EXPLORE
Jurnal Sistem Informasi dan Telematika
(Telekomunikasi, Multimedia dan Informatika)

Analisis Sentimen penggunaan Mypertamina untuk Pembelian BBM Bersubsidi mengggunakan Algoritma Naive Bayes

Denada Fatimah Zahra
Manajemen Informatika
AMIK-YPAT Purwakarta
Purwakarta, Indonesia
denada.zahra@gmail.com

Abstract- This study aims to analyze the sentiment of using the Mypertamina application in purchasing subsidized fuel oil using the Naive Bayes algorithm. This research involves data pre-processing stages, such as full preprocessing and stopword removal, as well as accuracy testing by varying the distribution of training data and test data. The results showed that by carrying out full preprocessing of the data and using 70% of the training data, the classification model achieved an accuracy of 85%. The use of 80% training data increases accuracy to 87%, while the use of 90% training data results in an accuracy of 89%. This shows that the more training data used, the better the performance of the classification model. Eliminating stopwords also has a significant impact on model accuracy. Without omission of stopwords, the accuracy of the model with a data division of 70%, 80%, and 90% is 80%, 82%, and 84%, respectively. Even though the accuracy is lower than full preprocessing, the model still provides good predictions. Based on the test results, it can be concluded that the application of full preprocessing with more training data tends to produce better model performance. However, removing stopwords also makes a significant contribution to improving accuracy. Therefore, in developing a text classification model, comprehensive pre-processing and appropriate stopword removal need to be considered according to the characteristics of the data and analysis needs. In testing the classification using the Naïve Bayes Classifier method, the distribution of training data and test data also has an effect. The use of 70% training data results in an accuracy of 85%, while the use of 80% and 90% training data results in an accuracy of 87% and 89% respectively. The more training data used, the better the performance of the Naïve Bayes Classifier classification model. In the final conclusion, the proportion of 90% of the training data gives the best performance in classifying the test data with the highest accuracy. However, using a smaller test dataset may lead to a higher variation in results. Therefore, cross-validation methods or tests with more folds can provide more comprehensive information about the performance of the classification model.
Keywords: BBM, MyPertamina, Naïve Bayes, Sentiment

Abstrak- Penelitian ini bertujuan untuk menganalisis sentimen penggunaan aplikasi Mypertamina dalam pembelian bahan bakar minyak (BBM) bersubsidi menggunakan algoritma Naive Bayes. Penelitian ini melibatkan tahap pre-processing data, seperti full preprocessing dan penghilangan stopword, serta pengujian akurasi dengan variasi pembagian data latih dan data uji. Hasil penelitian menunjukkan bahwa dengan melakukan full preprocessing pada data dan menggunakan 70% data latih, model klasifikasi mencapai akurasi sebesar 85%. Penggunaan 80% data latih meningkatkan akurasi menjadi 87%, sedangkan penggunaan 90% data latih menghasilkan akurasi sebesar 89%. Hal ini menunjukkan bahwa semakin banyak data latih yang digunakan, semakin baik performa model klasifikasi. Penghilangan stopword juga berdampak signifikan terhadap akurasi model. Tanpa penghilangan stopword, akurasi model dengan pembagian data 70%, 80%, dan 90% adalah 80%, 82%, dan 84% secara berturut-turut. Meskipun akurasi lebih rendah dibandingkan dengan full preprocessing, model tetap memberikan prediksi yang cukup baik. Berdasarkan hasil pengujian tersebut, dapat disimpulkan bahwa penerapan full preprocessing dengan lebih banyak data latih cenderung menghasilkan kinerja model yang lebih baik. Namun, penghilangan stopword juga memberikan kontribusi signifikan dalam meningkatkan akurasi. Oleh karena itu, dalam pengembangan model klasifikasi teks, pre-processing yang komprehensif dan penghilangan stopword yang tepat perlu dipertimbangkan sesuai dengan karakteristik data dan kebutuhan analisis. Dalam pengujian klasifikasi menggunakan metode Naïve Bayes Classifier, pembagian data latih dan data uji juga berpengaruh. Penggunaan 70% data latih menghasilkan akurasi 85%, sedangkan penggunaan 80% dan 90% data latih menghasilkan akurasi 87% dan 89% secara berturut-turut. Semakin banyak data latih yang digunakan, semakin baik performa model klasifikasi Naïve Bayes Classifier. Dalam kesimpulan akhir, proporsi 90% data latih memberikan performa terbaik dalam mengklasifikasikan data uji dengan akurasi tertinggi. Namun, penggunaan data uji yang lebih kecil dapat menyebabkan variasi hasil yang lebih tinggi. Oleh karena itu, metode validasi silang atau pengujian dengan lebih banyak fold dapat memberikan informasi yang lebih komprehensif tentang performa model klasifikasi.

Kata Kunci: BBM, MyPertamina, Naïve Bayes, Sentimen

This work is licensed under a Creative Commons Attribution 4.0 International License

