TUGAS PROJECT (A) AKUISISI DAN PRAPROSES DATA PENGOLAHAN BAHASA ALAMI

INFORMATION RETRIEVAL FOR BPJS SERVICES ARTICLE

KELOMPOK BPJS

Naura Jasmine Azzahra	5026211005
Ramadhanul Husna A. M.	5026211059
Fikri Septa Setiawan	5026211109

KELAS
PENGOLAHAN BAHASA ALAMI (A)

DEPARTEMEN SISTEM INFORMASI
FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SEMESTER GASAL 2024

DAFTAR ISI

Information Retrieval For BPJS ServicesArticle

ABSTRAK

BPJS Kesehatan merupakan program jaminan sosial di Indonesia yang bertujuan menyediakan akses layanan kesehatan bagi seluruh masyarakat, namun persepsi publik terkait kualitas layanan, aksesibilitas, dan fasilitas BPJS masih beragam. Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis sentimen publik terhadap layanan BPJS melalui berita online, dengan data yang diperoleh dari berbagai portal berita terpercaya menggunakan metode web scraping. Setelah praproses data, analisis sentimen dilakukan menggunakan metode TF-IDF, Bag of Words, dan Cosine Similarity untuk mengelompokkan berita dan mendeteksi sentimen positif, netral, dan negatif. Hasil analisis menunjukkan bahwa sentimen netral mendominasi berita tentang BPJS, sementara sentimen negatif sering kali muncul terkait kendala akses layanan dan keterbatasan fasilitas, serta sentimen positif ditemukan dalam ulasan terkait program atau kebijakan yang dinilai efektif. Temuan ini memberikan gambaran tentang tren persepsi publik dan dapat dijadikan dasar bagi BPJS untuk mengevaluasi dan meningkatkan kualitas layanan sesuai ekspektasi masyarakat.

ABSTRACT

BPJS Kesehatan is a social security program in Indonesia aimed at providing healthcare access to all citizens. However, public perception of the quality, accessibility, and facilities of BPJS services remains mixed. This study aims to identify and analyze public sentiment towards BPJS services through online news articles. Data was collected from various reputable news portals using web scraping, and preprocessed before sentiment analysis was conducted using TF-IDF, Bag of Words, and Cosine Similarity to cluster the articles and detect positive, neutral, and negative sentiments. The analysis results indicate that neutral sentiment dominates BPJS-related news, while negative sentiment often arises concerning access issues and limited facilities, and positive sentiment is found in reviews related to effective programs or policies. These findings provide insight into public perception trends and can serve as a foundation for BPJS to evaluate and improve service quality according to public expectations.

1. LATAR BELAKANG

BPJS (Badan Penyelenggara Jaminan Sosial) Kesehatan merupakan program jaminan sosial terbesar di Indonesia yang mencakup pelayanan kesehatan bagi seluruh masyarakat. Dalam beberapa tahun terakhir, layanan BPJS menjadi sorotan publik dan kerap dibahas dalam berbagai berita di media online. Pelayanan, ketersediaan fasilitas, serta kualitas administrasi BPJS sering kali menjadi topik yang diperbincangkan, baik dalam konteks keluhan maupun apresiasi. Persepsi publik terhadap BPJS, yang disampaikan melalui berbagai sumber berita, memberikan gambaran yang penting bagi pemerintah dan manajemen BPJS untuk terus meningkatkan kualitas layanan.

Meski berfungsi sebagai solusi bagi masalah akses kesehatan, persepsi publik terhadap BPJS kerap kali beragam, terutama terkait kualitas dan kecepatan pelayanan. Menurut Syafrizal dan Fitri P. Sari (2015), persepsi masyarakat memainkan peran penting dalam keberhasilan BPJS, karena persepsi positif dapat meningkatkan partisipasi dan dukungan publik terhadap program ini, sementara persepsi negatif sering kali muncul dari pengalaman yang tidak memuaskan di fasilitas kesehatan atau masalah administratif.

Analisis sentimen terhadap berita yang berasal dari media terpercaya seperti Indozone, CNBC, dan media lain yang serupa merupakan metode efektif untuk memahami opini publik terhadap layanan BPJS. Melalui teknik web scraping untuk mengumpulkan data dan analisis sentimen untuk menganalisis opini yang terkandung di dalamnya, penelitian ini dapat memberikan wawasan tentang persepsi masyarakat terhadap BPJS tanpa dipengaruhi bias opini yang mungkin muncul di media sosial. Analisis ini penting untuk menilai tingkat kepuasan atau ketidakpuasan masyarakat terhadap BPJS secara objektif berdasarkan liputan media.

Penelitian sebelumnya telah menunjukkan bahwa persepsi masyarakat terhadap layanan publik dapat diukur secara efektif melalui analisis sentimen teks berita. Yuliantini dan Sukarno (2023) menunjukkan bahwa analisis sentimen pada layanan publik seperti BPJS dapat memberikan wawasan yang mendalam terkait persepsi masyarakat terhadap inovasi pelayanan dan efektivitas kinerja BPJS. Dalam studi ini, publikasi media tentang layanan BPJS dianalisis untuk mengidentifikasi sentimen positif, negatif, atau netral yang berkembang di masyarakat (Yuliantini & Sukarno, 2023).

Selain itu, studi oleh Kusumaningtyas et al. (2023) menemukan bahwa analisis terhadap berita terkait layanan kesehatan mental yang didukung BPJS mengungkapkan adanya respons negatif publik terhadap keterbatasan fasilitas kesehatan yang ditanggung BPJS. Analisis sentimen disini membantu mengidentifikasi aspek-aspek layanan BPJS yang perlu diperbaiki dari sudut pandang masyarakat (Kusumaningtyas et al., 2023).

Priatna dan rekan-rekannya (2023) juga menemukan bahwa peristiwa pelanggaran data BPJS oleh peretas memunculkan sentimen negatif yang signifikan terhadap BPJS di media. Analisis ini menunjukkan pentingnya kepercayaan publik terhadap BPJS

sebagai lembaga yang memiliki akses ke data pribadi masyarakat luas. Hal ini menunjukkan bahwa keamanan data merupakan salah satu faktor kunci dalam membangun persepsi positif terhadap BPJS (Priatna et al., 2023).

Penelitian lain yang dilakukan oleh Inan dan Juita (2024) mencerminkan respons masyarakat yang terbagi terhadap layanan BPJS, di mana sentimen negatif cenderung muncul dari liputan tentang ketidakpuasan terkait aksesibilitas dan keterbatasan fasilitas kesehatan BPJS. Studi ini penting karena menunjukkan persepsi masyarakat yang tidak sepenuhnya positif terhadap layanan BPJS, terutama pada saat layanan dianggap kurang memenuhi harapan (Inan & Juita, 2024).

