JURNAL SISFOTENIKA
Vol. 15 No. 2, Juli 2025
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/sisfotenika.v15i1.553

Analisis Sentimen Program Tabungan Perumahan Rakyat Menggunakan Metode Naïve Bayes

Sentiment Analysis of the Public Housing Savings Program Using the Naïve Bayes Method

Novi Amaliah[1], Rudi Kurniawan[2], Tati Suprapti[3]
[1,3]Program Studi Teknik Informatika, STMIK IKMI Cirebon, Cirebon
[2]Program Studi Rekayasa Perangkat Lunak, STMIK IKMI Cirebon, Cirebon
e-mail: ***[1]noviamaliah005@gmail.com**, [2] rudi226ikmi@gmail.com ,
[3]tatisuprapti112004@gmail.com

Abstrak

Tabungan Perumahan Rakyat (Tapera) adalah program pemerintah yang mewajibkan pekerja berpenghasilan sebesar upah minimum untuk menyisihkan 3% dari gajinya untuk iuran kepada BP Tapera. Program ini memicu beragam tanggapan masyarakat yang diungkapkan melalui media sosial, khususnya Twitter. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap Tapera dan menentukan rasio dari training dan testing data yang menghasilkan nilai akurasi terbaik menggunakan algoritma Naïve Bayes. Sebanyak 3.284 tweet dikumpulkan melalui crawling dari Twitter dengan Tweet-Harvest, kemudian diproses dengan tahap pre-processing, lalu dilabeli secara manual dengan bantuan ahli ke dalam sentimen positif, netral, dan negatif, dan dilakukan oversampling agar data seimbang. Kemudian, visualisasi data dan pengujian model Naïve Bayes dengan tiga rasio yaitu 70:30, 80:20, dan 90:10. Hasil penelitian menunjukkan bahwa pembagian data 80:20 memberikan kinerja terbaik, dengan akurasi sebesar 87%, precision 87%, recall 87%, dan f1-score 87%. Dengan demikian, rasio 80:20 lebih optimal karena memberikan keseimbangan jumlah data latih dan data uji sehingga model lebih andal dan generalisasi lebih baik. Dengan begitu, model ini berpotensi membantu instansi dalam memantau tren kepuasan dan keluhan pelanggan secara lebih efektif dan real-time. Penelitian ini menegaskan pentingnya distribusi data yang seimbang dan pemilihan rasio pembagian data yang tepat dalam meningkatkan performa model Naïve Bayes pada analisis sentimen.

Kata kunci— *Analisis Sentimen, Tabungan Perumahan Rakyat, Naive Bayes,*

Abstract

The Public Housing Savings (Tapera) is a government program that requires workers earning the minimum wage to set aside 3% of their salary for contributions to BP Tapera. This program triggers a variety of public responses that are expressed through social media, especially Twitter. This study aims to analyze public sentiment towards Tapera and determine the ratio of training and testing data that produces the best accuracy value using the Naïve Bayes algorithm. A total of 3,284 tweets were collected through crawling from Twitter with Tweet-Harvest, then processed with the pre-processing stage, then manually labeled with the

help of experts into positive, neutral, and negative sentiments, and oversampling is done so that the data is balanced. Then, data visualization and Naïve Bayes model testing with three ratios, namely 70:30, 80:20, and 90:10. The results showed that the 80:20 data division gave the best performance, with an accuracy of 87%, precision 87%, recall 87%, and f1-score 87%. Thus, the 80:20 ratio is more optimal because it provides a balance in the amount of training data and test data so that the model is more reliable and generalizes better. Thus, this model has the potential to help agencies in monitoring trends in customer satisfaction and complaints more effectively and in real-time. This research confirms the importance of a balanced data distribution and the selection of an appropriate data sharing ratio in improving the performance of Naïve Bayes models in sentiment analysis.

Keywords— *Sentiment Analysis, Public Housing Savings, Naïve Bayes*

1. PENDAHULUAN

Pada era sekarang, kemajuan teknologi telah mengubah cara manusia untuk menerima dan menyampaikan informasi atau opini, salah satunya melalui media sosial. Media sosial memungkinkan informasi tersebar secara cepat dan menjangkau berbagai lapisan masyarakat. Salah satu media sosial yang paling populer adalah media sosial X atau lebih dikenal dengan Twitter [1]. Twitter adalah media sosial yang membuat penggunanya dapat mengirim opini secara bebas dengan cara mengunggah pesan singkat atau tweet dalam berbagai bentuk seperti teks, foto, audio, video, dan sebagainya [2]. Hal tersebut menjadikan Twitter sebagai salah satu sumber data yang berlimpah, dimana terdapat banyak sentimen publik terhadap isu-isu terkini dapat dikumpulkan dan dianalisis dengan menggunakan teknik *text mining* atau analisis sentimen.

