5th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
7 September 2024 – Jakarta, Indonesia
Volume 3, Nomor 2, September 2024 - ISSN 2962-8628 (*online*)

IMPLEMENTASI TEXT MINING PADA ANALISIS SENTIMEN OPINI MASYARAKAT TERHADAP PEMINDAHAN IBUKOTA KE IKN NUSANTARA

Ramandhanu Yuchnan Utomo[1], Arief Wibowo[2*]

[1,2] Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1]2011500333@student.budiluhur.ac.id, [2*]arief.wibowo@budiluhur.ac.id
(* : corresponding author)

Abstrak-Urgensi untuk memindahkan ibukota negara Indonesia ke luar pulau Jawa sudah dicetuskan oleh Presiden Soekarno sejak tahun 1957. Ide tersebut baru dapat direalisasikan pada masa pemerintahan Presiden Jokowi dan sudah berada dalam masa pembangunan sejak tahun 2020 dan diperkirakan akan selesai pada tahun 2025. Beragam sentimen masyarakat bermunculan terhadap isu tersebut termasuk pengguna media sosial *Twitter*. *Twitter* merupakan salah satu platform media sosial yagn memungkinkan penggunanya untuk melakukan kegiatan interaksi dengan pengguna lain secara daring. Pada penelitian ini, Twitter juga akan menjadi sumber dari data yang akan dikumpulkan. Penelitian ini akan menganalisis sentimen masyarakat dengan memberi label kelas sentimen menjadi tiga kategori yaitu positif, negatif dan netral. Dari hasil yang plelabelan tersebut, akan diimplementasikan *text mining* dengan algoritma *Multinomial Naïve Bayes*. Proses tersebut akan menghasilkan prediksi label dari dokumen selanjutnya. Data cuitan yang berhasil dikumpulkan berjumlah 2146 data mentah yang kemudian dilakukan pelabelan otomatis dengan memanfaatkan kamus *postive* dan *negative*, adapun proses pelabelan alternatif yang dilakukan pada penelitian ini yang hasil pelabelannya sudah divalidasi oleh seorang pakar. Setelah semua data sudah mempunyai label, data dibersihkan dengan proses *preprocessing* yang mencakup proses *case folding*, *cleansing*, tokenisasi, normalisasi, *stopwords*, *stemming* dan memperoleh 1040 data bersih. Data yang sudah bersih tersebut akan dibelah menjadi dua dan diperoleh data latih dengan perbandingan 80% sebanyak 832 data, dan data uji dengan perbandingan 20% sebanyak 208 data. Setelahnya akan dilakukan pembobotan menggunakan TF-IDF pada data latih serta menghitung nilai probabilitas priors setiap kelas pada data latih. Nilai TF-IDF dan nilai probabilitas priors akan diimplementasikan untuk penghitungan nilai likelihood untuk setiap kata terhadap setiap kelas pada data uji. Dengan menggunakan aplikasi yang dikembangkan pada penelitian ini, diperoleh data hasil klasifikasi sebanyak 23 data positif, 115 data negatif, dan 70 data netral. Serta akan diperoleh hasil *confusion matrix* dengan nilai akurasi sebesar 57,21%.

Kata Kunci: analisis sentimen, *multinomial naïve bayes*, probabilitas posterior.

IMPLEMENTATION OF TEXT MINING IN THE ANALYSIS OF PUBLIC OPINION SENTIMENT TOWARDS THE MOVING OF THE CAPITAL TO IKN NUSANTARA

Abstract-*The urgency to move Indonesia's capital outside Java was initiated by President Soekarno in 1957. This idea was only realized during President Jokowi's administration and has been under construction since 2020 and is expected to be completed in 2025. Various public sentiments have emerged. regarding this issue including Twitter social media users. Twitter is a social media platform that allows its users to interact with other users online. In this research, Twitter will also be the source of the data that will be collected. This research will analyze public sentiment by labeling sentiment classes into three categories, namely positive, negative and neutral. From the labeling results, text mining will be implemented with the Multinomial Naïve Bayes algorithm. This process will produce a predicted label for the next document. The tweet data that was collected amounted to 2146 raw data which was then automatically labeled using positive and negative dictionaries. There is an alternative labeling process carried out in this research where the labeling results have been validated by an expert. After all the data has labels, the data is cleaned using a preprocessing process which includes case folding, cleansing, tokenization, normalization, stop words, stemming and obtaining 1040 clean data. The clean data will be split into two and obtain training data with a ratio of 80%, totaling 832 data, and test data with a comparison of 20%, totaling 208 data. After that, weighting will be carried out using TF-IDF on the training data and calculating the prior probability values for each class on the training data. The TF-IDF value and prior probability values will be implemented to calculate the likelihood value for each word for each class in the test data. By using the application developed in this research, the classification data obtained was 23 positive data, 115 negative data and 70 neutral data. And you will get confusion matrix results with an accuracy value of 57.21%.*

Keywords*: sentiment analysis, multinomial naïve bayes, posterior probability*

1. PENDAHULUAN

Text mining merupakan suatu metde penggalian informasi atau data yang berupa teks yang bersumber dari kumpulan dokumen dengan menggunakan suatu alat atau program analisis, contoh *text mining* seperti klasifikasi data, klasterisasi data, dan ekstraksi informasi[1]. *Text mining* digunakan untuk memecahkan permasalahan informasi dengan menggunakan teknik dari *data mining*, *machine learning*, *Natural Language Processing* (NLP), *Information Retrieval* (IR), dan *knowledge management*. Pada kebanyakan kasus, *text mining* biasanya melibatkan proses *preprocessing*, ektraksi *dataset*, penyimpanan representasi perantara data, pengelompokkan data seperti analisis tren, dan visualisasi hasil[2].

