Analisis Sentimen Komentar Video Youtube Terhadap Visi Indonesia Emas 2045 Menggunakan Metode Naïve Bayes

1st Aldi Yoga Setiawan
Fakultas Rekayasa Industri
Universitas Telkom
Purwokerto, Indonesia
aldiiyg@student.telkomuniversity.ac.id

2nd Sena Wijayanto, S.Pd., M.T
Fakultas Rekayasa Industri
Universitas Telkom
Purwokerto, Indonesia
senawijayanto@telkomuniversity.ac.id

***Abstrak* — Visi Indonesia Emas 2045 merupakan gagasan pemerintah indonesia dalam mengusung momentum 100 tahun kemerdekaan Indonesia. Visi Indonesia Emas 2045 dirancang oleh Kementerian Bappenas dalam Rencana Pembangunan Jangka Panjang Nasional (RPJPN) untuk periode tahun 2025 hingga 2045. Perancangan visi Indonesia Emas 2045 memiliki tujuan mewujudkan indonesia menjadi negara maju dengan peradaban masyarakat modern. Berbagai upaya sosialisasi terkait Visi Indonesia Emas 2045 memunculkan respon masyarakat yang bervariasi untuk menyampaikan pendapat melalui komentar media sosial, seperti *Youtube*. Analisis sentimen pada penelitian ini dilakukan untuk mengklasifikasikan opini masyarakat menjadi tiga kategori sentimen yaitu positif, negatif, dan netral. Penelitian ini menggunakan metode *Naïve Bayes* untuk dapat menganalisis sentimen masyarakat pada komentar video *Youtube* terhadap Visi Indonesia Emas 2045. Pengumpulan data diambil dari 7 konten video *Youtube* diperoleh 5520 komentar. Setelah melalui preprocessing terdapat 5346 data yang digunakan untuk tahap pelabelan dengan kategori sentimen terdiri dari 3644 sentimen positif, 1167 sentimen negatif, dan 535 sentimen netral. Metode *Naïve Bayes* menghasilkan performa akurasi sebesar 79% dengan nilai *precision* sebesar 81%, *recall* sebesar 79%, dan *f1-score* sebesar 80%.**

Kata kunci*— Analisis Sentimen, Visi Indonesia Emas 2045, *Naïve Bayes*, *Youtube*, *Preprocessing

I. PENDAHULUAN

Visi Indonesia Emas 2045 merupakan gagasan yang dirancang dalam menyambut momentum menuju 100 tahun kemerdekaan Indonesia. Visi ini dirumuskan oleh Badan Perencanaan Pembangunan Nasional (Bappenas) dalam Rencana Pembangunan Jangka Panjang Nasional (RPJPN) untuk rentang waktu dari tahun 2025 hingga 2045. Sasaran dari visi ini untuk mempercepat Indonesia bertransformasi menuju peradaban masyarakat modern [1]. Pada Tahun 2045, Indonesia mengusung target menjadi negara maju dengan masyarakat yang kompeten, sejahtera, dan berdaya saing tinggi [2]. Dalam realisasi visi tersebut, sinergi antara masyarakat dan pemerintah menjadi faktor yang penting.

Upaya sosialisasi dari pemerintah terkait Visi Indonesia Emas 2045 memunculkan harapan yang besar untuk masa depan indonesia 20 tahun ke depan, namun dibalik itu terdapat tantangan yang harus dihadapi, hal ini dapat memicu opini masyarakat. Pendapat-pendapat ini menimbulkan beragam perspektif di masyarakat indonesia bagaimana cara mereka memahami dan merespon hal tersebut [3]. Opini yang muncul menggambarkan pandangan masyarakat berupa aspirasi, dukungan, dan kritik terhadap Visi Indonesia Emas 2045 mendatang. Penyampaian pendapat antar masyarakat, saat ini banyak dilakukan di platform digital sebagai media untuk mengekspresikan pendapat, salah satunya melalui *Youtube* [4]. *Youtube* menyediakan kolom komentar yang dapat menjadi ruang diskusi yang interaktif dan informatif. Kemudahan akses dan ketersediaan konten video yang beragam dapat menjadikan menjadikan Youtube sebagai sumber data untuk menganalisis opini publik [5].

Analisis Sentimen merupakan teknik dalam *Natural Language Processing* (*NLP*) yang dimanfaatkan untuk mengidentifikasi, memproses, mengklasifikasikan, dan memahami kata dari sebuah teks. Data komentar pada video *Youtube* dapat dimanfaatkan untuk analisis sentimen untuk memetakan pandangan masyarakat dan mengidentifikasi sentimen positif, negatif, dan netral [6]. Memahami sentimen publik terhadap Visi Indonesia Emas 2045 melalui analisis sentimen menjadi hal yang penting. Reaksi serta opini yang disampaikan masyarakat dapat menjadi informasi yang berguna bagi pemerintah dalam memetakan pandangan masyarakat yang berkembang mengenai visi tersebut. Dengan analisis sentimen dapat mempermudah untuk mengidentifikasi kecenderungan opini masyarakat yang beragam secara sistematis.

Penelitian ini menggunakan metode *Naïve Bayes* dalam melakukan analisis sentimen. *Naïve Bayes* menerapkan pendekatan komputasi yang digunakan untuk mengklasifikasikan sentimen ke dalam kategori positif, negatif, dan netral [7]. Metode ini dapat menjadi pilihan yang sesuai untuk klasifikasi karena metode ini mudah diimplementasikan, memiliki kecepatan pada saat pelatihan model, serta memberikan akurasi yang baik dalam beragam jenis data dan klasifikasi [8]. Penelitian terdahulu yang dilakukan oleh [9] telah menerapkan metode *Naïve Bayes* pada ulasan pengguna aplikasi Jamsostek Mobile di *Google Playstore*, *Naïve Bayes* menghasilkan tingkat akurasi

mencapai 95%. Selain itu, penelitian lainnya oleh [10] telah menganalisis sentimen pada media sosial *Youtube* mengenai respon masyarakat terhadap wacana kenaikan biaya haji pada tahun 2023. Penelitian tersebut menerapkan metode klasifikasi *Naïve Bayes*, *Decision Tree*, dan *Random Forest* dengan memperoleh akurasi *Naïve Bayes* mencapai 90%, *Decision Tree* mencapai akurasi sebesar 83% dan akurasi *Random Forest* mencapai 87%.

