LAPORAN AKHIR

Review Sentiment Analysis berbasis Lexicon Based menggunakan VADER

KELOMPOK 5

1. Lanyta Setyani Gunawan (672021047)
2. Axell Amadeus Siagian (672021213)
3. Rio Arya Andika (672021166)
4. Ibrahim (672021178)
5. Gilang Hidayatullah (672021171)
6. Muhammad Ja'far (672021150)

Dosen Pengampu : Hendry, S.Kom., M.Kom., Ph.D

PROGRAM STUDI S1 TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN SATYA WACANA

2024

1. Pendahuluan

1.1 Latar Belakang

Dalam dunia bisnis yang semakin kompetitif, Yelp sebagai platform yang menyediakan layanan untuk berbagai mitra seperti restoran, bar, hotel, dan lainnya, perlu terus mengembangkan strategi untuk meningkatkan nilai tambah dan menjaga daya saing di pasar. Salah satu langkah inovatif yang dapat dilakukan adalah dengan mengimplementasikan sistem penilaian performa berbasis Machine Learning untuk setiap mitranya. Menggunakan teknologi analisis sentimen terhadap ulasan pelanggan, Yelp dapat memberikan wawasan mendalam mengenai persepsi pelanggan terhadap layanan mitra. Analisis sentimen ini mampu mengidentifikasi dan mengkategorikan sentimen dari setiap ulasan sebagai positif, negatif, atau netral. Dengan informasi ini, Yelp dapat menyusun laporan performa secara berkala yang menggambarkan secara komprehensif bagaimana pelanggan menilai layanan mitra selama periode tertentu.

Setiap tiga bulan, Yelp akan mengumpulkan dan mengolah data ulasan pelanggan yang telah dianalisis. Laporan performa ini tidak hanya memberikan gambaran menyeluruh mengenai kepuasan pelanggan, tetapi juga memberikan umpan balik yang terstruktur dan berbasis data. Dengan adanya laporan ini, mitra akan mendapatkan wawasan berharga untuk mengevaluasi dan meningkatkan kualitas layanan mereka. Selain itu, Yelp juga dapat memperoleh keuntungan lebih lanjut dengan memberlakukan tarif untuk layanan penyediaan laporan performa tersebut. Melalui penerapan analisis sentimen dan penyusunan laporan performa yang komprehensif, Yelp tidak hanya membantu mitra dalam meningkatkan kualitas layanan, tetapi juga memperkuat posisinya sebagai platform yang inovatif dan berorientasi pada nilai tambah di pasar yang kompetitif.

1.2 Ruang Lingkup (Batasan)

- Menggunakan 1 dataset sekunder Yelp yang berjumlah 1.569.264 baris dan 10 kolom yang mencakup berbagai informasi terkait ulasan pelanggan.
- Rekomendasi menggunakan model VADER (Valence Aware Dictionary and sEntiment Reasoner) untuk analisis sentimen dengan fokus pada fitur utama business_id dan text.
- Sistem analisis sentimen akan diimplementasikan sebagai laporan performa berkala yang akan diberikan kepada mitra-mitra Yelp untuk membantu mereka dalam meningkatkan kualitas layanan.

1.3 Identifikasi menggunakan 4W Canvas (Who, What, Where, Why)

Para [stakeholders]	• **Yelp**: Platform penyedia layanan untuk berbagai mitra seperti restoran, bar, hotel, dan lainnya. • **Pelanggan**: Pengguna yang memberikan ulasan terhadap layanan mitra di platform Yelp. • **Mitra-mitra Yelp**: Seperti restoran, bar, hotel, dan lainnya. • **Tim internal Yelp**: Tim yang bekerja di Yelp dan bertanggung jawab dalam pengembangan dan pemeliharaan sistem.	WHO
Masalah yang dihadapi [isu, masalah, kebutuhan]	Yelp ingin meningkatkan nilai tambahnya di pasar yang kompetitif dengan memperbaiki performa mitranya.	WHAT
Ketika [konteks dan situasi]	Ketika sebuah platform seperti Yelp menghadapi persaingan yang semakin ketat di pasar, di mana kepuasan pelanggan menjadi kunci untuk mempertahankan dan menarik mitra baru.	WHERE
Solusi yang diharapkan	Mengembangkan sistem analisis sentimen terhadap ulasan pelanggan berbasis Machine Learning untuk memberikan laporan performa berkala kepada mitra-mitra Yelp.	WHY

1.4 Tujuan

Mengembangkan sistem analisis sentimen ulasan yang efektif, yang dapat memberikan laporan performa berkala kepada mitra berdasarkan sentimen dan perilaku pengguna, sehingga membantu mitra dalam meningkatkan kualitas layanan mereka dan meningkatkan nilai tambah Yelp di pasar yang kompetitif.

2. Data Collection and Data Understanding

2. 1 Deskripsi Dataset

Dataset "yelp_academic_dataset_review.json" terdiri dari 1569264 baris dan 10 kolom yang mencakup berbagai informasi terkait review.