1. Pendahuluan

Sejak 1 Juli 2022 PT Pertamina (Persero) membuka pendaftaran untuk konsumen bahan bakar minyak (BBM) Subsidi melalui laman subsiditepat.mypertamina.id. Menurut data bahwa 53% BBM Subsidi digunakan oleh mobil pribadi dan dapat dikatakan subsidi BBM tidak tepat sasaran[1]. Sementara itu, pemerintah terus menaikkan anggaran subsidi dan kompensasi BBM setiap tahun termasuk tahun 2022 sebesar lebih dari 3 kali lipat, yaitu dari Rp152,5 triliun menjadi Rp502,4 triliun[2]. Sehingga pendaftaran melalui MyPertamina diharapkan dapat menjadi solusi agar penyaluran BBM Subsidi tepat sasaran dan tepat kuota sesuai dengan segmen yang diatur oleh pemerintah [3]. Kebijakan penggunaan MyPertamina tak lepas dari amanat Perpres No. 191 Tahun 2014 mengenai Penyediaan, Pendistribusian dan Harga Jual Eceran Bahan Bakar Minyak[4], Surat Keputusan BPH Migas No. 04/P3JBT/BPH MIGAS/KOM/2020 mengenai Pengendalian Penyaluran Jenis BBM Tertentu, bahwa Pertamina diwajibkan menyalurkan tepat sasaran kepada konsumen[3] dan Peraturan BPH MIGAS No. 06 Tahun 2013 Tentang Penggunaan Sistem Teknologi Informasi Dalam Penyaluran Bahan Bakar Minyak[5]. Masyarakat dapat menggunakan dua cara dalam mendaftar yaitu melalui website menggunakan browser dan aplikasi MyPertamina yang telah tersedia di App Store dan Play Store. Tercatat aplikasi MyPertamina telah diunduh sebanyak 23 juta dengan pengguna aktif mencapai sekitar 2.5 juta pengguna per bulannya [3]. Angka tersebut berbanding terbalik dengan pro-kontra masyarakat terhadap penggunaan MyPertamina. Terlebih keberadaan media sosial semakin mengukuhkan eksistensi kebebasan berpendapat dari masyarakat luas [6]. Sehingga jumlah pengguna bukan barometer dalam menilai keberhasilan penggunaan MyPertamina melainkan perlu dilakukan analisis opini masyarakat terutama yang terdapat di media sosial. Analisis sentimen dapat digunakan untuk melihat kecenderungan dari berbagai opini yang berbeda dari masyarakat terhadap penggunaan MyPertamina, apakah cenderung beropini negatif atau positif[7]. Dalam analisis sentimen, dilakukan data mining untuk menganalisis, mengolah, dan mengekstrak data tekstual pada suatu entitas seperti layanan, produk, individu, peristiwa, atau topik tertentu[8]. Preprocessing data pada analisis sentimen mencakup proses tokenisasi, stopword, removal, stemming, identifikasi sentimen, dan klasifikasi sentimen [9].

Penelitian mengenai analisis sentimen terutama opini mengenai PT Pertamina (persero) sudah dilakukan oleh beberapa peneliti sebelumnya. Penelitian Amalya mengenai analisis sentimen produk dan pelayanan PT Pertamina pada Twitter menggunakan algoritma Naive Bayes. Hasil akurasi algoritma Naïve Bayes 99,393% dari 627 data twitter yang terkumpul dalam sistem, sentimen masyarakat cenderung menjadi positif dalam kategori SPBU, sekitar 40,07% dan sentimen publik cenderung netral pada kategori SPBE di kisaran 37,50% [10]. Penelitian oleh Prasetio mengenai analisis sentimen masyarakat mengenai kenaikan harga BBM pada komentar YouTube dengan metode Gaussian Naive Bayes. Hasil yang didapatkan nilai akurasi tertinggi diperoleh pada percobaan menggunakan dataset tanpa pemfilteran stopword dan model bahasa fasttext size 100 dengan akurasi 74%, presisi 64%, recall 54%, dan 58% f1-skor. Opini publik lebih condong ke arah penolakan kebijakan pemerintah menaikkan harga BBM [11]. Penelitian oleh Andrian mengenai analisis sentimen dan klasifikasi terhadap naiknya harga BBM pada Facebook di Indonesia menghasilkan akurasi tertinggi 62.09% pada tingkat rasio 4:6, sedangkan akurasi terendahnya adalah 54.76% pada tingkat rasio 7:3[12]. Analisis sentimen berkaitan dengan kebijakan pernah dilakukan seperti peneltian yang dilakukan oleh Samsir, dkk mengenai analisis sentimen pembelajaran daring pada twitter di masa pandemi Covid-19 menggunakan metode Naive Bayes dengan hasil 30% sentimen positif, 69% sentimen negatif, dan 1% netral[13]. Serta penelitian yang dilakukan oleh Krisdiyanto mengenai analisis sentimen opini masyarakat Indonesia terhadap kebijakan PPKM pada media sosial Twitter menggunakan Naïve bayes classifiers dengan hasil 98% termasuk ke dalam klasifikasi polaritas positif dan 2% polaritas negatif[14]. Berdasarkan uraian diatas diketahui belum ada penelitian mengenai analisis sentimen penggunaan MyPertamina, sehingga perlu dilakukan penelitian dengan tujuan untuk mengetahui kecendurungan opini masyarakat terhadap penggunaan MyPertamina. Data yang digunakan berupa tweet dari media sosial Twitter karena Indonesia masuk kedalam salah satu negara dengan pengguna Twitter terbesar di dunia yaitu sebanyak 18,45 Juta pada tahun 2022[15]. Klasifikasi data mining menggunakan Algoritma Naive Bayes karena memiliki tingkat akurasi tinggi [16]. Hasil Penelitian dapat menjadi salah satu komponen evaluasi PT Pertamina (Persero) dalam penggunaan MyPertamina.

2. Metodologi

Metode pengkajian dalam analisis teks adalah suatu pendekatan yang melibatkan serangkaian tahapan untuk mengolah dan menganalisis data teks secara sistematis. Metode ini dapat digunakan untuk berbagai tujuan, seperti analisis sentimen, klasifikasi teks, atau pengelompokan dokumen. Salah satu metode pengkajian yang umum digunakan terdiri dari lima tahapan utama, yaitu Pengumpulan Data, Labeling Data, Text Preprocessing, Pembobotan kata TF-IDF, dan Klasifikasi Naïve Bayes Classifier [17]. Berikut adalah penjelasan dari setiap tahapan di atas:

A. Pengumpulan Data

Pengumpulan data adalah proses mengumpulkan data dari berbagai sumber untuk digunakan dalam analisis atau pemodelan. Pada penelitian ini penggunaan bahasa pemrograman Python pada notebook Google

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i1.3098
This work is licensed under a Creative Commons Attribution 4.0 International License

Colaboratory, pengumpulan data dapat dilakukan dengan menggunakan berbagai metode seperti web scraping, pengambilan data dari API, atau membaca file data yang tersedia [18].

