JAMBURA JOURNAL OF INFORMATICS
Volume 6, No. 1 | April 2024: 26-39
ISSN 2685-4244, doi: 10.37905/jji.v6i1.24718

Analisis sentimen komentar youtube terhadap pemindahan ibu kota negara menggunakan metode *Naïve Bayes*

Shafira Faira Huwaida, Rosita Kusumawati, Bayutama Isnaini

Universitas Negeri Yogyakarta, Indonesia

Riwayat Artikel:
Diterima 12 Maret 2024
Direvisi 21 April 2024
Disetujui 21 April 2024
Diterbitkan 30 April 2024

Kata Kunci:
Analisis Sentimen
Ibu Kota Nusantara
Naïve Bayes
SMOTE

ABSTRACT. The relocation of Indonesia's capital city in the form of the construction of the IKN Nusantara in North Penajam Paser Regency, East Kalimantan, began in mid-March 2022 with a target of gradual relocation from 2024 to 2045. This invited various reactions from the public and enlivened social media, including YouTube. This research uses the naïve Bayes algorithm with and without SMOTE techniques to determine the sentiment of YouTube users about the transfer of the national capital. The data was in the comment section of the YouTube videos. The top three videos were selected based on the criteria of relevance to the keyword and the number of comments. Stages of pre-processing data include dropping duplicates, case folding, tokenizing, cleaning, steaming, converting slang words, removing stop words, and dropping missing values. Text data labeling was divided into three sentiment classes, Positive, Neutral, and Negative, using the help of the Vader Lexicon dictionary. The data was split into train and test with the ratio 70%:30%, 80%:20%, and 90%:10%. The research results show that the best performance was obtained from implementing SMOTE Naive Bayes on a data comparison train of 90% with a data test of 10%. *Naïve Bayes* with SMOTE model could provide *balanced accuracy* 76,01%, AUC Score 0,8711, dan G-mean 0,8089.

ABSTRAK. Pemindahan Ibu Kota Indonesia berupa pembangunan IKN Nusantara di Kabupaten Penajam Paser Utara, Kalimantan Timur dimulai pada pertengahan Maret 2022 dengan target pemindahan bertahap mulai tahun 2024 hingga 2045. Hal tersebut mengundang berbagai reaksi dari masyarakat hingga meramaikan media sosial termasuk YouTube. Penelitian ini menggunakan algoritma *naïve bayes* dengan dan tanpa teknik SMOTE untuk mengetahui sentimen pengguna YouTube tentang pemindahan ibu kota negara. Data yang digunakan berupa komentar dari video yang ada di YouTube. Terpilih tiga video teratas berdasarkan kriteria relevansi dengan *keyword* dan jumlah komentar. Tahapan *pre-processing* data meliputi *drop duplicates, case folding, tokenizing, cleaning, stemming, convert slangword, removal stopword,* dan *drop missing value.* Pelabelan data teks terbagi menjadi 3 kelas sentimen yaitu Positif, Netral, dan Negatif menggunakan bantuan kamus vader lexicon. Pembagian data *train* dan *test* dilakukan mengggunakan tiga perbandingan, yaitu 70%:30%, 80%:20%, dan 90%:10%. Hasil penelitian menunjukkan bahwa performa terbaik diperoleh dari penerapan SMOTE *naive bayes* pada perbandingan data *train* 90% dengan data *test* 10%. Model SMOTE *naïve bayes* mampu memberikan nilai *balanced accuracy* 76,01%, AUC Score 0,8711, dan G-mean 0,8089.

Penulis Korespondensi:
Shafira Faira Huwaida,
Universitas Negeri Yogyakarta,
Jl. Colombo No.1 Karangmalang, Yogyakarta 55281, Indonesia.
Email: shafirafaira.2019@student.uny.ac.id

PENDAHULUAN

Pemindahan Ibu Kota Negara (IKN) Nusantara telah melalui perjalanan panjang hingga saat ini. Pada tanggal 26 Agustus 2019, Presiden Joko Widodo mengumumkan ibu kota baru akan dibangun di wilayah administratif Kabupaten Penajam Paser Utara dan Kabupaten Kutai Kartanegara, Kalimantan Timur (*Profil Penajam Paser Utara, Lokasi Ibu Kota Baru Di Kalimantan Timur – DPRD KABUPATEN PENAJAM PASER UTARA*, 2021). Pembangunan IKN dimulai pada pertengahan Maret 2022 dengan target pemindahan bertahap mulai tahun 2024 hingga 2045 (Nugroho, 2022). Pemindahan Ibu Kota Negara Indonesia menimbulkan pro dan kontra. Hal tersebut disampaikan masyarakat salah satunya dengan meramaikan media sosial YouTube yang ditunjukkan dengan berbagai reaksi pada unggahan video terkait pemindahan ibu kota. Opini yang diberikan oleh para pengguna YouTube dalam kolom komentar seringkali mencapai jumlah yang cukup besar dan sering digunakan untuk membuat penilaian yang merujuk pada suatu topik atau tujuan tertentu dikenal dengan istilah analisis sentimen. Cara kerja analisis sentimen dalam mengambil data dapat dibagi menjadi tiga langkah, yaitu klasifikasi, evaluasi, dan visualisasi hasil (LP2M UMA, 2022). Teknik klasifikasi sentimen terdiri dari *machine learning, lexicon based*, dan *hybrid* (Sharma dan Singh, 2018). *Machine learning* merupakan proses pelatihan komputer sehingga mampu mengambil keputusan sendiri yang melibatkan *supervised learning* dan *unsupervised learning*, sedangkan *lexicon based* merupakan metode analisis teks yang menggunakan kata-kata atau kamus yang telah diklasifikasikan sebelumnya. *Lexicon based* terdiri dari dua metode, yaitu *dictionary based* dan *corpus based* (Wikarsa et al., 2022). Penggabungan *machine learning* dengan *lexicon-based* dikenal dengan pendekatan *hybrid.* Berdasarkan hal tersebut, maka metode yang digunakan untuk analisis sentimen komentar tentang Ibu Kota Negara (IKN) Nusantara adalah *naïve bayes classifier* yang merupakan salah satu metode klasifikasi *supervised learning*.