Berdasarkan penjelasan tersebut, penelitian ini bertujuan untuk melakukan analisis terhadap distribusi sentimen positif, negatif, dan netral serta mengetahui performa akurasi metode *Naïve Bayes* dalam mengklasifikasikan sentimen. Hasil penelitian ini diharapkan dapat memberikan gambaran berupa informasi sentimen masyarakat terkait Visi Indonesia Emas 2045 melalui penerapan metode *Naïve Bayes*.

II. KAJIAN TEORI

A. Analisis Sentimen

Analisis Sentimen bertujuan untuk melakukan analisis terhadap pendapat serta perasaan dari informasi yang disampaikan dalam bentuk teks [11]. Sentimen yang dianalisis diambil dari banyaknya pendapat atau pandangan publik mengenai suatu topik tertentu. Proses analisis sentimen mencakup klasifikasi teks ke dalam kategori positif, negatif, dan netral [6].

B. Visi Indonesia Emas 2045

Tahun 2045 menjadi momen bersejarah Indonesia, karena menandai tepatnya 100 tahun usia kemerdekaan. Dalam menyongsong momentum tersebut, Indonesia perlu mempersiapkan generasi untuk tahun 2045 yang siap menghadapi tuntutan ilmu pengetahuan, teknologi, dan informasi yang semakin masif di masa mendatang [12]. Tantangan tersebut menyebabkan pemerintah merancang visi besar yang disebut Visi Indonesia Emas 2045.

C. *Youtube*

Youtube telah menjadi salah satu platform media sosial dalam berinteraksi dan berbagi informasi, terutama saat ini masyarakat membutuhkan informasi yang berkualitas dan relevan. Platform ini menyediakan berbagai jenis konten yang dapat diakses oleh pengguna [5]. Aktivitas penayangan video dalam *Youtube* menunjukkan lebih dari satu miliar video yang ditonton setiap harinya [13]. *Youtube* menyediakan fitur kolom komentar untuk setiap videonya, yang telah menjadi sarana bagi pengguna untuk menyampaikan pendapat, reaksi, serta berdiskusi terkait konten video yang ditayangkan.

D. *Naïve Bayes*

Naive Bayes adalah metode klasifikasi yang didasarkan pada Teorema Bayes dengan berasumsi setiap fitur pada data merupakan independen satu sama lain. Metode ini digunakan guna memprediksi label dari suatu data berdasarkan fitur-fitur yang tersedia dengan menghitung kemunculan probabilitas fitur-fitur pada label [14]. Berikut bentuk umum rumus perhitungan *Naïve Bayes*.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Keterangan :
X = Data label yang belum diketahui
H = Hipotesis terhadap data X
P(H|X) = Probabilitas hipotesis H berdasarkan data X.
P(H) = Probabilitas hipotesis H
P(X|H) = Probabilitas data X berdasarkan hipotesis H
P(X) = Probabilitas data X

Multinomial Naïve Bayes yaitu salah satu jenis algoritma Naïve Bayes yang cocok untuk merepresentasikan tingkat kemunculan kata, seperti pada data teks. Penggunaan *Multinomial Naïve Bayes* diperlukan dalam membantu prediksi klasifikasi [15]. Perhitungan Multinomial *Naïve Bayes* dapat diperhatikan pada rumus berikut.

$$P(c) = \frac{N_c}{N_{doc}} \quad (2)$$

Keterangan :
N_c = Jumlah kategori
N_{doc} = Jumlah dokumen

Selanjutnya menghitung probabilitas kata-I ke dalam kategori tertentu.

$$P(w|c) = \frac{count(w_i, c) + 1}{\sum w\ count\ (w, c) + |V|} \quad (3)$$

Penggunaan nilai 1 untuk mencegah hasil probabilitas yang bernilai nol. Kemudian dilakukan tahap klasifikasi data baru berdasarkan perhitungan berikut.

$$P(c, d) = P(c)\prod 1 \le i \le count(v, d)p(w\_i, c) \quad (4)$$

Keterangan :
count(v, d) = Jumlah kata unik pada dokumen

E. *Preprocessing*

Preprocessing adalah tahap pengolahan data teks guna mempersiapkannya agar dapat dianalisis lebih lanjut. Pada tahap ini, dataset diolah menjadi lebih bersih dan optimal. *Preprocessing* mentransformasikan data mentah menjadi data yang siap digunakan [16]. Beberapa tahapan preprocessing diantaranya : *Cleaning*, *Case Folding*, *Normalization*, *Tokenization*, *Stopword Removal*, dan *Stemming*.

F. *Term Frequency-Inverse Document Frequency*

Term Frequency-Inverse Document Frequency (*TF-IDF*) merupakan teknik yang digunakan untuk menghitung sejauh mana kata (*term*) mempunyai keterkaitan dengan dokumen melalui pemberian bobot pada kata. Teknik *TF-IDF* menghitung tingkat kemunculan kata dalam sebuah dokumen (*TF*) serta memperhitungkan seberapa jarang kata tersebut muncul pada seluruh kumpulan dokumen (*IDF*) [17]. Perhitungan TF-IDF dapat dilakukan melalui rumus berikut.