2. RUMUSAN MASALAH

Berdasarkan latar belakang tersebut maka rumusan masalah dari penelitian ini adalah sebagai berikut.

2.1. Bagaimana persepsi publik terhadap layanan BPJS Kesehatan yang tercermin dalam berita yang dimuat di media online?

2.2. Bagaimana tren sentimen publik terhadap BPJS Kesehatan berubah seiring dengan waktu?

2.3. Bagaimana pengelompokan berita mengenai BPJS Kesehatan?

3. AKUISISI DATA

Pengumpulan data (data acquisition) dilakukan untuk membentuk kumpulan data (dataset) yang diperlukan dalam pengembangan teknologi NLP untuk analisis sentimen terhadap layanan BPJS Kesehatan. Data yang diperlukan dalam konteks ini adalah konten berita yang berasal dari portal berita atau situs web terpercaya yang membahas topik layanan BPJS Kesehatan dalam bahasa Indonesia. Proses pengumpulan data dilakukan menggunakan metode web scraping dengan menggunakan library Python, yaitu Selenium, Dateparser, dan Newspaper. Proses akuisisi data dilakukan melalui dua tahap:

3.1. Pengidentifikasian Sumber Data

Menentukan portal berita terpercaya yang relevan untuk topik BPJS Kesehatan, seperti Indozone, CNBC, AntaraNews, detik, kompas, CNN serta mengidentifikasi artikel-artikel yang memuat informasi penting mengenai layanan BPJS Kesehatan dengan cara mengambil berita dengan judul yang memuat kata BPJS

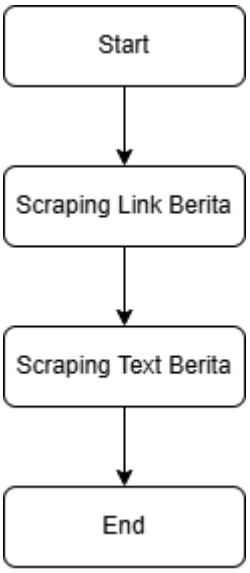
3.2. Ekstraksi Data

Ekstraksi Data dilakukan untuk membentuk kumpulan data (dataset) yang diperlukan dalam pengembangan teknologi NLP untuk analisis sentimen terhadap layanan BPJS Kesehatan. Data yang diperlukan dalam konteks ini adalah konten berita yang berasal dari portal berita atau situs web terpercaya yang membahas topik layanan BPJS Kesehatan dalam bahasa Indonesia. Proses

pengumpulan data dilakukan menggunakan metode web scraping dengan menggunakan library Python, yaitu Selenium, Dateparser, dan Newspaper.

Gambar 3.1 menunjukkan alur pengumpulan data yang dilakukan dalam penelitian ini. Langkah pertama dalam proses ini adalah melakukan pengambilan atau scraping link dari artikel yang relevan di setiap portal berita. Dengan menggunakan Selenium, kita dapat mengakses halaman web dan melakukan navigasi untuk menemukan artikel yang sesuai. Setiap link yang berhasil diambil akan disimpan dalam CSV.

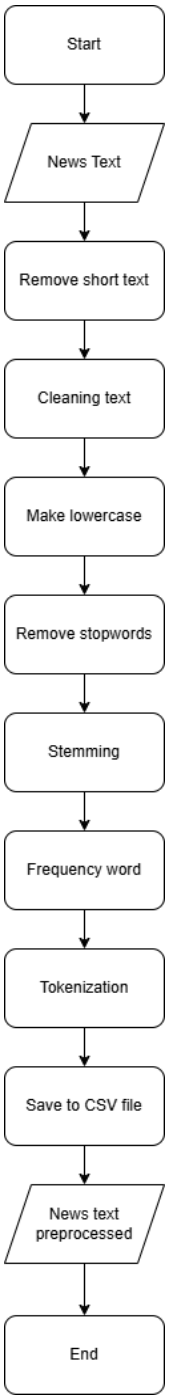
Setelah semua link terkumpul, langkah selanjutnya adalah mengekstrak data dari setiap artikel yang terdapat pada link tersebut. Menggunakan library Newspaper, kita dapat mengekstrak konten teks serta informasi penting lainnya, seperti tanggal publikasi dari artikel. Konten yang diekstrak kemudian akan disimpan dalam format CSV atau Excel, sehingga memudahkan analisis lebih lanjut dan pengolahan data menggunakan teknik NLP. Dengan cara ini, kita dapat mengumpulkan dan menganalisis opini publik terhadap layanan BPJS Kesehatan secara sistematis dan terstruktur.



Gambar 3.1 ***Alur Pengumpulan Data***

4. PRAPROSES DATA

Dataset yang telah berhasil dibuat dari hasil scraping teks berita artikel di internet masih bersifat mentah dan tidak terstruktur. Oleh karena itu, dilakukan praproses data (data preprocessing) untuk menjaga akurasi dan konsistensi struktur data sebelum dilakukan pemodelan. Kombinasi library Python seperti Pandas, String, Regex, NLTK, dan Sastrawi digunakan dalam proses ini.



Gambar 4.1 ***Alur Praproses Data***

Tahapan pertama adalah menghapus teks yang terlalu pendek untuk memastikan hanya teks yang relevan dan memiliki panjang tertentu yang diproses lebih lanjut. Setelah itu, dilakukan pembersihan teks (cleaning text) untuk menghilangkan karakter tidak diperlukan seperti angka, tanda baca, simbol, dan tautan, yang dapat mengganggu analisis. Selanjutnya, teks diubah menjadi huruf kecil (lowercase) untuk menjaga konsistensi penulisan kata yang sama, menghindari perbedaan akibat penggunaan huruf kapital.

Kemudian, dilakukan penghapusan stopwords atau kata-kata umum yang sering muncul tetapi tidak memberikan makna penting, seperti "dan", "di", "yang", dan lain-lain. Setelah itu, teks melalui proses stemming, yaitu mengubah kata-kata menjadi bentuk dasar (root word) untuk menyederhanakan teks. Hal ini memastikan bahwa kata-kata dengan bentuk yang berbeda tetapi makna yang sama, seperti "berlari" dan "berlari-lari", dihitung sebagai satu kata.

Tahap berikutnya adalah menghitung frekuensi kata (word frequency), yang membantu memahami seberapa sering kata-kata tertentu muncul dalam teks. Setelah itu, dilakukan tokenisasi (tokenization), yaitu pemecahan teks menjadi unit-unit yang lebih kecil seperti kata-kata atau frasa, yang merupakan langkah penting untuk analisis teks.

Setelah semua tahapan tersebut selesai, hasilnya disimpan ke dalam file CSV agar dapat digunakan dalam analisis atau pemodelan machine learning selanjutnya. Pada akhir proses ini, teks berita yang telah di praproses siap untuk digunakan dalam tahap analisis atau pemodelan, dengan data yang lebih bersih, terstruktur, dan siap untuk memberikan hasil yang lebih akurat.