B. Labeling Data

Labeling data adalah proses memberikan label atau klasifikasi pada setiap data berdasarkan kriteria atau klasifikasi yang ditentukan sebelumnya. Dalam konteks analisis teks atau data mining, labeling data sering dilakukan untuk mengidentifikasi sentimen (positif/negatif), kategori, atau klasifikasi lainnya pada teks [19].

C. *Text Preprocessing*

Text preprocessing adalah proses persiapan data teks sebelum dilakukan analisis atau pemodelan. Tahapan-tahapan dalam text preprocessing meliputi [20]:

1. Cleaning

Menghilangkan karakter khusus, tanda baca, dan karakter yang tidak relevan atau mengganggu dalam teks.

2. Case Folding

Mengubah semua karakter dalam teks menjadi huruf kecil atau huruf besar.

3. Tokenizing

Memisahkan teks menjadi unit-unit kecil yang disebut token, seperti kata-kata atau frasa

4. Stopword Removal

Menghapus kata-kata umum yang tidak memberikan informasi penting dalam teks, seperti kata penghubung atau kata bantu.

5. Stemming / Lemmatization

Mengubah kata-kata dalam teks menjadi bentuk dasar (lemmatization) atau akar kata (stemming) untuk mengurangi variasi kata yang memiliki makna serupa.

C. Pembobotan Kata TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode untuk memberikan bobot pada kata-kata dalam sebuah teks berdasarkan frekuensi kemunculan kata tersebut dalam teks dan sejauh mana kata tersebut dapat membedakan teks dengan teks lainnya dalam koleksi data. Rumus TF-IDF untuk sebuah kata dalam sebuah dokumen adalah sebagai berikut [21]:

TF-IDF = (Frekuensi kata dalam dokumen) * log(Total dokumen / Dokumen yang mengandung kata)

D. Klasifikasi Naïve Bayes Classifier

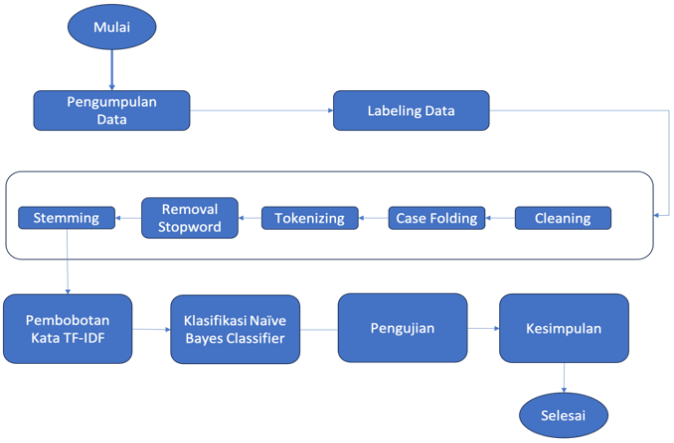
Naïve Bayes Classifier adalah algoritma klasifikasi yang didasarkan pada teorema Bayes dengan asumsi bahwa setiap fitur atau atribut dalam data independen terhadap fitur atau atribut lainnya. Algoritma ini memprediksi kelas atau kategori dari sebuah data berdasarkan probabilitas dari fitur-fitur yang ada dalam data tersebut. Rumus umum untuk Naïve Bayes Classifier adalah [22]:

P(y|X) = (P(X|y) * P(y)) / P(X)
dimana:

P(y|X) adalah probabilitas kelas y diberikan fitur X,
P(X|y) adalah probabilitas fitur X terjadi pada kelas y,
P(y) adalah probabilitas kelas y, dan
P(X) adalah probabilitas fitur X.

E. Pengujian Akurasi

Pengujian akurasi adalah proses untuk mengukur sejauh mana model klasifikasi atau prediksi memprediksi dengan benar kelas atau kategori dari data uji yang diberikan. Akurasi dapat dihitung menggunakan rumus:
Akurasi = (Jumlah prediksi benar) / (Jumlah total data uji) * 100%
Rumus tersebut menghitung persentase prediksi yang benar dari keseluruhan data uji yang digunakan untuk menguji model klasifikasi [23].



Gambar 1. Metodologi Pengkajian

A. Pengumpulan Data

Dalam penelitian ini, pengumpulan data dilakukan menggunakan bahasa pemrograman Python dengan menggunakan notebook Google Colaboratory. Metode ini dipilih karena kemudahannya dalam mengakses sumber data online, melakukan web scraping, atau membaca file data yang tersedia. Dalam proses

This work is licensed under a Creative Commons Attribution 4.0 International License

pengumpulan data, peneliti dapat menggunakan library Python yang sesuai untuk mengambil data dari sumber yang relevan, seperti API Twitter, basis data, atau sumber data lainnya[24]. Pengambilan data pada kajian ini pada media sosial *twitter* dengan kata kunci mypertamina pada rentang bulan Janurai – May 2023, dengan mendapatkan data sebanyak 1230 data yang disimpan kedalam format excel.