1) Menghitung *TF*

$$Tf_{t,d} = f(d, t) \quad (5)$$

Keterangan :
$Tf_{t,d}$ = *Term Frequency*
f(d,t) = banyaknya kemunculan kata pada dokumen

2) Menghitung *IDF*

$$Idf_{t,d} = \log\frac{D}{df_t} \quad (6)$$

Keterangan :
Idf_t = *Inverse Document-Frequency*
D = jumlah dokumen
df_t = jumlah dokumen yang terdapat kata$_t$

3) Menghitung *TF-IDF*

$$W_{t,d} = Tf_{t,d} \times Idf_t \quad (7)$$

Keterangan :
$W_{t,d}$ = nilai *TF-IDF*
$Tf_{t,d}$ = *Term Frequency*
Idf_t = *Inverse Document-Frequency*

G. *Confusion Matrix*

Confusion Matrix merupakan tahapan dalam menghitung kinerja atau performa akurasi ketika klasifikasi. Pada *Confusion Matrix* divisualisasikan dengan tabel yang menunjukkan jumlah data uji yang berhasil diklasifikasikan dan jumlah data uji yang salah diklasifikasikan [18].

TABEL 1
(CONFUSION MATRIX)

Aktual	**Prediksi**	
	Positif	**Negatif**
Positif	*True Positive*	*False Negative*
Negatif	*False Positive*	*True Negative*

True Positive (TP) berarti data aktual positif diprediksi sebagai nilai positif. *True Negative* (TN) berarti data aktual negatif kemudian diprediksi sebagai nilai negatif. *False Positive* (FP) berarti data aktual negatif diprediksi sebagai nilai positif. *False Negative* (FN) berarti data aktual positif diprediksi sebagai nilai negatif.

Dalam menghitung hasil *Confusion Matrix* menggunakan nilai *accuracy*, *precision*, *recall*, dan *f1-score* [18].

1. *Accuracy*

Accuracy merupakan rasio antara data yang berhasil diklasifikasikan dengan benar terhadap seluruh data hasil klasifikasi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (8)$$

2. *Precision*

Precision merupakan rasio antara jumlah data positif yang berhasil diklasifikasikan dengan benar terhadap total data yang diklasifikasikan positif.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (9)$$

3. *Recall*

Recall adalah rasio data dengan nilai positif yang berhasil diklasifikasikan dengan benar terhadap seluruh data bernilai positif.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (10)$$

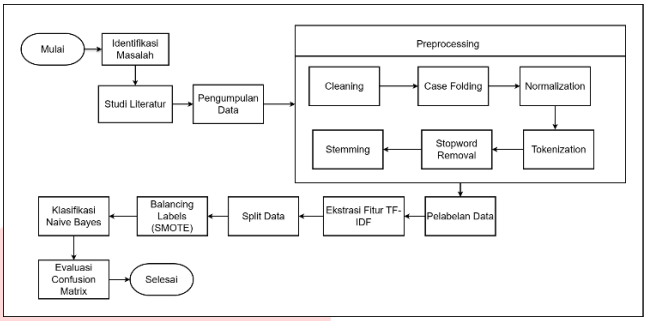
4. *F1-Score*

F1-Score menunjukkan performa model klasifikasi berdasarkan tingkat keseimbangan antara nilai *precision* serta *recall*.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (11)$$

III. METODE

Penelitian ini melakukan analisis sentimen pada opini masyarakat pada *Youtube* terkait Visi Indonesia Emas 2045 menggunakan metode *Naïve Bayes*. Sistematika penelitian yang dilakukan yaitu identifikasi masalah, studi literatur, pengumpulan data, *preprocessing*, pelabelan data, ekstrasi fitur *TF-IDF*, *split* data, *Balancing Labels* (*SMOTE*), Klasifikasi *Naïve Bayes*, dan Evaluasi *Confusion Matrix*.



GAMBAR 1
(SISTEMATIKA PENELITIAN)

Berdasarkan Sistematika Penelitian pada Gambar 1, Tahapan awal penelitian melakukan identifikasi masalah yang menjadi fokus penelitian mengenai komentar pada Youtube mengenai visi Indonesia Emas 2045, kemudian dilanjutkan dengan studi literatur untuk mengumpulkan dan mengkaji referensi ilmiah dari penelitian terdahulu, seperti jurnal yang relevan dengan topik penelitian. Setelah tahapan awal, langkah berikutnya melakukan pengumpulan data berupa *crawling* data yaitu mengakses situs web untuk mengumpulkan informasi yang digunakan sebagai sumber data. Data yang dikumpulkan dari komentar video *Youtube* dibantu dengan *Youtube API* pada *Google Cloud* untuk *crawling* data. Setelah pengumpulan data, kemudian data dilakukan preprocessing untuk mempersiapkan data teks hasil *crawling* yang masih mentah supaya lebih terstruktur dan bersih sehingga siap untuk dilakukan analisis. Tahap ini dengan melakukan *Cleaning*, *Case Folding*, *Normalization*, *Tokenization*, *Stopword Removal*, dan *Stemming*.

Tahap berikutnya melakukan pelabelan data dengan memberikan kategori sentimen terhadap data. Proses pelabelan penelitian ini menggunakan kamus Senticnet, berisi daftar kata beserta nilai bobotnya untuk membantu menentukan sentimen tersebut memiliki kategori positif, negatif, dan netral. Setelah data dilabeli, berikutnya yaitu ekstrasi fitur *TF-IDF* dalam membantu mengukur tingkat relevansi sebuah kata pada dokumen. Selanjutnya tahap split data yaitu pembagian data menjadi data latih (*training* data) serta data uji (*testing* data). Dengan rasio perbandingan 80:20, data latih sebesar 80% dan data uji sebesar 20%. Setelah data dibagi, dilakukan proses penyeimbangan terhadap distribusi data yang timpang melalui teknik *Synthetic Minority Over-Sampling Technique* (*SMOTE*).

