data Mining*

Penelitian ini menggunakan Framework CRISP-DM *(Cross-Industry Standard Process for Data Mining)* untuk melakukan analisis sentimen terhadap pengguna media sosial *X*. Tahapan proses CRISP-DM yang akan dilakukan, meliputi *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment. Framework* ini dibatasi hingga tahap *evaluation* untuk memastikan kesesuaian dan efektivitas metode yang digunakan dalam penelitian. Tahapan evaluasi dalam CRISP-DM mencakup pengujian model, pengukuran performa menggunakan metrik seperti *accuracy, precision, recall,* dan *f1-score*, serta penilaian terhadap kelayakan hasil analisis sentimen yang telah dilakukan.

C. Alur Penelitian



Gambar 1 Alur Kerja Penelitian

Dari gambar 1 diatas dapat diuraikan penjelasan mengenai alur kerja penelitian sebagai berikut:

1. Identifikasi Masalah

Pada tahap ini, peneliti mengidentifikasi sentimen dan pandangan masyarakat terhadap proyek Ibu Kota Nusantara (IKN) melalui analisis *tweet* dari pengguna media sosial *X* dan mencari algoritma pengklasifikasian yang terbaik untuk dataset *X*, sehingga diperlukan suatu tindakan pengujian atau seleksi.

2. Studi Literatur

Pada tahap ini, Studi literatur yang diperoleh melalui jurnal dan buku terkait metode dan algoritma analisis sentimen untuk melakukan riset mengenai pemahaman suatu masalah untuk menghasilkan sebuah solusi yang tepat.

3. Perumusan Masalah

- Bagaimana melakukan perbandingan klasifikasi performa pada tiap algoritma dari tanggapan pengguna media sosial *X* terhadap Ibu Kota Nusantara (IKN)

menggunakan algoritma *Naïve Bayes, SVM* dan *Random* Forest meliputi a*ccuracy, precision, recall* dan *f-1 score?*

- Agoritma manakah yang memiliki performa terbaik dalam klasifikasi data sentiment pengguna media sosial *X?*

4. Business Understanding

Pada penelitian yang akan diteliti terdapat kebutuhan pengklasifikasian terhadap analisis sentimen pengguna media sosial *X.* Sehingga menggunakan beberapa metode untuk mencari performa algoritma terbaik dari algoritma *Naïve Bayes, SVM* dan *Random Forest* terhadap konteks IKN melalui data media sosial *X.*

5. Data Understanding

Pada tahap ini data yang tersedia pada media sosial *X* terkait Ibu Kota Nusantara (IKN) akan dikumpulkan dengan teknik *crawling data* menggunakan *python* yang dijalankan dengan *Library Python* melalui *code editor google collaboration* dan melakukan seleksi data.

6. Data Preparation

Pada tahap ini data yang dikumpulkan dan telah melewati proses seleksi data akan melalui proses *text preprocessing* dengan tahap *Case Folding, Cleansing, Tokenizing, Normalisasi, Stopword Removal,* dan *Stemming*, lalu dilakukan labeling data dengan *lexicon based* dan menghitung bobot tiap kata menggunakan TF-IDF.

7. Modeling

Pada tahap ini data yang telah melewati proses *text pre-processing, lexicon based* dan TF-IDF akan melewati proses pengujian *data training* dan *data testing* dengan perbandingan akurasi 60:40, 70:30, 80:20, 90:10 dengan split data dan menggunakan *k-fold cross validation* dengan SMOTE dan tanpa SMOTE terhadap dataset menggunakan algoritma *Naïve Bayes, SVM* dan *Random Forest*.

8. Evaluation

Pada tahap ini dataset yang didapatkan pada proses *modeling* akan dilakukan pengujian dengan implementasi *confusion matrix,* jika hasil evaluasi *confusion matrix* valid sesuai dengan data yang digunakan maka penelitian selesai, namun jika sebaliknya maka proses akan kembali ke *data preparation.*

III. HASIL DAN PEMBAHASAN

A. Identifikasi Masalah

Pada tahap ini, fokus utama permasalahan adalah menentukan algoritma pengklasifikasi yang terbaik untuk melakukan analisis sentimen terhadap dataset *X* yang terkait dengan opini dan sentimen publik mengenai Ibu Kota Nusantara (IKN).