Setiap baris mewakili satu review terhadap sebuah bisnis, sementara kolom-kolomnya adalah sebagai berikut:

1. user_id : ID unik akun pengguna
2. review_id : ID unik ulasan
3. stars : Jumlah bintang untuk ulasan
4. date : Tanggal ulasan
5. text : Isi ulasan
6. type : Jenis ulasan
7. business_id : ID unik sebuah bisnis
8. votes.funny : Jumlah orang yang menganggap ulasannya lucu
9. votes.useful : Jumlah orang yang menganggap ulasannya berguna
10. votes.cool : Jumlah orang yang menganggap ulasannya keren

	user_id	review_id	stars	date	text	type	business_id	votes.funny	votes.useful	votes.cool
0	Xqd0DzHaiyRqVH3WRG7hzg	15SdjuK7DmYqUAj6rjGowg	5	2007-05-17	dr. goldberg offers everything i look for in a...	review	vcNAWiLM4dR7D2nwwJ7nCA	0	2	1
1	H1kH6QZV7Le4zqTRNxoZow	RF6UnRTtG7tWMcrO2GEoAg	2	2010-03-22	Unfortunately, the frustration of being Dr. Go...	review	vcNAWiLM4dR7D2nwwJ7nCA	0	2	0
2	zvJCcrpm2yOZrxKffwGQLA	-TsVN230RCkLYKBeLsuz7A	4	2012-02-14	Dr. Goldberg has been my doctor for years and ...	review	vcNAWiLM4dR7D2nwwJ7nCA	0	1	1
3	KBLW4wJA_fwoWmMhiHRVOA	dNocEAyUucjT371NNND41Q	4	2012-03-02	Been going to Dr. Goldberg for over 10 years. ...	review	vcNAWiLM4dR7D2nwwJ7nCA	0	0	0
4	zvJCcrpm2yOZrxKffwGQLA	ebcN2aqmNUuYNoyvQErgnA	4	2012-05-15	Got a letter in the mail last week that said D...	review	vcNAWiLM4dR7D2nwwJ7nCA	0	2	1

Gambar 1. Dataset dengan 5 baris pertama

2. 2 Metode Pengumpulan Dataset

Dataset "yelp_academic_dataset_review.json" diambil dari platform Flearn yang diberikan. Dataset ini dikumpulkan dari sumber-sumber yang sah dan kemudian diunggah ke Kaggle untuk digunakan oleh para peneliti, ilmuwan data, dan pengembang perangkat lunak. Maka dari itu, jenis dataset ini adalah dataset sekunder. Dataset sekunder merupakan kumpulan data yang dikumpulkan oleh pihak lain atau untuk tujuan lain, dan kemudian digunakan kembali untuk analisis atau

penelitian baru. Dataset sekunder merupakan sumber data yang sangat berharga karena dapat memungkinkan kami untuk menjawab pertanyaan-pertanyaan baru atau menguji hipotesis tanpa harus melakukan pengumpulan data dari awal. Data ini bersumber dari scraping website YELP.

2. 3 Features yang akan digunakan

Kolom business_id digunakan sebagai penanda tempat atau toko yang diulas, sedangkan kolom text berisi isi ulasan dari pelanggan terhadap tempat atau toko tersebut. Kedua fitur ini dipilih untuk dianalisis karena business_id memungkinkan kita mengidentifikasi dan mengelompokkan ulasan berdasarkan tempat atau toko tertentu, sementara text menyediakan data yang diperlukan untuk melakukan analisis sentimen guna memahami pendapat dan pengalaman pelanggan. Berikut adalah *feature* yang digunakan:

- text
- business_id

2. 4 Parameter Tipe Data

No.	Parameter	Status	Tipe Data
1.	text	Input	Object
2.	Sentiment	Output	Object
3.	Score	Output	Object

3. Eksplorasi Data (Exploratory Data Analysis)

3.1 Menampilkan Ringkasan Deskriptif Kolom pada Dataset Film

Tahap pertama dalam EDA (Exploratory Data Analysis) adalah menampilan ringkasan deskriptif kolom. Tujuan dari tahap ini adalah memberikan gambaran komprehensif mengenai struktur dan karakteristik dataset yang sedang dianalisis. Gambar 2 menampilkan informasi penting mengenai setiap kolom dalam dataset, termasuk jumlah entri non-null, jenis data (Dtype), dan tipe informasi yang direpresentasikan oleh masing-masing kolom.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1569264 entries, 0 to 1569263
Data columns (total 10 columns):
 #   Column        Non-Null Count    Dtype
---  ------        --------------    -----
 0   user_id       1569264 non-null  object
 1   review_id     1569264 non-null  object
 2   stars         1569264 non-null  int64
 3   date          1569264 non-null  object
 4   text          1569264 non-null  object
 5   type          1569264 non-null  object
 6   business_id   1569264 non-null  object
 7   votes.funny   1569264 non-null  int64
 8   votes.useful  1569264 non-null  int64
 9   votes.cool    1569264 non-null  int64
dtypes: int64(4), object(6)
memory usage: 119.7+ MB
None
```

Gambar 2. Informasi Kolom Dataset

Pada dataset film, terdapat dua kategori utama jenis data, yaitu numerikal dan kategorikal. Jenis data numerikal mencakup data integer (int64) sedangkan, jenis data kategorikal direpresentasikan oleh data object.

Data integer (int64) merepresentasikan bilangan bulat, pada dataset ini adalah kolom “stars” yang mengidentifikasi berapa banyak bintang dalam review bisnis di dataset, kemudian diikuti vote.funny, votes.useful, votes.cool yang berisi jumlah orang yang menyetujui bahwa sebuah komentar tersebut lucu, berguna, atau keren. Data object merepresentasikan teks atau kombinasi teks dan angka, pada dataset ini adalah kolom "user_id”, “review_id”, “date”, “text”, “type”, “business_id”.

Selanjutnya, *non-null count* disini merepresentasikan jumlah entri dalam dataset yang memiliki nilai yang tidak kosong atau tidak null. Dengan kata lain, ini menunjukkan berapa banyak entri dalam setiap kolom yang memiliki nilai yang diisi.

Dalam analisis data, informasi tentang non-null count sangat penting karena ini memberikan gambaran tentang seberapa lengkap atau seberapa banyak data yang tersedia untuk setiap atribut dalam suatu dataset. Kolom dengan non-null count yang tinggi menunjukkan bahwa data untuk atribut tersebut relatif lengkap, sedangkan kolom dengan non-null count yang rendah menunjukkan adanya kekurangan data atau missing values.

Dengan mempertimbangkan non-null count ini, kami dapat memutuskan langkah selanjutnya dalam mengelola missing values, melakukan pemfilteran data, atau menerapkan teknik analisis data yang sesuai. Dengan demikian, *non-null count* membantu memandu proses pengolahan data dan analisis selanjutnya serta memastikan bahwa kesimpulan yang diambil dari analisis data ini didasarkan pada data yang tepat dan lengkap.

3.2 Menampilkan nilai Missing Value pada Dataset

Pada tahap ini kami akan mengevaluasi hasil dari pemeriksaan *missing values* pada dataset film yang telah dilakukan sebelumnya. Gambar 3 menunjukkan jumlah *missing values* untuk setiap kolom di dataset film kami.