Tahap selanjutnya dilakukan tahap klasifikasi pemodelan *Naïve Bayes* dengan menggunakan jenis *Multinomial Naïve Bayes*. Setelah pemodelan, dilakukan pengujian performa *Naïve Bayes* mampu dengan optimal atau tidak optimal divisualisasikan menggunakan *Confusion Matrix*. Parameter pengujian yang terdapat pada *Confusion Matrix* diantaranya *accuracy*, *precision*, *f1-score*, dan *recall*.

IV. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data dikumpulkan dengan melibatkan penggunaan *Youtube API* pada *Google Cloud* menggunakan *API key* dan video id *Youtube* dengan *tools Google Colab* serta Bahasa Pemrograman *Python* untuk melakukan proses Crawling data. Pengumpulan data dengan rentang waktu 26 Oktober 2024 – 4 Mei 2025. Hasil pengumpulan data terdiri dari 7 konten video *Youtube* dan memperoleh 5520 data yang disimpan dalam format *csv*. Hasil pengumpulan data ditampilkan pada Gambar 2.

	publishedAt	authorDisplayName	textDisplay	likeCount
0	2024-09-18T07:41:56Z	@skwrealme1383	Gk nyambung apa yg di bicarakan	1
1	2024-06-06T07:27:55Z	@YulMan-gh8td	Bangsa Indonesia memfunyal kekayaan Indonesia ...	0
2	2024-06-06T07:16:55Z	@YulMan-gh8td	Afakah pemerinta yang sudah dibintingi olih ke...	0
3	2024-06-06T07:04:01Z	@YulMan-gh8td	Assalamualikum warah matullahi wabarakatuh. Ka...	0
4	2024-03-23T18:31:40Z	@apuakchannel	Kok bisa banyak juga yg milih bapak ini ya?	0
...	...	...	...	...
5515	2024-05-13T07:13:22Z	@sumsago	negara maju kayak apa? kayak Amerika Inggris a...	0
5516	2024-05-13T00:53:03Z	@Orangkayanih	Gak perbaiki dulu sistem nya	0
5517	2024-05-12T14:30:16Z	@waluyosk8636	Yg jelas Indonesia akan maju klo di lihat apa ...	1
5518	2024-05-12T12:12:45Z	@aladadarnya	😮😮😮😮😮	0
5519	2024-05-12T11:14:49Z	@YUSRIZALSENSEICH	Ga akan bisa menjadi negara maju, bila para pa...	1

5520 rows × 4 columns

GAMBAR 2
(HASIL PENGUMPULAN DATA)

B. *Preprocessing*

Preprocessing dilakukan untuk mempersiapkan dan meningkatkan konsistensi data supaya dapat dilakukan analisis. Berikut tahapan dalam preprocessing :

1) *Cleaning*

Tahap *cleaning* dilakukan untuk membersihkan komentar dengan menghapus elemen dalam data yang dapat menganggu proses analisis seperti nama pengguna, tanda baca, simbol, emoji, url, angka dan entitas pada HTML. Tabel 2 menampilkan hasil *cleaning*.

TABEL 2
(CLEANING)

Komentar	***Cleaning***
negara maju kayak apa? kayak Amerika Inggris atau kayak china?	negara maju kayak apa kayak Amerika Inggris atau kayak china

2) *Case Folding*

Tahap mengubah seluruh karakter pada komentar yang mengandung huruf besar (*uppercase*) menjadi huruf kecil (*lowercase*). Sebagai contoh, kata “Amerika” dan “Inggris” yang awalnya terdapat huruf kapital diubah menjadi “amerika” dan “inggris” dimana menjadi huruf kecil. Hasil *Case Folding* dapat dilihat pada Tabel 3.

TABEL 3
(CASE FOLDING)

Cleaning	***Case Folding***
negara maju kayak apa kayak Amerika Inggris atau kayak china	negara maju kayak apa kayak amerika inggris atau kayak china

3) *Normalization*

Tahap ini mengoreksi kesalahan kata, mengganti singkatan, dan mengubah kata tidak baku menjadi baku. Sebagai contoh, kata “kayak” dilakukan normalisasi menjadi kata “seperti”. Tabel 4 menunjukkan hasil tahap *normalization*.

TABEL 4
(NORMALIZATION)

Case Folding	***Normalization***
negara maju kayak apa kayak amerika inggris atau kayak china	negara maju seperti apa seperti amerika inggris atau seperti china

4) *Tokenization*

Proses memecah kalimat menjadi kata-kata tunggal dengan menguraikan setiap kalimat dalam data menjadi daftar kata. Hasil *tokenization* ditampilkan pada Tabel 5 berikut.

TABEL 5
(TOKENIZATION)

Normalization	***Tokenization***
negara maju seperti apa seperti amerika inggris atau seperti china	['negara', 'maju', 'seperti', 'apa', 'seperti', 'amerika', 'inggris', 'atau', 'seperti', 'china']

5) *Stopword Removal*

Tahap ini bertujuan untuk menghilangkan kata yang dianggap tidak relevan dan tidak memiliki kontribusi untuk dianalisis. Sebagai contoh, kata “seperti”, “apa”, serta “atau” yang telah dihapus. Tabel 6 merupakan hasil *stopword removal*.

TABEL 6
(STOPWORD REMOVAL)

Tokenization	***Stopword Removal***
['negara', 'maju', 'seperti', 'apa', 'seperti', 'amerika', 'inggris', 'atau', 'seperti', 'china']	negara maju amerika inggris china

6) *Stemming*

Proses dalam mengubah kata menjadi kembali ke bentuk dasar kata tersebut. Berikut hasil *stemming* terdapat pada Tabel 7.