Analisis sentimen adalah teknik yang digunakan untuk mengekstrak dan menilai opini, evaluasi, serta perasaan yang terkandung di dalam teks, yang dapat diterapkan dalam berbagai aspek, seperti ekonomi, isu politik, kepuasan produk atau layanan, dan sebagainya [3], [4]. Analisis sentimen juga merupakan teknik yang digunakan untuk mengidentifikasi polaritas (positif, negatif dan netral) dari tweet tertentu, menggunakan pendekatan *machine learning* atau *lexicon sentiment* [5].

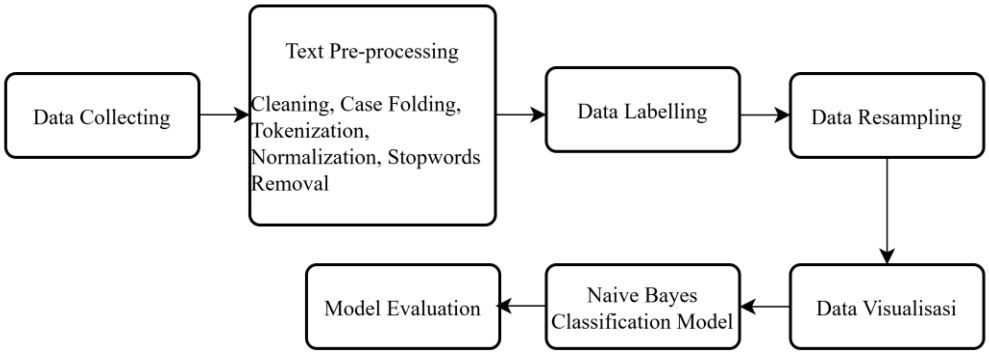
Salah satu isu yang hangat dibicarakan oleh pengguna Twitter yaitu mengenai program Tabungan Perumahan Rakyat atau Tapera. Tapera adalah penyimpanan yang dilakukan oleh peserta secara periodik dalam jangka waktu tertentu yang hanya dimanfaatkan untuk pembiayaan perumahan dan/atau dikembalikan berikut hasil pemupukannya setelah kepesertaan berakhir [6]. Atau dengan kata lain program Tapera didirikan untuk menghimpun dan menyediakan dana murah jangka panjang yang berkelanjutan untuk pembiayaan perumahan bagi masyarakat. Namun, program tersebut menimbulkan pro dan kontra dari masyarakat. Perbedaan pendapat tersebut menjadikan analisis sentimen terhadap program Tapera menjadi relevan. Analisis sentimen menghadapi berbagai tantangan, di antaranya penentuan rasio data pelatihan dan data pengujian yang optimal untuk memperoleh akurasi terbaik dalam klasifikasi sentimen [7]. Selain itu, penggunaan bahasa informal, bahasa gaul, serta singkatan oleh pengguna Twitter, yang menyulitkan algoritma analisis sentimen dalam memahami bahasa tidak konvensional secara akurat [8].

Penelitian ini menggunakan metode Naïve Bayes karena kemampuannya dalam mengklasifikasikan dokumen dengan sederhana dan efektif, metode ini terbukti efisien pada berbagai masalah klasifikasi sentimen [9]. Penelitian terdahulu terkait analisis sentimen dengan Naïve Bayes oleh [10] mendapatkan akurasi sebesar 80% dengan menggunakan data testing sebanyak 20%. Lalu penelitian oleh [11] menggunakan K-Means untuk klastering data latih dan Naïve Bayes untuk klasifikasi data testing diperoleh akurasi rata-rata 93,35% dan error rate sebesar 6,66%. Selain itu, penelitian oleh [12] mengenai penggunaan emotikon untuk mengidentifikasi sentimen, dengan hasil akurasi sebesar 96,3%.

Berdasarkan penjelasan di atas, penelitian ini bertujuan untuk melakukan analisis sentimen mengenai Tapera dengan menggunakan Naïve Bayes. Penelitian serupa yang secara khusus mengangkat program Tapera dan mengeksplorasi rasio optimal antara data pelatihan dan data pengujian masih terbatas. Oleh karena itu, penelitian ini diharapkan dapat mengisi kesenjangan tersebut dan memberikan kontribusi dalam pengembangan model analisis sentimen yang lebih akurat, serta membantu memahami respons publik terhadap kebijakan pemerintah secara lebih mendalam.

2. METODE PENELITIAN

Penelitian ini melewati beberapa tahapan yang dimulai dari tahap *data collecting* dengan *crawling data* dari Twitter. Tahap berikutnya adalah *text pre-processing*, lalu *data labelling* yang dilakukan secara manual dengan bantuan seorang ahli bahasa Indonesia. Setelah itu, dilakukan *resampling* agar data pada kelas sentimen seimbang, lalu masuk ke tahap visualisasi. Selanjutnya, model dibangun dengan metode Naïve Bayes dan terakhir model dievaluasi untuk mengukur kinerja model menggunakan confusion matrix. Metode penelitian dapat dilihat pada Gambar 1.