B. *Business Understanding*

Pada tahap ini, penelitian memfokuskan pemahaman terhadap data dari media sosial *X* mengenai Ibu Kota Nusantara (IKN), yang akan menjadi Ibu Kota Nusantara. Tujuan utamanya adalah untuk menganalisis beragam pendapat dari masyarakat (baik positif, negatif, atau netral) yang diungkapkan dalam *tweet*. *Business understanding* ini juga bertujuan untuk menentukan metode analisis sentimen yang paling sesuai, dengan membandingkan *Naïve Bayes, SVM*, dan *Random Forest*. Penelitian ini mencari algoritma yang paling terbaik untuk mengklasifikasikan opini serta sentimen terkait proyek IKN dari data media sosial *X*.

C. *Data Understanding*

Pada tahap data understanding merupakan proses memahami data yang akan digunakan dan akan menjadi bahan penelitian yang akan diteliti terkait opini publik yang bersifat positif, negatif atau bahkan pun netral melalui *tweet* mengenai Ibu Kota Nusantara (IKN). Proses pengumpulan data dari awal bulan Februari sampai dengan bulan April melalui proses *crawling data* dengan *Library* Python yakni *Harvest* dan *Node.js* melalui *code editor google collaboration*. Data disimpan dengan format *Comma Separated Values* (CSV) yang berhasil memperoleh 10.000 tweet.

D. *Data Preparation*

Pada tahap ini ini membahas proses persiapan data menggunakan beberapa teknik, yaitu *text preprocessing, lexicon based*, dan TF-IDF. Berikut ini adalah hasil dari setiap teknik yang digunakan:

1. Text Preprocessing

Text Preprocessing merupakan bagian dari suatu proses persiapan data yang akan dilakukan analisis lebih lanjut. *Text Preprocessing* memiliki beberapa tahap yaitu proses mengubah semua kata menjadi

huruf kecil (*case folding),* menghilangkan tanda baca, angka, tautan URL, nama pengguna, dan symbol (*cleaning),* kalimat akan diproses dengan menguraikannya menjadi token-token atau kata-kata yang berdiri sendiri (*tokenizing),* perbaikan kata-kata pada dokumen teks yang mengalami salah dalam pengejaan *(Normalization),* proses menghapus kata-kata yang tidak penting seperti "di", "akan", dan sejenisnya *(stopword removal),* dan perubahan kata yang berimbuhan menjadi kata dasar *(stemming).* Dengan dilakukan tahap tersebut dataset mudah dipahami oleh model analisis dan siap untuk dilakukan analisis lebih lanjut. Tabel 1 dibawah merupakan hasil dari tahap *Text Preprocessing.*