Analisis sentimen atau dapat juga disebut sebagai *opinion mining* adalah salah satu metode dari studi komputasi yang dilakukan dengan mengestrak data opini, memahami dan memproses data tekstual secara otomatis untuk mengidentifikasi sentimen yang terkandung dalam sebuah opini. Analisis sentimen juga dilakukan untuk menentukan apakah opini atau komentar terhadap suatu permasalahan atau isu terkait topik tertentu memiliki kecenderungan positif, negatif, atau netral dan dapat dijadikan sebagai acuan dalam meningkatkan suatu pelayanan, ataupun meningkatkan suatu kualitas produk[3].

Pemindahan ibu kota Indonesia ke luar Pulau Jawa, pertama dicetuskan oleh Presiden Soekarno pada 1957 dengan Kota Palangkaraya sebagai pilihan, kini direalisasikan sebagai IKN Nusantara di masa pemerintahan Presiden Joko Widodo, yang pembangunannya dimulai pada 2020 dan diperkirakan selesai pada 2025. Dengan bebasnya masyarakat beropini di media sosial seperti Twitter, pastinya akan muncul berbagai macam opini terkait topik pemindahan ibukota tersebut. Analisis sentimen, adalah metode untuk memprediksi sentimen dalam cuitan Twitter menjadi positif, negatif, dan netral. Untuk itu, penelitian ini akan mengembangkan sebuah sistem yang dapat memprediksi sentimen masyarakat bedasarkan komentar yang didapat dari cuitan di Twitter dengan memanfaatkan algoritma Multinomial Naive Bayes. penelitian ini diharapkan dapat memberikan informasi dan pandangan kepada masyarakat lain terhadap isu pemindahan ibukota Indonesia ke IKN Nusantara.

Algoritma multinomial naïve bayes adalah sebuah metode *supervised learning* yang biasanya digunakan untuk mengklasifikasi teks. Algoritma ini bekerja dengan konsep *term frequency*, yang berarti berapa kali suatu kata muncul dalam sebuah dokumen. Ada dua fakta yang dapat dijelaskan dalam model ini, yaitu apakah suatu kata muncul atau tidak dalam sebuah dokumen serta frekuensi kemunculannya dalam dokumen tersebut[4]. Karena algoritma ini menggunakan adalah salah satu algoritma yang bermetode supervised learning, setiap data yang akan diklasifikasi harus mempunyai label. Penelitian ini menggunakan metode pelabelan *Lexicon Based*, yaitu metode pelabelan berbasis kamus dan melakukan perbandingan setiap kata. Kata-kata yang terdapat di dalam kamus akan digunakan untuk mengidentifikasi apakah data teks yang kita analisis berisikan opini atau tidak. Setelah diidentifikasi, *machine learning* akan melakukan klasifikasi secara otomatis terhadap teks yang berisikan opini, dengan melihat data latih yang telah dilakukan klasifikasi sebelumnya[5].

Dalam penelitian sebelumnya telah dilakukan analisis sentimen data Twitter terkait pemindahan ibukota negara Indonesia menggunakan algoritma *Naïve Bayes*. Dalam penelitian tersebut, diperoleh total 1200 data yang berisi 654 data ganda dan 546 data yang tidak ganda (unik). Proses pelabelan data dilakukan secara manual dengan menggunakan 2 kategori "Positif" dan "Negatif". Dengan menggunakan rapidminer dan melakukan pengulangan sebanyak 10 kali, proses klasifikasi menggunakan cross validation menghasilkan akurasi 90.48%[6]. Pada penelitian ini, terdapat perbedaan pada jumlah data, jumlah kategori kelas dan metode pelabelan.

Penelitian ini bertujuan untuk mendapatkan evaluasi terkait sentimen masyarakat terhadap topik pemindahan ibukota negara Indonesia ke IKN Nusantara, serta mengetahui hasil akurasi dari model algoritma *Multinomial Naïve Bayes*.

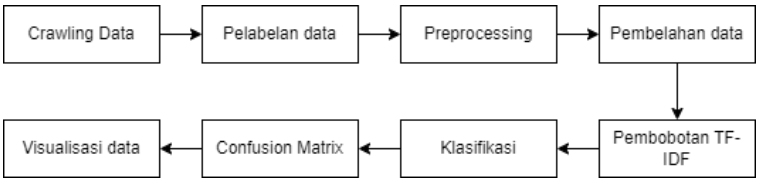
2. METODE PENELITIAN.

2.1 Data Penelitian

Pada penelitian ini, data yang digunakan bersumber dari media sosial *Twitter* (sekarang bernama X). Data tersebut dikumpulkan dengan cara *crawling* dengan memanfaatkan pustaka *tweet-harvest*. Kata kunci yang digunakan pada proses ini adalah "ibukota", "ikn", "IKN", "ibukota nusantara", dan "pindah ibukota". Proses pengumpulan data dilakukan dalam rentang waktu 4 hari, yaitu pada tanggal 5 April 2024 sampai tanggal 8 April 2024.