5. DEFINISI DATASET

Pada bagian ini akan dijelaskan struktur beserta temuan *dataset* yang telah didapatkan melalui proses akuisisi dan praproses data.

5.1. Dataset Link Artikel Berita

Dataset ini merupakan data tautan (link) artikel berita hasil dari melakukan akuisisi data dengan metode web scraping pada situs-situs seperti Indozone, CNBC, AntaraNews, Detik, Kompas, CNN, serta Google News dengan topik BPJS. Dataset ini terdiri dari beberapa kolom dengan definisi setiap kolom tertera pada Tabel 5.1.

Tabel 5.1 ***Dataset Link Artikel Berita***

No.	Kolom	Definisi
1.	*title* / Judul	Judul dari artikel berita.
2.	*url / Link*	Tautan (*link*) dari artikel berita.
3.	*date / Tanggal*	Tanggal publikasi artikel berita.

Proses akuisisi data dan pembuatan dataset link artikel berita ini berhasil mengumpulkan lebih dari 9.000 link artikel terkait topik BPJS dari berbagai sumber tersebut.

5.2. Dataset Konten Artikel Berita

Dataset ini berisi konten artikel berita dengan topik BPJS, beserta beberapa data lainnya, sebagai hasil dari melakukan akuisisi konten berita dengan metode web *scraping* pada seluruh link artikel yang ada pada *dataset* link artikel berita. Dataset ini terdiri dari beberapa kolom dengan definisi setiap kolom tertera pada Tabel 5.2.

Tabel 5.2 ***Dataset Konten Artikel Berita***

No.	Kolom	Definisi
1.	judul	Judul artikel berita.
2.	*link*	Tautan (*link*) dari artikel berita.
3.	*text_berita*	Isi konten artikel berita
4.	tanggal	Tanggal publikasi artikel berita.
5.	*tags*	Label atau kategori artikel berita.
6.	portal_berita	Sumber media artikel berita.

Proses akuisisi data dan pembuatan dataset konten artikel berita berhasil mengumpulkan total kurang lebih sebanyak 5.000 konten berita dari lebih dari 9.000 link artikel terkait topik BPJS. Sehingga, terdapat sejumlah konten berita yang tidak berhasil diakuisisi oleh library Selenium.

5.3. Dataset Hasil Praproses

Dataset ini memuat teks artikel dari berbagai situs berita atau platform serupa yang membahas topik BPJS dalam bahasa Indonesia. Artikel-artikel ini telah melalui proses pra proses data, termasuk pembersihan teks, stemming, serta perhitungan jumlah kata. Dataset ini memiliki beberapa kolom, dengan definisi masing-masing kolom dijelaskan pada Tabel 5.3

Tabel 5.3 ***Dataset Konten Artikel Berita Hasil Praproses***

No.	Kolom	Definisi
1.	judul	Judul artikel berita.
2.	*link*	Tautan (*link*) dari artikel berita.
3.	*text_berita*	Isi konten artikel berita
4.	tanggal	Tanggal publikasi artikel berita.
5.	*tags*	Label atau kategori artikel berita.
6.	portal_berita	Sumber media artikel berita.
7.	*text_berita_clean*	Isi konten artikel berita yang telah di praproses

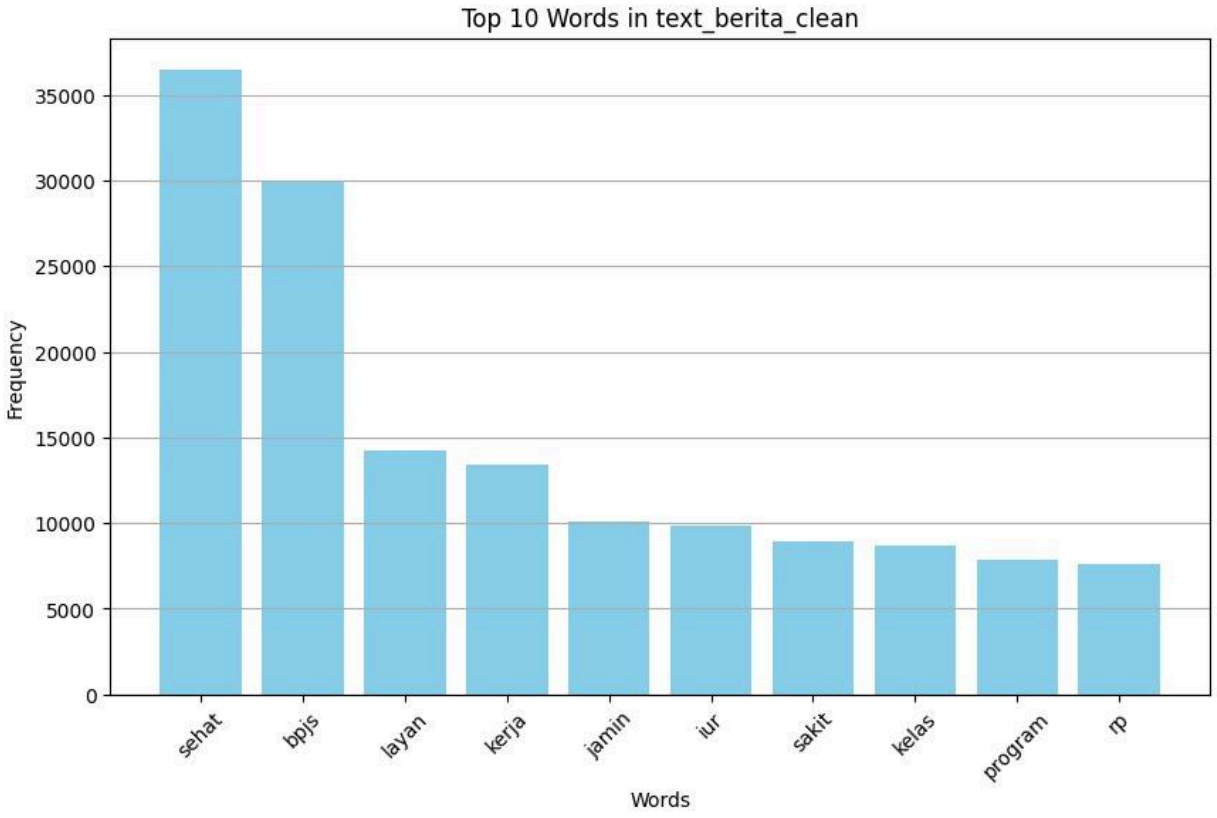
Hasil dataset dari praproses data menghasilkan data artikel berita sejumlah kurang lebih 5.000, tanpa ada pengurangan signifikan dari jumlah awal. Seluruh artikel berhasil diproses karena tidak ada artikel yang memiliki isi konten yang terlalu sedikit, dan setiap konten terkait dengan topik BPJS sesuai dengan hasil ekstraksi konten menggunakan library Newspaper.