Tabel 1 Tahap *Text Preprocessing*

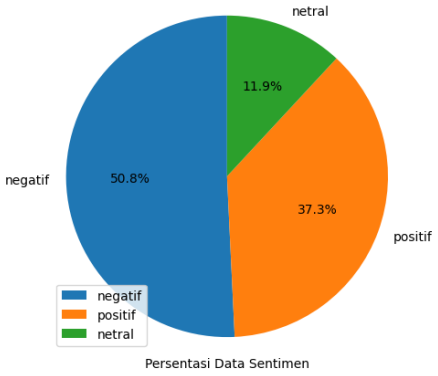
Text Preprocessing	Sebelum	Sesudah
Case Folding	@KompasTV Saran: Buat kota itu mbok yang mateng-teng Jangan Grusak-grusuk Mengejar ego Pribadi Demi Legacy Buat Studinya Minta saran masyarakat dan pakar @ikn_id	@kompastv saran : buat kota itu mbok yang mateng-teng jangan grusak-grusuk mengejar ego pribadi demi legacy buat studinya minta saran masyarakat dan pakar @ikn_id
Cleaning	@kompastv saran : buat kota itu mbok yang mateng-teng jangan grusak-grusuk mengejar ego pribadi demi legacy buat studinya minta saran masyarakat dan pakar @ikn_id	saran buat kota itu mbok yang matengteng jangan grusakgrusuk mengejar ego pribadi demi legacy buat studinya minta saran masyarakat dan pakar
Tokenizing	saran buat kota itu mbok yang matengteng jangan grusakgrusuk mengejar ego pribadi demi legacy buat studinya minta saran masyarakat dan pakar	‘saran’, ‘buat’, ‘kota’, ‘itu’, ‘mbok’, ‘yang’, ‘matengteng’, ‘jangan’, ‘grusakgrusuk’, ‘mengejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘buat’, ‘studinya’, ‘minta’, ‘saran’, ‘masyarakat’, ‘dan’, ‘pakar’.
Normalization	‘saran’, ‘buat’, ‘kota’, ‘itu’, ‘mbok’, ‘yang’, ‘matengteng’, ‘jangan’, ‘grusakgrusuk’, ‘mengejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘buat’, ‘studinya’, ‘minta’, ‘saran’, ‘masyarakat’, ‘dan’, ‘pakar’.	‘saran’, ‘buat’, ‘kota’, ‘itu’, ‘mbok’, ‘yang’, ‘matengteng’, ‘jangan’, ‘grusakgrusuk’, ‘mengejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘buat’, ‘studinya’, ‘minta’, ‘saran’, ‘masyarakat’, ‘dan’, ‘pakar’.
Stopword Removal	‘saran’, ‘buat’, ‘kota’, ‘itu’, ‘mbok’, ‘yang’, ‘matengteng’, ‘jangan’, ‘grusakgrusuk’, ‘mengejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘buat’, ‘studinya’, ‘minta’, ‘saran’, ‘masyarakat’, ‘dan’, ‘pakar’.	‘saran’, ‘kota’, ‘mbok’, ‘matengteng’, ‘grusakgrusuk’, ‘mengejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘studinya’, ‘saran’, ‘masyarakat’, ‘pakar’.
Stemming	‘saran’, ‘kota’, ‘mbok’, ‘matengteng’, ‘grusakgrusuk’, ‘mengejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘studinya’, ‘saran’, ‘masyarakat’, ‘pakar’.	‘saran’, ‘kota’, ‘mbok’, ‘matengteng’, ‘grusakgrusuk’, ‘kejar’, ‘ego’, ‘pribadi’, ‘demi’, ‘legacy’, ‘studi’, ‘saran’, ‘masyarakat’, ‘pakar’.

2. Lexicon Based

Pada tahap ini dilakukan teknik pelabelan data berbasis *lexicon based*, yaitu dengan melakukan labeling dengan 3 kategori, yakni positif, negatif dan netral. Data yang telah melalui proses *text preprocessing* akan dilakukan klasifikasi otomatis dengan kamus *Inset Lexicon* untuk kamus opini. Tabel 2 Dibawah ini merupakan hasil dari tahap *lexicon* pada dataset tweet dan Gambar 2 diagram visualisasi dari presentasi data sentiment.

Tabel 2 Tahap Lexicon Based

Tweet	Polaritas	Sentimen
‘penandatanganan’, ‘nota’, ‘paham’, ‘otorita’, ‘ikn’, ‘kolaborasi’, ‘kembang’, ‘nusantara’, ‘ke depan’, ‘yuk’, ‘ikut’, ‘kembang’, ‘otorita’, ‘ikn’, ‘sobat’, ‘website’, ‘sosial’, ‘media’, ‘ikn’	11	positif
'saran', 'kota', 'mbok', 'matengteng', 'grusakgrusuk', 'kejar', 'ego', 'pribadi', 'legacy', 'studi', 'saran', 'masyarakat', 'pakar'	-6	Negatif
'hukum', 'utama'	0	Netral