TABEL 7
(STEMMING)

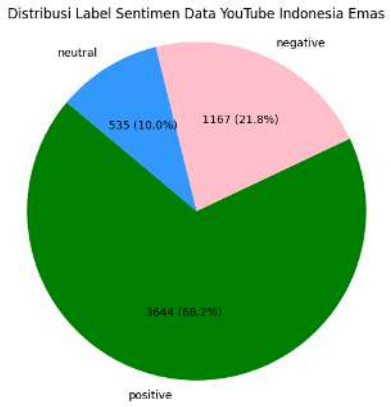
Stopword Removal	***Stemming***
negara maju amerika inggris china	negara maju amerika inggris china

C. Pelabelan Data

Tahap pelabelan data penelitian ini dilakukan untuk menentukan sentimen dari data setelah *preprocessing* dengan kategori positif, negatif, dan netral. Pelabelan data menggunakan kamus Senticnet yang berisi daftar kata beserta nilai bobotnya. Setiap kata dalam komentar dicocokan dengan daftar kata yang terdapat pada kamus Senticnet. Penentuan label melalui pemberian skor dimana positif bernilai lebih dari 0, negatif bernilai kurang dari 0, dan netral bernilai sama dengan 0. Berikut contoh pelabelan data dapat diperhatikan pada Tabel 8.

TABEL 8
(CONTOH PELABELAN DATA)

Hasil *Preprocessing*	**Label**	**Skor *Polarity***
perlu revolusi mental tuju indonesia emas	positif	2.027
indonesia cemas tidak mungkin emas	negatif	-0.909
sudah maklum	netral	0



GAMBAR 3
(DISTRIBUSI PELABELAN DATA)

Pada Gambar 3 menunjukkan hasil distribusi pelabelan data. Hasil pelabelan menggunakan data hasil dari preprocessing didapat sebanyak 5346 data, lalu memperoleh distribusi label sentimen positif sebanyak 3644 (68,2%) data, label sentimen negatif sebanyak 1167 (21.8%) data, dan label sentimen netral sebanyak 535 (10%) data.

D. Ekstrasi Fitur *TF-IDF*

Pada Tahap ini mengukur seberapa penting kata-kata yang muncul dengan memberikan bobot nilai pada kata tersebut dalam setiap dokumen. Tahap ini melakukan inisialisasi *TfidVectorizer* untuk mengubah kumpulan data teks menjadi numerik berdasarkan pentingnya kata dalam dokumen. Inisialisasi dilakukan dengan memasukkan beberapa parameter seperti max_df = 0.7, untuk membatasi kata yang sering muncul maksimal pada 70% dokumen, min_df=4, guna mempertahankan kata-kata yang muncul minimal pada 4 dokumen, max_features=4400, untuk membatasi jumlah kata maksimal sebesar 4400, ngram_range(1, 1), untuk penggunaan unigram, dan sublinear_tf=True, untuk menyesuaikan nilai frekuensi kata-kata yang sering muncul. Berikut contoh dokumen serta perhitungan TF-IDF terdapat pada Tabel 9 dan Tabel 10.

TABEL 9
(CONTOH DOKUMEN TF-IDF)

Dokumen	Komentar *Preprocessing*
1	perlu revolusi mental tuju indonesia emas
2	indonesia cemas tidak mungkin emas
3	sudah maklum

TABEL 10
(PERHITUNGAN TF-IDF)

Term	TF			IDF	TF-IDF		
	D1	D2	D3		D1	D2	D3
perlu	1	0	0	0.477	0.477	0	0
revolusi	1	0	0	0.477	0.477	0	0
mental	1	0	0	0.477	0.477	0	0
tuju	1	0	0	0.477	0.477	0	0
indonesia	1	1	0	0.176	0.176	0.176	0
emas	1	1	0	0.176	0.176	0.176	0
cemas	0	1	0	0.477	0	0.477	0
tidak	0	1	0	0.477	0	0.477	0
mungkin	0	1	0	0.477	0	0.477	0
sudah	0	0	1	0.477	0	0	0.477
maklum	0	0	1	0.477	0	0	0.477

E. *Split* Data

Proses *split* data dilakukan dengan menentukan data yang menjadi fitur (x) dan label (y). Fitur merupakan hasil dari TF-IDF dan label merupakan kategori sentimen positif, negatif, dan netral. Kemudian pembagian data menjadi data latih dan data uji. Perbandingan proporsi split data yaitu 80:20, dengan data latih sebesar 80% dan data uji sebesar 20%. Rasio ini memberikan data yang representatif untuk data latih dan data uji. Hasil distribusi label dapat dilihat pada Tabel 11 dan hasil split data pada Tabel 12.

TABEL 11
(DISTRIBUSI LABEL SPLIT DATA)

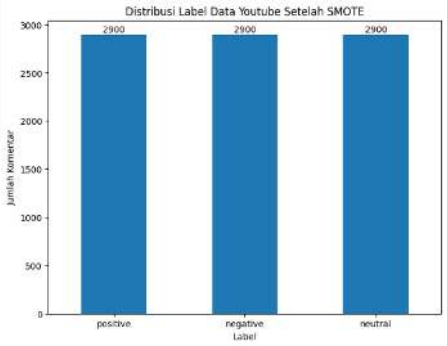
Distribusi Label *Split* Data		
Positif	**Negatif**	**Netral**
2900	941	435

TABEL 12
(SPLIT DATA)