Penelitian yang dilakukan oleh Novendri et al. (2020) menyebutkan bahwa analisis sentimen komentar YouTube terhadap trailer untuk musim keempat Money Heist dengan metode *naïve bayes classifier* menghasilkan akurasi sebesar 81%. Pada penelitian yang dilakukan oleh Rahman et al. (2020) terkait penggunaan metode *naïve bayes* untuk menganalisis akurasi sentimen komentar YouTube memberikan hasil akurasi sebesar 78,17%. Berdasarkan beberapa penelitian di atas, *naïve bayes* terbukti dapat memberikan hasil yang cukup baik untuk analisis sentimen sehingga penelitian ini menggunakan metode *naïve bayes.* Namun, terdapat salah satu masalah yang dihadapi dalam analisis sentimen, yaitu *imbalance* data pada jumlah sampel positif, negatif, dan netral (Setiawan, 2023). Pernyataan tersebut didukung dengan pendapat dari Magnolia et al. (2023) yang menyatakan bahwa *imbalance* data umum ditemukan pada pengambilan data secara langsung. Algoritma klasifikasi akan mengalami penurunan performa jika menghadapi kelas *imbalance* data (García et al., 2012). Salah satu cara untuk mengatasi hal tersebut adalah dengan *resampling.*

Resampling merupakan salah satu cara untuk mengatasi *imbalance* data dengan memodifikasi jumlah individu di kelas mayoritas dan minoritas menjadi data yang seimbang (Indrawati et al., 2020). Secara umum, *resampling* dikelompokkan menjadi tiga, yaitu *undersampling* kelas mayor, *oversampling* kelas minor, dan kombinasi teknik *over-* dan *under-sampling* (Longadge dan Dongre, 2013). Salah satu teknik *oversampling* yang sering digunakan adalah *synthetic minority oversampling technique* (SMOTE). Oleh karena itu, teknik oversampling SMOTE diimplementasikan pada algoritma *naïve bayes* apabila data komentar YouTube tentang Ibu Kota Negara (IKN) Nusantara *imbalance.* Penelitian oleh Sulistiyowati dan Jajuli (2020) pada data nasabah kredit di Koperasi Guru Rawamerta menunjukkan bahwa kombinasi SMOTE dengan *naïve bayes* efektif untuk menangani *imbalance* data dan menghasilkan tingkat akurasi 94.015% dan G-mean 0.948. Penelitian lain oleh Kurnia et al. (2023) menggunakan algoritma *naïve bayes*, SMOTE, dan AdaBoost dalam analisis sentimen pada

twitter Bank BTN menunjukkan bahwa SMOTE efektif untuk mengatasi kondisi *imbalance* dibuktikan dengan kenaikan *accuracy* dari 75,40% menjadi 85,8%.

Perbandingan kinerja performa klasifikasi dari metode *naive bayes* dengan dan tanpa SMOTE dilakukan untuk mengetahui efektifitas penerapan SMOTE pada metode *naïve bayes* dalam mengklasifikasikan kelas sentimen dengan menggunakan model pembagian data train dan data test 70%:30%, 80%:20%, dan 90%:10% berdasarkan nilai *balanced accuracy,* AUC Score, dan G-mean. Penelitian ini bertujuan untuk memperoleh informasi tentang sentimen pengguna YouTube tentang Ibu Kota Negara Nusantara serta bagaimana efektifitas penerapan SMOTE pada metode *naïve bayes* mengklasifikasikan sentimen tersebut.

METODE

Data

Data komentar penonton diperoleh menggunakan API YouTube sebanyak 7728 baris komentar dari 3 video. Judul video dan nama *channel* YouTube yang digunakan sebagai sumber pengambilan data dipaparkan melalui Tabel 1. Data komentar penonton diperoleh sejak tanggal unggah masing-masing video hingga 31 Agustus 2023.

Tabel 1. Daftar video sumber data

No.	*Channel* YouTube	Judul Video	Tanggal Unggah	Jumlah komentar
1	IKN Indonesia	Marketing Sounding IKN | Sejarah Baru Peradaban Baru	22 Oktober 2022	889
2	KIKI KHOTO	Super Bangga! Mengintip Wujud IKN Nusantara 2024 Nan Megah, 4 Kali Luas Jakarta 3 Kali Luas Singapura	6 Februari 2023	1498
3	Sekretariat Presiden	Sambutan Presiden Jokowi pada IKN: Sejarah Baru Peradaban Baru, 18 Oktober 2022	19 Oktober 2022	5341