Dengan membandingkan jumlah data pada setiap tahap pra proses, kita dapat menilai seberapa efektif proses yang dilakukan dan memastikan bahwa data yang dihasilkan telah bersih serta siap digunakan untuk langkah berikutnya dalam pengembangan teknologi NLP, seperti *Information Retrieval*. Tabel 4.4 dan Gambar 4.1 menampilkan contoh hasil analisis frekuensi kata yang ditemukan dalam dataset.

Tabel 5.4 ***Top 10 Frekuensi Kata pada Dataset***

Kata	Frekuensi
sehat	35596
bpjs	22373
serta	18830

kerja	13305
jamin	10102
layan	9014
kelas	8511
program	7840
rp	7625
perintah	7228



Gambar 5.1 ***Bar-chart 10 Kata yang Paling Sering Muncul***

Dari hasil most frequent words pada Gambar diatas, terlihat bahwa kata-kata yang paling sering muncul dalam dataset artikel berita adalah "sehat", "bpjs", dan lainnya. Hal ini menunjukkan bahwa pra proses data telah berhasil membersihkan dataset dari *stopwords*, tanda baca, dan elemen tidak relevan lainnya. Oleh karena itu, dataset ini sudah siap untuk dimanfaatkan pada tahap berikutnya, yaitu pemodelan dan analisis data.

6. ANALISIS DATA DAN PEMODELAN

6.1. Clustering

Clustering adalah teknik dalam analisis data yang digunakan untuk mengelompokkan sekumpulan objek atau data berdasarkan kesamaan tertentu. Dalam konteks analisis sentimen terhadap berita BPJS Kesehatan, clustering memungkinkan kita untuk mengelompokkan artikel-artikel berita ke dalam kategori yang relevan berdasarkan konten dan tema yang diangkat. Dengan demikian, setiap kelompok dapat memberikan wawasan yang lebih mendalam tentang persepsi publik terhadap layanan BPJS Kesehatan. Metode clustering yang akan dibahas selanjutnya meliputi TF-IDF, Bag of Words (BoW), dan Cosine Similarity, yang masing-masing memiliki pendekatan berbeda dalam mengolah dan menganalisis data teks**.**

a. Metode Clustering

Dalam laporan ini, tiga metode utama digunakan untuk melakukan clustering pada dataset berita BPJS Kesehatan:

a. TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization
b. Bag of Words (BoW) Vectorization
c. Cosine Similarity

Ketiga metode ini diimplementasikan menggunakan algoritma K-Means untuk mengelompokkan artikel berita berdasarkan kesamaan kontennya. Pemilihan jumlah cluster optimal dilakukan menggunakan metode Silhouette Score dan WCSS (Within-Cluster Sum of Squares).

TF-IDF Vectorization memberikan bobot pada kata-kata berdasarkan frekuensi kemunculannya dalam dokumen dan inverse frekuensi dokumen. Metode ini efektif untuk menangkap kata-kata yang penting dan unik dalam setiap dokumen.

Bag of Words (BoW) Vectorization menghitung frekuensi kemunculan setiap kata dalam dokumen tanpa mempertimbangkan urutan kata. Metode ini sederhana namun efektif untuk menangkap konten utama dari dokumen.

Cosine Similarity mengukur kesamaan antara dua vektor dengan menghitung cosinus sudut antara mereka. Dalam konteks text clustering, ini digunakan untuk mengukur kesamaan antar dokumen berdasarkan representasi vektor mereka.

b. Implementasi Clustering

1) Persiapan Data

Implementasi clustering dimulai dengan mempersiapkan dataset. Dataset berisi artikel berita tentang BPJS Kesehatan yang telah melalui tahap preprocessing. Kolom yang digunakan untuk clustering adalah 'text_berita_clean', yang berisi teks berita yang telah dibersihkan. Library yang digunakan meliputi pandas, numpy, sklearn, dan matplotlib.

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer,
CountVectorizer
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score,
calinski_harabasz_score, davies_bouldin_score
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
```

Dataset dimuat menggunakan pandas dan kolom 'text_berita_clean' diekstrak:

```
file_path = '/content/data_ready_final.xlsx'
df = pd.read_excel(file_path)
documents = df['text_berita_clean'].dropna()
```

2) Fungsi Clustering

Fungsi perform_clustering didefinisikan untuk melakukan clustering dan menghitung metrik evaluasi. Fungsi ini menggunakan KMeans untuk clustering dan menghitung Silhouette Score, Calinski-Harabasz Score, dan Davies-Bouldin Score untuk evaluasi:

```
def perform_clustering(X, num_clusters):
    kmeans = KMeans(n_clusters=num_clusters, random_state=42,
n_init=10)
    kmeans.fit(X)
    cluster_labels = kmeans.labels_
    silhouette_avg = silhouette_score(X, cluster_labels)
    CH = calinski_harabasz_score(X.toarray(), cluster_labels)
    DB = davies_bouldin_score(X.toarray(), cluster_labels)
    return cluster_labels, silhouette_avg, CH, DB
```

3) Penentuan Jumlah Cluster Optimal

Dalam laporan ini, penentuan jumlah cluster optimal dilakukan menggunakan tiga metode utama yang dikombinasikan dengan algoritma KMeans. Metode-metode tersebut adalah:

a. TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization
b. Bag of Words (BoW) Vectorization

c. Cosine Similarity

Ketiga metode ini digunakan untuk mengubah teks berita BPJS Kesehatan menjadi representasi vektor, yang kemudian dikelompokkan menggunakan algoritma KMeans. Untuk menentukan jumlah cluster optimal, dua teknik evaluasi utama digunakan:

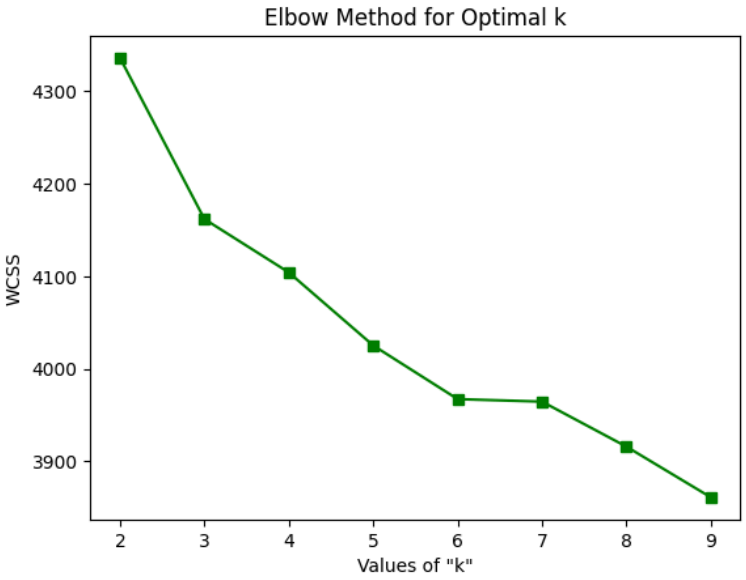
Metode WCSS (Within-Cluster Sum of Squares) menilai performa dengan mengukur seberapa kompak cluster yang terbentuk. Sedangkan silhouette Score menilai performa dengan mengukur seberapa baik objek dalam cluster terpisah dari cluster lainnya.

4) Implementasi Elbow Method untuk Menentukan Jumlah Cluster Optimal

a) TF-IDF Vectorization

Pada metode TF-IDF, data teks diubah menjadi representasi vektor menggunakan TfidfVectorizer. Untuk menentukan jumlah cluster optimal, digunakan metode Elbow berdasarkan nilai WCSS.