Gambar 1. Metode Penelitian

2.1 Data Collecting

Tahap pengumpulan data dilakukan dengan teknik *crawling* dari media sosial Twitter menggunakan *Tweet-Harvest* yang dikembangkan oleh Helmi Satria. Proses *crawling* dilakukan terhadap tweet yang dipublikasikan pada periode Mei hingga Oktober 2024 dan mengandung kata kunci “Tapera”, “tapera”, dan “Tabungan Perumahan Rakyat” serta dari *threads* mengenai Tabungan Perumahan Rakyat. Sebanyak 3.284 tweet berhasil dikumpulkan dan disimpan dalam format CSV.

2. 2 Text Pre-processing

Text pre-processing memegang peranan penting dalam teknik dan aplikasi *text mining*, serta merupakan langkah awal yang wajib dilakukan dalam proses *text mining* [13]. *Text Pre-processing* digunakan untuk mengolah data mentah menjadi data yang siap dianalisis lebih lanjut. Tahap ini diperlukan karena data biasanya tidak lengkap dan kompleks, sehingga memerlukan penanganan khusus [14]. Tahap *text pre-processing* pada penelitian ini akan dimulai dari *cleaning, case folding, tokenization, normalization,* dan *stopwords removal*.

2. 3 Data Labelling

Setiap kalimat akan diberikan label dengan tiga kategori sentimen, yaitu positif, netral dan negatif. Proses *labelling* dilakukan secara manual dengan dibantu oleh seorang ahli. Kalimat dengan sentimen positif umumnya mengandung kata kerja dan kata sifat yang bermakna positif, serta menunjukkan ekspresi seperti kepuasan, persetujuan, dukungan, apresiasi, atau pujian. Sementara itu, kalimat dengan sentimen negatif biasanya memuat kata

kerja dan kata sifat yang bermakna negatif, serta mencerminkan ekspresi seperti ketidakpuasan, kecurigaan, ketidaksetujuan, kritik, atau sikap meremehkan.

Adapun kalimat dengan sentimen netral tidak mengandung ekspresi emosional yang kuat, baik positif maupun negatif. Kalimat netral cenderung hanya menyampaikan fakta tanpa opini, dapat berupa pertanyaan tanpa muatan emosi, tidak mengandung kata-kata bernada positif atau negatif, serta tidak menunjukkan dukungan maupun ketidaksetujuan. Kalimat netral juga dapat berupa saran atau usulan yang disampaikan secara objektif

2.4 Data Resampling

Data resampling bertujuan untuk memastikan data pada kelas sentimen terdistribusi secara adil atau seimbang. Ketidakseimbangan data terjadi pada saat suatu kelas atau kategori tertentu memiliki data yang lebih banyak dibandingkan dengan kategori yang lainnya.

2.5 Data Visualisasi

Selanjutnya yaitu visualisasi data dengan *word cloud* untuk melihat kata-kata yang sering muncul pada sentimen. Semakin besar ukuran font pada *word cloud* maka semakin sering juga topik tersebut dibicarakan.

2.6 Naïve Bayes Classification Model

Algoritma Naïve Bayes adalah salah satu metode analisis statistik yang menggunakan probabilitas Bayesian untuk memproses data numerik [15]. Penelitian ini menggunakan salah satu model Naïve Bayes yang sering digunakan dalam klasifikasi teks yaitu Multinomial Naïve Bayes (MultinomialNB). Pengklasifikasian kelas dari suatu dokumen pada MultinomialNB tidak hanya ditentukan berdasarkan pada jumlah kata yang terdapat pada dokumen tetapi juga ditentukan oleh frekuensi kemunculan kata tersebut [16]. Notasi klasifikasi Naïve Bayes ditampilkan pada Persamaan (1).

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

Keterangan dari Persamaan (1), yaitu :

A	: Hipotesis data merupakan suatu class spesifik.
B	: Data dengan kelas yang masih belum diketahui.
P(A|B)	: Probabilitas hipotesis berdasarkan kondisi.
P(A)	: Probabilitas hipotesis.
P(B|A)	: Probabilitas berdasarkan kondisi pada hipotesis.
P(B)	: Probabilitas B

Sebelum membangun model klasifikasi, dilakukan pembagian data dan pembobotan kata dengan TF-IDF terlebih dahulu. Pembagian data dilakukan untuk membagi data menjadi dua kategori, yaitu *training data* (data latih) dan *testing data* (data uji). Pada penelitian ini akan menggunakan tiga rasio pembagian data, yaitu 70:30, 80:20, dan 90:10. Pemilihan ketiga rasio ini bertujuan untuk mengeksplorasi pengaruh proporsi data latih dan data uji terhadap performa model, khususnya dalam mencapai akurasi terbaik dalam klasifikasi sentimen. selanjutnya akan dilakukan pembobotan kata dengan TF-IDF. Tahap pembobotan kata dilakukan untuk menghitung setiap kata yang muncul pada data berdasarkan pada *term frequency*. *Term Frequency* digunakan untuk menghitung frekuensi kemunculan kata dari suatu bobot kata (*term*) dalam dokumen. *Inverse Document Frequency* adalah perhitungan bagaimana *term* disebarkan secara luas dalam dokumen [7].