Split Data		
Rasio	**Data Latih**	**Data Uji**
80:20	4276	1070

F. *Balancing Labels* (*SMOTE*)

Tahap *balancing labels* menyeimbangkan adanya data yang tidak seimbang (*imbalanced*) setelah hasil pelabelan. Ketidakseimbangan terjadi ketika data pada satu label misalnya netral dan negatif jauh lebih sedikit datanya dibandingkan positif. Hal ini dapat mempengaruhi performa model dalam menguji kelas yang lebih sedikit (kelas minoritas) diabaikan dan cenderung dengan kelas mayoritas. Untuk mengatasi ketidakseimbangan, diterapkan metode *SMOTE* (*Synthetic Minority Over-Sampling Technique*) yaitu metode untuk *balancing* yang dapat menyamakan jumlah data minoritas (*oversampling*) menjadi seimbang dengan mayoritas dalam data latih pada setiap label. Distribusi label hasil *balancing* didapatkan sebanyak 2900 data untuk setiap labelnya (positif, negatif, dan netral), hal ini menjadikan data label lebih seimbang. Selanjutnya, jumlah data training setelah dilakukan SMOTE sebanyak 8700 bertambah lebih banyak dari yang sebelumnya sebanyak 4276. Sementara jumlah data uji tidak ada yang berubah. Berikut distribusi hasil *SMOTE* pada Gambar 4 dan jumlah hasil *balancing* pada Tabel 13.



GAMBAR 4
(DISTRIBUSI LABEL HASIL SMOTE)

TABEL 13
(DATA HASIL SMOTE)

Data Hasil Balancing SMOTE	
Data Latih	**Data Uji**
8070	1070

G. Klasifikasi *Naïve Bayes*

Implementasi *Naïve Bayes* dengan menggunakan jenis *Multinomial Naïve Bayes*. Jenis ini umum digunakan untuk melakukan klasifikasi data teks. Tahap klasifikasi ini dilakukan setelah dilakukan *split* data dan *balancing SMOTE*, dengan tujuan melatih dan menguji performa label sentimen dari data yang telah melalui *preprocessing*, pelabelan, *split data*, *balancing SMOTE* dan *TF-IDF*. Proses perhitungan *Naïve Bayes* berdasarkan pada teorema bayes, perhitungan yang dilakukan diantaranya menghitung probabilitas prior pada setiap label, menghitung probabilitas term pada dokumen ke- n terhadap dokumen yang telah memiliki label, dan menentukan label data dokumen ke- n. Berikut contoh perhitungan Naïve Bayes.

TABEL 14
(KOMENTAR YANG DIKLASIFIKASI)

Dokumen	*Preprocessing*	Label
1	perlu revolusi mental tuju indonesia emas	positif
2	indonesia cemas tidak mungkin emas	negatif
3	sudah maklum	netral
4	semangat tuju indonesia emas	?

1) Probabilitas prior setiap label

a) positif

$$P(positif) = \frac{1}{3} = 0.333$$

b) negatif

$$P(negatif) = \frac{1}{3} = 0.333$$

c) netral

$$P(netral) = \frac{1}{3} = 0.333$$

2) Probabilitas term pada dokumen ke-4 terhadap dokumen label

a) positif

$$P(semangat|positif) = \frac{0+1}{6+11} = \frac{1}{17} = 0.058$$

$$P(tuju|positif) = \frac{1+1}{6+11} = \frac{2}{17} = 0.117$$

$$P(indonesia|positif) = \frac{1+1}{6+11} = \frac{2}{17} = 0.117$$

$$P(emas|positif) = \frac{1+1}{6+11} = \frac{2}{17} = 0.117$$

b) negatif

$$P(semangat|negatif) = \frac{0+1}{5+11} = \frac{1}{16} = 0.062$$

$$P(tuju|negatif) = \frac{0+1}{5+11} = \frac{1}{16} = 0.062$$

$$P(indonesia|negatif) = \frac{1+1}{5+11} = \frac{2}{16} = 0.125$$

$$P(emas|negatif) = \frac{1+1}{5+11} = \frac{2}{16} = 0.125$$

c) netral

$$P(semangat|netral) = \frac{0+1}{2+11} = \frac{1}{13} = 0.076$$

$$P(tuju|netral) = \frac{0+1}{2+11} = \frac{1}{13} = 0.076$$

$$P(indonesia|netral) = \frac{0+1}{2+11} = \frac{1}{13} = 0.076$$

$$P(emas|netral) = \frac{0+1}{2+11} = \frac{1}{13} = 0.076$$

3) Menentukan label dokumen ke-4

a) positif

$$P(positif|D4) = 0.333 \times (0.058 \times 0.117 \times 0.117 \times 0.117) = 0.0000309$$

b) negatif

$$P(negatif|D4) = 0.333 \times (0.062 \times 0.062 \times 0.125 \times 0.125) = 0.0000200$$

c) netral

$$P(netral|D4) = 0.333 \times (0.076 \times 0.076 \times 0.076 \times 0.076) = 0.0000111$$

Berdasarkan hasil perhitungan *Naïve Bayes*, dokumen ke-4 merupakan label positif karena nilainya lebih tinggi dibandingkan label negatif dan netral.

H. Evaluasi *Confusion Matrix*

Setelah menginisialisasi *Naïve Bayes*, dilakukan tahap perhitungan metrik evaluasi. Pada tahap ini mencakup perhitungan *accuracy*, *precision*, *recall*, dan *f1-score*. Hasil perhitungan metrik evaluasi terdapat pada Gambar 5.