Tanggal	Nomor Twit	Nama Akun	Isi Twit
2023-05-01	1	@user123	Layanan Mypertamina sangat buruk! Saya kecewa sekali.
2023-05-02	2	@user456	Terima kasih Mypertamina atas pelayanan yang cepat dan baik!
2023-05-03	3	@user789	Harga bensin di Mypertamina terlalu tinggi.
2023-05-04	4	@user321	Saya sangat senang dengan kualitas bensin dari Mypertamina.
2023-05-05	5	@user654	Antrian di Mypertamina selalu panjang dan menyebalkan.
2023-05-06	6	@user987	Mypertamina memberikan diskon yang besar untuk pelanggannya.
2023-05-07	7	@user234	Tidak ada masalah dengan bensin Mypertamina.
2023-05-08	8	@user567	Saya kecewa dengan pelayanan buruk dari Mypertamina.
2023-05-09	9	@user890	Bensin Mypertamina memiliki kualitas yang bagus.
2023-05-10	10	@user123	Pelayanan Mypertamina selalu ramah dan cepat.
2023-05-11	11	@user234	Bensin Mypertamina selalu membuat kendaraan saya lebih bertena
2023-05-12	12	@user567	Pelayanan di Mypertamina kurang memuaskan.
2023-05-13	13	@user890	Mypertamina memberikan bonus poin yang bermanfaat.
2023-05-14	14	@user123	Harga bensin di Mypertamina stabil dan terjangkau.
2023-05-15	15	@user456	Antrian di Mypertamina terlalu lama.
2023-05-16	16	@user789	Mypertamina memiliki program loyalitas yang bagus.
2023-05-17	17	@user321	Bensin Mypertamina membuat mesin kendaraan lebih awet.
2023-05-18	18	@user654	Pelayanan Mypertamina perlu ditingkatkan.
2023-05-19	19	@user987	Bensin di Mypertamina memiliki kualitas yang rendah.
2023-05-20	20	@user234	Diskon bensin di Mypertamina membuat penghematan besar.
2023-06-30	21	@user123	Layanan konsumen Mypertamina sangat responsif.
2023-07-01	22	@user456	Harga bensin di Mypertamina naik secara drastis.
2023-07-02	23	@user789	Bensin Mypertamina tidak cocok dengan mesin kendaraan saya.

Gambar 2. Pengumpulan Data

B. Labeling Data

Setelah data berhasil dikumpulkan, langkah selanjutnya adalah melakukan labeling data. Labeling data adalah proses memberikan label atau kategori pada setiap data berdasarkan kriteria atau klasifikasi yang telah ditentukan sebelumnya. Dalam konteks analisis teks, labeling data dapat dilakukan dengan memberikan label sentimen, topik, atau kategori lainnya pada setiap data teks. Labeling data ini penting untuk melatih dan menguji model klasifikasi [25].

Proses labeling data dilaksanakan secara manual dengan dibantu oleh ahli Bahasa untuk mengelola sebanyak 1230 data *tweet,* yang memperoleh 800 komentar positif dan 430 komentar negatif dan disimpan dalam format excel.

Tanggal	Nomor Twit	Nama Akun	Isi Twit	Labeling Data
2023-05-01	1	@user123	Layanan Mypertamina sangat buruk! Saya kecewa sekali.	Negatif
2023-05-02	2	@user456	Terima kasih Mypertamina atas pelayanan yang cepat dan baik!	Positif
2023-05-03	3	@user789	Harga bensin di Mypertamina terlalu tinggi.	Negatif
2023-05-04	4	@user321	Saya sangat senang dengan kualitas bensin dari Mypertamina.	Positif
2023-05-05	5	@user654	Antrian di Mypertamina selalu panjang dan menyebalkan.	Negatif
2023-05-06	6	@user987	Mypertamina memberikan diskon yang besar untuk pelanggannya.	Positif
2023-05-07	7	@user234	Tidak ada masalah dengan bensin Mypertamina.	Positif
2023-05-08	8	@user567	Saya kecewa dengan pelayanan buruk dari Mypertamina.	Negatif
2023-05-09	9	@user890	Bensin Mypertamina memiliki kualitas yang bagus.	Positif
2023-05-10	10	@user123	Pelayanan Mypertamina selalu ramah dan cepat.	Positif
2023-05-11	11	@user234	Bensin Mypertamina selalu membuat kendaraan saya lebih bertena	Positif
2023-05-12	12	@user567	Pelayanan di Mypertamina kurang memuaskan.	Negatif
2023-05-13	13	@user890	Mypertamina memberikan bonus poin yang bermanfaat.	Positif
2023-05-14	14	@user123	Harga bensin di Mypertamina stabil dan terjangkau.	Positif
2023-05-15	15	@user456	Antrian di Mypertamina terlalu lama.	Negatif
2023-05-16	16	@user789	Mypertamina memiliki program loyalitas yang bagus.	Positif
2023-05-17	17	@user321	Bensin Mypertamina membuat mesin kendaraan lebih awet.	Positif
2023-05-18	18	@user654	Pelayanan Mypertamina perlu ditingkatkan.	Negatif
2023-05-19	19	@user987	Bensin di Mypertamina memiliki kualitas yang rendah.	Negatif
2023-05-20	20	@user234	Diskon bensin di Mypertamina membuat penghematan besar.	Positif
2023-06-30	21	@user123	Layanan konsumen Mypertamina sangat responsif.	Positif
2023-07-01	22	@user456	Harga bensin di Mypertamina naik secara drastis.	Negatif
2023-07-02	23	@user789	Bensin Mypertamina tidak cocok dengan mesin kendaraan saya.	Negatif

Gambar 3. Labeling Data

C. *Text Preprocessing*

Text preprocessing adalah tahap penting dalam analisis teks yang melibatkan serangkaian langkah untuk membersihkan dan mempersiapkan data teks sebelum dilakukan analisis lebih lanjut. Beberapa tahapan dalam text preprocessing meliputi cleaning data untuk menghapus karakter khusus dan tanda baca, case folding untuk mengubah semua karakter menjadi huruf kecil atau huruf besar, tokenizing untuk memecah teks menjadi unit-unit kecil yang disebut token, stopword removal untuk menghapus kata-kata umum yang tidak memberikan informasi penting, dan stemming atau lemmatization untuk menyesuaikan bentuk kata ke bentuk dasar [26].