2.7 Model Evaluation

Tahap terakhir, yaitu evaluasi kinerja model, yang digunakan untuk mengukur seberapa baik model dalam memprediksi data baru dengan menggunakan confusion matrix. Confusion matrix ditunjukkan pada Tabel 1.

Tabel 1. Confusion Matrix

Confusion Matrix		**Kelas Aktual**	
		Positif	**Negatif**
Kelas Prediksi	**Positif**	True Positive(TP)	False Positive(FP)
	Negatif	False Negative (FN)	True Negative (TN)

Keterangan:

a) *True Positive* (TP) yaitu jumlah data kelas positif yang diklasifikasikan sebagai kelas positif.

b) *True Negative* (TN) yaitu jumlah data kelas negatif yang diklasifikasikan sebagai kelas negatif.

c) *False Positive* (FP) yaitu jumlah data kelas negatif yang diklasifikasikan sebagai kelas positif.

d) *False Negative* (FN) yaitu jumlah data kelas positif yang diklasifikasikan sebagai kelas negatif.

Penghitungan kinerja model dapat dilakukan dengan menghitung nilai *accuracy*, *precission*, *recall*, dan *f1-score* berdasarkan rumus yang terlihat pada Persamaan (2), (3), (4), dan (5).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$precission = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4)$$

$$f1 - score = \frac{2 \times precission \times recall}{precission + recall} \quad (5)$$

3. HASIL DAN PEMBAHASAN

3.1 Data Collecting

Dari hasil crawling data Twitter menggunakan Tweet-harvest, diperoleh sebanyak 3.284 data. Detail data hasil crawling dengan Tweet-harvest mengenai Tapera ditampilkan pada Gambar 2. Dan detail kolom pada data ditampilkan pada Tabel 2.

conversat	created_a	favorite_c	full_text	id_str	image_url	in_reply_t	lang	location	quote_cou	reply_cou	retweet_c	tweet_url	user_id_st	username
179535505	Tue May 2	0	Para buzz	179535505	https://pbs.twimg.cc		in	Moskow, I	0	0	0	https://x.c	998800626	BeBuzzerNKRI
179535500	Tue May 2	0	Masih me	1795355007614894541			in		0	0	0	https://x.c	69539974	adhiparama
179535498	Tue May 2	0	Tapera ga	1795354981660516624			in		0	0	0	https://x.c	111783360	gudangfactos
179533262	Tue May 2	0	@Marahlc	1795354963323109 85	Marahlchs		in		0	0	0	https://x.c	141046695	fsaidhamzah
179535493	Tue May 2	0	Tapera ad	179535493	https://pbs.twimg.cc		in	ðŸ‡°ðŸ‡©	0	0	0	https://x.c	5.06E+08	Urrangawak
179529746	Tue May 2	0	@fajarnug	1795354928237719 73	fajarnugrc		in	ThePlaceN	0	0	0	https://x.c	2.57E+09	dveir4u

Gambar 2. Data Hasil Crawling

Tabel 2. Detail Kolom Data

Kolom	Tipe Data	Deskripsi
Conversation_id_str	Int64	ID konversasi unik
created_at	Object	Timestamp pembuatan *tweet*
favorite_count	Int64	Jumlah favorite

full_text	Object	Teks lengkap *tweet*
id_str	Int64	ID *tweet* unik
image_url	Object	URL gambar *tweet*
in_reply_to_screen_name	Object	Nama layar pengguna yang dijawab
lang	Object	Bahas *tweet*
location	Object	Lokasi pengguna
quote_count	Int64	Jumlah kutipan
reply_count	Int64	Jumlah balasan
re*tweet*_count	Int64	Jumlah re*tweet*
*tweet*_url	Object	URL *tweet*
user_id_str	Int64	ID pengguna unik
username	Object	Nama pengguna *tweet*

Data tersebut memiliki 15 kolom, tetapi hanya satu kolom yang digunakan, yaitu kolom full_text karena penelitian ini difokuskan pada penentuan sentimen dalam tweet. Setelah pengumpulan data selesai, selanjutnya akan masuk ke tahap *text pre-processing*.

3.2 Text Pre-processing

Selanjutnya tahap *text pre-processing* yang dimulai dari proses *cleaning data*. Pada proses ini dilakukan penghapusan *missing value*, data duplikat, *mention*, URL, spasi berlebih, non alfa numerik, retweet, dan hashtag. Hasil dari *cleaning data* dapat dilihat pada Tabel 3 yang berisi data mentah sebelum dan sesudah cleaning data.