Gambar 2 Diagram Presentasi Data Sentimen

3. TF-IDF

Pada tahap ini, *dataset* pengguna media sosial *X* dilakukan *plotting data* untuk menampilkan 10 kata pada data *tweet* yang sering muncul. Tabel 3 Dibawah ini merupakan hasil kata yang sering muncul pada *dataset tweet*

Tabel 3 Hasil Frequently Occurring Words

No.	Frequently Occurring Words
1.	Ikn
2.	Kota
3.	nusantara
4.	bangun
5.	kalimantan
6.	indonesia
7.	negara
8.	timur
9.	presiden
10.	pindah

Dalam langkah ini, metode *Term Frequency* (TF) diperlukan untuk memastikan seberapa sering kata-kata muncul dalam sebuah dokumen. TF berfokus pada hasil jumlah kemunculan atau frekuensi kata dalam file dokumen tersebut. Tabel 4 ini merupakan hasil kemunculan kata pada *dataset tweet*

Tabel 4 Hasil Dari Proses Word Occurrence

No.	*Frequently Occurring Words*	*Word Occurrence*
1.	Ikn	10609
2.	Kota	3786
3.	nusantara	3376
4.	bangun	2559
5.	kalimantan	2442
6.	indonesia	2083
7.	negara	1952
8.	timur	1669
9.	presiden	1493
10.	pindah	1488

Langkah selanjutnya, metode *Inverse Document Frequency* (IDF) dipakai untuk mengukur keumuman atau kelangkaan suatu kata dalam semua dokumen. IDF menilai jumlah suatu kata yang jarang muncul dalam seluruh dokumen dan mempengaruhi bobotnya. Nilai IDF yang tinggi menunjukkan jika kata tersebut jarang muncul pada dokumen lainnya. Tabel 5 merupakan hasil bobot kata pada *dataset tweet*

Tabel 5 Hasil *Word Weighting*

No.	*Frequently Occurring Words*	*Word Weighting*
1.	Ikn	0.067394
2.	Kota	0.041205
3.	nusantara	0.040775
4.	bangun	0.034315
5.	kalimantan	0.028178
6.	indonesia	0.039134
7.	negara	0.024078
8.	timur	0.020820
9.	presiden	0.021043
10.	pindah	0.025925

E. *Modeling*

Kemudian dataset yang telah melalui persiapan data dengan *preprocessing* dan ekstraksi fitur, akan masuk ke tahap *modeling* dengan menggunakan algoritma yang telah ditetapkan sebelumnya yaitu *Naïve Bayes, SVM* dan *Random Forest*. Pada tahap ini, data dibagi menjadi data pelatihan *(training)* dan data pengujian *(testing)* dengan metode *split validation.* Pembagian data dilakukan dalam empat model perbandingan: 60:40, 70:30, 80:20, dan 90:10, untuk mempermudah proses klasifikasi dan pengujian.

Lalu dilakukan pengujian menggunakan *k-fold validation* untuk menguji performa

model dengan variasi pengaturan *5-fold* serta menerapkan teknik SMOTE (*Synthetic Minority Over-sampling Technique)* untuk menangani ketidakseimbangan kelas, dan tanpa SMOTE untuk perbandingan. Hasil dari tahap ini memberikan prediksi dari ketiga model untuk masing-masing perbandingan data, memungkinkan untuk melakukan evaluasi kembali dan memilih model yang paling cocok untuk penyelesaian masalah klasifikasi yang diteliti.

F. *Evaluation*

Dalam penelitian ini, akan dilakukan evaluasi kinerja dari hasil model prediksi yang dihasilkan dari tiga algoritma yaitu *Naïve Bayes, SVM* dan *Random Forest* dengan tujuan untuk mengukur tingkat performa dari masing-masing algoritma

1. *Naïve Bayes*

Tabel 6 dan Tabel 7 merupakan hasil pengujian akurasi pada algoritma *Naïve Bayes* dengan SMOTE dan tanpa SMOTE, ditampilkan dalam bentuk hasil dan rata-rata akurasi dari *k-fold.*