```
df.isnull().sum()
```

```
user_id         0
review_id       0
stars           0
date            0
text            0
type            0
business_id     0
votes.funny     0
votes.useful    0
votes.cool      0
dtype: int64
```

Gambar 3. Informasi Missing Value Dataset

Dari gambar di atas, terlihat bahwa tidak ada data yang null atau kosong. Sehingga setelah evaluasi ini kita dapat melanjutkan ke langkah selanjutnya.

3.3 Menampilkan Data Duplikat pada Dataset

Pada tahap ini kami menghitung jumlah baris duplikat dalam dataframe. Gambar 4 menunjukan jumlah baris yang terduplikasi dalam dataframe.

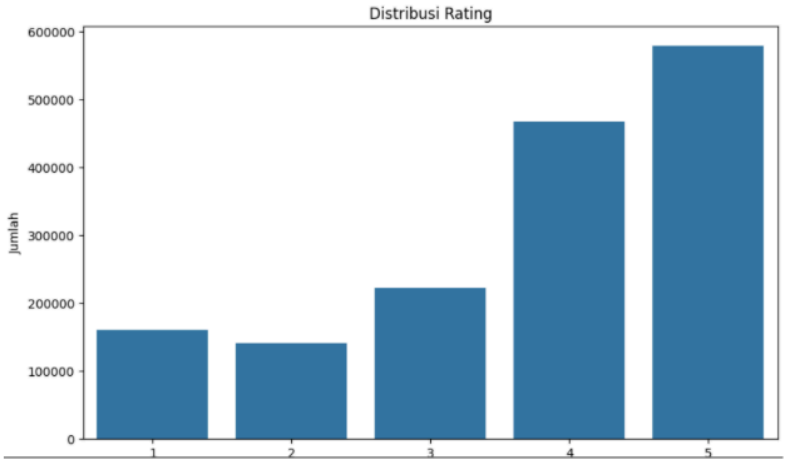
```
df.duplicated().sum()
```