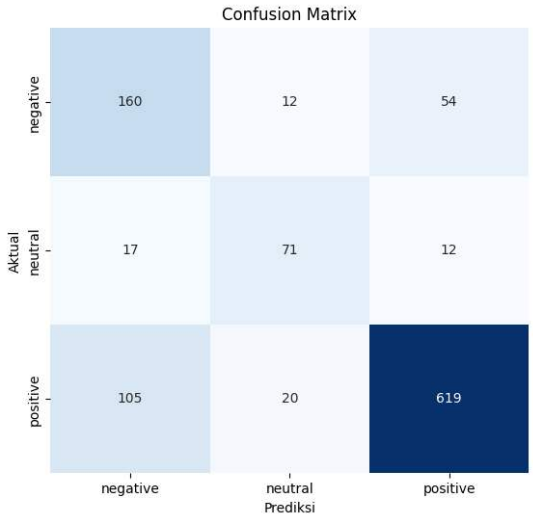
```
Performa Naive Bayes: 0.794392523364486
              precision    recall  f1-score   support

    negative       0.57      0.71      0.63       226
     neutral       0.69      0.71      0.70       100
    positive       0.90      0.83      0.87       744

    accuracy                           0.79      1070
   macro avg       0.72      0.75      0.73      1070
weighted avg       0.81      0.79      0.80      1070
```

GAMBAR 5
(HASIL METRIK EVALUASI)

Pada label positif menunjukkan performa terbaik model dengan nilai *precision* sebesar 0.90, *recall* 0.83, dan *f1-score* 0.87, hal ini menandakan bahwa metode *Naïve Bayes* cukup optimal dalam mengenali dan mengklasifikasikan komentar sentimen positif secara konsisten. Untuk label netral nilai *precision* sebesar 0.69, nilai *recall* sebesar 0.71, dan nilai *f1-*

score 0.70. Nilai-nilai ini menunjukkan kemampuan *Naïve Bayes* cukup baik dalam mengklasifikasikan sentimen netral meskipun masih terdapat kesalahan dalam mengenali maupun memprediksi sentimen yang seharusnya netral. Kemudian label negatif hasil *precision* sebesar 0.57, nilai *recall* sebesar 0.71, dan *f1-score* sebesar 0.63. Hasil ini mengindikasikan masih terdapat kesalahan dalam prediksi terhadap data negatif mengacu pada nilai *precision* 57% meskipun sebagian besar data negatif berhasil dikenali berdasarkan nilai *recall* sebesar 71%. *F1-score* sebesar 63% merupakan tingkat keseimbangan *precision* dan *recall* pada sentimen negatif. Rekap perhitungan metrik evaluasi didapat *accuracy* memperoleh 79%, *precision* sebanyak 81%, *recall* sebesar 79%, dan *f1-score* mencapai 80%.

Setelah metrik evaluasi dihitung, hasil pengujian ditampilkan menggunakan *Confusion Matrix*. Visualisasi ini menunjukkan distribusi pengujian model terhadap label aktual dan label prediksi. Berikut hasil dari *Confusion Matrix* dapat dilihat pada Gambar 6.



GAMBAR 6
(CONFUSION MATRIX)

Berdasarkan Gambar 6, hasil *Confusion Matrix* untuk label positif menampilkan model berhasil mengklasifikasikan 619 data positif diprediksi dengan benar walaupun masih terdapat kesalahan klasifikasi sebanyak 105 data positif diprediksi sebagai negatif dan sebanyak 20 data positif diprediksi sebagai netral. Pada label netral model berhasil mengklasifikasikan 71 data netral yang dapat diprediksi secara benar meskipun masih terdapat 17 data aktual netral diprediksi sebagai label negatif, dan 12 data aktual netral diprediksi sebagai label positif. Kemudian untuk label negatif berhasil mengklasifikasikan 160 data negatif diprediksi secara benar namun masih terdapat 12 data negatif diprediksi sebagai netral serta 54 data negatif diprediksi sebagai positif.

I. *Wordcloud*

Persebaran data label dapat divisualisasikan melalui *wordcloud* yang menampilkan kata-kata yang sering muncul sebagai representasi respon sentimen masyarakat pada setiap kategori sentimen.

GAMBAR 7
WORDCLOUD POSITIF

Gambar 7 menampilkan kata-kata yang sering muncul pada label positif mengenai Visi Indonesia Emas 2045. Kata-kata pada *wordcloud* positif seperti “indonesia”, “prabowo”, “baik”, “jelas”, “hilirisasi” dan “visi”. Kemunculan beberapa kata-kata ini mencerminkan adanya harapan dan dukungan dari masyarakat dalam mewujudkan Visi Indonesia Emas 2045.

GAMBAR 8
(WORDCLOUD NEGATIF)

Kata-kata yang muncul antara lain “kurang”, “korupsi”, “masalah”, “prabowo”, dan “gagal”. Berdasarkan *wordcloud* pada Gambar 8 beberapa kata negatif merepresentasikan adanya kekhawatiran dan keraguan masyarakat terkait hambatan dalam mewujudkan visi Indonesia Emas 2045 misalnya terhadap pengelolaan dan implementasi kebijakan menuju Visi Indonesia Emas 2045 yang tidak tepat sasaran.

GAMBAR 9
(WORDCLOUD NETRAL)

Gambar 9 memvisualisasikan kata-kata yang sering muncul pada label netral terkait Visi Indonesia Emas 2045. Kata-kata pada label netral diantaranya “prabowo”, “pimpin”, “negara”, “banyak”, dan “anies”. Berdasarkan *wordcloud* pada label netral didominasi oleh kata yang tidak

secara jelas menyatakan dukungan maupun kekhawatiran terkait Visi Indonesia Emas 2045.