Tabel 3. Cleaning Data

Sebelum	Sesudah
Para buzzer pembenci Pak Presiden @jokowi lg ngamuk2 soal Tapera pdhal Tapera itu bkn dana yg kemudian akn dimakan Pak Jokowi utk kepentingan pribadinya. Stlh kepesertaan berakhir dana simpanan Tapera akn dikembalikan kpd pesertanya. https://t.co/83X74iczHF	Para buzzer pembenci Pak Presiden lg ngamuk2 soal Tapera pdhal Tapera itu bkn dana yg kemudian akn dimakan Pak Jokowi utk kepentingan pribadinya Stlh kepesertaan berakhir dana simpanan Tapera akn dikembalikan kpd pesertanya

Setelah itu, proses *case folding* yang digunakan untuk mengubah semua karakter dalam data menjadi huruf kecil agar datanya konsisten. Hasil proses *case folding* dapat dilihat pada Tabel 4.

Tabel 4. Case Folding

Sebelum	Sesudah
Para buzzer pembenci Pak Presiden lg ngamuk2 soal Tapera pdhal Tapera itu bkn dana yg kemudian akn dimakan Pak Jokowi utk kepentingan pribadinya Stlh kepesertaan berakhir dana simpanan Tapera akn dikembalikan kpd pesertanya	para buzzer pembenci pak presiden lg ngamuk2 soal tapera pdhal tapera itu bkn dana yg kemudian akn dimakan pak jokowi utk kepentingan pribadinya stlh kepesertaan berakhir dana simpanan tapera akn dikembalikan kpd pesertanya

Setelah semua karakter menjadi huruf kecil, selanjutnya masuk ke proses *tokenization*, yaitu proses mengubah teks menjadi token atau unit lebih kecil, seperti kata atau frasa dengan menghilangkan spasi. Hasil proses *tokenize* dapat dilihat pada Tabel 5.

Tabel 5. Tokenization

Sebelum	Sesudah
para buzzer pembenci pak presiden lg ngamuk2 soal tapera pdhal tapera itu bkn dana yg kemudian akn dimakan	[‘para’, ‘buzzer’, ‘pembenci’, ‘pak’, ‘presiden’, ‘lg’, ‘ngamuk2’, ‘soal’, ‘tapera’, ‘pdhal’, ‘tapera’, ‘itu’, ‘bkn’, ‘dana’, ‘yg’, ‘kemudian’, ‘akn’,

pak jokowi utk kepentingan pribadinya stlh kepesertaan berakhir dana simpanan tapera akn dikembalikan kpd pesertanya	‘dimakan’, ‘pak’, ‘jokowi’, ‘utk’, ‘kepentingan’, ‘pribadinya’, ‘stlh’, ‘kepesertaan’, ‘berakhir’, ‘dana’, ‘simpanan’, ‘tapera’, ‘akn’, ‘dikembalikan’, ‘kpd’, ‘pesertanya’]

Kemudian lakukan normalization untuk mengubah kata-kata ke dalam bentuk yang sesuai dengan KBBI (Kamus Besar Bahasa Indonesia) atau mengubah kata-kata yang disingkat. Misalnya, kata “kpd” menjadi “kepada”, “adlh” menjadi “adalah”, “akn” menjadi “akan”, dan sebagainya. Hasil proses *normalization* dapat dilihat pada Tabel 6.

Tabel 6. Normalization

Sebelum	Sesudah
[‘para’, ‘buzzer’, ‘pembenci’, ‘pak’, ‘presiden’, ‘lg’, ‘ngamuk2’, ‘soal’, ‘tapera’, ‘pdhal’, ‘tapera’, ‘itu’, ‘bkn’, ‘dana’, ‘yg’, ‘kemudian’, ‘akn’, ‘dimakan’, ‘pak’, ‘jokowi’, ‘utk’, ‘kepentingan’, ‘pribadinya’, ‘stlh’, ‘kepesertaan’, ‘berakhir’, ‘dana’, ‘simpanan’, ‘tapera’, ‘akn’, ‘dikembalikan’, ‘kpd’, ‘pesertanya’]	[‘para’, ‘buzzer’, ‘pembenci’, ‘pak’, ‘presiden’, ‘lagi’, ‘ngamuk-ngamuk’, ‘soal’, ‘tapera’, ‘padahal’, ‘tapera’, ‘itu’, ‘bukan’, ‘dana’, ‘yang’, ‘kemudian’, ‘akan’, ‘dimakan’, ‘pak’, ‘jokowi’, ‘untuk’, ‘kepentingan’, ‘pribadinya’, ‘setelah’, ‘kepesertaan’, ‘berakhir’, ‘dana’, ‘simpanan’, ‘tapera’, ‘akan’, ‘dikembalikan’, ‘kepada’, ‘pesertanya’]

Proses selanjutnya yaitu stopwords removal, yang berguna untuk menghilangkan kata-kata yang sering muncul dalam data, tetapi tidak memberikan arti yang signifikan dan tidak berpengaruh secara semantik. Misalnya, kata “dan”, “juga”, “yang”, dan sebagainya. Hasil proses stopwords removal disajikan pada Tabel 7.