```
0
```

Gambar 4. Informasi Jumlah Data Duplicated

3. 4 Menganalisis Distribusi Rating

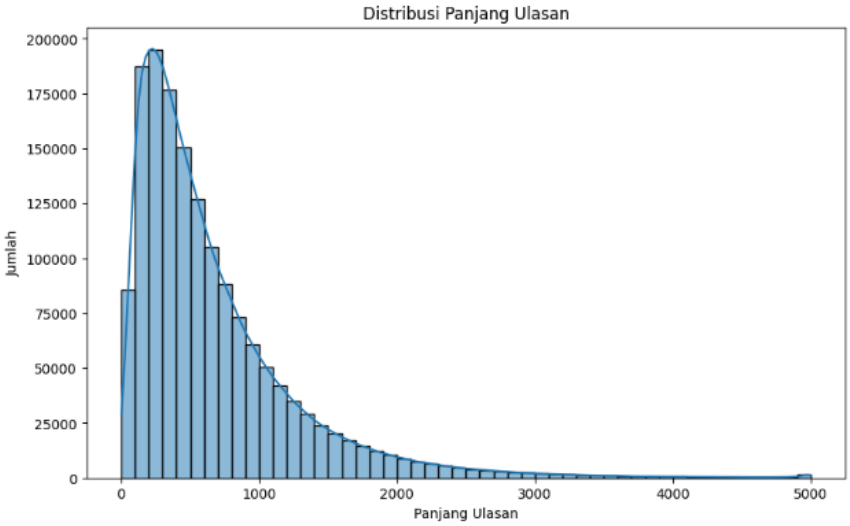
Visualisasi distribusi rating untuk menampilkan sebaran peringkat ulasan dapat membantu untuk melihat seberapa sering masing-masing peringkat muncul dalam dataset. Berikut ini adalah visualisasi dari distribusi rating:



Gambar 5. Distribusi Rating pada Dataset Yelp

3.5 Menganalisis Distribusi Panjang Ulasan

Memvisualisasikan distribusi panjang ulasan dalam dataset memberikan wawasan awal tentang sebaran panjang ulasan, yang dapat membantu dalam beberapa hal misalnya panjang teks dapat menjadi fitur yang relevan untuk analisis atau pemodelan lebih lanjut. Misalnya, dalam analisis sentimen, panjang ulasan dapat mempengaruhi cara model memproses teks. Dengan memahami distribusi panjang ulasan, kita dapat membuat keputusan tentang bagaimana menangani fitur ini dalam pemodelan. Berikut ini adalah gambar hasil analisis distribusi panjang teks ulasan:



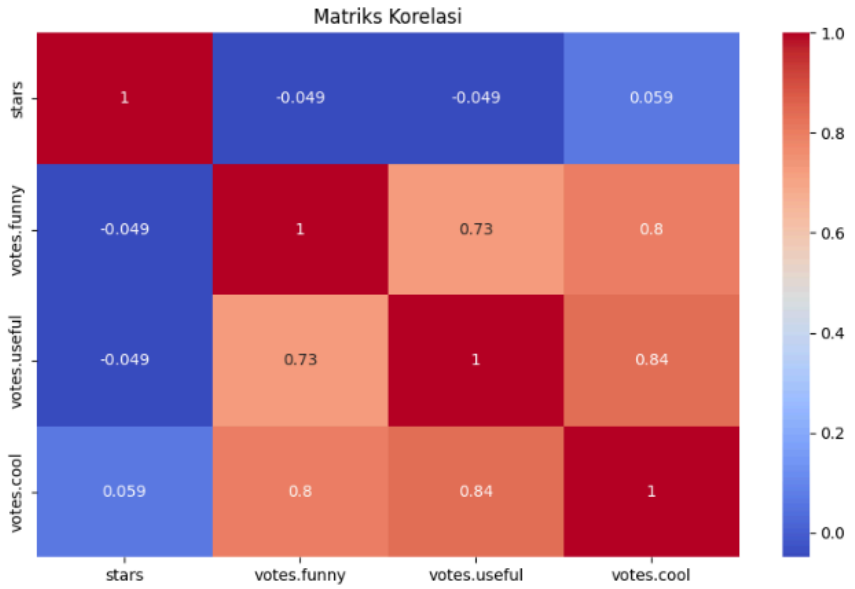
Gambar 6. Distribusi Panjang Teks Ulasan

3.6 Menghitung Korelasi Antar Kolom Numerik

Dalam kasus ini, fungsi EDA matriks korelasi antara 'stars', 'votes.funny', 'votes.useful', dan 'votes.cool' bertujuan untuk memahami hubungan korelasi antara peringkat ulasan (stars) dengan jumlah suara lucu (votes.funny), bermanfaat (votes.useful), dan keren (votes.cool) yang diberikan

oleh pengguna. Analisis ini bisa memberikan wawasan tentang bagaimana perilaku pengguna (seperti memberikan suara lucu, bermanfaat, atau keren) berkorelasi dengan peringkat ulasan yang diberikan yang berguna dalam memahami faktor-faktor apa yang mungkin mempengaruhi peringkat ulasan dalam platform seperti YELP, dan dapat menjadi dasar untuk analisis lebih lanjut atau pengambilan keputusan.

Di bawah ini adalah analisa dari beberapa kolom numerik dengan hasil perhitungan korelasi paling besar, atau yang paling mendekati positif:

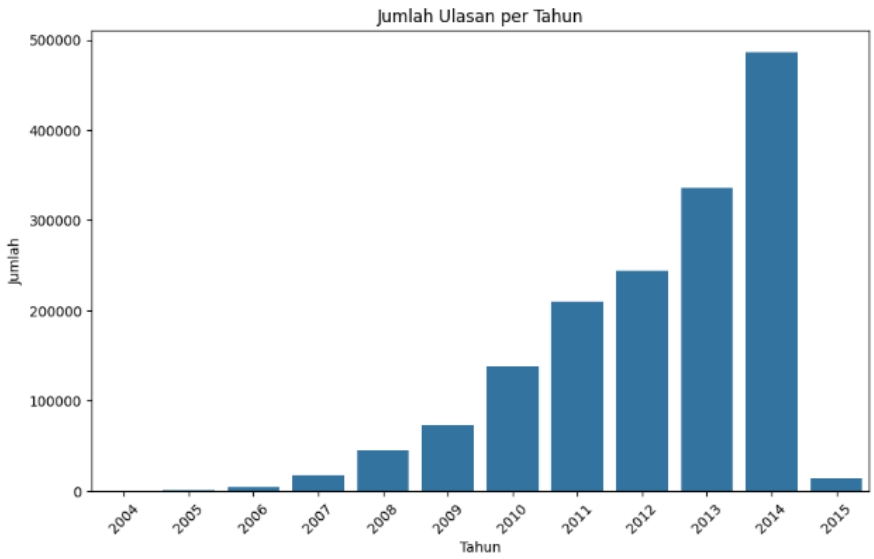


Gambar 7. Matriks Korelasi dari Data Numerik

3.7 Analisis Jumlah Ulasan Tiap Tahun