Tabel 6 Akurasi Naïve Bayes Tanpa SMOTE

Split Data	Akurasi K-Fold Validation					Rata-Rata Akurasi K-Fold Validation
	1	2	3	4	5	
(60:40)	68.4%	68.3%	68.7%	69.3%	67.2%	68.3%
(70:30)	70.4%	68.9%	68.1%	69.7%	67.8%	69.9%
(80:20)	70.7%	69.8%	69.1%	69.4%	69.6%	69.7%
(90:10)	69.8%	71.5%	69.6%	70.5%	69.2%	70.1%

Tabel 7 Akurasi Naive Bayes Dengan SMOTE

Split Data	Akurasi K-Fold Validation					Rata-Rata Akurasi K-Fold Validation
	1	2	3	4	5	
(60:40)	71.2%	71.7%	73.5%	73.5%	76.9%	73.3%
(70:30)	70.6%	71.4%	74%	73.9%	79.5%	73.8%
(80:20)	71.2%	71.7%	73.4%	74.4%	77.9%	73.7%
(90:10)	72%	72.5%	73.8%	74.7%	77.9%	73.9%

Dapat disimpulkan penggunaan SMOTE meningkatkan secara signifikan akurasi rata-rata *Naïve Bayes* pada berbagai proporsi pembagian data. Tanpa SMOTE, akurasi tertinggi adalah 70.1% pada pembagian data 90:10, sementara dengan SMOTE, akurasi tertinggi mencapai 73.9% pada proporsi yang sama. Kombinasi optimal untuk mencapai akurasi tertinggi adalah menggunakan SMOTE dengan pembagian data 90:10 untuk model *Naïve Bayes.*

2. *Support Vector Machine*

Tabel 8 dan Tabel 9 merupakan hasil pengujian akurasi pada algoritma *support vector machine (SVM)* dengan SMOTE dan tanpa SMOTE, ditampilkan dalam bentuk hasil dan rata-rata akurasi dari *k-fold.*

Tabel 8 Akurasi SVM Tanpa SMOTE

Split Data	Akurasi K-Fold Validation					Rata-Rata Akurasi K-Fold Validation
	1	2	3	4	5	
(60:40)	79.3%	78.4%	79.9%	76.5%	76.5%	78.1%
(70:30)	76.8%	79.8%	80.1%	80.1%	80%	79.3%
(80:20)	80.4%	80.4%	78.7%	78.8%	79.6%	79.5%
(90:10)	81.1%	80.5%	79.1%	80.4%	82.4%	80.7%

Tabel 9 Akurasi SVM Dengan SMOTE

Split Data	Akurasi K-Fold Validation					Rata-Rata Akurasi K-Fold Validation
	1	2	3	4	5	
(60:40)	86.5%	87%	85.5%	85.9%	87.5%	86.4%
(70:30)	86.3%	87%	85.8%	85.2%	87%	86.2%

(80:20)	87.5%	86.8%	86.7%	86.4%	87.6%	87.0%
(90:10)	86.6%	86%	86.4%	86.2%	87%	86.4%

Dapat disimpulkan penggunaan SMOTE signifikan meningkatkan akurasi rata-rata algoritma *SVM* pada berbagai proporsi pembagian data. Tanpa SMOTE, akurasi tertinggi adalah 80.7% pada pembagian data 90:10, sedangkan dengan SMOTE, akurasi tertinggi mencapai 87.0% pada pembagian data 80:20. Penggunaan SMOTE pada pembagian data 80:20 memberikan performa terbaik untuk algoritma *SVM*, menunjukkan bahwa kombinasi ini efektif dalam mencapai akurasi tertinggi.