2.2 Penerapan Metode

Dalam penerapan metode dan pengembangan aplikasi analisis sentimen untuk menganalisis opini masyarakat tentang pemindahan ibukota Indonesia ke IKN Nusantara ini terdapat beberapa tahapan proses dilakukan sesuai Gambar 1.



Gambar 1. Penerapan Metode

2.2.1 *Crawling* Data

Crawling data merupakan suatu untuk mengumpulkan dan megindeks data yang bisa didapatkan dari berbagai sumber seperti dokumen, situs web, media sosial, dan database[7]. Pada tahap *crawling* data, dilakukan proses pengumpulan *dataset* yang nantinya akan digunakan dalam penelitian ini. Proses ini dilakukan dengan memanfaatkan pustaka yang bernama *tweet-harvest*.

2.2.2 Pelabelan Data

Dalam proses ini pelabelan akan dilakukan secara otomatis dengan menggunakan metode *Lexicon Based Labelling*. Pada proses ini akan memanfaatkan dua kamus yaitu kamus *positive* dan kamus *negative*. Kamus tersebut berisi kumpulan kata yang masing-masingnya memiliki bobot berupa angka. Kamus ini akan menjadi dasar pengkategorian label pada *dataset* mentah. Bahasa dalam teks yang mengandung kalimat atau argumen positif seperti dukungan atau pujian akan mengandung bobot positif atau lebih dari 0 dan dinyatakan sebagai data positif. Sedangkan bahasa dalam teks yang mengandung kalimat atau argumen negatif seperti menentang, makian atau cacian dan mengandung bobot negatif atau kurang dari 0 dinyatakan sebagai data negatif. Karena ada beberapa data yang mempunyai bobot 0, label dari data tersebut akan diidentifikasi sebagai netral.

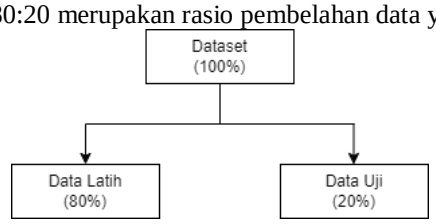
2.2.3 *Preprocessing*

Tahap *preprocessing* merupakan proses pembersihan data supaya data akan menjadi lebih bersih dan terstruktur. Proses ini mencakup penghapusan karakter dan istilah yang tidak berarti dan tidak perlu ada di data. Proses ini mencakup beberapa tahap yang masing-masingnya dilakukan sesuai dengan konten yang didapat dari proses pelabelan data.

a. *Case Folding*: Dalam proses *case folding* dilakukan penyetaraan kata pada kata yang berhuruf besar ataupun kapital. Kata tersebut akan diubah ke dalam bentuk huruf kecil (*lowercase*) secara keseluruhannya. Misalnya kata ‘IKN’ menjadi ‘ikn’, ‘Indonesia’ menjadi ‘indonesia’.
b. *Cleansing:* Dalam proses cleansing akan dilakukan penyaringan kata pada teks dengan cara membuang karakter selain a sampai Z atau komponen-komponen yang tidak memiliki hubungan dengan konten yang ada pada data. Proses cleansing memiliki beberapa tahapan yaitu: menghapus karakter selain huruf seperti *emoticon* dan simbol, menghapus *mention* (@...), menghapus URL (https://t.co/...), menghapus *hashtag*, dan menghapus spasi berlebih. Pada akhir proses *cleansing* juga dilakukan penghapusan data duplikat.
c. Tokenisasi: Tokenisasi atau tokenisasi adalah proses membagi teks menjadi potongan-potongan kata yang disebut sebagai token yang tiap tokennya akan dipisahkan oleh tanda baca koma atau spasi. Dalam bentuk token, teks akan lebih mudah diolah dan dianalisis.
d. Normalisasi: Normalisasi adalah proses yang bertujuan untuk memperbaiki kata-kata yang kurang tepat seperti kata singkatan, kata tidak baku, ataupun kesalahan pengetikan (*typo*)[8]. Misalnya ‘mmg’ menjadi ‘memang’, ‘bocah’ menjadi ‘anak’. Pada penelitian ini, proses normalisasi bergantung pada suatu kamus yang berisi kumpulan kata tidak baku atau kesalahan pengetikan beserta versi kata bakunya.
e. *Stopwords*: *Stopwords* adalah proses penghapusan kata yang kurang mempunyai makna dan hubungan pada teks tetapi banyak dijumpai. Beberapa kata yang dihapus misalnya: ‘lah’, ‘yang’, ‘wkwk’, ‘dong’. Pada penelitian ini, proses *stopwords* bergantung pada pustaka pustaka *nltk* dari *python*.
f. *Stemming*: Dalam proses stemming dilakukan penguraian kata-kata yang berimbuhan menjadi kata dasarnya. Proses stemming dilakukan dengan bergantung pada pustaka dari *python* yaitu *Sastrawi stemmer factory*. Misalnya adalah ‘berdamai’ menjadi ‘damai’, ‘menyelesaikan’ menjadi ‘selesai’.