Langkah selanjutnya dalam proses adalah melakukan analisis jumlah ulasan pada tiap tahunnya. Analisis ini dapat membantu dalam memahami bagaimana popularitas atau aktivitas ulasan telah berkembang seiring waktu, yang dapat bermanfaat untuk pengambilan keputusan dan perencanaan strategis di platform seperti Y. Misalnya, melihat peningkatan jumlah ulasan dari tahun ke tahun mungkin menunjukkan pertumbuhan penggunaan platform, sementara penurunan mungkin memicu perubahan dalam kebijakan atau kualitas layanan.

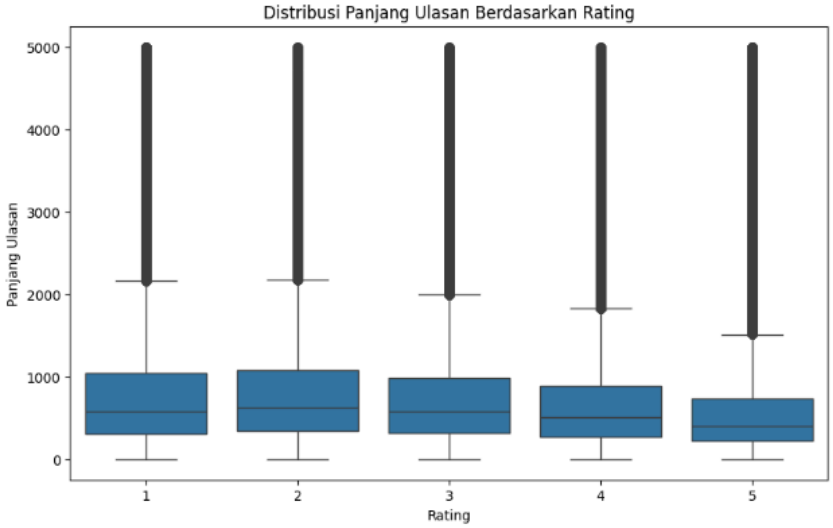
Di bawah ini merupakan graf hasil analisis jumlah ulasan dari tahun ke tahun yang menunjukkan peningkatan jumlah ulasan yang cukup signifikan:



Gambar 8. Visualisasi Jumlah Ulasan tiap Tahun

3.8 Distribusi Panjang Ulasan Berdasarkan Rating

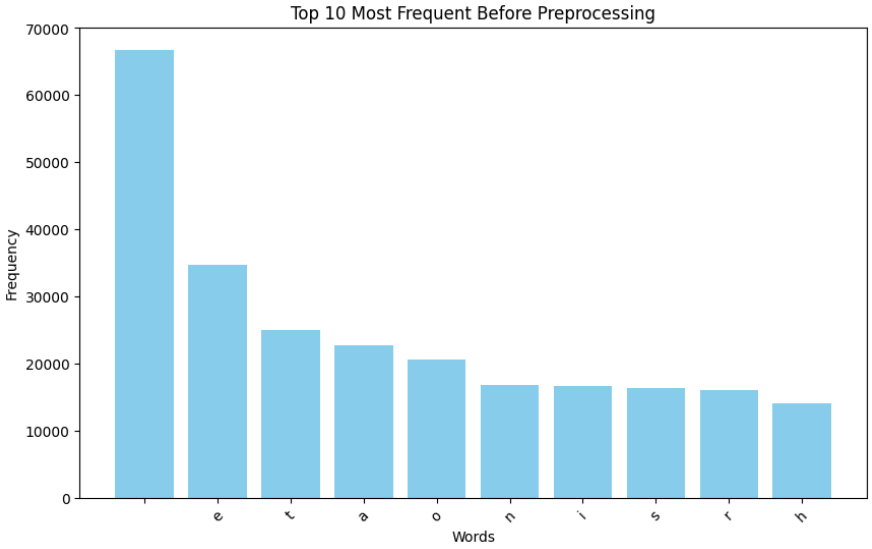
Langkah selanjutnya dalam proses adalah melakukan analisis untuk melihat distribusi panjang ulasan berdasarkan rating. Analisis ini membantu kita memahami pola atau tren dalam panjang ulasan tergantung pada peringkat ulasan yang diberikan. Hal ini dapat memberikan wawasan tentang bagaimana perilaku atau preferensi pengguna dalam memberikan ulasan berbeda tergantung pada pengalaman mereka, yang dapat berguna dalam analisis lebih lanjut atau pengambilan keputusan. Misalnya, pemilik bisnis dapat menggunakan informasi ini untuk memahami apa yang dianggap penting oleh pelanggan dengan peringkat tertentu dan meresponsnya dengan cara yang sesuai.



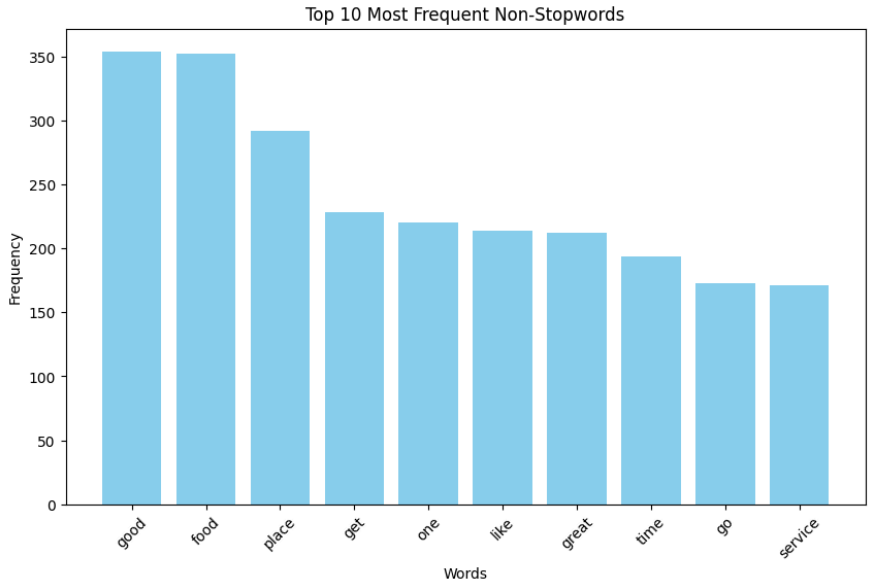
Gambar 9. Graf Distribusi Ulasan Berdasarkan Rating

3.9 Analisis *Frequent Word*

Tahap lain dalam Exploratory Data Analysis dataset ini adalah penggunaan Frequent Word Analysis. Frequent Word Analysis adalah representasi visual dari kata-kata di mana ukuran kata tersebut berbanding dengan frekuensinya dalam teks yang dianalisis. Ini adalah alat yang berguna untuk memvisualisasikan data teks dengan menyoroti kata kunci secara yang intuitif dan mudah dipahami.

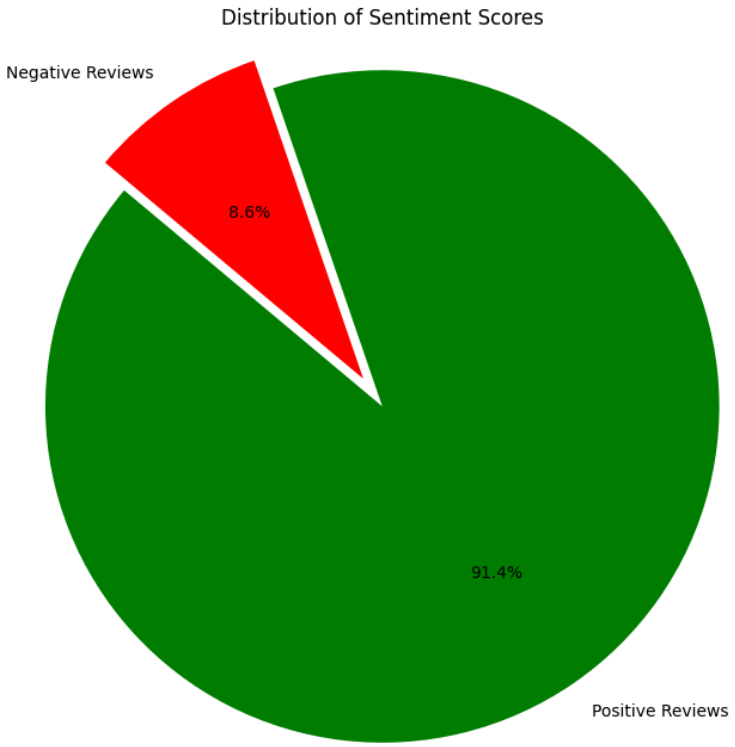


Gambar 10. Frequent Word Sebelum Preprocessing



Gambar 11. Frequent Word Setelah Preprocessing

3.10 Analisis Sentiment Score



4. Data Preprocessing

4.1 Melakukan inisialisasi dataframe

Setelah melakukan EDA, kami melakukan Data Preprocessing. Langkah ini bertujuan untuk meningkatkan akurasi pengolahan data karena data NLP sangat sensitif. Hal pertama yang kami lakukan adalah melakukan inisialisasi pada dataframe. Langkah ini kami lakukan untuk