V. KESIMPULAN

Hasil Penelitian ini menunjukkan distribusi label dari data yang dikumpulkan sebanyak 5520 kemudian dilakukkan *preprocessing* menjadi 5346 data. Data tersebut diterapkan untuk dilakukan pelabelan, pada proses ini terdapat distribusi sentimen ke dalam tiga kategori, yaitu sentimen positif sebesar 68%, sentimen negatif sebesar 22%, dan sentimen netral sebesar 10%. Performa *Naïve Bayes* setelah melalui perbandingan split data 80:20 lalu dilakukan ekstrasi fitur TF-IDF dan balancing SMOTE menghasilkan nilai *accuracy* sebesar 79%, dengan nilai *precision* sebesar 81%, *recall* sebesar 79%, dan *f1-score* sebesar 80%. Berdasarkan hasil performa akurasi *Naïve Bayes*, menunjukkan kemampuan model memiliki performa yang cukup baik dan konsisten dalam mengklasifikasikan sentimen data komentar *Youtube* terkait Visi Indonesia Emas 2045.

REFERENSI

[1] R. Wardani, "Perkembangan Arah Non-Governmental Organization (NGO) serta Civil Society di Indonesia: Periode 2024-2025," *J. Sos. dan Teknol.*, vol. 4, 2024.

[2] I. F. Budiman, "Peran Pancasila Sebagai Ideologi Negara Dalam Mewujudkan Indonesia Emas 2045," *Cendikia J. Pendidik. dan Pengajaran*, vol. 2, no. 3, pp. 47–54, 2024.

[3] A. Suryasuciramdhan, B. Ramadhan, and D. Deden, "Analisis Framing Komunikasi Politik Jokowi tentang Indonesia Emas 2045 di Media Online detik.com dan Kompas," *Filos. Publ. Ilmu Komunikasi, Desain, Seni Budaya*, vol. 1, no. 3, pp. 66–74, 2024.

[4] A. Hendrawan and E. I. Sela, "Analisis Sentimen Komentar Youtube Tentang Resesi Global 2023 Menggunakan LSTM," *J. Indones. Manaj. Inform. dan Komun.*, vol. 5, no. 1, pp. 587–593, 2024.

[5] F. Setiawan and P. F. Ariyani, "Analisis Sentimen Putusan MK Sengketa Pilpres 2024 Pada Youtube Berbasis Web Dengan Naive Bayes," *Semin. Nas. Mhs. Fak. Teknol. Inf.*, vol. 3, no. September, pp. 769–778, 2024.

[6] A. W. Davita, S. M. Agustin, and N. Astagini, "Sentimen Media melalui Social Network Analysis pada Kanal YouTube 'Harian Kompas,'" *J. Ris. Jurnalistik dan Media Digit.*, vol. 4, pp. 69–78, 2024.

[7] Rasiban and S. Riyadi, "Analisis Sentimen Opini Masyarakat Terhadap Stadion Jakarta Internasional Stadium (Jis) Pada Twitter Dengan Perbandingan Metode Naive Bayes Dan Support Vector Machine," *J. Sains dan Teknol.*, vol. 5, no. 3, pp. 951–960, 2024.

[8] L. A. Fudholi, N. Rahaningsih, and R. D. Dana, "Sentimen Analisis Perilaku Penggemar Coldplay Di Media Sosial Twitter Menggunakan Metode Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 4150–4159, 2024.

[9] S. A. Rizaldi, S. Alam, and I. Kurniawan, "Analisis Sentimen Pengguna Aplikasi JMO (Jamsostek Mobile) Pada Google Play Store Menggunakan Metode Naive Bayes," *STORAGE J. Ilm. Tek. dan Ilmu Komput.*, vol. 2, no. 3, pp. 109–117, 2023.

[10] M. Yasir and R. Suraji, "Perbandingan Metode Klasifikasi Naive Bayes, Decision Tree, Random Forest Terhadap Analisis Sentimen Kenaikan Biaya Haji 2023 pada Media Sosial Youtube," *J. Cahaya Mandalika*, vol. 3, no. 2, pp. 180–192, 2023.

[11] V. Chandradev, I. M. A. D. Suarjaya, and I. P. A. Bayupati, "Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT," *J. Buana Inform.*, vol. 14, no. 02, pp. 107–116, 2023.

[12] M. Permatasari, Z. B. Hubi, H. Mulyani, N. N. Insani, and M. L. Bribin, "Membangun Karakter Warga Negara Digital dan Pendidikan Hukum Global Menuju Indonesia Emas 2045," *Nomos J. Penelit. Ilmu Huk.*, vol. 4, no. 2, pp. 46–56, 2024.

[13] S. Hadiani, Y. S. Zamil, and L. Rafianti, "Aspek Tanggung Jawab Youtube Dalam Penyelenggaraannya Di Indonesia Berdasarkan Hukum Penyiaran, Telekomunikasi, Dan Hukum ITE," *J. Indones. Sos. Sains*, vol. 2, no. 3, p. 494, 2021.

[14] B. R. Atmadja, "Analisis Sentimen Bahasa Indonesia Pada Tempat Wisata Di Kabupaten Sukabumi Dengan Naive Bayes," *J. Elektron. dan Komput.*, vol. 15, no. 2, pp. 371–382, 2022.

[15] M. A. Akbar and A. Solichin, "Perbandingan Sentimen Ulasan Pengguna Aplikasi Ride-Hailing Gojek dan Grab Menggunakan Algoritma Multinomial Naïve Bayes," *KRESNA J. Ris. dan Pengabdi. Masy.*, vol. 4, no. 1, pp. 1–11, 2024.

[16] D. E. Saputra and A. R. Isnain, "Implementasi Algoritma Convolutional Neural Network Untuk Analisis Sentimen Bacapres 2024 Pada Kolom Komentar Youtube Mata Najwa," *J. Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 3, pp. 1431–1441, 2024.

[17] I. S. Wibowo, A. Witanti, and I. Susilawati, "Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 1, pp. 99–111, 2024.

[18] N. Blesyova and F. N. Hasan, "Analisis Sentimen Terhadap Bea Cukai Menggunakan Support Vector Machine Dan K-Fold Cross Validation," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 12051–12056, 2024.