Tabel 7. Stopwords Removal

Sebelum	Sesudah
[‘para’, ‘buzzer’, ‘pembenci’, ‘pak’, ‘presiden’, ‘lagi’, ‘ngamuk-ngamuk’, ‘soal’, ‘tapera’, ‘padahal’, ‘tapera’, ‘itu’, ‘bukan’, ‘dana’, ‘yang’, ‘kemudian’, ‘akan’, ‘dimakan’, ‘pak’, ‘jokowi’, ‘untuk’, ‘kepentingan’, ‘pribadinya’, ‘setelah’, ‘kepesertaan’, ‘berakhir’, ‘dana’, ‘simpanan’, ‘tapera’, ‘akan’, ‘dikembalikan’, ‘kepada’, ‘pesertanya’]	buzzer pembenci pak presiden ngamuk-ngamuk soal tapera padahal tapera bukan dana kemudian dimakan pak jokowi kepentingan pribadinya kepesertaan berakhir dana simpanan tapera dikembalikan pesertanya

Setelah tahap *text pre-processing* selesai, jumlah data yang sebelumnya 3.284 berkurang menjadi 2.413 data. Pengurangan data ini terjadi karena beberapa tweet yang kosong, atau tidak lagi relevan setelah melalui proses *cleaning, normalization,* dan *stopwords removal*. Selain itu, tweet duplikat dan tweet yang hanya mengandung simbol, URL, atau karakter tidak bermakna juga dihapus karena tidak memberikan kontribusi terhadap analisis sentimen.

3.3 Data Labelling

Tahap selanjutnya yaitu labelling secara manual, setiap kalimat akan diberi label dengan tiga kategori sentimen, yaitu positif, negatif, dan netral. Tabel 8 memperlihatkan data hasil *labelling*.

Tabel 8. Data Labelling

Data	Sentimen
mending kamu tapera doang lah sama korpri tidak jelas uang diapain mana besar potongannya	Negatif
tapera 671 triliun	Netral

ribut ribut tapera habis ngobrol sama bokap ternyata pegawai negeri sipil ada zaman dulu namanya tabungan wajib perumahan duitnya diambil rumah dicairin kamu bokap kebeli rumah bandung terus sisa juga 26jt buat dicairin	Positif

Setelah dilakukan pelabelan, jumlah data pada setiap kelas sentimen ditampilkan pada Tabel 9.

Tabel 9. Distribusi Kelas Sentimen

Kelas Sentimen	Jumlah Data
Positif	209
Negatif	1596
Netral	607

3.4 Data Resampling

Dari Tabel 9 terlihat bahwa jumlah data pada setiap kelas sentimen tidak seimbang, maka dilakukan proses *resampling* dengan metode *random oversampling* agar data pada setiap kelas sentimen seimbang. *Random oversampling* adalah proses sampling yang dilakukan secara acak untuk meningkatkan jumlah sampel pada kelas minoritas. Hasil distribusi kelas sentimen setelah diresampling ditampilkan pada Tabel 10.

Tabel 10. Hasil Resampling

Kelas Sentimen	Jumlah Data
Positif	1596
Negatif	1596
Netral	1596

3.5 Data Visualisasi

Selanjutnya tahap visualisasi dengan *word cloud* dilakukan dengan bantuan *library* WordCloud. Hasil dari *word cloud* dapat dilihat pada Gambar 3.

Gambar 3. Visualisasi Word Cloud

Dari visualisasi *word cloud* pada Gambar 3. terlihat bahwa kata-kata seperti “Tapera”, “Rakyat”, “Tabungan”, “Perumahan”, dan “Pemerintah” memiliki frekuensi kemunculan yang tinggi. Hal ini mengindikasikan bahwa kata-kata tersebut sering dibicarakan dalam isu-isu terkait Tabungan Perumahan Rakyat (Tapera).

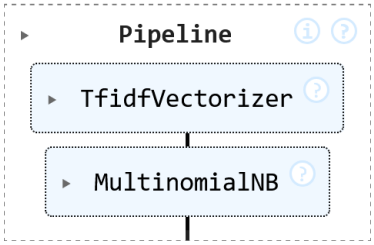
3.6 Naïve Bayes Classification Model

Selanjutnya, proses berlanjut ke tahap klasifikasi dengan Naïve Bayes. Sebelum itu, dataset harus masuk ke proses *splitting data*. Pada penelitian ini dilakukan tiga rasio pembagian *training data* dan *testing data*, yaitu 70:30, 80:20, dan 90:10. Hasil dari proses splitting data dengan tiga rasio diperlihatkan pada Tabel 11.