```
# Cutting df
data2 = dat.loc[0:500]
data2.to_csv('data2.csv')

data = pd.read_csv('data2.csv')
```

```
data
```

	Unnamed: 0	text	business_id
0	0	dr. goldberg offers everything i look for in a...	vcNAWiLM4dR7D2nwwJ7nCA
1	1	Unfortunately, the frustration of being Dr. Go...	vcNAWiLM4dR7D2nwwJ7nCA
2	2	Dr. Goldberg has been my doctor for years and ...	vcNAWiLM4dR7D2nwwJ7nCA
3	3	Been going to Dr. Goldberg for over 10 years. ...	vcNAWiLM4dR7D2nwwJ7nCA
4	4	Got a letter in the mail last week that said D...	vcNAWiLM4dR7D2nwwJ7nCA
...	...	...	...
496	496	This is easily one of the best Giant Eagles I'...	eT5Ck7Gg1dBJobca9VFovw
497	497	This place has amazing selection when it comes...	eT5Ck7Gg1dBJobca9VFovw
498	498	oddly enough, of all the giant eagles i have v...	eT5Ck7Gg1dBJobca9VFovw
499	499	This is a absolutely beautiful Giant Eagle. I...	eT5Ck7Gg1dBJobca9VFovw
500	500	I'm with Rachel C. on this one. You get an ext...	eT5Ck7Gg1dBJobca9VFovw

Gambar 12. Implementasi inisialisasi dataframe

4.2 Melakukan regex dan lowercase

Setelah melakukan Inisialisasi pada dataframe, kita perlu melakukan beberapa proses terhadap text agar pada tahap analisis model dapat bekerja dengan baik, proses yang akan kita lakukan meliputi Regex dan Lowercase.

Regex adalah tahap dimana kita perlu penghapusan tanda baca,unsur angka dan simbol-simbol non-alfabet lainnya, agar model dapat menganalisis text dengan kondisi bersih tanpa adanya tambahan - tambahan simbol yang dapat mempengaruhi nilai sentimen yang diberikan model.

Lowercase adalah tahap dimana kita perlu mengganti setiap huruf kapital yang terdapat pada text menjadi huruf kecil. Tahap ini dilakukan karena jika terdapat huruf kapital pada sebuah kata. Kemudian kata tersebut akan memberikan makna atau sentimen yang berbeda dengan kata lain yang serupa namun tanpa huruf kapital.

```python
import re

def remove_punctuation(text):
    # Define a regex pattern that matches all common punctuation marks
    pattern = r'[.,"\'?!:;()\[\]{}—\-…/\\|@#\$%^&*_=+~`<>]'
    return re.sub(pattern, '', text)

# Apply the function to the 'text' column of the DataFrame and lower the case
data['clear_text'] = data['text'].apply(lambda x: remove_punctuation(x).lower())
```