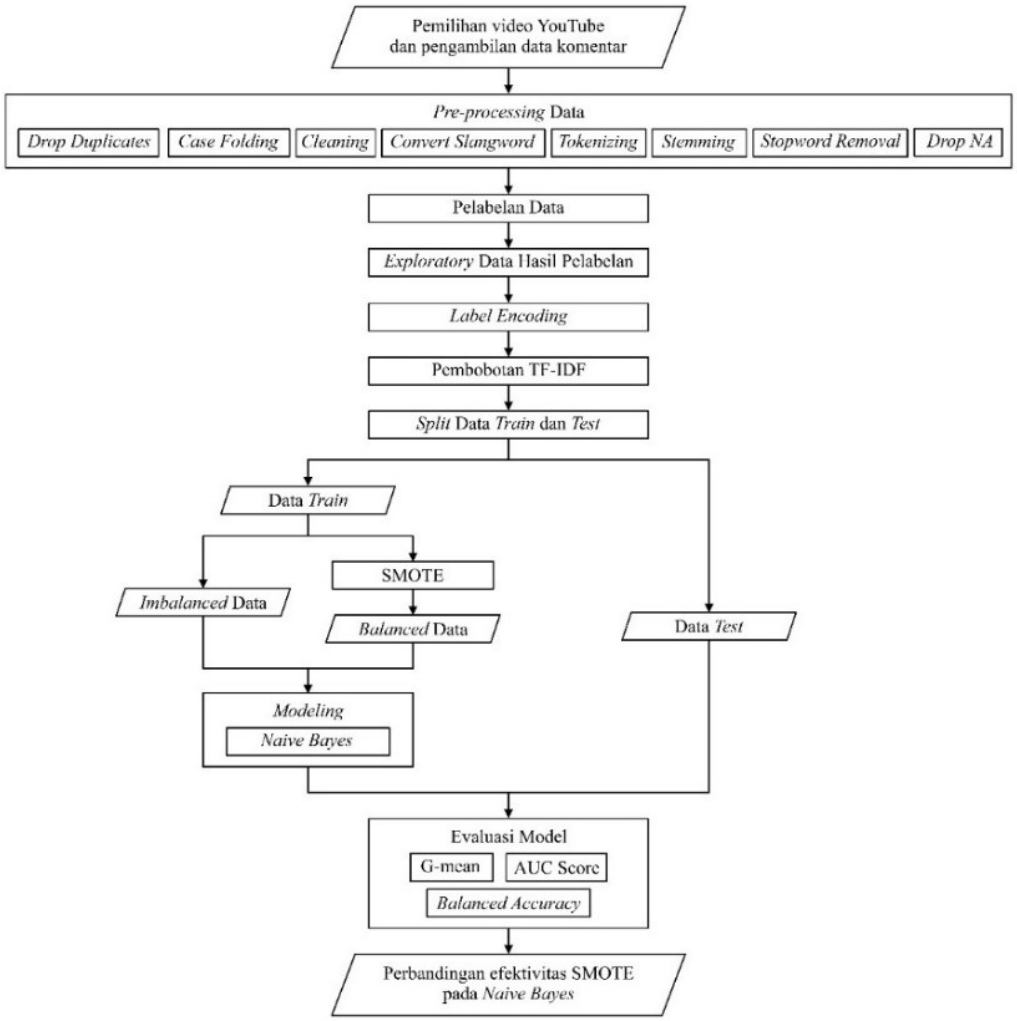
Tahapan Analisis Data

Dalam pembuatan model klasifikasi, metode yang digunakan yakni *naïve bayes* dengan dan tanpa SMOTE. Pada data, dilakukan *pre-processing* (*drop duplicates, case folding, cleaning, convert slangword, tokenizing, stemming, stopword removal,*dan *drop* NA), pelabelan data menggunakan VADER Lexicon, *exploratory* data hasil pelabelan, pembagian data *train* dan *test, balancing* data, *modeling,* dan evaluasi model. Langkah-langkah untuk melakukan penelitian ini ditunjukkan dalam Gambar 1.

Pre-processing

Proses *pre-processing* merupakan tahapan memperbaiki data teks yang tidak terstruktur menjadi data yang dapat dianalisis. Tahapan *pre-processing* pada data adalah sebagai berikut.

1. *Drop Duplicates* untuk menghilangkan duplikat komentar yang sama dari user yang sama.
2. *Case Folding* untuk mengubah semua teks yang mengandung huruf kapital menjadi huruf kecil.
3. *Cleaning-Remove Unnecessary Character* untuk menghilangkan username yang dimulai dengan *mention,* karakter simbol dan URL.
4. *Cleaning-Remove Emoji* untuk menghilangkan emoji.
5. *Cleaning-Remove Non-Alpha Numeric* untuk menghapus tanda baca.
6. *Convert Slangwords* untuk mengubah kata *slang* menjadi kata baku menurut Kamus Besar Bahasa Indonesia.
7. *Tokenizing* untuk memecah teks menjadi kata per kata.
8. *Stemming* untuk menghilangkan kata imbuhan sehingga diperoleh kata dasar.
9. *Stopwords Removal* untuk menghilangkan kata-kata yang tidak penting dalam teks dengan menghapus kata pada komentar yang terdapat dalam daftar *stopwords.*
10. *Drop NA* untuk menghilangkan baris komentar kosong.



Gambar 1. Diagram alur tahapan analisis data

Pelabelan Data

Pabelan data dilakukan dengan menggunakan pelabelan data otomatis *VADER* (*Vader Aware Sentiment Reasoner*) pada data teks yang terdiri dari tiga label kelas, yaitu label positif, negatif, dan netral.

Exploratory Data Hasil Pelabelan

Exploratory data hasil pelabelan dilakukan dengan membuat visualisasi untuk mengetahui pola atau tren frekuensi komentar setiap bulan pada kelas sentimen.

TF-IDF

TF-IDF merupakan metode pembobotan kata dimana kata yang telah melalui tahap *pre-processing* akan dihitung bobotnya. TF-IDF bekerja dengan melibatkan perkalian antara *term frequency (TF)* dengan *inverse document frequency (IDF). Term frequency* ($tf_{i,j}$) merupakan jumlah kemunculan *term* ke-*i* pada dokumen ke-*j,* sedangkan *Inverse Document Frequency* dapat dihitung dengan menggunakan formula sebagai berikut (Joachims, 1997).

$$idf_{(j)} = \log\left(\frac{|D|}{df_j}\right) \quad (1)$$

keterangan:
$|D|$: jumlah semua dokumen
df_j : jumlah dokumen yang mengandung term pada kelas *j*

$$W_{i,j} = tf_{(i,j)} \times idf_{(j)} \quad (2)$$

dengan:
$W_{i,j}$: bobot term i terhadap dokumen *j*
$tf_{(i,j)}$: jumlah kemunculan term i dalam dokumen *j*

a. Pembagian data *train* dan *test*

Pada tahap ini dilakukan pembagian dataset menjadi data pelatihan (data *train*) dan pengujian (data *test*) untuk tiga perbandingan, yaitu 70%:30%, 80%:20%, dan 90%: 10%.

b. *Naïve Bayes Classifier*

Algoritma *naive bayes* menjadi salah satu metode yang popular digunakan dalam analisis sentimen. *Naive bayes* termasuk kedalam metode klasifikasi sederhana dengan menghitung semua probabilitas berdasar pada teorema Bayes yang dikombinasikan dengan kombinasi nilai frekuensi database. Konsep peluang bersyarat menjadi acuan dari teorema *naive bayes*. Secara umum, peluang *posterior* memiliki persamaan sebagai berikut (Raschka, 2014):

$$P(w_j|x_i) = \frac{P(x_i|w_j)P(w_j)}{P(x_i)} \quad (3)$$

dengan:
$P(w_j|x_i)$: Nilai posterior atau peluang kategori *j* ketika terdapat kemunculan kata ke-i
$P(x_i|w_j)$: Nilai likelihood atau peluang kata ke-i masuk kedalam kategori
$P(w_j)$: Nilai prior atau peluang kemunculan kategori *j*
$P(x_i)$: Nilai evidence atau peluang kemunculan kata