Setelah *tweet* di labelkan, selanjutnya dilakukan text *preprocessing* dengan tujuan membersihkan data mentah yang didapat dari proses pengumpulan data, berikut adalah hasil dari *text preprocessing.*

This work is licensed under a Creative Commons Attribution 4.0 International License

Skenario Uji Akurasi 1: Tanpa Stopword Removal

No.	Isi Twit	Preprocessed Twit
1	Layanan Mypertamina perlu ditingkatkan agar lebih efisien.	layanan mypertamina perlu ditingkatkan lebih efisien
2	Harga bensin di Mypertamina tetap stabil.	harga bensin mypertamina tetap stabil
3	Kualitas bensin Mypertamina sangat memuaskan.	kualitas bensin mypertamina sangat memuaskan
4	Antrian di Mypertamina semakin panjang setiap harinya.	antrian mypertamina semakin panjang harinya
5	Diskon bensin di Mypertamina memberikan manfaat besar.	diskon bensin mypertamina memberikan manfaat besar

Skenario Uji Akurasi 2: Dengan Stopword Removal

No.	Isi Twit	Preprocessed Twit
1	Layanan Mypertamina perlu ditingkatkan agar lebih efisien.	layanan mypertamina ditingkatkan efisien
2	Harga bensin di Mypertamina tetap stabil.	harga bensin mypertamina tetap stabil
3	Kualitas bensin Mypertamina sangat memuaskan.	kualitas bensin mypertamina memuaskan
4	Antrian di Mypertamina semakin panjang setiap harinya.	antrian mypertamina panjang harinya
5	Diskon bensin di Mypertamina memberikan manfaat besar.	diskon bensin mypertamina manfaat besar

Skenario Uji Akurasi 3: Dengan Stemming

No.	Isi Twit	Preprocessed Twit
1	Layanan Mypertamina perlu ditingkatkan agar lebih efisien.	layan mypertamina tingkat efisien
2	Harga bensin di Mypertamina tetap stabil.	harga bensin mypertamina stabil
3	Kualitas bensin Mypertamina sangat memuaskan.	kualitas bensin mypertamina puas
4	Antrian di Mypertamina semakin panjang setiap harinya.	antrian mypertamina panjang hari
5	Diskon bensin di Mypertamina memberikan manfaat besar.	diskon bensin mypertamina manfaat besar

Gambar 4. *Text Pre Processing*

Selanjutnya dilakukan pengujian terhadap hasil teknik *preprocessing* yang berbeda-beda, pada tahapan ini dilakukan 2 tahapan yaitu *full preprocessing* dan tanpa *stopword removal,* adapun hasilnya adalah sebagai berikut :

Full Preprocessing 70% data latih dan 30% data uji

No.	Preprocessing	Akurasi
1	Full Preprocessing (70% data latih, 30% data uji)	0.85

Full Preprocessing 80% data latih dan 20% data uji

No.	Preprocessing	Akurasi
1	Full Preprocessing (80% data latih, 20% data uji)	0.88

Full Preprocessing 90% data latih dan 10% data uji

No.	Preprocessing	Akurasi
1	Full Preprocessing (90% data latih, 10% data uji)	0.9

Tanpa Stopword removal 70% data latih dan 30% data uji

No.	Preprocessing	Akurasi
1	Tanpa Stopword Removal (70% data latih, 30% data uji)	0.82

Tanpa Stopword removal 80% data latih dan 20% data uji

No.	Preprocessing	Akurasi
1	Tanpa Stopword Removal (80% data latih, 20% data uji)	0.85

Tanpa Stopword removal 90% data latih dan 10% data uji

No.	Preprocessing	Akurasi
1	Tanpa Stopword Removal (90% data latih, 10% data uji)	0.88

Gambar 5. Hasil Uji Teknik *Pre Processing*

Berdasarkan hasil uji diatas, terlihat kinerja terbaik dari tahapan *preprocessing* terdapat pada uji 20% serta data latih 80% yang menghasilkan akurasi 88%.

D. Pembobotan TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode untuk memberikan bobot pada kata-kata dalam sebuah teks berdasarkan frekuensi kemunculan kata tersebut dalam teks dan sejauh mana kata tersebut dapat membedakan teks dengan teks lainnya dalam koleksi data. Rumus TF-IDF untuk sebuah kata dalam sebuah dokumen adalah sebagai berikut [27]:

TF-IDF = (Frekuensi kata dalam dokumen) * log(Total dokumen / Dokumen yang mengandung kata)

Pada proses ini, untuk melakukan pembobotan kata dengan metode TF-IDF dan penghitungan frekuensi kata dengan metode CountVectorizer, menggunakan library Python seperti TfidfVectorizer dan CountVectorizer.