```python
data.head()
```

	Unnamed: 0	text	business_id	clear_text
0	0	dr. goldberg offers everything i look for in a...	vcNAWiLM4dR7D2nwwJ7nCA	dr goldberg offers everything i look for in a ...
1	1	Unfortunately, the frustration of being Dr. Go...	vcNAWiLM4dR7D2nwwJ7nCA	unfortunately the frustration of being dr gold...
2	2	Dr. Goldberg has been my doctor for years and ...	vcNAWiLM4dR7D2nwwJ7nCA	dr goldberg has been my doctor for years and i...
3	3	Been going to Dr. Goldberg for over 10 years. ...	vcNAWiLM4dR7D2nwwJ7nCA	been going to dr goldberg for over 10 years i ...
4	4	Got a letter in the mail last week that said D...	vcNAWiLM4dR7D2nwwJ7nCA	got a letter in the mail last week that said d...

Gambar 8. Implementasi regex dan lowercase

4.3 Melakukan tokenisasi, stopwords, dan stemming

Tokenisasi adalah proses memecah teks menjadi unit-unit yang lebih kecil, seperti kata-kata atau frasa, untuk mempermudah analisis teks. Tujuannya adalah untuk mengubah teks mentah menjadi serangkaian token yang dapat diproses lebih lanjut secara komputasional. Di sisi lain, stopwords merujuk pada kata-kata umum yang sering muncul dalam teks namun tidak memberikan informasi yang signifikan dalam analisis konten, seperti "and", "or", dan "with". Penghapusan stopwords membantu mengurangi noise dalam data teks dan meningkatkan relevansi hasil analisis dengan fokus pada kata-kata yang lebih bermakna secara kontekstual.

Langkah berikutnya dalam pengolahan data kami adalah Stemming. Stemming adalah metode yang digunakan untuk mengubah kata-kata menjadi bentuk dasar atau akar kata mereka, dengan tujuan mempermudah analisis teks, mengurangi kompleksitas, dan meningkatkan efisiensi pemrosesan data. Kami menggunakan Porter Stemmer karena alat ini dapat mempercepat analisis teks dengan mengurangi variasi kata yang memiliki arti serupa, sehingga memungkinkan kami untuk fokus pada informasi yang relevan dalam data yang kami proses.

```
[ ]  1  import spacy
     2
     3  data['tokens'] = data['clear_text'].apply(lambda x: x.split())
```

```
1  stop_words = set(stopwords.words('english'))
2
3  def remove_stopwords(text):
4    token = word_tokenize(text)
5
6    stopwords_deleted = [word for word in token if word.lower() not in stop_words]
7
8    return stopwords_deleted
9
10 data['non_stopwords'] = data['clear_text'].apply(remove_stopwords)
```

```
1  from nltk.stem import PorterStemmer
2
3  stemmer = PorterStemmer()
4
5  def stem_text(text):
6
7    stemmed_token = [stemmer.stem(token) for token in text]
8
9    join = ' '.join(stemmed_token)
10   return join
11
12
13 data['stemming'] = data['non_stopwords'].apply(stem_text)
```

Gambar 9. Implementasi tokenisasi, stopwords, dan stemming

Dengan hasil analisis, model analisis sentimen berbasis Naive Bayes ini sudah menunjukkan kinerja yang memuaskan. Selama proses EDA, terlihat bahwa distribusi skor sentimen pada data pelatihan lebih banyak yang positif daripada netral dan negatif. Namun, saat dilakukan prediksi, model mampu menangani ketidakseimbangan tersebut dengan menghasilkan prediksi yang cenderung lebih netral daripada positif. Ini menunjukkan bahwa algoritma Naive Bayes efektif dalam menangani ketidakseimbangan dataset ini. Namun, untuk meningkatkan kualitas model, perlu dilakukan penambahan data pelatihan yang seimbang antara jenis ulasan. Disarankan untuk mengimbangi data pelatihan dengan proporsi yang setara untuk ulasan positif, negatif, dan netral, yakni masing-masing 1/3 dari total data.

5. Modeling dan Evaluation

5.1 Modeling

5.1.1 Vader

Metode ini menggunakan *kamus* sebagai patokan untuk menentukan nilai pada tiap kata yang terdapat pada ulasan, sebelum nantinya dikalkulasi dan memberikan nilai pada masing" label: Positive, Negative atau Netral. dan dari nilai pada label tsb akan dihitung kembali. Dari perhitungan tersebut akan dihasilkan *score* keseluruhan sentimen yaitu compound. sebagai patokan untuk menentukan apakah text tersebut dapat dikatakan positif, negative atau netral.

5.1.2 Kamus

Kamus yang dimaksud adalah, kumpulan kata dan frasa yang telah diberikan nilai sentimen tertentu. kamus ini pula yang menjadi inti dari **vader** dalam menentukan nilai sentimen dari teks.

5.1.3 Perhitungan Compound

$$compound = \frac{total\ score}{\sqrt{total\ score^{2} + \varphi}}$$

5.1.4 Penentuan Label (pos, neg, neut) Berdasarkan Compound

Untuk menentukan apakah teks dapat dikatakan positif, negative atau netral, kita perlu menetapkan standar pada nilai compound. yang pada umumnya yaitu compound >= 0.05 : positive, compound <= -0.05 : negative, compound antara -0.05 sampai 0.05 : netral.

Pemodelan dalam Review Sentiment Analysis ini mengadopsi pendekatan Lexicon Based, karena kesesuaiannya dengan business understanding kami. Metode ini dirancang untuk memudahkan analisis sentimen bagi individu yang mungkin tidak memiliki latar belakang dalam Machine Learning namun memiliki kepedulian tinggi terhadap umpan balik yang diterima oleh bisnis mereka. Dalam hal ini, Lexicon Based Method menyediakan solusi yang mudah dipahami dan diimplementasikan untuk mengidentifikasi dan mengevaluasi sentimen dalam ulasan. Selain itu, kami memilih untuk menggunakan VADER (Valence Aware Dictionary and sEntiment Reasoner) sebagai alat analisis sentimen karena kemampuannya dalam menangani kata-kata kapital dengan efektif, serta kemampuannya untuk menangani ekspresi sentimen yang lebih kompleks seperti slang, emotikon, dan bentuk-bentuk informal lainnya.

VADER menawarkan pendekatan yang terukur dan dapat diandalkan dalam mengklasifikasikan sentimen, sehingga memberikan analisis yang lebih akurat dan relevan dalam konteks ulasan bisnis. Dengan memanfaatkan kekuatan VADER dan

pendekatan Lexicon Based, kami dapat menyediakan wawasan yang mendalam dan bermanfaat untuk pengambilan keputusan strategis bagi bisnis.