2.2.4 Pembelahan Data

Pada Gambar 2 tahapan pembelahan data atau *data splitting* ini data yang sudah mempunyai label dari proses pelabelan otomatis akan dibelah menjadi data latih dan data uji yang dengan 80% data latih dan 20% data uji. Referensi [9] menunjukkan bahwa 80:20 merupakan rasio pembelahan data yang optimal dan biasa digunakan.



Gambar 2. Pembelahan Data

2.2.5 *Pembobotan* TF-IDF

Dalam analisis sentimen, penting untuk melakukan pembobotan kata yang disebut dengan *Term Weighting* dengan persamaan (1), yang mana dalam penelitian ini mencakup *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF merupakan proses yang dapat menghitung seberapa penting suatu kata dalam suatu kalimat atau dokumen dalam suatu kumpulan dokumen. TF memiliki fungsi untuk memonitor kata terhadap seberapa sering kemunculannya di dalam teks atau dokumen. IDF merupakan metode untuk mengukur keunikan suatu kata yang muncul dalam dokumen tersebut, pada penelitian ini dilakukan pendekatan *smoothing* yaitu dilakukan penambahan nilai satu pada hasil IDF supaya mencegah *underflow*[7].

$$W(c,t) = TF_{ct} \times IDF_t \tag{1}$$

2.2.6 *Klasifikasi Multinomial Naïve Bayes*

Tahapan klasifikasi menggunakan algoritma *Multinomial Naïve Bayes* untuk analisis sentimen menghitung peluang terhadap kelas tertentu dan peluang setiap kata pada dokumen terhadap suatu kelas, yang artinya algoritma ini digunakan untuk memprediksi kelas pada suatu dokumen. Tahap ini akan menghasilkan keluaran berupa prediksi kelas positif, negatif, dan netral. Referensi [10] menyatakan bahwa sebelum beralih ke penghitungan *Multinomial Naïve Bayes* untuk memprediksi suatu dokumen, ada beberapa tahapan yang harus dilakukan yaitu menghitung nilai *priors* dan *likelihood*. *Priors* merupakan nilai probabilitas kemunculan dari setiap kelas pada data latih. Formula *priors* adalah persamaan (2):

$$P(C_i) = \frac{N_c}{N_{doc}} \tag{2}$$

Sedangkan *likelihood* merupakan suatu probabilitas yang mengidentifikasi kata tertentu yang muncul pada dokumen baru. Formula *likelihood* adalah persamaan (3):

$$P(W_i, C_i) = \frac{W_{ct} + 1}{(\sum W'€VW'ct) + B'} \tag{3}$$

Setelah mendapatkan nilai kemunculan kata pada setiap dokumen, untuk memprediksi kelas dapat dilakukan perhitungan probabilitas *posterior* yaitu peluang kategori kelas i terdapat kata I, dengan persamaan (4) [11]:

$$Pr(Y|X1, X2, \dots, Xn) = Pr(Y)\prod(Xi|Y) \tag{4}$$

Yang formulanya akan disederhanakan dalam penelitian ini menjadi persamaan (5):

$$P(C_i, W_i) = P(C_i) \times P(W_1, C_i) \times P(W_2, C_i) \times \dots \times P(W_n, C_i) \tag{5}$$

2.2.7 *Confusion Matrix*

Setelah mendapatkan hasil prediksi label dari proses klasifikasi, selanjutnya adalah menghitung confusion matrix beserta akurasi. *Confusion matrix* merupakan suatu metode yang digunakan untuk menghitung kinerja atau tingkat kebenaran akurasi proses klasifikasi. Karena penelitian ini menggunakan 3 kelas label, persamaan yang digunakan adalah persamaan (6):

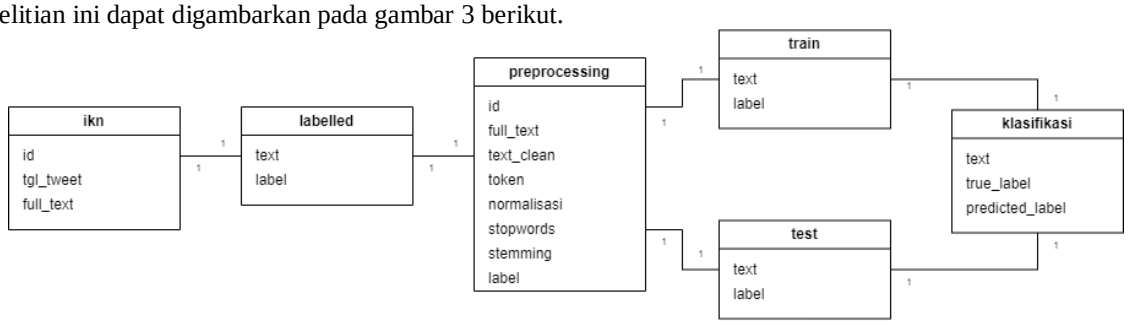
$$Akurasi = \frac{TP + TN + TNet}{Total\ keseluruhan\ prediksi} \times 100\% \tag{6}$$

2.2.8 Visualisasi Data

Pada tahap akhir yaitu visualisasi, akan ditampilkan rangkuman data yang didapat dari hasil tahapan sebelumnya. Hasil yang akan ditampilkan antara lain perbandingan label positif, negatif, dan netral sebelum dan sesudah proses klasifikasi yang berupa grafik batang, serta *wordcloud* data positif dan *wordcloud* data negatif.