```
# Convert the text data to TF-IDF features
vectorizer = TfidfVectorizer(max_features=1000)
X_tfidf = vectorizer.fit_transform(documents).toarray()

# Determining the maximum number of clusters using a simple
method
wcss = {}
for k in range(2, 10):
    model = KMeans(n_clusters=k)
    model.fit(X_tfidf)
    wcss[k] = model.inertia_

# Plotting the WCSS values to find the elbow point
plt.plot(list(wcss.keys()), list(wcss.values()), 'gs-')
plt.xlabel('Values of "k"')
plt.ylabel('WCSS')
plt.title('Elbow Method for Optimal k (TF-IDF)')
plt.show()
```



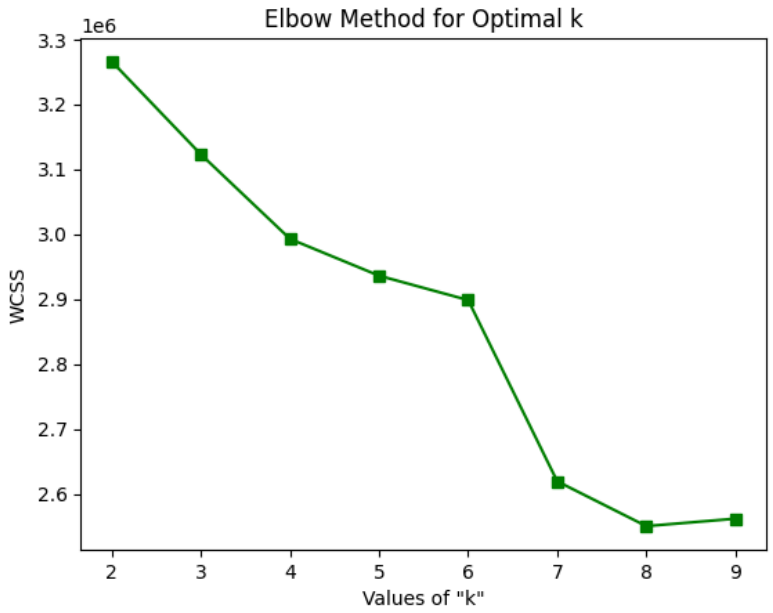
Gambar 6.1 Grafik Elbow Method TF-IDF Vectorization

Hasil TF-IDF bisa dilihat berdasarkan grafik Elbow Method, titik elbow terlihat pada k = 3, sehingga jumlah cluster optimal untuk TF-IDF adalah 3

b) Bag of Words (BoW) Vectorization

Pada metode BoW, data teks diubah menjadi representasi vektor menggunakan CountVectorizer. Sama seperti TF-IDF, jumlah cluster optimal ditentukan menggunakan Elbow Method berdasarkan nilai WCSS.

```
# Convert the text data to Bag-of-Words features
vectorizer = CountVectorizer(max_features=1000)
X_bow = vectorizer.fit_transform(documents).toarray()

# Determining the maximum number of clusters using a simple
method
wcss = {}
for k in range(2, 10):
    model = KMeans(n_clusters=k)
    model.fit(X_bow)
    wcss[k] = model.inertia_

# Plotting the WCSS values to find the elbow point
plt.plot(list(wcss.keys()), list(wcss.values()), 'gs-')
plt.xlabel('Values of "k"')
plt.ylabel('WCSS')
plt.title('Elbow Method for Optimal k (BoW)')
plt.show()
```



Gambar 6.2 Grafik Elbow Method Bag of Words (BoW) Vectorization

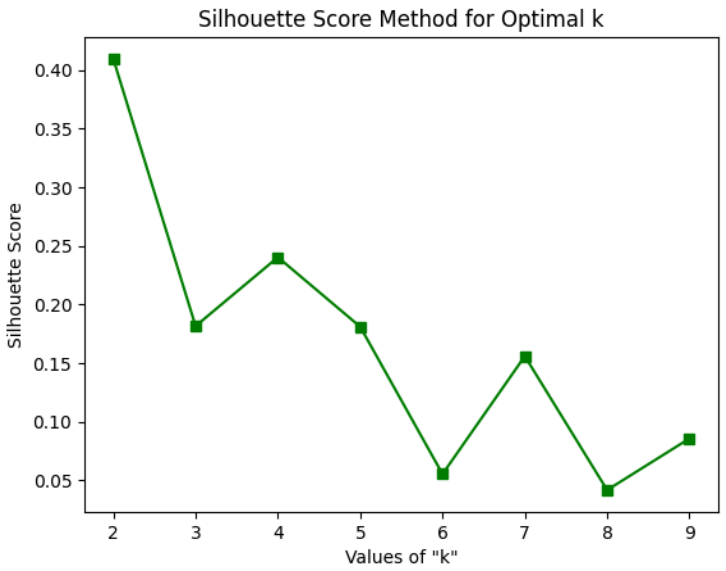
Hasil BoW dapat diambil berdasarkan grafik Elbow Method, titik elbow terlihat pada k = 2, sehingga jumlah cluster optimal untuk BoW adalah 2.

c) Cosine Similarity Clustering

Pada metode Cosine Similarity, representasi BoW digunakan untuk menghitung matriks kesamaan cosinus antar dokumen. Jumlah cluster optimal ditentukan menggunakan Silhouette Score.