Naïve bayes multinomial merupakan model pengembangan dari *naïve bayes* yang cocok untuk klasifikasi teks dan dokumen. *Naïve bayes* multinomial lebih umum digunakan dalam klasifikasi teks karena dapat menangani data teks yang direpresentasikan sebagai vektor frekuensi data. *Naïve bayes* multinomial digunakan ketika fitur-fitur mewakili frekuensi atau jumlah kemunculan, seperti dalam klasifikasi teks dimana setiap fitur mewakili jumlah kata dalam dokumen. Hal ini sesuai dengan pernyataan oleh Destuardi dan Sumpeno (2009) dimana algoritma *naïve bayes* multinomial memperhitungkan jumlah kata yang muncul dalam dokumen. *Term frequency* digunakaan untuk menghitung banyaknya sebuah kata muncul dalam kalimat atau dokumen dalam konsep *naïve bayes* multinomial. Nilai probabilitas suatu dokumen *d* berada di kelas *c* dapat dihitung menggunakan persamaan seperti berikut (Manning et al., 2009).

$$P(c|d) \propto P(c) \prod_{k=1}^{n_d} P(t_k|c) \quad (4)$$

dengan:
$P(c|d)$: peluang dokumen d berada di kelas c
$P(c|d)$: prior peluang suatu dokumen berada di kelas c
$\{t_1, t_2, \dots, t_{n_d}\}$: token dalam dokumen d yang merupakan bagian dari kosa kata (*vocabulary*) dengan jumlah n
$P(t_k|c)$: probabilitas bersyarat term t_k berada di dokumen pada kelas c

Nilai $P(v_j)$ dan probabilitas kata a_i untuk setiap kategori $P(a_i|v_j)$ dapat dihitung saat *training* dengan formula pada persamaan berikut.

$$P(v_j) = \frac{|doc_j|}{|training|} \quad (5)$$

$$P(a_i|v_j) = \frac{n_i+1}{|n+\text{kosakata}|} \quad (6)$$

dengan:

$\|doc_j\|$	: jumlah dokumen kelas j
$\|training\|$	: jumlah dokumen yang digunakan dalam proses *training*
$P(a_i\|v_j)$	: Probabilitas kondisional bahwa fitur a_i muncul dalam dokumen yang termasuk dalam kategori atau kelas v_j
n_i	: Jumlah kemunculan fitur a_i dalam dokumen yang termasuk dalam ketegori atau kelas v_j
$\|n\|$	: Jumlah total kemunculan semua fitur dalam dokumen yang termasuk dalam kategori atau kelas v_j
|kosakata|	: Jumlah total fitur unik dalam seluruh dataset atau kumpulan data

c. SMOTE (*Synthetic Minority Oversampling Technique*) *Naïve Bayes Classifier*

Metode kedua yang digunakan adalah penerapan SMOTE pada *naïve bayes*. Prinsip dasar dari SMOTE adalah menambah jumlah data pada kelas minoritas agar seimbang dengan kelas mayoritas dengan cara membangkitkan data sintetis berdasarkan k-tetangga terdekat (*k-nearest neighbor*). Persentase replikasi data minoritas untuk data sintesis dapat dirumuskan pada persamaan berikut (Chawla et al., 2002).

$$N\% = \frac{\text{Jumlah data mayoritas}}{\text{Jumlah data minoritas}} \times 100\% \quad (7)$$

Rumus untuk membangkitkan data sintesis dengan SMOTE sebagai berikut:

$$x_{syn} = x_i + (x_{knn} - x_i)\delta \quad (8)$$

dengan

x_{syn}	: data sintesis hasil replikasi
x_i	: data ke$-i$ yang akan direplikasi (dari kelas minoritas)
x_{knn}	: data yang memiliki jarak terdekat dengan x_i
δ	: nilai acak antara 0 dan 1
i	: komentar ke 1, 2, …, n

d. Evaluasi Model

Menghitung ketepatan klasifikasi dan membandingkan performa kinerja dari metode *naïve bayes* dengan SMOTE dan tanpa SMOTE berdasarkan nilai *balanced accuracy,* AUC *score,* dan G-mean.

HASIL DAN DISKUSI

Pre-processing Data

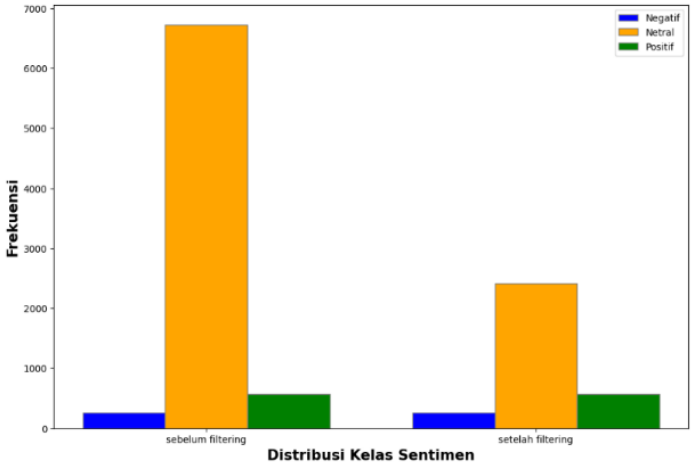
Data komentar YouTube tentang pemindahan ibu kota negara yang memiliki jumlah sebesar 7728 komentar memasuki tahap *pre-processing* terlebih dahulu untuk menyeragamkan isi dari data yang diperoleh. Adapun hasil dari setiap tahap *pre-processing* data dapat dilihat pada Tabel 2.