2.3 Rancangan Basis Data

Penelitian ini memanfaatkan MySQL untuk menyimpan segala data yang diproses, rancangan basis data dalam penelitian ini dapat digambarkan pada gambar 3 berikut.



Gambar 3. Rancangan Basis Data

3. HASIL DAN PEMBAHASAN

3.1 Implementasi Metode

Pada sebuah sistem, tentu diperlukan langkah-langkah yang harus diikuti agar sistem tersebut dapat dijalankan dengan baik tanpa kendala. Alur implementasi metode penelitian ini yaitu:

3.1.1 Pengumpulan *Dataset*

Tahap pengumpulan *dataset* pada penelitian ini dilakukan dengan metode *crawling* yang bersumber dari media sosial *Twitter*. Proses *crawling* sendiri dilakukan dengan memanfaatkan pustaka yang bernama *tweet-harvest*. Penggunaan *tweet-harvest* memerlukan token yang bernama *auth-token* yang didapat dari *cookies Twitter* web. Kata kunci yang digunakan untuk *crawling* data pada tahap ini adalah: “ikn”, “IKN”, “ibukota”, “ibukota nusantara”, “pindah ibukota” dan berhasil mengumpulkan data sebanyak 2147 baris yang disimpan dalam format csv sesuai Gambar 4.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	1.78E+18	Fri Apr 05 1:	0	@TxtdariHI	1.78E+18		TxtdariHI	in	????	0	0	0	https://twit	1.06E+18	DittoRadhitya	
2	1.78E+18	Fri Apr 05 1:	0	Prabowo-Gi	1.78E+18			in	Balikpapan,	0	0	0	https://twit	103551225	tribunkaltim	
3	1.78E+18	Fri Apr 05 1(	0	Terungkap f	1.78E+18			in	Balikpapan,	0	0	0	https://twit	103551225	tribunkaltim	
4	1.78E+18	Fri Apr 05 1(	0	Sebagai tam	1.78E+18		inang_hasih	in		0	1	0	https://twit	8.03E+17	inang_hasiholan	
5	1.78E+18	Fri Apr 05 1(	0	Rencananya	1.78E+18		inang_hasih	in		0	1	0	https://twit	8.03E+17	inang_hasiholan	
6	1.78E+18	Fri Apr 05 1(	0	@prabowo	1.78E+18		itspragibtim	in		0	1	0	https://twit	1.71E+18	itspragibtime	
7	1.78E+18	Fri Apr 05 1(	0	@prabowo	1.78E+18		itspragibtim	in		0	1	0	https://twit	1.71E+18	itspragibtime	
8	1.78E+18	Fri Apr 05 1(	0	Wah ini terr	1.78E+18	https://pbs.twimg.com/		in		1	1	0	https://twit	1.71E+18	itspragibtime	
9	1.78E+18	Fri Apr 05 0{	0	Sebagai tam	1.78E+18		Rizwariryb	in	Senayan Jak	0	1	0	https://twit	2.429E+09	Rizwariryb	
10	1.78E+18	Fri Apr 05 0{	0	Rencananya	1.78E+18		Rizwariryb	in	Senayan Jak	0	1	0	https://twit	2.429E+09	Rizwariryb	
11	1.78E+18	Fri Apr 05 0{	0	Progres lanc	1.78E+18	https://pbs.twimg.com/		in	East Bornec	0	0	0	https://twit	312852705	Nomahbirashn	
12	1.78E+18	Fri Apr 05 0{	0	OIKN Komit	1.78E+18	https://pbs.twimg.com/		in	East Bornec	0	0	0	https://twit	312852705	Nomahbirashn	
13	1.78E+18	Fri Apr 05 0{	0	Walaupun U	1.78E+18		revmen_id	in	Indonesia	0	1	0	https://twit	3.681E+09	revmen_id	
14	1.78E+18	Fri Apr 05 0:	0	OIKN Komit	1.78E+18	https://pbs.twimg.com/		in	East Bornec	0	0	0	https://twit	1.61E+18	VanessaPadmana7	
15	1.78E+18	Fri Apr 05 0:	0	Progres lanc	1.78E+18	https://pbs.twimg.com/		in	East Bornec	0	0	0	https://twit	1.61E+18	VanessaPadmana7	
16	1.78E+18	Fri Apr 05 0:	0	Progres lanc	1.78E+18	https://pbs.twimg.com/		in	East Bornec	0	0	0	https://twit	1.50E+18	BarronIndra	
17	1.78E+18	Fri Apr 05 0:	0	OIKN Komit	1.78E+18	https://pbs.twimg.com/		in	East Bornec	0	0	0	https://twit	1.50E+18	BarronIndra	
18	1.78E+18	Fri Apr 05 0:	0	OIKN Komit	1.78E+18	https://pbs.twimg.com/		in		0	0	0	https://twit	1.56E+18	Inggridv112	
19	1.78E+18	Fri Apr 05 0:	0	Progres lanc	1.78E+18	https://pbs.twimg.com/		in		0	0	0	https://twit	1.56E+18	Inggridv112	
20	1.78E+18	Fri Apr 05 0:	0	Warga sete	1.78E+18	https://pbs.twimg.com/		in		0	0	0	https://twit	1.56E+18	Inggridv112	
21	1.78E+18	Fri Apr 05 0:	1	#IKN Nusan	1.78E+18	https://pbs.twimg.com/		in	South Suma	0	0	0	https://twit	901441285	Vigo__alfindo	
22	1.78E+18	Fri Apr 05 0(	0	Gus Yahya n	1.78E+18	https://pbs.	SatuNusant	in	Sumatera Se	0	0	0	https://twit	1.56E+18	SatuNusantara3	
23	1.78E+18	Fri Apr 05 0(	0	Semua haru	1.78E+18	https://pbs.twimg.com/		in	Kota Paleml	0	0	0	https://twit	9.70E+17	AhmadAhlan10	
24	1.78E+18	Fri Apr 05 0(	0	Semua haru	1.78E+18	https://pbs.twimg.com/		in	Kota Paleml	0	0	0	https://twit	9.70E+17	AhmadAhlan10	
25	1.78E+18	Fri Apr 05 0(	0	Semua haru	1.78E+18	https://pbs.twimg.com/		in	Kota Paleml	0	0	0	https://twit	9.70E+17	AhmadAhlan10	
26	1.78E+18	Fri Apr 05 0(	0	Tokoh adat	1.78E+18	https://pbs.	SatuNusant	in	Sumatera Se	0	1	0	https://twit	1.56E+18	SatuNusantara3	
27	1.78E+18	Fri Apr 05 0(	0	Pembangun	1.78E+18	https://pbs.twimg.com/		in		0	0	0	https://twit	1.07E+18	PolitikLingkar	
28	1.78E+18	Fri Apr 05 0(	0	IKN Nusanta	1.78E+18	https://pbs.	Renatasama	in	Kota Batam	0	1	0	https://twit	1.41E+18	Renatasamasa	
29	1.78E+18	Fri Apr 05 0(	0	IKN Nusanta	1.78E+18	https://pbs.twimg.com/		in	Kota Batam	0	1	0	https://twit	1.41E+18	Renatasamasa	
30	1.78E+18	Fri Apr 05 0(	0	IKN Nusanta	1.78E+18	https://pbs.twimg.com/		in	Sumatera Se	0	1	0	https://twit	1.56E+18	SatuNusantara3	
31	1.78E+18	Fri Apr 05 0(	0	Pembangun	1.78E+18	https://pbs.	gungaia	in	South Suma	0	0	0	https://twit	1.40E+18	gungaia	