```
# Convert the text data to Bag of Words (BOW) features
vectorizer = CountVectorizer(max_features=1000)
X_bow = vectorizer.fit_transform(documents)

# Calculate cosine similarity matrix
cosine_sim_matrix = cosine_similarity(X_bow)

# Determine optimal k using silhouette score
silhouette_scores = []
for k in range(2, 10):
    model = KMeans(n_clusters=k)
    labels = model.fit_predict(cosine_sim_matrix)
    silhouette_avg = silhouette_score(cosine_sim_matrix,
labels, metric="cosine")
    silhouette_scores.append((k, silhouette_avg))

# Extracting the optimal k value based on the maximum
silhouette score
optimal_k_cosine = max(silhouette_scores, key=lambda x:
x[1])[0]
```

```
# Plotting silhouette scores to find the optimal number of
clusters
plt.plot([k for k, _ in silhouette_scores], [score for _,
score in silhouette_scores], 'gs-')
plt.xlabel('Values of "k"')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score Method for Optimal k (Cosine
Similarity)')
plt.show()

print(f"The optimal number of clusters based on silhouette
score is: {optimal_k_cosine}")
```



Gambar 6.3 Grafik Elbow Method Cosine Similarity Clustering

Hasil Cosine Similarity dapat disimpulkan berdasarkan grafik Silhouette Score, nilai tertinggi ditemukan pada k = 3, sehingga jumlah cluster optimal untuk Cosine Similarity adalah 3

c. Hasil dan Evaluasi Clustering

Setelah melakukan clustering menggunakan metode KMeans dengan tiga pendekatan vektorisasi teks (TF-IDF, Bag of Words, dan Cosine Similarity), langkah selanjutnya adalah memvisualisasikan hasil clustering. Visualisasi ini dilakukan untuk memahami distribusi dokumen dalam setiap cluster yang terbentuk. Salah satu teknik yang digunakan untuk visualisasi adalah PCA (Principal Component Analysis), yang mereduksi dimensi data menjadi dua komponen utama agar dapat divisualisasikan dalam bentuk scatter plot.

1) Performa dan Visualisasi Clustering dengan TF-IDF

Tabel 6.1 ***Performa Clustering TF-IDF***

Silhouette Score	0.0237
Calinski-Harabasz Score	133.16
Davies-Bouldin Score	5.28

Kinerja clustering menggunakan metode TF-IDF menunjukkan Silhouette Score yang sangat rendah, yang mengindikasikan bahwa dokumen dalam cluster tidak terpisah dengan baik dari cluster lainnya. Calinski-Harabasz Score yang rendah juga menunjukkan bahwa cluster yang terbentuk tidak kompak, sedangkan Davies-Bouldin Score yang tinggi menandakan bahwa ada kesamaan yang signifikan antar cluster.

Hasil Top Terms di Setiap Cluster:

Tabel 6.2 ***Top Terms Cluster TF-IDF***

Cluster 0	ketenagakerjaan, kerja, rp, juta, bpjs, jht, terima, serta, usaha
Cluster 1	kelas, iur, sehat, standar, kris, inap, rawat, ruang, serta, rp
Cluster 2	sehat, bpjs, layan, serta, sakit, jkn, program, rumah, jamin, masyarakat

Cluster 0 didominasi oleh kata-kata seperti ketenagakerjaan, kerja, dan bpjs, yang menunjukkan fokus pada aspek ketenagakerjaan dalam konteks BPJS. Sementara itu, Cluster 1 terdiri dari kata-kata seperti kelas, iur, dan standar, yang mencerminkan topik terkait layanan kesehatan dan standar. Di sisi lain, Cluster 2 menampilkan kata-kata seperti sehat, bpjs, dan masyarakat, yang menunjukkan perhatian pada layanan kesehatan secara umum.

Jumlah Dokumen di Tiap Cluster:

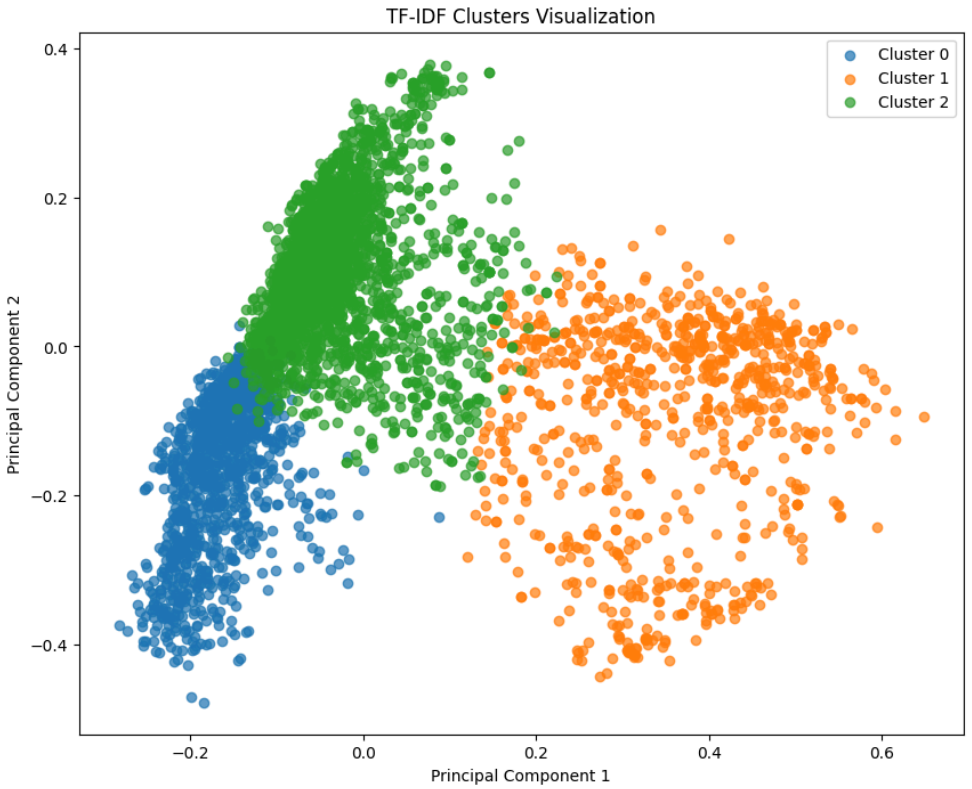
Tabel 6.3 ***Jumlah Dokumen tiap Cluster TF-IDF***

Cluster 0	1243 dokumen
Cluster 1	858 dokumen
Cluster 2	2934 dokumen

Jumlah dokumen yang paling banyak terdapat di Cluster 2, yang mungkin mencerminkan bahwa banyak berita berkaitan dengan layanan kesehatan BPJS yang lebih umum diterima oleh masyarakat.

Dalam visualisasi pada metode TF-IDF, data teks diubah menjadi representasi vektor menggunakan *TfidfVectorizer*. Setelah melakukan clustering, hasilnya divisualisasikan menggunakan PCA.