Tabel 11. Splitting Data

	Rasio Splitting Data		
	70:30	**80:20**	**90:10**
Training Data (X)	3351	3828	4308
Testing Data (Y)	1437	960	480

Setelah data di bagi menjadi training data dan testing data, selanjutnya masuk tahap modelling dengan memanfaatkan teknik pipeline untuk menggabungkan tahap TF-IDF dan tahap klasifikasi dengan MultinomialNB. Gambar 4. menunjukkan representasi visual dari pipeline dengan dua tahapan yang sudah dilakukan, yaitu TfidfVectorizer dan MultinomialNB.



Gambar 4. Pipeline Model

3.7 Model Evaluation

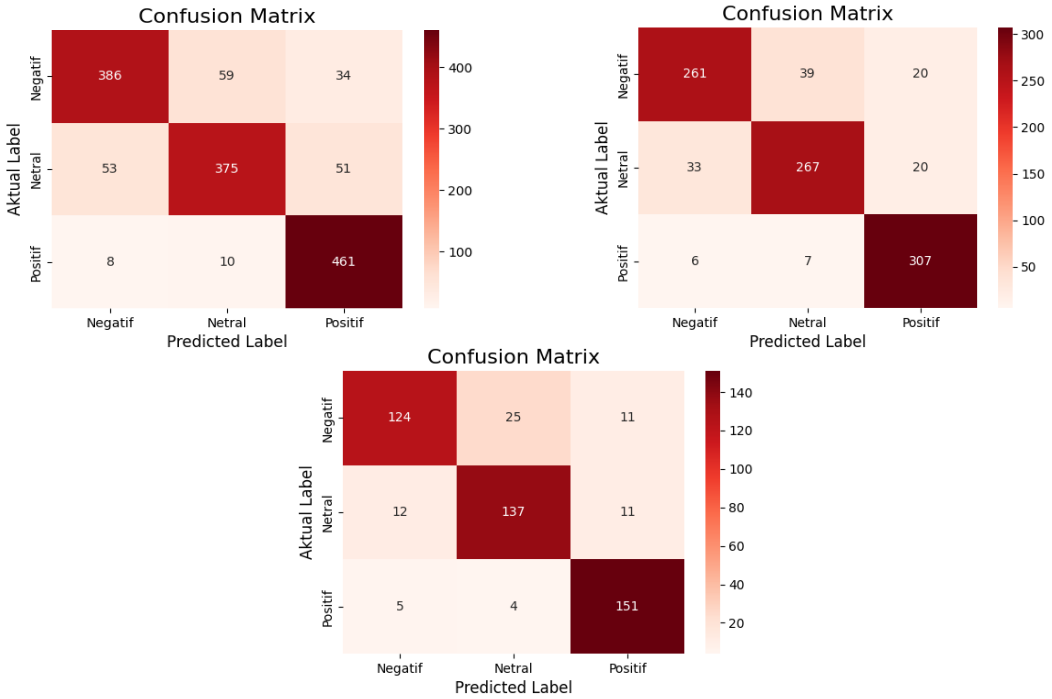
Terakhir, hasil evaluasi kinerja model dengan confusion matrix untuk setiap rasio pembagian data didapatkan nilai *accuracy, precision, recall,* dan *f1-score* yang disajikan pada Tabel 12.

Tabel 12. Hasil Confusion Matrix

	Rasio Splitting Data		
	70:30	**80:20**	**90:10**
Accuracy	85%	87%	86%
Precision	85%	87%	86%
Recall	85%	87%	86%
F1-score	85%	87%	86%

Dari Tabel 12. Menunjukkan bahwa nilai akurasi meningkat seiring dengan bertambahnya rasio data latih, dari 85% pada rasio 70:30 menjadi 87% pada rasio 80:20, kemudian sedikit menurun menjadi 86% pada rasio 90:10. Hal ini mengindikasikan bahwa meskipun penggunaan data latih yang lebih besar umumnya dapat meningkatkan kinerja model, terlalu sedikitnya data uji pada rasio 90:10 dapat menyebabkan evaluasi kurang representatif terhadap performa model secara keseluruhan. Selain itu, nilai metrik seperti precision, recall, dan f1-score yang relatif stabil pada setiap rasio pembagian data menunjukkan bahwa distribusi prediksi model cukup seimbang di antara kelas sentimen. Namun, untuk mendapatkan gambaran yang lebih mendalam mengenai kesalahan klasifikasi, visualisasi confusion matrix untuk ketiga rasio pembagian data ditampilkan pada Gambar 5.

Gambar 5. Hasil Confusion Matrix



Gambar 5 menunjukkan hasil visualisasi confusion matrix untuk ketiga rasio pembagian data, yakni 70:30 pada bagian kiri atas, 80:20 pada bagian kanan atas, dan 90:10 di bagian bawah. Terlihat bahwa model mampu mengenali kelas Positif dengan baik, sedangkan kelas Negatif dan Netral lebih sering tertukar. Dengan demikian, perbaikan model perlu difokuskan untuk mengurangi kesalahan prediksi diantara kedua kelas tersebut.