Gambar 4. *Dataset* Hasil *Crawling*

3.1.2 Pelabelan Data

Pada Penelitian ini, tahap pelabelan dilakukan dengan 2 metode yaitu pelabelan otomatis dan *import* file dengan data berlabel. Dalam pelabelan otomatis, akan dilakukan dengan memanfaatkan 1 kamus *positive* dan 1 kamus *negative*. Kamus tersebut berisi 3609 kata positif dan 6609 kata negatif beserta bobotnya. Setelah membaca isi dari kamus, pengidentifikasian sentimen dari teks sudah bisa dilakukan. Untuk mengidentifikasi sentimen, perlu dilakukan perhitungan skor atau bobot dari setiap teks pada setiap baris data. Tahap ini dilakukan dengan cara membandingkan setiap kata dalam teks dengan kamus positif dan negatif. Proses ini akan menghasilkan bobot positif dan negatif dari setiap teks pada setiap baris data. Pada persamaan (7), Kedua bobot tersebut akan dijumlahkan untuk menghitung skor total dan menentukan sentimen dari teks tersebut.

$$Skor = \ Skor\ Positif + Skor\ Negatif \tag{7}$$

Jika sudah dilakukan perhitungan skor, selanjutnya dilakukan pengidentifikasian label sentimen menggunakan formula persamaan (8):

$$\begin{matrix} Skor > 0, sentimen\ adalah\ Positif \\ Skor < 0, sentimen\ adalah\ Negatif \\ Skor = 0, sentimen\ adalah\ Netral \end{matrix} \tag{8}$$

Data yang sudah mempunyai label sentimen juga sudah divalidasi oleh seorang pakar yang membuat proses pelabelan data otomatis menjadi lebih kredibel.

3.1.3 Preprocessing

Tahap pembersihan teks atau *preprocessing* merupakan tahap dimana dilakukan proses pembersihan terhadap *dataset* mentah yang berhasil dikumpulkan agar data dapat lebih terorganisir dan lebih mudah diproses pada tahap selanjutnya. Pada penelitian ini, *dataset* mentah yang sudah dikumpulkan dan disimpan ke dalam basis data akan mengalami proses pengolahan yaitu integrasi, *cleaning*, tranformasi dan reduksi data. Proses pembersihan teks ini mempunyai beberapa tahapan yaitu: *case folding*, *cleansing*, *tokenisasi*, *normalisasi*, *stopwords*, *stemming*. Proses pembersihan teks ini menghasilkan teks yang lebih terstruktur dan terorganisir agar dapat diolah, serta pengurangan jumlah baris yang awalnya 2147 baris menjadi 1040 baris.