Tabel 2. Hasil *pre-processing*

Tahap	Komentar
Kalimat awal	Sejarah Nusantara baru masa depan nanti 👍 Tks pemimpin pak presiden RI bpk Joko Widodo... kami masyarakat Indonesia sayang pakde www.iknjaya.com
Case Folding (Lowercase)	sejarah nusantara baru masa depan nanti 👍 tks pemimpin pak presiden ri bpk joko widodo... kami masyarakat indonesia sayang pakde www.iknjaya.com
Cleaning	sejarah nusantara baru masa depan nanti tks pemimpin pak presiden ri bpk joko widodo kami masyarakat indonesia sayang pakde
Convert Slangwords	sejarah nusantara baru masa depan nanti terima kasih pemimpin bapak presiden ri bapak joko widodo kami masyarakat indonesia sayang pakde
Tokenizing	['sejarah', 'nusantara', 'baru', 'masa', 'depan', 'nanti', 'terima', 'kasih', 'pemimpin', 'bapak', 'presiden', 'ri', 'bapak', 'joko', 'widodo', 'kami', 'masyarakat', 'indonesia', 'sayang', 'pakde']
Stemming	['sejarah', 'nusantara', 'baru', 'masa', 'depan', 'nanti', 'terima', 'kasih', 'pimpin', 'bapak', 'presiden', 'ri', 'bapak', 'joko', 'widodo', 'kami', 'masyarakat', 'indonesia', 'sayang', 'pakde']
Removal Stopword	sejarah nusantara baru masa depan nanti terima kasih pimpin presiden ri joko widodo kami masyarakat indonesia sayang pakde

Pelabelan Data

Setelah melalui tahap *preprocessing,* kemudian dilakukan pelabelan data. Pelabelan data teks terbagi menjadi 3 kelas sentimen yaitu Positif untuk sentimen score $\geq 0{,}05$, Netral untuk sentimen score antara $-0{,}05$ hingga $0{,}05$, dan Negatif untuk sentimen score$\leq -0{,}05$ menggunakan bantuan kamus *Lexicon* VADER.

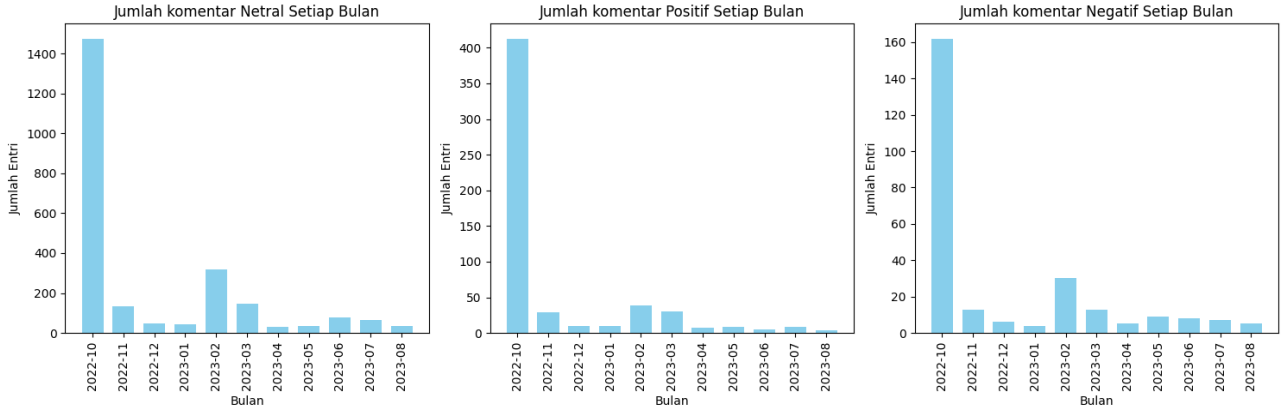


Gambar 2. Distribusi kelas sentimen sebelum dan sesudah *filtering*

Berdasarkan Gambar 2 diketahui bahwa hasil dari *pre-processing* dan pelabelan kelas sentimen kelas netral mempunyai frekuensi terbanyak yaitu sebesar 6724 komentar, sedangkan kelas sentimen negatif sebanyak 262 komentar dan sentimen positif sebanyak 565 komentar. Oleh karena tingginya proporsi komentar kelas netral, dilakukan *filtering* komentar secara manual pada komentar kelas

netral untuk menurunkan rasio ketidakseimbangan kelas sentimen. *Filtering* komentar kelas netral dilakukan dengan hanya mengambil komentar yang memuat salah satu kata dari ikn, nusantara, ibu kota, ibukota, negara, nkri, ri, dan indonesia. Berdasarkan proses tersebut, diperoleh frekuensi kelas netral sebesar 2410 komentar.

Exploratory Data Hasil Pelabelan



Gambar 3. Frekuensi komentar setiap bulan

Gambar 3 menunjukkan bahwa pada ketiga kelas komentar setelah tahap pelabelan, mayoritas komentar diunggah pada bulan pertama yaitu Oktober 2022. Sejumlah 1476 komentar netral diunggah pada bulan Oktober 2022, dari total 2410 komentar. Pada bulan yang sama, terdapat 413 komentar positif dan 162 komentar negatif yang diunggah. Setelah bulan Oktober 2022, unggahan komentar netral, positif, dan negatif memiliki pola cenderung naik-turun secara periodik. Hal ini dapat menunjukkan sifat kecenderungan tren jangka pendek.

Gambar 4. *Wordcloud* setiap kelas sentimen

Berdasarkan Gambar 4 visualisasi dalam bentuk *Wordcloud* dapat menujukkan lebih jelas kata-kata yang sering muncul pada setiap kelas sentimen terkait Ibu Kota Negara Nusantara. Terlihat bahwa dalam ketiga kelas sentimen, kata “ikn”, “moga”, dan “maju” merupakan kata dengan frekuensi penggunaan cukup tinggi. Pada kelas negatif dapat dilihat bahwa penggunaan kata “doa” banyak digunakan dan kata “hoax” juga cukup sering digunakan, sedangkan pada kelas positif terdapat kata “bangga” dan “best” yang memiliki frekuensi penggunaan cukup tinggi.

Label Encoding

Digunakan pembuat encode Label Encoder untuk mengubah jenis data kategorikal menjadi data numerik yang dapat dipahami model (Gambar 5).

	komen	Sentiment
0	hormat sangatberharap jalan umum ibu kota ikn ...	1
1	punya usul supaya tidak terlalu banyak pinjam ...	1
2	bagus ibukotanya pindah karna jakarta suda sem...	1
3	persiden asli utus allah atau islam aullah tuh...	1
4	kawan kawan rabbani mari kita jaga intan berli...	1
...	...	...