```python
# inisialisasi vader
vader = SentimentIntensityAnalyzer()

# Func untuk menentukan label (Pos / Neg) berdasarkan nilai vader : neg,neu,pos
# nilai score = nilai compound

def vader_sentimen(text):
  score = vader.polarity_scores(text)
  compound = score['compound']

  if compound >= 0.05:
    label = 'positive'
  elif compound <= -0.05:
    label = 'negative'
  else:
    label = 'netral'
  return label,score

data[['sentiment','Score']] = data['text'].apply(lambda x: pd.Series(vader_sentimen(x)))
data = pd.concat([data.drop(['Score'],axis=1), data['Score'].apply(pd.Series)], axis = 1)
```

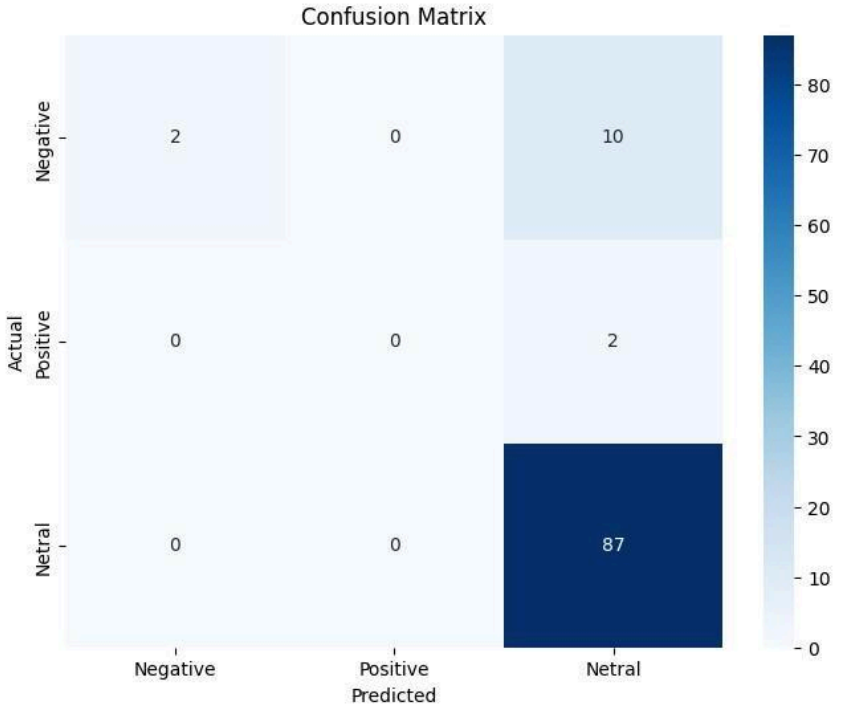
Gambar 10. Implementasi Vader dan labeling teks

	stemming	sentiment	neg	neu	pos	compound
	dr goldberg offer everyth look gener practitio...	positive	0.019	0.925	0.056	0.3708
	unfortun frustrat dr goldberg patient repeat e...	negative	0.117	0.860	0.023	-0.8997
	dr goldberg doctor year like ive found offic f...	positive	0.000	0.772	0.228	0.9298
	go dr goldberg 10 year think one 1st patient s...	positive	0.000	0.959	0.041	0.6249
	got letter mail last week said dr goldberg mov...	negative	0.091	0.866	0.043	-0.4075

Gambar 11. Hasil Vader dan labeling teks

5.2 Evaluasi

Evaluasi performa sistem dilakukan menggunakan metrik standar seperti akurasi, presisi, recall, dan F1-score untuk memastikan keakuratan rekomendasi yang diberikan. Dengan hasil evaluasi sebagai berikut: **Akurasi** sebesar 0.881, **Presisi** sebesar 0.876, **Recall** sebesar 0.876, dan **F1-score** sebesar 0.840. Sedangkan pada confusion matrix hasilnya adalah sebagai berikut:



Gambar 12. Hasil Confusion Matrix

Dengan demikian, model analisis sentimen berbasis Naive Bayes ini sudah cukup baik. Pada proses EDA, terlihat bahwa distribusi skor sentimen pada data pelatihan lebih dominan skor positif dibandingkan netral dan negatif. Namun, saat melakukan prediksi, model dapat mengatasi ketidakseimbangan tersebut, ditunjukkan dengan hasil prediksi yang cenderung netral dibandingkan positif. Hal ini menunjukkan bahwa algoritma Naive Bayes cukup baik dalam mengatasi ketidakseimbangan pada dataset ini. Namun, data pelatihan masih perlu ditambah dan diseimbangkan antara jenis ulasan. Sebaiknya, data pelatihan diseimbangkan dengan proporsi masing-masing 1/3 untuk ulasan positif, negatif, dan netral.

Sumber Data: https://drive.google.com/file/d/1jEP2JkL-69DBbV8vYwu5kZjTYbhDLu7c/view

Sumber Literatur

[1] Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big data, 2, 1-14.

[2] Aung, K. Z., & Myo, N. N. (2017, May). Sentiment analysis of students' comment using lexicon based approach. In 2017 IEEE/ACIS 16th international conference on computer and information science (ICIS) (pp. 149-154). IEEE.

[3] Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. SMU Data Science Review, 1(4), 7.