3.1.4 Pembelahan Data

Proses pembelahan data dilakukan bersamaan dengan proses *preprocessing*. Total baris pada *dataset* berjumah 1040 baris, lalu dibagi menjadi data latih dan data uji dengan rasio 80:20 dengan nilai *random state* 30. Proses ini dilakukan menggunakan *stratified sampling* untuk memastikan distribusi label pada data uji dan data latih menjadi lebih seimbang. Proses pembelahan data ini menghasilkan 832 baris data latih dan 208 baris data uji pada Tabel 1.

Tabel 1. Pembelahan Data

Data	Jumlah Baris
Dataset bersih	1040 baris (100%)
Data latih	832 baris (80%)
Data uji	208 baris (20%)

3.1.5 Klasifikasi *Multinomial Naïve Bayes*

Setelah dilakukan proses *preprocessing*, pelabelan otomatis dan pembelahan data, proses selanjutnya adalah analisis sentimen. Tahap pertama yang harus dilakukan dalam melakukan proses ini adalah dengan mengambil sampel data latih Tabel 2 dari *dataset* bersih, sedangkan data uji Tabel 3 merupakan contoh data *dummy* yang hanya akan digunakan pada pengujian kali ini.

Tabel 2. Sampel Data Latih

Text	Label
tokoh adat dukung ikn	Positif
sedih warga jakarta dukung pindah ibukota	Netral
mari dukung bangun ikn nusantara	Positif
maklum bro dukung hilang lapak demo ibukota pindah ikn narasi	Negatif

Tabel 3. Sampel Data Uji

Text	Label
warga sedih ibukota jakarta pindah ikn	?

Dengan adanya data latih dan data uji, selanjutnya adalah melakukan pembobotan TF-IDF untuk setiap kata yang ada di dalam dokumen data yangdapat dilihat pada Tabel 4.

Tabel 4. TF-IDF

Kata	TF-IDF			
	D1	D2	D3	D4
Tokoh	0.40075	0	0	0
Adat	0.40075	0	0	0
Dukung	0,25	0	0,2	0,1
Ikn	0.28125	0	0,225	0,1
Sedih	0	0.26722	0	0
Warga	0	0.26722	0	0
Jakarta	0	0.26722	0	0
Pindah	0	0,21687	0	0,1301
Ibukota	0	0,21687	0	0,1301
Mari	0	0	0,3206	0
Bangun	0	0	0,3206	0
Nusantara	0	0	0,3206	0
Maklum	0	0	0	0,1603
Bro	0	0	0	0,1603
Hilang	0	0	0	0,1603
Lapak	0	0	0	0,1603
Demo	0	0	0	0,1603
Narasi	0	0	0	0,1603

Setelah diperoleh nilai TF-IDF dari setiap kata dari setiap dokumen, selanjutnya menghitung nilai probabilitas *priors* dari setiap kelas yang ada pada data latih yaitu positif, negatif dan netral, nilai *likelihood* dari setiap kata dari masing-masing kelas pada data uji, serta nilai probabilitas *posterior* dari setiap kelas. Nilai probabilitas *priors* dihitung menggunakan persamaan (2), yaitu dengan membagi kemunculan suatu kelas dari total yang ada pada data latih. Nilai *likelihood* dihitung menggunakan persamaan (3). Nilai probabilitas *posterior* dihitung dengan menggunakan persamaan (5), yaitu dengan mengkalikan semua nilai probabilitas *priors* dan nilai likelihood semua kata dari masing-masing kelas. Hasil perhitungan dapat dilihat pada Tabel 5.

Tabel 5. Probabilitas Posterior

Kelas	Kata	*Priors*	*Likelihood*	Probabilitas *Posterior*
Positif	Warga	2/4 = 0,5	0,0333	$0{,}5 \times 0{,}0333 \times 0{,}0333 \times 0{,}0333 \times 0{,}0333 \times 0{,}0333 \times 0{,}0502 = 1.02776752e-9$
	Sedih		0,0333	
	Ibukota		0,0333	
	Jakarta		0,0333	
	Pindah		0,0333	
	Ikn		0,0502	
Netral	Warga	1/4 = 0,25	0,0528	$0{,}25 \times 0{,}0528 \times 0{,}0528 \times 0{,}0507 \times 0{,}0528 \times 0{,}0507 \times 0{,}0416 = 4.09804751e-9$
	Sedih		0,0528	
	Ibukota		0,0507	
	Jakarta		0,0528	
	Pindah		0,0507	
	Ikn		0,0416	
Negatif	Warga	1/4 = 0,25	0,0357	$0{,}25 \times 0{,}0357 \times 0{,}0357 \times 0{,}0403 \times 0{,}0357 \times 0{,}0403 \times 0{,}0397 = 7.3340735e-10$
	Sedih		0,0357	
	Ibukota		0,0403	
	Jakarta		0,0357	
	Pindah		0,0403	
	Ikn		0,0397	

Setelah diperoleh nilai probabilitas *posterior* dari setiap kelas, akan dilakukan perbandingan antara nilai-nilai tersebut untuk menentukan kelas mana yang mempunyai nilai probabilitas *posterior* terbesar. Berdasarkan hasil di atas, kelas netral mempunyai nilai probabilitas *posterior* terbesar, maka dapat dipastikan kelas yang diprediksi oleh model pada data uji adalah "Netral".