Gambar 5. Data setelah proses *Label Encoding*

Term Weighting (TF-IDF)

TF-IDF merupakan metode pembobotan kata dimana kata yang telah melalui tahap *pre-processing* akan dihitung bobotnya. Proses ini diawali dengan mengubah data komentar YouTube ke dalam bentuk kata hingga didapatkan frekuensi kemunculan setiap kata hingga teks kemudian diubah dalam bentuk numerik. Contoh perhitungan TF-IDF pada sebuh dokumen, dapat dilihat pada Tabel 3 berikut ini: Dokumen 1 = "jaya indonesia jaya jokowi kita senang presiden bukti kerja nyata smoga sukses laksana", Dokumen 2 = "wow luar biasa ide jokowi tidak ada tanding saya bangga bos aku".

Tabel 3. Perhitungan TF-IDF

doc	*term*		TF		IDF	TF*IDF	
	d1	*d2*	*d1*	*d2*		*d1*	*d2*
kita	1	0	0,0769	0,0	0,301	0,0232	0,0
jaya	2	0	0,1538	0,0	0,301	0,0463	0,0
saya	0	1	0,0	0,0833	0,301	0,0	0,0251
sukses	1	0	0,0769	0,0	0,301	0,0232	0,0
bangga	0	1	0,0	0,0833	0,301	0,0	0,0251
kerja	1	0	0,0769	0,0	0,301	0,0232	0,0
tidak	0	1	0,0	0,0833	0,301	0,0	0,0251
biasa	0	1	0,0	0,0833	0,301	0,0	0,0251
smoga	1	0	0,0769	0,0	0,301	0,0232	0,0
ada	0	1	0,0	0,0833	0,301	0,0	0,0251
senang	1	0	0,0769	0,0	0,301	0,0232	0,0
presiden	1	0	0,0769	0,0	0,301	0,0232	0,0
jokowi	1	1	0,0769	0,0833	0,0	0,0	0,0
bukti	1	0	0,0769	0,0	0,301	0,0232	0,0
luar	0	1	0,0	0,0833	0,301	0,0	0,0251
ide	0	1	0,0	0,0833	0,301	0,0	0,0251
wow	0	1	0,0	0,0833	0,301	0,0	0,0251
laksana	1	0	0,0769	0,0	0,301	0,0232	0,0
tanding	0	1	0,0	0,0833	0,301	0,0	0,0251
aku	0	1	0,0	0,0833	0,301	0,0	0,0251
indonesia	1	0	0,0769	0,0	0,301	0,0232	0,0
bos	0	1	0,0	0,0833	0,301	0,0	0,0251
nyata	1	0	0,0769	0,0	0,301	0,0232	0,0

Distribusi data antar komentar pada Tabel 4 menunjukkan perbandingan rasio label positif 17,45% (565 komentar), netral 74,45% (2410 komentar) dan negatif 8,09% (262 komentar) dari total 3237 komentar. Persentase kelas netral memiliki persentase yang jauh lebih tinggi daripada label positif dan negatif mengidentifikasikan *imbalance data*.

Split *Data* Train *dan* Test

Tabel 4. Hasil *split* data *train* dan *test*

Kelas Sentimen	Perbandingan 70%:30%		Perbandingan 80%:20%		Perbandingan 90%:10%	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*
Negatif	190	72	217	45	239	23
Netral	1692	718	1934	476	2168	242
Positif	383	182	438	127	506	59
Total	2245	972	2589	648	2913	324

SMOTE

Tabel 5. Hasil *balancing* data dengan SMOTE

Kelas Sentimen	70% Data *Train*		80% Data *Train*		90% Data *Train*	
	Tanpa SMOTE	Dengan SMOTE	Tanpa SMOTE	Dengan SMOTE	Tanpa SMOTE	Dengan SMOTE
Negatif	190	1692	217	1934	239	2168
Netral	1692	1692	1934	1934	2168	2168
Positif	383	1692	438	1934	506	2168
Total	2245	5076	2589	5802	2913	6504

Pada Tabel 5 menunjukkan terjadi penambahan masing-masing data untuk kelas minoritas (negatif dan positif) pada data train 70% menjadi 1692, data train 80% menjadi 1934, dan pada data train 90% menjadi 2168 data.

Modeling Naïve Bayes

Berdasarkan pembagian data *train* dan data *test* yang telah dilakukan sebelumnya, proses klasifikasi selanjutnya dilakukan dengan metode *naïve bayes*.

Tabel 6. Contoh komentar data *train* dan *test*

doc	Label	Dokumen
1	positif	wow benar benar kota impi suara kecil dari masyarakat pinggir salam sukses damai sejahtera indonesia maju
2	netral	bagus sekali lanjut ikn agar indonesia makin rata bangun kurang macet dki
3	negatif	ikn no comen untuk siapa
4	?	sukses bangun ikn indonesia

Tabel 6 merupakan contoh data yang telah diberikan label dengan doc 1, 2, dan 3 merupakan data *train* dan doc 4 merupakan data *test*. Perhitungan dilakukan untuk mengklasifikasikan apakah data *test* tersebut termasuk ke dalam sentimen negatif, netral, atau positif. Digunakan metode klasifikasi

naïve bayes untuk memberikan label pada doc 4. Tahapan klasifikasi setelah pembobotan kata adalah sebagai berikut:

1. Menghitung probabilitas *prior* setiap kelas sentimen yang terdiri dari kelas positif (v_2), netral (v_1), dan negatif (v_0).

   $P(v_2) = \frac{1}{3} = 0{,}333 \qquad P(v_1) = \frac{1}{3} = 0{,}333 \qquad P(v_0) = \frac{1}{3} = 0{,}333$

2. Menghitung peluang kemunculan kata untuk setiap *term* pada data *train.* Berikut contoh perhitungan probabilitas kata "sukses":

   $P(sukses|v_2) = \frac{1+1}{16+30} = 0{,}0435$

   $P(sukses|v_1) = \frac{0+1}{12+30} = 0{,}0238$

   $P(sukses|v_0) = \frac{0+1}{5+30} = 0{,}286$

3. Setelah didapat nilai *prior* dan peluang kemunculan kata setiap *term* dilakukan pengujian data *test* dengan mencari nilai peluang tertinggi.