4. KESIMPULAN

Dari eksperimen yang telah dilakukan terhadap data tweet mengenai program Tapera, diperoleh bahwa pembagian data 80:20 terbukti paling sesuai untuk melatih model Naïve Bayes, dengan capaian akurasi tertinggi sebesar 87%. Hasil ini menunjukkan bahwa keseimbangan jumlah data latih dan data uji berpengaruh signifikan terhadap kemampuan model dalam mengenali pola sentimen secara andal. Dengan pembagian data ini, model lebih mampu mendeteksi perbedaan kelas sentimen dan meminimalkan kesalahan prediksi, khususnya antara kelas Netral dan Negatif. Oleh karena itu, rasio 80:20 direkomendasikan sebagai rasio pembagian data yang optimal untuk memaksimalkan performa sekaligus menjaga validitas evaluasi, sehingga model lebih siap digunakan dalam pemantauan opini publik secara praktis dan efisien.

5. SARAN

Untuk penelitian selanjutnya, disarankan untuk menggunakan metode algoritma *machine learning* yang berbeda seperti SVM, Decision Trees atau bisa juga menggunakan metode berbasis neural network.

DAFTAR PUSTAKA

[1] D. M. Y. Sinurat, D. E. Ratnawati, and D. W. Brata, “Analisis Sentimen Terhadap Kenaikan Cukai Rokok pada Media Sosial Twitter menggunakan Algoritma Naïve Bayes Classifier,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 1, pp. 17–25,

2023.
[2] A. Safira and F. N. Hasan, "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier," *Zo. J. Sist. Inf.*, vol. 5, no. 1, pp. 59–70, 2023, doi: 10.31849/zn.v5i1.12856.
[3] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020.
[4] D. F. Sjoraida, B. W. K. Guna, and D. Yudhakusuma, "Analisis Sentimen Film Dirty Vote Menggunakan BERT (Bidirectional Encoder Representations from Transformers)," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 2, pp. 393–404, 2024, doi: 10.35870/jtik.v8i2.1580.
[5] D. Purnamasari *et al.*, *Pengantar Metode Analisis Sentimen*, 1st ed. Jakarta: Penerbit Gunadarma, 2023.
[6] Peraturan pemerintah RI, "Peraturan Pemerintah Republik Indonesia No.21 Tahun 2024 Tentang Perubahan Atas Peraturan Pemerintah No.25 Tahun 2020 Tentang Penyelenggaraan Tabungan Perumahan Rakyat." Jakarta, 2024.
[7] Z. Firmansyah and N. F. Puspitasari, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Berdasarkan Opini Pada Twitter Menggunakan Algoritma Naive Bayes," *J. Tek. Inform.*, vol. 14, no. 2, pp. 171–178, 2021, [Online]. Available: https://doi.org/10.15408/jti.v14i2.24024
[8] L. Chaudhary, N. Girdhar, D. Sharma, J. Andreu-Perez, A. Doucet, and M. Renz, "A Review of Deep Learning Models for Twitter Sentiment Analysis: Challenges and Opportunities," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 3, pp. 3550–3579, 2024, doi: 10.1109/TCSS.2023.3322002.
[9] J. C. Tesoro, M. J. M. Buen, R. C. S. Jr, and M. V. Aborde, "A Semantic Approach of the Naïve Bayes Classification Algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3287–3294, 2020, doi: 10.30534/ijatcse/2020/125932020.
[10] D. D. Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 10, no. 1, pp. 34–40, 2022, doi: 10.23960/jitet.v10i1.2262.
[11] I. Kurniawan and A. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," *Eksplora Inform.*, vol. 9, no. 1, pp. 1–10, 2019, doi: 10.30864/eksplora.v9i1.237.
[12] K. A. Nugraha, "Analisis Sentimen Berbasis Emoticon pada Komentar Instagram Bahasa Indonesia Menggunakan Naïve Bayes," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 3, pp. 715–721, 2021, doi: 10.28932/jutisi.v7i3.4094.
[13] D. Rifaldi, Abdul Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, 2023, doi: 10.51454/decode.v3i2.131.
[14] Y. A. Wijaya, N. Suarna, Iin, R. Hamonangan, and R. Nining, "Comparison of machine learning algorithm for Santander dataset," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012032, 2021, doi: 10.1088/1757-899x/1088/1/012032.
[15] A. D. Zendrato, S. S. Berutu, Y. P. Sumihar, and H. Budiati, "Pengembangan Model Klasifikasi Sentimen Dengan Pendekatan Vader dan Algoritma Naive Bayes Terhadap Ulasan Aplikasi Indodax," *J. Inf. Syst. Res.*, vol. 5, no. 3, pp. 755–764, 2024, doi: 10.47065/josh.v5i3.5050.
[16] N. Samrin and M. N. Akbar, "Analisis Sentimen Komentar Pengguna Aplikasi Threads Pada Google Play Store Menggunakan Algoritma Multinomial Naïve Bayes," *J. Artif. Intell. Data Sciene*, vol. 3, no. 2, pp. 1–9, 2023.