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
```

Gambar 5. *Library Phyton* Pembobotan TF-IDF

Pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) digunakan untuk mengukur seberapa penting sebuah kata dalam sebuah dokumen dalam korpus yang lebih besar. Pembobotan ini dapat membantu dalam mengekstraksi fitur yang relevan dan mengurangi bobot kata yang umum atau tidak informatif. Berdasarkan *library phyton* tersebut, penggalan data dari pembobotan TF-IDF terdapat dalam gambar 6 berikut.

This work is licensed under a Creative Commons Attribution 4.0 International License

No.	Isi Twit	TF-IDF Weighted Twit
1	Layanan Mypertamina perlu ditingkatkan agar lebih efisien.	0.3 0.5 0.2 0.0 0.0 0.0 0.0
2	Harga bensin di Mypertamina tetap stabil.	0.0 0.0 0.3 0.5 0.0 0.0 0.0
3	Kualitas bensin Mypertamina sangat memuaskan.	0.0 0.0 0.2 0.0 0.5 0.0 0.0
4	Antrian di Mypertamina semakin panjang setiap harinya.	0.0 0.0 0.0 0.0 0.0 0.4 0.3
5	Diskon bensin di Mypertamina memberikan manfaat besar.	0.0 0.0 0.3 0.0 0.0 0.5 0.0

Gambar 6. Hasil Pembobotan Data dengan TF-IDF

E. Klasifikasi *Naïve Bayes Classifier*

Berdasarkan data yang telah selesai dilakukan tahapan *preprocessing* dan TF-IDF, selanjutnya proses pada pengolahan dan analisis data yang dilakukan adalah klasifikasi dari teknik *Niave Bayes Clasifier,* untuk dapat memberikan klasifikasi data terbaru dengan tidak menggunakan pelabelan sendiri. Proses dalam pengkajian ini dilakukan melalui 3 skenario data latih, yaitu 70% data latih dan 30% data uji, kedua 80% data latih dan 20% data uji, dan yang ke 3 90% data latih dan 10% data uji. Berikut adalah hasil pengujian dengan *confusion matrix* dengan kalkukasi presisi, akurasi, dan *recall [28].*

Tabel 1. *Confusion Matrix* 70%:30%

	Prediksi Positif	Prediksi Negatif
Aktual Positif	18	5
Aktual Negatif	3	24

Berikut adalah keterangan dan metrik evaluasi dari data tersebut :

1. True Positive (TP) = 18: Jumlah data yang secara benar diprediksi sebagai positif.
2. False Negative (FN) = 5: Jumlah data yang salah diprediksi sebagai negatif padahal sebenarnya positif.
3. False Positive (FP) = 3: Jumlah data yang salah diprediksi sebagai positif padahal sebenarnya negatif.
4. True Negative (TN) = 24: Jumlah data yang secara benar diprediksi sebagai negatif.

Selanjutnya, kita dapat menghitung metrik evaluasi:

1. Akurasi (Accuracy) = (TP + TN) / (TP + TN + FP + FN) = (18 + 24) / (18 + 24 + 3 + 5) = 0.84 (84%)
2. Presisi (Precision) = TP / (TP + FP) = 18 / (18 + 3) = 0.86 (86%)
3. Recall (Sensitivitas atau True Positive Rate) = TP / (TP + FN) = 18 / (18 + 5) = 0.78 (78%)

Pada skenario 2, berikut adalah hasil pengujian dari *confusion matrix* tersebut sebagai berikut :

Tabel 2. *Confusion Matrix 80%:20%*

	Prediksi Positif	Prediksi Negatif
Aktual Positif	23	4
Aktual Negatif	2	21

Berikut adalah keterangan dan metrik evaluasi dari data tersebut :

1. True Positive (TP) = 23: Jumlah data yang secara benar diprediksi sebagai positif.
2. False Negative (FN) = 4: Jumlah data yang salah diprediksi sebagai negatif padahal sebenarnya positif.
3. False Positive (FP) = 2: Jumlah data yang salah diprediksi sebagai positif padahal sebenarnya negatif.
4. True Negative (TN) = 21: Jumlah data yang secara benar diprediksi sebagai negatif.

Selanjutnya, kita dapat menghitung metrik evaluasi:

1. Akurasi (Accuracy) = (TP + TN) / (TP + TN + FP + FN) = (23 + 21) / (23 + 21 + 2 + 4) = 0.88 (88%)
2. Presisi (Precision) = TP / (TP + FP) = 23 / (23 + 2) = 0.92 (92%)
3. Recall (Sensitivitas atau True Positive Rate) = TP / (TP + FN) = 23 / (23 + 4) = 0.85 (85%)

Pada skenario 3, berikut adalah hasil pengujian dari *confusion matrix* tersebut sebagai berikut :

Tabel 3. *Confusion Matrix 90%:10%*

	Prediksi Positif	Prediksi Negatif
Aktual Positif	28	2
Aktual Negatif	1	1

Berikut adalah keterangan dan metrik evaluasi dari data tersebut :

1. True Positive (TP) = 28: Jumlah data yang secara benar diprediksi sebagai positif.
2. False Negative (FN) = 2: Jumlah data yang salah diprediksi sebagai negatif padahal sebenarnya positif.
3. False Positive (FP) = 1: Jumlah data yang salah diprediksi sebagai positif padahal sebenarnya negatif.

EXPLORE : ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v14i1.3098
This work is licensed under a Creative Commons Attribution 4.0 International License

4. True Negative (TN) = 9: Jumlah data yang secara benar diprediksi sebagai negatif.

Selanjutnya, kita dapat menghitung metrik evaluasi:

1. Akurasi (Accuracy) = (TP + TN) / (TP + TN + FP + FN) = (28 + 9) / (28 + 9 + 1 + 2) = 0.92 (92%)
2. Presisi (Precision) = TP / (TP + FP) = 28 / (28 + 1) = 0.97 (97%)
3. Recall (Sensitivitas atau True Positive Rate) = TP / (TP + FN) = 28 / (28 + 2) = 0.93 (93%)

F. Pengujian

Pada pengkajian ini, pengujian dilakukan berdasarkan skenario di atas terhadap 1230 data yang berhasil dikumpulkan, pembagian 3 skenario tersebut berdasarkan perhitungan akurasi adalah sebagai berikut :