   $P(v_0)\prod_i P(a_i|v_0)$

   $= (0{,}333)(P(a_{sukses}|v_0) \times P(a_{bangun}|v_0) \times P(a_{ikn}|v_0) \times P(a_{indonesia}|v_0)$

   $= 4{,}44814 \times 10^{-7}$

   $P(v_1)\prod_i P(a_i|v_1)$

   $= (0{,}333)(P(a_{sukses}|v_1) \times P(a_{bangun}|v_1) \times P(a_{ikn}|v_1) \times P(a_{indonesia}|v_1)$

   $= 8{,}54756 \times 10^{-7}$

   $the\ P(v_2)\prod_i P(a_i|v_2)$

   $= (0{,}333)(P(a_{sukses}|v_2) \times P(a_{bangun}|v_2) \times P(a_{ikn}|v_2) \times P(a_{indonesia}|v_2)$

   $= 2{,}96717 \times 10^{-7}$

   Nilai probabilitas kata pada doc 4 yang terbesar pada setiap *term* adalah probabilitas setiap kata pada sentimen netral sehingga doc 4diklasifikasikan kedalam sentimen netral.

Evaluasi Model (Perbandingan Efektivitas SMOTE pada *Naïve Bayes*)

Tabel 7. Hasil nilai *TP (True Positive), TN (True Negative), FP (False Positive), dan FN (False Negative).*

Kriteria	*Naïve Bayes*			SMOTE *Naïve Bayes*		
	70%:30%	80%:20%	90%:10%	70%:30%	80%:20%	90%:10%
TP	718	476	241	624	423	213
TN	41	35	20	172	128	60
FP	213	137	62	82	44	22
FN	0	0	1	94	53	29

Berdasarkan Tabel 7 didapatkan hasil klasifikasi algoritma *naïve bayes* tanpa SMOTE lebih sensitif untuk sentimen netral yang ditunjukkan dengan banyaknya nilai *true positif* (TP) atau sentimen netral yang diklasifikasikan dalam kelas netral. Hasil *True Positive* (TP) menunjukkan komentar terklasifikasi sentimen netral lebih banyak dibandingkan sentimen positif dan negatif baik

menggunakan algoritma *naïve bayes* baik dengan penerapan SMOTE maupun tanpa penerapan SMOTE. Namun jika ditinjau dari banyaknya nilai TP, klasifikasi *naïve bayes* dengan menerapkan SMOTE memberikan nilai TP lebih rendah daripada tanpa SMOTE. Di sisi lain, klasifikasi algoritma *naïve bayes* dengan SMOTE lebih sensitif terhadap sentimen negatif dan positif yang ditunjukkan dengan banyaknya nilai *true negative* (TN) atau sentimen negatif yang diklasifikasikan negatif dan positif diklasifikasikan positif.

Tabel 8. Kinerja model *Naïve Bayes* dan SMOTE *Naïve Bayes*

Kriteria	*Naïve Bayes*			SMOTE *Naïve Bayes*		
	70%:30%	80%:20%	90%:10%	70%:30%	80%:20%	90%:10%
Balanced Accuracy	41,4%	42,99%	43,93%	70,3%	75,41%	76,01%
AUC Score	0,7598	0,7775	0,7744	0,8544	0,8663	0,8711
G-mean	0,5451	0,572	0,59	0,7772	0,8182	0,8089

Tabel 8 menunjukkan model *naïve bayes* menghasilkan score AUC, *balanced accuracy,* dan G-mean lebih rendah dibandingkan model SMOTE *naïve bayes.* Hasil dari kedua skenario percobaan, menunjukkan bahwa SMOTE mampu meningkatkan nilai *balanced accuracy* dan G-Mean pada setiap perbandingan data. Pada model tanpa SMOTE, ketiga perbandingan data *train* dan *test* menunjukkan nilai AUC 0,7598 hingga 0,7775 yang menunjukkan bahwa kebaikan model dikategorikan baik. Setelah penerapan SMOTE, nilai AUC model meningkat sebesar 0,0833 hingga 0,1065 sehingga membuat ketiga model termasuk dalam kategori sangat baik dan menjadi lebih akurat. Hal ini sesuai dengan penelitian yang dilakukan oleh Barro et al. (2013) yang membahas tentang penerapan SMOTE pada *imbalance data* pada pembuatan model komposisi jamu yang menyebutkan bahwa model dengan SMOTE lebih akurat karena nilai AUC yang dihasilkan lebih tinggi daripada model tanpa SMOTE. Hasil dari kedua skenario percobaan menunjukkan bahwa SMOTE mampu meningkatkan nilai *balanced accuracy* dan G-Mean pada setiap perbandingan data. Hasil ini sejalan dengan penelitian dari Sulistiyono et al. (2021) yang mengimplementasikan algoritma SMOTE untuk menangani ketidakseimbangan kelas pada dataset klasifikasi. Penelitian tersebut menyimpulkan bahwa penanganan distribusi kelas *imbalance* pada dataset menggunakan algoritma SMOTE dapat meningkatkan nilai akurasi maupun G-mean pada algoritma *naïve bayes, SVM, KNN,* dan *Decision Tree.*

Berdasarkan hasil yang diperoleh dari penerapan SMOTE pada algoritma *naïve bayes* untuk klasifikasi sentimen menunjukkan bahwa penanganan *imbalance data* menggunakan SMOTE mampu memperbaiki kinerja klasifikasi dalam memprediksi kelas sentimen minoritas (positif dan negatif). Hasil pengujian membuktikan bahwa teknik SMOTE efektif meningkatkan kinerja algoritma *naïve bayes* untuk nilai *balanced accuracy,* score AUC, dan nilai G-mean. Perbandingan *balanced accuracy* dan AUC *score* pada ketiga model SMOTE *naïve bayes* menunjukkan bahwa klasifikasi pada perbandingan data *train* dan *test* 90%:10% terbukti lebih efektif dibandingkan dengan klasifikasi pada rasio perbandingan lainnya.