3.1.6 *Confusion Matrix*

Dengan menggunakan data penelitian yang asli dan bukan data dummy yang berisi sebanyak 832 data latih dan 208 data uji dalam proses klasifikasi *multinomial naïve bayes*, diperoleh kelas prediksi dan kelas actual Tabel 6. Data dari kelas prediksi dan kelas aktual akan dimasukkan ke dalam tabel *confusion matrix* untuk dievaluasi serta dihitung akurasi dari model yang sudah dibuat.

Tabel 6. Nilai *Confusion Matrix*

Confusion Matrix		**Prediksi**		
		Negatif	Netral	Positif
Aktual	Negatif	58	16	1
	Netral	27	44	2
	Positif	21	22	17

Tabel 6 *confusion matrix* di atas merepresentasikan label prediksi dan label aktual berupa matrix. Dari tabel tersebut dapat dilakukan perhitungan akurasi dengan persamaan (6):

$$Akurasi = \frac{58 + 44 + 17}{208} \times 100\% = 57{,}21\%$$

Dari proses perhitungan dengan persamaan (6) di atas, diperoleh nilai akurasi dari model yang sudah dibuat pada penelitian ini adalah sebesar 57,21%.

4. KESIMPULAN

Penggunaan bahasa pemrograman *python* dan framework *Flask* dan database MySQL dapat digunakan untuk mengembangkan suatu aplikasi web analisis sentimen yang menerapkan algoritma *multinomial naïve bayes* yang memanfaatkan nilai probabilitas *priors*, nilai *likelihood*, nilai probabilitas *posterior*. Dari 208 data uji yang digunakan pada penelitian ini, diperoleh data prediksi yang dominan bersentimen "Negatif", dengan hasil yang diperoleh: 20 data positif (9,6%), 82 data netral (39,4%), 106 data negatif (50,9%). Model yang digunakan pada penelitian ini untuk pengujian analisis sentimen menggunakan algoritma *multinomial naïve bayes* menghasikan *confusion matrix* dan diperoleh nilai akurasi sebesar 57,21 %,

DAFTAR PUSTAKA

[1] S. Supriyatna, E. Fahrudin, S. Informasi, F. I. Komputer, U. Pamulang, and T. Selatan, "Pemanfaatan Algoritma Text Mining Dalam Pengetahuan Kebencanaan Dari Dokumen Kajian," vol. 2, no. 1, pp. 35–42, 2024.

[2] A. N. Nurkalyisah, A. Triayudi, and I. D. Sholihati, "Analisis Sentimen pada Twitter Berbahasa Indonesia Terhadap Penurunan Performa Layanan Indihome dan Telkomsel," *J. Sist. dan Teknol. Inf.*, vol. 10, no. 4, p. 387, 2022, doi: 10.26418/justin.v10i4.50858.

[3] Y. Nooryuda Prasetya, D. Winarso, and Syahril, "Penerapan Lexicon Based Untuk Analisis Sentimen Pada TwiterTerhadap Isu Covid-19," *J. Fasilkom*, vol. 11, no. 2, pp. 97–103, 2021.

[4] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 4, pp. 820–826, 2021, doi: 10.29207/resti.v5i4.3146.

[5] N. Habibah, E. Budianita, M. Fikry, and I. Iskandar, "Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan Lexicon Based dan K-Nearest Neighbor," *J. Ris. Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, [Online]. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

[6] R. K. Septiani, S. Anggraeni, and S. D. Saraswati, "Klasifikasi Sentimen Terhadap Ibu Kota Nusantara (IKN) pada Media Sosial Menggunakan Naive Bayes," *J. Tek.*, vol. 16, no. 2, pp. 245–254, 2022.

[7] F. A. Ramadhan, S. H. Sitorus, and T. Rismawan, "Penerapan Metode Multinomial Naïve Bayes untuk Klasifikasi Judul Berita Clickbait dengan Term Frequency - Inverse Document Frequency," *J. Sist. dan Teknol. Inf.*, vol. 11, no. 1, p. 70, 2023, doi: 10.26418/justin.v11i1.57452.

[8] A. C. Kamilla, N. Priyani, R. Priskila, and V. H. Pranatawijaya, "Analisis Sentimen Film Agak Laen Dengan Kecerdasan Buatan : Text Mining Metode Naïve Bayes Classifier," vol. 8, no. 3, pp. 2923–2928, 2024.

[9] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.

[10] L. Lesmana, Mukrodin, and F. Nabyla, "Analisis Sentimen Pengguna Twitter PPDB Menggunakan Algoritma

Multinominal Naive Bayes," *J. Sist. Inf. dan Teknol. Perad.*, vol. 1, no. 1, 2020, [Online]. Available: https://journal.peradaban.ac.id/index.php/jsitp/article/view/604

[11] A. Padilah, H. Perdana, and S. A. Intisari, "Implementasi Metode Naïve Bayes Dalam Prediksi Tingkat Kemenangan Pada Game Mobile Legends," *Bul. Ilm. Math. Stat. dan Ter.*, vol. 13, no. 4, pp. 437–446, 2024.