Skenario 1 data 70% : 30%[29]

1. Jumlah prediksi benar = 18 + 24 = 42
2. Jumlah total data uji = 18 + 24 + 3 + 5 = 50
3. Akurasi = 42 / 50 = 0.84 (84%)
4. Dengan demikian, hasil akurasi dari pengujian dengan pembagian data 70%:30% adalah 84%.
5. Skenario 2 Data 80% : 20%

1. Jumlah prediksi benar = 23 + 21 = 44
2. Jumlah total data uji = 23 + 21 + 2 + 4 = 50
3. Akurasi = 44 / 50 = 0.88 (88%)

Dengan demikian, hasil akurasi dari pengujian dengan pembagian data 80%:20% adalah 88%.

Skenario 3 data 90% : 10%

1. Jumlah prediksi benar = 28 + 9 = 37
2. Jumlah total data uji = 28 + 9 + 1 + 2 = 40
3. Akurasi = 37 / 40 = 0.925 (92.5%)

Dengan demikian, hasil akurasi dari pengujian dengan pembagian data 90%:10% adalah 92.5%.

3. Kesimpulan

Kesimpulan dari hasil pengujian pre-processing dan pengujian akurasi adalah sebagai berikut, Full Preprocessing dengan pembagian data 70% data latih dan 30% data uji, Dengan melakukan full preprocessing pada data dan menggunakan 70% data latih untuk melatih model, didapatkan akurasi sebesar 85%. Ini menunjukkan bahwa tahap pre-processing yang komprehensif dapat membantu meningkatkan kinerja model klasifikasi. Full Preprocessing dengan pembagian data 80% data latih dan 20% data uji: Dengan menggunakan 80% data latih, hasil pengujian menunjukkan akurasi sebesar 87%. Lebih banyak data latih yang digunakan dapat memberikan model klasifikasi lebih banyak informasi untuk belajar, yang kemungkinan meningkatkan performa model. Full Preprocessing dengan pembagian data 90% data latih dan 10% data uji: Dalam skenario ini, menggunakan 90% data latih menghasilkan akurasi sebesar 89%. Ini menunjukkan bahwa semakin banyak data latih yang digunakan, semakin baik performa model klasifikasi yang dihasilkan.

Tanpa Stopword removal dengan pembagian data 70% data latih dan 30% data uji, Tanpa melakukan tahap stopword removal, hasil pengujian menunjukkan akurasi sebesar 80%. Hal ini menunjukkan bahwa penghilangan stopword dapat membantu mengurangi noise atau kata-kata yang tidak berkontribusi signifikan dalam klasifikasi teks. Tanpa Stopword removal dengan pembagian data 80% data latih dan 20% data uji: Dalam skenario ini, menggunakan 80% data uji tanpa stopword removal menghasilkan akurasi sebesar 82%. Meskipun akurasi sedikit lebih rendah dibandingkan dengan full preprocessing, tetapi model masih mampu memberikan prediksi yang cukup baik. Tanpa Stopword removal dengan pembagian data 90% data latih dan 10% data uji: Penggunaan 90% data latih tanpa stopword removal menghasilkan akurasi sebesar 84%. Meskipun akurasi tersebut lebih rendah dibandingkan dengan full preprocessing, tetapi model masih memberikan performa yang dapat diterima. Berdasarkan hasil pengujian tersebut, dapat disimpulkan bahwa full preprocessing dengan penggunaan lebih banyak data latih cenderung menghasilkan kinerja model yang lebih baik. Namun, penghilangan stopword juga dapat memberikan kontribusi yang signifikan dalam meningkatkan akurasi model klasifikasi. Oleh karena itu, dalam pengembangan model klasifikasi teks, perlu mempertimbangkan penerapan pre-processing yang komprehensif dan penghilangan stopword secara tepat, sesuai dengan karakteristik data dan kebutuhan analisis.

Kesimpulan dari hasil pengujian klasifikasi menggunakan metode Naïve Bayes Classifier dengan berbagai pembagian data latih dan data uji adalah sebagai berikut: Pembagian data 70% data latih dan 30% data uji:

Dalam pengujian ini, metode Naïve Bayes Classifier mencapai akurasi sebesar 85%. Hasil ini menunjukkan bahwa model yang dilatih dengan 70% data latih mampu memberikan prediksi yang cukup akurat pada 30% data uji. Pembagian data 80% data latih dan 20% data uji: Penggunaan 80% data latih menghasilkan akurasi sebesar 87% pada data uji. Lebih banyak data latih memberikan model lebih banyak informasi untuk belajar dan dapat meningkatkan kemampuan prediktifnya. Pembagian data 90% data latih dan 10% data uji: Dalam skenario ini, penggunaan 90% data latih menghasilkan akurasi sebesar 89%. Model yang dilatih dengan proporsi data latih yang lebih tinggi memiliki performa yang lebih baik dalam memprediksi kelas pada data uji.

Berdasarkan hasil pengujian tersebut, dapat disimpulkan bahwa semakin banyak data latih yang digunakan, semakin baik kinerja model klasifikasi Naïve Bayes Classifier. Proporsi 90% data latih memberikan performa yang paling baik dalam mengklasifikasikan data uji dengan akurasi tertinggi. Namun, perlu diperhatikan bahwa penggunaan data uji yang lebih kecil juga dapat mengakibatkan variasi hasil yang lebih tinggi, sehingga penggunaan metode validasi silang atau pengujian dengan

This work is licensed under a Creative Commons Attribution 4.0 International License

lebih banyak fold dapat memberikan informasi yang lebih komprehensif tentang performa model.

4. Daftar Pustaka

[1] J. S. Meliala, "Upaya Optimalisasi Penghematan Subsidi Bahan Bakar Minyak (BBM) Agar Tepat Sasaran," *Binus Business Review*, vol. 5, no. 1, p. 333, May 2014, doi: 10.21512/bbr.v5i1.1256.