3. *Random Forest*

Tabel 10 dan Tabel 11 merupakan hasil pengujian akurasi pada algoritma *Random Forest* dengan SMOTE dan tanpa SMOTE, ditampilkan dalam bentuk hasil dan rata-rata akurasi dari *k-fold.*

Tabel 10 Akurasi Random Forest Tanpa SMOTE

Split Data	Akurasi K-Fold Validation					Rata-Rata Akurasi K-Fold Validation
	1	2	3	4	5	
(60:40)	74.7%	73.6%	73%	70.7%	71.7%	72.7%
(70:30)	72.6%	72.3%	76.5%	76.2%	74.6%	74.4%
(80:20)	74.4%	74.6%	75.2%	75.2%	73.7%	74.6%
(90:10)	75%	75.6%	73.9%	74.7%	75.7%	75.3%

Tabel 11 Akurasi Random Forest Dengan SMOTE

Split Data	Akurasi K-Fold Validation					Rata-Rata Akurasi K-Fold Validation
	1	2	3	4	5	
(60:40)	85.5%	85.2%	84.7%	84.1%	84.6%	84.8%
(70:30)	85.7%	85.1%	84.9%	84%	85.1%	85.0%
(80:20)	84.6%	85.4%	83%	85.1%	84.7%	84.5%
(90:10)	84.9%	84.4%	83.9%	84.5%	85.5%	84.6%

Dapat disimpulkan penggunaan SMOTE secara signifikan meningkatkan akurasi rata-rata algoritma *Random Forest* pada berbagai proporsi pembagian data. Tanpa SMOTE, akurasi tertinggi adalah 75.3% pada pembagian data 90:10, sedangkan dengan SMOTE, akurasi tertinggi mencapai 85.0% pada pembagian data 70:30. Penggunaan SMOTE pada pembagian data 70:30 memberikan performa terbaik untuk algoritma *Random Forest*, menunjukkan bahwa kombinasi ini efektif dalam mencapai akurasi tertinggi.

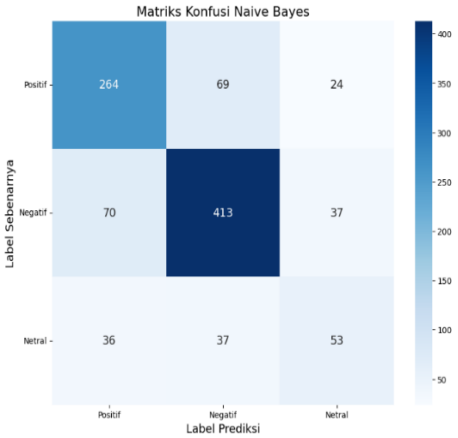
Berdasarkan hasil penelitian yang dilakukan, performa terbaik dari masing-masing algoritma yang diuji (*Naive Bayes, SVM*, dan *Random Forest*) adalah sebagai berikut pada Tabel 12

Tabel 12 Akurasi Terbaik dari Algoritma Naïve Bayes, SVM, dan Random Forest

Algoritma	Split Data	Teknik	Akurasi Tertinggi
Naïve Bayes	*(90:10)*	*k-fold dan SMOTE*	*73,9%*
SVM	*(80:20)*	*k-fold dan SMOTE*	*87,0%*
Random Forest	*(70:30)*	*k-fold dan SMOTE*	*85.0%*

Dari hasil pengujian tersebut, secara keseluruhan dapat disimpulkan bahwa penggunaan teknik *k-fold* dan SMOTE secara signifikan meningkatkan akurasi rata-rata untuk semua algoritma yang diuji. Algoritma yang memberikan performa paling terbaik adalah *SVM* dengan penerapan *k-fold Validation* dan SMOTE menggunakan pembagian data 80:20, dimana algoritma ini mencapai rata-rata akurasi *k-fold* tertinggi sebesar 87,0%.