```
# Visualize clusters using PCA (for TF-IDF)
pca_tfidf =
PCA(n_components=2).fit_transform(X_tfidf.toarray())
plt.figure(figsize=(10, 8))
for i in range(3):  # Assuming 3 clusters based on optimal k
    plt.scatter(pca_tfidf[labels_tfidf == i][:, 0],
pca_tfidf[labels_tfidf == i][:, 1], label=f'Cluster {i}',
alpha=0.7)
plt.title('TF-IDF Clusters Visualization')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend()
plt.show()
```



Gambar 6.4 Visualisasi Cluster TF-IDF

Pada visualisasi ini, setiap titik mewakili sebuah dokumen dalam dataset, dan warnanya menunjukkan cluster tempat dokumen tersebut berada. Hasil visualisasi TF-IDF menunjukkan bagaimana dokumen-dokumen dikelompokkan berdasarkan kesamaan kata-kata unik yang jarang muncul di seluruh dokumen.

2) Performa dan Visualisasi Clustering dengan Bag of Words (BoW)

Tabel 6.4 ***Performa Clustering Bag of Word***

Silhouette Score	0.094
Calinski-Harabasz Score	295.11
Davies-Bouldin Score	2.82

Hasil clustering menggunakan metode BoW menunjukkan peningkatan dibandingkan dengan TF-IDF, dengan Silhouette Score yang lebih tinggi, yang berarti cluster lebih terpisah. Calinski-Harabasz Score juga lebih baik, menunjukkan bahwa cluster lebih kompak, sementara Davies-Bouldin Score yang lebih rendah menunjukkan perbedaan antar cluster yang lebih jelas..

Hasil *Top Terms* di Setiap Cluster:

Tabel 6.5 ***Top Terms Cluster Bag of Word***

Cluster 0	bpjs, sehat, kerja, serta, ketenagakerjaan, jamin, program, rp, perintah, layan
Cluster 1	kelas, sehat, iur, bpjs, serta, standar, rp, rawat, perintah, ruang
Cluster 2	sehat, bpjs, layan, serta, sakit, jkn, jamin, program, masyarakat, tingkat

Cluster 0, yang terdiri dari kata-kata seperti bpjs, sehat, dan kerja, menandakan fokus pada program BPJS dan kesehatan. Sementara itu, Cluster 1 menunjukkan perhatian pada aspek layanan kesehatan tertentu dengan kata-kata seperti kelas, iur, dan rawat. Di sisi lain, Cluster 2 mencerminkan fokus pada kebijakan kesehatan masyarakat, yang ditunjukkan oleh kata-kata seperti sehat, jkn, dan masyarakat.

Jumlah Dokumen di Tiap Cluster:

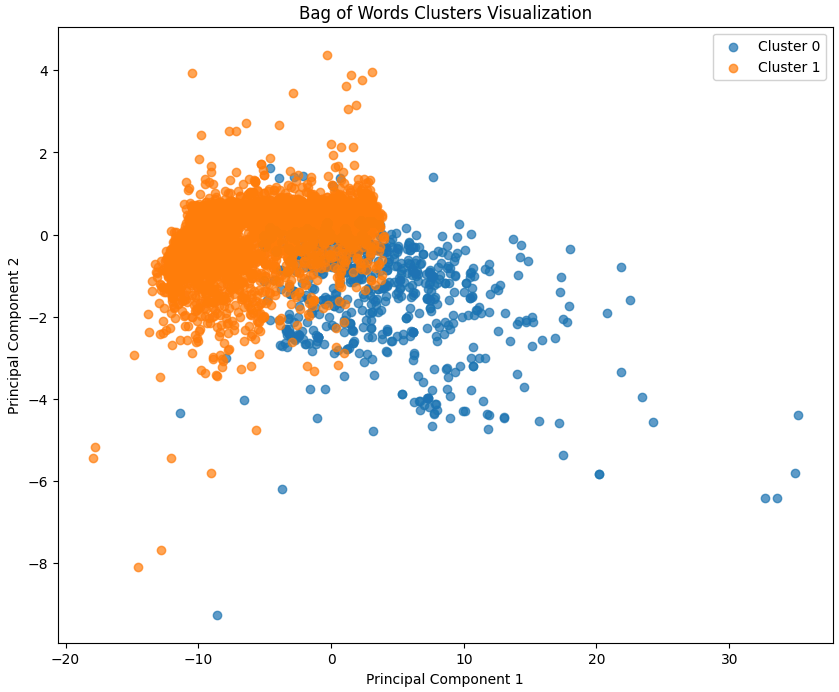
Tabel 6.6 ***Jumlah Dokumen tiap Cluster Bag of Word***

Cluster 0	686 dokumen
Cluster 1	3183 dokumen
Cluster 2	1166 dokumen

Jumlah dokumen terbanyak berada di Cluster 1, mencerminkan banyaknya berita terkait standar layanan kesehatan.

Dalam visualisasi pada metode BoW, data teks diubah menjadi representasi vektor menggunakan CountVectorizer. Setelah clustering dilakukan dengan KMeans, hasilnya divisualisasikan dengan PCA.