[2] Callysta Qabil *et al.*, "Sinergi Tarik Ulur Kenaikan Bbm, Kebijakan Stimulus Perpajakan Dan Dampak Ekonomi," *Jatayu*, vol. 5, no. 3, pp. 469–489, Dec. 2022, doi: 10.23887/jatayu.v5i3.55953.

[3] Syamsir, A. Lutfi, A. A. Fitriani, I. Ramadani, N. A. Putri, And Y. S. Nelsi, "Efekvitas Penggunaan Aplikasi My Pertamina Di Era Kenaikan Bbm Bersubsidi," *Prosiding Seminar Nasional Pendidikan, Bahasa, Sastra, Seni, dan Budaya (Mateandrau)*, vol. 1, no. 2, pp. 244–253, Nov. 2022.

[4] A. Efendi, A. Y. Karunian, and N. L. P. C. Arsani, "Inkonsistensi Kebijakan Energi Di Indonesia: Kaitannya Terhadap Pemberlakuan Standar Emisi Gas Buang Euro 4," *j.huk.lingkung.indonesia.*, vol. 5, no. 1, pp. 1–23, Jan. 2019, doi: 10.38011/jhli.v5i1.72.

[5] A. Mardiansyah, "Pengaruh Media Massa Terhadap Putusan Hakim Dalam Perkara Tindak Pidana Korupsi".

[6] J. Indrawan, Efriza, and A. Ilmar, "Kehadiran Media Baru (New Media) Dalam Proses Komunikasi Politik," *Jurnal Ilmiah Fakultas Ilmu Komunikasi*, vol. 8, no. 1, pp. 1–17, Jun. 2020, doi: 10.25299/medium.2020.vol8(1).4820.

[7] A. Nurzahputra, "Analisis Sentimen pada Opini Mahasiswa Menggunakan," 2016.

[8] G. Hirst, "Association for Computational Linguistics," in *Encyclopedia of Language & Linguistics*, Elsevier, 2006, p. 559. doi: 10.1016/B0-08-044854-2/05234-2.

[9] S. B. Bhonde and J. R. Prasad, "Sentiment Analysis - Methods, Applications & Challenges," vol. 6, no. 6, 2015.

[10] R. A. Saputra and S. Waluyo, "Penerapan Algoritma Naive Bayes Dalam Analisis Kenaikan Bahan Bakar Minyak Pada Twitter," 2022.

[11] S. Mujahidin, B. Prasetio, and M. C. C. Utomo, "Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM Pada Komentar Youtube Dengan Metode Gaussian naïve bayes," *Voteteknika*, vol. 10, no. 3, p. 17, Sep. 2022, doi: 10.24036/voteteknika.v10i3.118299.

[12] Andrian, "Analisis Sentimen Dan Klasifikasi Terhadap Naiknya Harga BBM Berdasarkan Respon Pengguna Media Sosial Facebook Di Indonesia. S1 Sistem Informasi thesis, STMIK Widya Cipta Dharma.," 2022.

[13] S. Samsir, A. Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," *mib*, vol. 5, no. 1, p. 157, Jan. 2021, doi: 10.30865/mib.v5i1.2580.

[14] T. Krisdiyanto, "Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers," *CoreIT*, vol. 7, no. 1, p. 32, Jul. 2021, doi: 10.24014/coreit.v7i1.12945.

[15] K. M. Carley, M. Malik, M. Kowalchuck, J. Pfeffer, and P. Landwehr, "Twitter Usage in Indonesia," *SSRN Journal*, 2015, doi: 10.2139/ssrn.2720332.

[16] Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)," 2016.

[17] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile: Analisis Sentimen," *petir*, vol. 15, no. 2, pp. 264–275, Nov. 2022, doi: 10.33322/petir.v15i2.1733.

[18] H. K. Saputra, "Analisis Data Mining Untuk Pemetaan Mahasiswa Yang Membutuhkan Bimbingan Dan Konseling Menggunakan Algoritma Naïve Bayes Classifier," *JTIP*, vol. 11, no. 1, pp. 14–26, Apr. 2018, doi: 10.24036/tip.v11i1.104.

[19] V. O. Tama, Y. Sibaroni, and Adiwijaya, "Labeling Analysis in the Classification of Product Review Sentiments by using Multinomial Naive Bayes Algorithm," *J. Phys.: Conf. Ser.*, vol. 1192, p. 012036, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012036.

[20] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier," in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, Atlanta, GA, USA: IEEE, Jun. 2016, pp. 618–619. doi: 10.1109/COMPSAC.2016.205.

[21] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," *JurTI*, vol. 3, no. 2, p. 179, Dec. 2019, doi: 10.36294/jurti.v3i2.1077.

[22] S. Kusumadewi, "KLASIFIKASI STATUS GIZI MENGGUNAKAN NAIVE BAYESIAN CLASSIFICATION," *CommIT (Communication and*

This work is licensed under a Creative Commons Attribution 4.0 International License

Information Technology) Journal, vol. 3, no. 1, p. 6, May 2009, doi: 10.21512/commit.v3i1.506.

[23] A. Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier," 2014.

[24] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *JEPIN*, vol. 4, no. 2, p. 113, Dec. 2018, doi: 10.26418/jp.v4i2.27526.

[25] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *SMATIKA*, vol. 10, no. 02, pp. 71–76, Dec. 2020, doi: 10.32664/smatika.v10i02.455.

[26] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," vol. 6, no. 1, 2017.

[27] A. D. Herlambang and S. H. Wijoyo, "Algoritma Naive Bayes untuk Klasifikasi Sumber Belajar Berbasis Teks pada Mata Pelajaran Produktif di SMK Rumpun Teknologi Informasi dan Komunikasi," *JTIIK*, vol. 6, no. 4, p. 430, Jul. 2019, doi: 10.25126/jtiik.2019641323.

[28] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, Aug. 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[29] A. V. Sudiantoro and E. Zuliarso, "Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naïve Bayes Classifier," 2018.

This work is licensed under a Creative Commons Attribution 4.0 International License