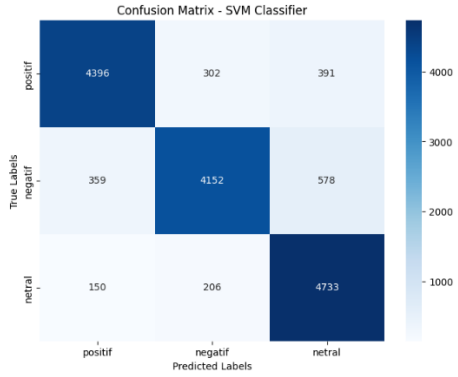
Setelah dilakukan proses *evaluation* pada model *Naïve Bayes, SVM* dan *Random Forest* tahap ini akan memberikan *output* berupa tabel *multiclass confusion matrix* dan metrik seperti *accuracy, precision, recall* dan *f-1 score.*



	precision	recall	f1-score	support
1	0.71	0.74	0.73	357
2	0.80	0.79	0.79	520
3	0.46	0.42	0.44	126
accuracy			0.73	1003
macro avg	0.66	0.65	0.65	1003
weighted avg	0.72	0.73	0.73	1003

Gambar 3 Hasil Evaluasi Confusion Matrix Naïve Bayes (90:10)

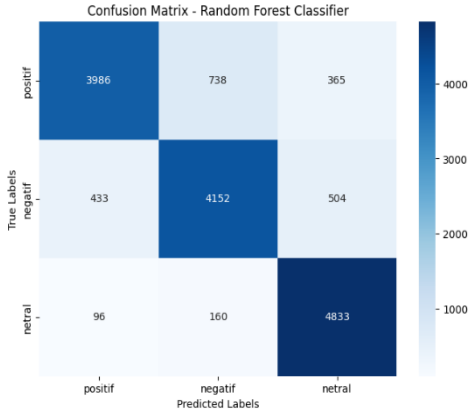
Dari gambar 3 diatas merupakan hasil evaluasi pada model *Naïve Bayes* dengan tingkat rata rata *accuracy* 73%, *precision* 66%, *recall* 65% dan *f-1 score* 65%.



	precision	recall	f1-score	support
1	0.90	0.86	0.88	5089
2	0.89	0.82	0.85	5089
3	0.83	0.93	0.88	5089
accuracy			0.87	15267
macro avg	0.87	0.87	0.87	15267
weighted avg	0.87	0.87	0.87	15267

Gambar 4 Hasil Evaluasi Confusion Matrix SVM (80:20)

Dari gambar 4 diatas merupakan hasil evaluasi pada model *SVM* dengan tingkat rata-rata *accuracy* 87%, *precision* 87%, *recall* 87% dan *f-1 score* 87%.



	precision	recall	f1-score	support
1	0.88	0.78	0.83	5089
2	0.82	0.82	0.82	5089
3	0.85	0.95	0.90	5089
accuracy			0.85	15267
macro avg	0.85	0.85	0.85	15267
weighted avg	0.85	0.85	0.85	15267

Gambar 5 Hasil Evaluasi Confusion Matrix Random Forest (70:30)

Dari gambar 5 diatas merupakan hasil evaluasi pada model Random Forest dengan tingkat rata-rata *accuracy* 85%, *precision* 85%, *recall* 85% dan *f-1 score* 85%.

IV. SIMPULAN

Dari hasil dari seleksi dan evaluasi, dapat disimpulkan bahwa Algoritma *SVM* menunjukkan performa terbaik untuk analisis sentimen pada 10.000 data *tweet* IKN dengan penerapan *k-fold validation* dan SMOTE pada rasio pembagian data (80:20), dengan nilai rata-rata tiap *accuracy, precision, recall*, dan *f1-Score* sejumlah 87% sehingga *SVM* mampu memberikan prediksi yang lebih konsisten dan akurat terhadap berbagai kelas sentimen (positif, negatif dan netral).

Algoritma *Random Forest* juga mengalami peningkatan signifikan dari 75.3% menjadi 85.0% pada pembagian data 70:30 dengan *k-fold* dan SMOTE, menunjukkan efektivitas kombinasi tersebut untuk meningkatkan akurasi. Meskipun *Naïve Bayes* menunjukkan peningkatan akurasi, performa pada kelas netral tetap rendah karena terjadi ketidakseimbangan data. Sehingga penggunaan *k-fold validation* dan SMOTE terbukti berhasil meningkatkan akurasi

untuk semua algoritma yang diuji, yaitu *Naïve Bayes, SVM* dan *Random Forest*. Dengan menerapkan validasi *k-fold validation* dan SMOTE, ditemukan peningkatan signifikan dalam performa model prediksi tersebut.