KESIMPULAN

Hasil sentimen pengguna YouTube terhadap Pemindahan Ibu Kota Negara menunjukkan sentimen netral lebih dominan dibandingkan komentar positif dan negatif. Metode klasifikasi SMOTE *naïve bayes* terbukti mampu meningkatkan nilai *balanced accuracy*, AUC *score,* dan G-mean. Hasil

menunjukkan klasifikasi sentimen menggunakan *naïve bayes* dengan SMOTE memberikan efektifitas yang baik dalam permasalahan *imbalance* dimana kelas netral lebih dominan. Hasil penerapan metode *naïve bayes* dengan teknik SMOTE dalam mengklasifikasikan data komentar YouTube terhadap Ibu Kota Negara Nusantara menjadi kelas negatif, netral, dan positif dengan menggunakan perbandingan data *train* dan data *test* sebesar 90%:10% diperoleh hasil klasifikasi sentimen sebanyak 266 komentar diklasifikasikan dengan benar yang terdiri dari 17 komentar negatif, 213 komentar netral, dan 29 komentar positif dari total data sebanyak 324 komentar pada data *test* Pada penelitian selanjutnya diharapkan dapat menambang data dari lebih banyak sumber video terkait topik dan menggunakan pustaka *google-translation* agar kata-kata dalam bahasa asing dapat terbobot dengan lebih baik.

REFERENSI

Barro, R. A., Sulvianti, I. D., & Afendi, F. M. (2013). Penerapan synthetic minority oversampling technique (SMOTE) terhadap data tidak seimbang pada pembuatan model komposisi jamu. *Xplore*, *1*(1). https://doi.org/10.29244/xplore.v1i1.12424

Chawla, N., Bowyer, K., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *ArXiv,* abs/1106.1813

Destuardi, I., & Sumpeno, S. (2009). Klasifikasi emosi untuk teks bahasa Indonesia menggunakan metode naive bayes. Surabaya: Institut Teknologi Sepuluh Nopember.

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*(1), 13–21. https://doi.org/10.1016/j.knosys.2011.06.013

Indrawati, A., Subagyo, H., Sihombing, A., Wagiyah, W., & Afandi, S. (2020). Analyzing the impact of a resampling method for imbalanced data text in Indonesian scientific articles categorization. *Jurnal Dokumentasi dan Informasi, 41*, 133-141.

Joachims, T. (1997). A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning. Guide Proceedings* (pp. 143–151). https://doi.org/10.5555/645526.657278

Kurnia, K., Purnamasari, I., & Saputra, D. D. (2023). Analisis sentimen dengan metode naïve bayes, SMOTE dan adaboost pada Twitter Bank BTN. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *7*(2), 235–242. https://doi.org/10.35870/jtik.v7i3.707

LP2M UMA. (2022). Analisis sentimen (sentiment analysis): definisi, tipe dan cara kerjanya. Retrieved from https://lp2m.uma.ac.id/2022/02/21/analisis-sentimen-sentiment-analysis-definisi-tipe-dan-cara-kerjanya/

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *International Journal of Computer Science and Network (IJCSN), 2*(1).

Magnolia, C., Nurhopipah, A., & Kusuma, B. A. (2023). Penanganan imbalanced dataset untuk klasifikasi komentar program kampus merdeka pada aplikasi twitter. *Edu Komputika Journal*, *9*(2), 105–113. https://doi.org/10.15294/edukomputika.v9i2.61854

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval. Choice Reviews Online*, *46*(05), 46–2715. https://doi.org/10.5860/choice.46-2715

Novendri, R., Callista, A. S., Pratama, D. N., & Puspita, C. E. (2020). Sentiment analysis of youtube movie trailer comments using naïve Bayes. *Bulletin of Computer Science and Electrical Engineering*, *1*(1), 26-32.

Nugroho, D. (2022). Bentuk dan kekhususan Ibu Kota Negara Nusantara dalam Negara Kesatuan Republik Indonesia. *The Indonesian Journal Of Politics And Policy (IJPP)*, *4*(1), 53–62. https://doi.org/10.35706/ijpp.v4i1.6527

Profil Penajam Paser Utara, Lokasi Ibu Kota Baru di Kalimantan Timur – DPRD KABUPATEN PENAJAM PASER UTARA. (2021, June 14). Diakses dari https://dprd.penajamkab.go.id/2021/06/14/profil-penajam-paser-utara-lokasi-ibu-kota-baru-di-kalimantan-timur/

Rahman, A., Rahmat, F., Fariqi, M. Y., & Adi, S. (2020). Metode Naïve Bayes untuk menganalisis akurasi sentimen komentar di YouTube. *Jurnal EECCIS. 14*(1), 31-34.

Raschka, S. (2014). Naive bayes and text classification I - Introduction and theory. *arXiv (Cornell University)*. Diakses dari https://arxiv.org/pdf/1410.5329v2

Setiawan, V. D. (2023). Kombinasi data augmentasi dan skema term weighting untuk analisis sentimen [Master's thesis, Universitas Gadjah Mada]

Sharma, S. & Singh, D. (2018). Study of sentiment classification techniques. *International Journal of Computer Sciences and Engineering, 6*(5), 479–783.

Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi algoritma synthetic minority over-sampling technique untuk menangani ketidakseimbangan kelas pada dataset klasifikasi. *Jurnal Sistem Informasi*, *10*(2), 445. https://doi.org/10.32520/stmsi.v10i2.1303

Sulistiyowati, N., & Jajuli, M. (2020). Integrasi naive bayes dengan teknik sampling SMOTE untuk menangani data tidak seimbang. *Nuansa Informatika: Jurnal Penelitian Dan Teknologi Informasi*, *14*(1), 34. https://doi.org/10.25134/nuansa.v14i1.2411

Wikarsa, L., Angdresey, A., & Kapantow, J. (2022). Implementasi metode naïve bayes dan lexicon-based approach untuk mengklasifikasi sentimen netizen pada tweet berbahasa Indonesia. *Jurnal Ilmiah Realtech*, *18*(1), 15–24. https://doi.org/10.52159/realtech.v18i1.5