```
# Visualize clusters using PCA (for BoW)
pca_bow = PCA(n_components=2).fit_transform(X_bow.toarray())
plt.figure(figsize=(10, 8))
for i in range(2):  # Assuming 2 clusters based on optimal k
    plt.scatter(pca_bow[labels_bow == i][:, 0],
pca_bow[labels_bow == i][:, 1], label=f'Cluster {i}',
alpha=0.7)
plt.title('Bag of Words Clusters Visualization')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend()
plt.show()
```



Gambar 6.5 Visualisasi Cluster Bag of Words

Visualisasi BoW menunjukkan bagaimana dokumen dikelompokkan berdasarkan frekuensi kemunculan kata-kata tanpa memperhatikan urutan atau bobotnya. Dengan dua cluster optimal yang ditemukan dari metode Elbow, distribusi dokumen terlihat lebih terpisah dibandingkan dengan TF-IDF.

3) Performa dan Visualisasi Clustering dengan Cosine Similarity

Tabel 6.7 ***Performa Clustering Cosine Similarity***

Silhouette Score	0.246
Calinski-Harabasz Score	2757.22
Davies-Bouldin Score	1.52

Metode Cosine Similarity menunjukkan kinerja terbaik dari ketiga metode, dengan Silhouette Score tertinggi yang menunjukkan pemisahan cluster yang baik. Calinski-Harabasz Score yang sangat tinggi menandakan bahwa cluster sangat kompak, sementara Davies-Bouldin Score terendah menunjukkan bahwa perbedaan antar cluster sangat jelas.

Hasil *Top Terms* di Setiap Cluster:

Tabel 6.8 ***Top Terms Cluster Cosine Similarity***

Cluster 0	base, batukbatuk, autoimun, absurd, abraham, abk, abrisam, diagnosed, batulicin, dmengerti
Cluster 1	dengan, african, elektro, anafilaksis, fajar, abses, abpd, buyar, argentina, dulu
Cluster 2	diskusi, alrahman, anafilaksis, anastomosis, depresi, elektro, direct, desa, eksploitasi, cso

Cluster 0 terdiri dari kata-kata seperti base, autoimun, dan absurd yang menunjukkan topik-topik kesehatan spesifik. Sementara itu, Cluster 1 mencakup kata-kata seperti dengan, elektro, dan anafilaksis yang mencerminkan isu medis tertentu. Di sisi lain, Cluster 2 berisi kata-kata seperti diskusi, depresi, dan eksploitasi yang menunjukkan tema-tema yang lebih luas dalam konteks kesehatan. Dengan demikian, setiap cluster memberikan wawasan yang berbeda mengenai fokus dan isu yang diangkat dalam berita terkait kesehatan.

Jumlah Dokumen di Tiap Cluster:

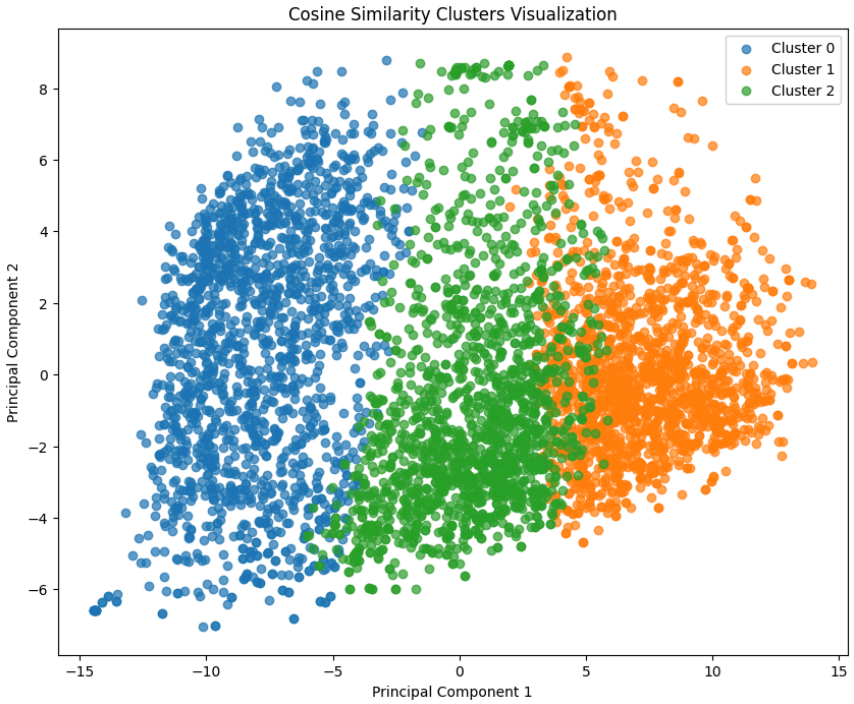
Tabel 6.9 ***Jumlah Dokumen tiap Cluster Cosine Similarity***

Cluster 0	1648 dokumen
Cluster 1	1669 dokumen
Cluster 2	1718 dokumen

Distribusi jumlah dokumen relatif merata di antara cluster, mencerminkan keberagaman topik dalam berita terkait BPJS Kesehatan.

Dalam visualisasi pada metode Cosine Similarity, representasi BoW digunakan untuk menghitung matriks kesamaan cosinus antar dokumen. Setelah clustering dilakukan dengan KMeans, hasilnya divisualisasikan menggunakan PCA.

```
# Visualize clusters using PCA (for Cosine Similarity)
pca_cosine =
PCA(n_components=2).fit_transform(cosine_sim_matrix)
plt.figure(figsize=(10, 8))
for i in range(3):  # Assuming 3 clusters based on optimal k
    plt.scatter(pca_cosine[labels_cosine == i][:, 0],
pca_cosine[labels_cosine == i][:, 1], label=f'Cluster {i}',
alpha=0.7)
plt.title('Cosine Similarity Clusters Visualization')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend()
plt.show()
```



Gambar 6.6 Visualisasi Cosine Similarity

Visualisasi ini menunjukkan bagaimana dokumen dikelompokkan berdasarkan kesamaan semantik antar dokumen yang diukur dengan Cosine Similarity. Cluster yang terbentuk lebih jelas terpisah dibandingkan dengan metode lainnya karena Cosine Similarity mampu menangkap hubungan semantik antar dokumen lebih baik.

4) Kesimpulan

Dari hasil analisis kinerja clustering menggunakan ketiga metode (TF-IDF, BoW, dan Cosine Similarity), dapat disimpulkan bahwa metode Cosine Similarity memberikan hasil performa terbaik dalam hal pemisahan dan kompaknya cluster. Sementara itu, meskipun BoW memberikan hasil clustering yang kurang memuaskan, analisis top terms dan distribusi jumlah dokumen lebih sesuai dan dapat diterima tentang tema-tema utama dalam berita terkait BPJS Kesehatan. Sehingga clustering BoW ini akan digunakan untuk melihat distribusi banyak dokumen pada analisis sentimen ini.

Visualisasi setiap metode memberikan gambaran yang jelas mengenai bagaimana dokumen dikelompokkan berdasarkan kesamaan konten mereka, dengan Cosine Similarity menghasilkan visualisasi paling terpisah dan kohesif dari ketiga metode tersebut.

DAFTAR PUSTAKA

Kusumaningtyas, K., Habibi, M., Dwijayanti, I., & Sumiyarini, R. (2023). Tweet analysis of mental illness using K-means clustering and support vector machine. *Telematika: Jurnal Informatika dan Teknologi Informasi*, 20(3), 295-308.

Priatna, W., A'ini, E. N., Warta, J., Hidayat, A., & Lestari, T. S. (2023, December). Sentiment analysis of Bjorka hacker using the Naive Bayes and C. 45 algorithms. In *IAIC International Conference Series* (Vol. 4, No. 1, pp. 79-87).

Syafrizal, S., & Sari, F. P. (2015). Persepsi masyarakat pengguna Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan mandiri dalam pelayanan RSUD Lubuk Basung Kabupaten Agam (Doctoral dissertation, Riau University).

Wajdi, M. F., Inan, D. I., Juita, R., & Sanglise, M. (2024). Study on the quality of service of the mobile-based JKN application: A sentiment analysis approach. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 9(3), 1506-1517.

Yuliantini, L. S., & Sukarno, M. (2023, November). The Twitter sentiment analysis of public service innovation: Pandawa BPJS health service. In *Proceedings Universitas Muhammadiyah Yogyakarta Undergraduate Conference* (Vol. 3, No. 1, pp. 146-153).