DAFTAR RUJUKAN

[1] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139084789.

[2] M. Izunnahdi, G. Aburrahman, and A. Eko Wardoyo, "Sentimen Analisis Pada Data Ulasan Aplikasi KAI Access Di Google PlayStore Menggunakan Metode Multinomial Naive Bayes Sentiment Analysis on KAI Access Application Review Data on Google PlayStore Using Multinomial Naive Bayes Method," *J. Smart Teknol.*, vol. 4, no. 2, pp. 2774–1702, 2023, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[3] R. S. Samosir, L. Abdillah, J. Gatc, and J. Gatc, "Pendampingan Pemanfaatan Smartphone Bagi Usia Dini di Lingkungan Perumahan XYZ," *ABDIMAS J. Pengabdi. Kpd. Masy.*, vol. 4, no. 1, pp. 67–71, 2023, doi: 10.53008/abdimas.v4i1.2109.

[4] A. Andreas and Y. D. Prabowo, "Analisis Sentimen Opini Publik dalam Bahasa Indonesia Terhadap UU Cipta Kerja Menggunakan Naïve Bayes," *KALBISIANA J. Mhs. Inst. Teknol. dan Bisnis Kalbis*, vol. 8, no. 2, pp. 2146–2161, 2022.

[5] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data Cogn. Comput.*, vol. 4, no. 1, pp. 1–34, 2020, doi: 10.3390/bdcc4010001.

[6] G. Erlangga, E. Sutoyo, and H. Fakhrurroja, "Analisis Sentimen Masyarakat Terhadap Penggunaan Aplikasi PeduliLindungi untuk Aktivitas Ruang Publik pada Media Sosial Twitter," *e-Proceeding Eng.*, vol. 10, no. 2, pp. 1376–1384, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/19898%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/19898/19263

[7] E. Benia and G. Nabilah, "Politik Hukum dalam Proses Pemindahan Ibu Kota Negara Melalui Pembentukan Undang-Undang Ibu Kota Negara (UU IKN)," *J. Huk. Lex Gen.*, vol. 3, no. 10, pp. 806–825, 2022, doi: 10.56370/jhlg.v3i10.323.

[8] F. Nurhuda, S. Widya Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," *J. Teknol. Inf. ITSmart*, vol. 2, no. 2, p. 35, 2016, doi: 10.20961/its.v2i2.630.

[9] A. Tiara Susilawati, A. H. Tiara Susilawati Universitas Muhammadiyah Kalimantan Timur Nur Anjeni Lestari Universitas Muhammadiyah Kalimantan Timur Puput Alpria Nina Universitas Muhammadiyah Kalimantan Timur Jl Ir Juanda No, K. Samarinda Ulu, K. Samarinda, and K. Timur, "Analisis Sentimen Publik Pada Twitter Terhadap Boikot Produk Israel Menggunakan Metode Naïve Bayes," *J. Ilm. Mhs.*, vol. 2, no. 1, pp. 26–35, 2024, [Online]. Available: https://doi.org/10.59603/niantanasikka.v2i1.240

[10] J. A.K. Suykens, M. Sigronetto, and A. Argyriou, *Regularization, Optimization, Kernels, and Support Vector Machines*. CRC Press, 2014.

[11] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[12] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[13] T. Fadiyah Basar, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen Pengguna Twitter terhadap Pembayaran Cashless menggunakan Shopeepay dengan Algoritma Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, pp. 1426–1433, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[14] Widodo, Steven Sondra Allen. Analisis Sentimen Penutupan Tiktokshop di Indonesia Menggunakan Algoritma Pengklasifikasi Naive Bayes. https://repository.uksw.edu

[15] Mariana. Analysis of Public Sentiment towards the Indonesian Presidential Election Based on Opinions from Online Media: A Literature Review.*Jurnal Sistem Informasi dan e-business.* Vol. 6, no. 01, 2024. https://doi.org/10.54650/jusibi.v6i1.541