Import Library

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import re
import nltk
from nltk.corpus import stopwords
from wordcloud import WordCloud, STOPWORDS
from sklearn.feature_extraction.text import CountVectorizer
from nltk.tokenize import RegexpTokenizer

from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,classification_report
from sklearn.naive_bayes import MultinomialNB

from wordcloud import WordCloud, STOPWORDS
```

View data

```
data = pd.read_csv('mobil_listrik.csv')
data.head()
```

	id_komentar	nama_akun	tanggal	text_cleaning	sentimen
0	UgzbII5eyrIy3-gdUUJ4AaABAg	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio ...	positif
1	UgzEDUiV3OTrV943p8p4AaABAg	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga	negatif

Next steps: View recommended plots

1. DATA UNDERSTANDING

```
data.dtypes
```

```
id_komentar      object
nama_akun        object
tanggal          object
text_cleaning    object
sentimen         object
dtype: object
```

```
data.shape
```

```
(1517, 5)
```

we have 1517 raw and 5 cloumn

```
data = data.dropna()
```

```
data.isnull().sum()
```

```
id_komentar      0
nama_akun        0
tanggal          0
text_cleaning    0
sentimen         0
dtype: int64
```

```
data.duplicated().sum()
```

```
0
```

```
data.describe()
```

	id_komentar	nama_akun	tanggal	text_cleaning	sentimen
count	1514	1514	1514	1514	1514
unique	1514	1513	1511	1513	3
top	Ugzbll5eyrly3-gdUUJ4AaABAg	Albert	2023-06-06 14:02:09+00:00	mahal	negatif
freq	1	2	2	2	868

2. PREPROCESSING TEKS

```
data.head()
```

	id_komentar	nama_akun	tanggal	text_cleaning	sentimen
0	Ugzbll5eyrly3-gdUUJ4AaABAg	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio ...	positif
1	UgzEDUiV3OTrV943p8p4AaABAg	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga	negatif

Next steps: View recommended plots

```
data = data.drop(columns='id_komentar')
data.head()
```

	nama_akun	tanggal	text_cleaning	sentimen
0	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio ...	positif
1	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga dinai...	negatif
2	Fatih Al-Ayyubi	2023-08-04 10:17:57+00:00	baik kualitas kembang dulu baik kualitas motor...	positif
		2023-08-04	model jelek kwalitas buruk harga	

Next steps: View recommended plots

```
# mengubah teks menjadi lower

data['text_cleaning'] = data['text_cleaning'].str.lower()
data.head()
```

	nama_akun	tanggal	text_cleaning	sentimen
0	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio ...	positif
1	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga dinai...	negatif
2	Fatih Al-Ayyubi	2023-08-04 10:17:57+00:00	baik kualitas kembang dulu baik kualitas motor...	positif
		2023-08-04	model jelek kwalitas buruk harga	

Next steps: View recommended plots

```
# normalisasi
df_alay = pd.read_csv('kamusalay.csv', encoding='ISO-8859-1', header=None)
df_alay = df_alay.rename(columns={0: 'alay', 1: 'formal'})
df_alay_dict = dict(zip(df_alay['alay'], df_alay['formal']))
```

```
def normalize_text(text):
    words = text.split()
    normalized_words = [df_alay_dict[word] if word in df_alay_dict else word for word in words]
    return ' '.join(normalized_words)

data['text_cleaning'] = data['text_cleaning'].apply(normalize_text)
```

```
data.head()
```

	nama_akun	tanggal	text_cleaning	sentimen
0	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio in...	positif
1	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga dinai...	negatif
2	Fatih Al-Ayyubi	2023-08-04 10:17:57+00:00	baik kualitas kembang dulu baik kualitas motor...	positif
		2023-08-04	model jelek kualitas buruk harga	

Next steps: View recommended plots

```
# stopword
!pip install Sastrawi

import Sastrawi
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory, StopWordRemover, ArrayDictionary
more_stop_word = []

stop_words = StopWordRemoverFactory().get_stop_words()
new_array = ArrayDictionary(stop_words)
stop_words_remover_new = StopWordRemover(new_array)

def stopword(str_text):
  str_text = stop_words_remover_new.remove(str_text)
  return str_text

data['text_cleaning'] = data['text_cleaning'].apply(lambda x: stopword(x))
data.head()
```

```
Collecting Sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 kB 5.2 MB/s eta 0:00:00
Installing collected packages: Sastrawi
Successfully installed Sastrawi-1.0.1
```

	nama_akun	tanggal	text_cleaning	sentimen
0	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio in...	positif
1	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga dinai...	negatif
2	Fatih Al-Ayyubi	2023-08-04 10:17:57+00:00	baik kualitas kembang dulu baik kualitas motor...	positif
3	vp office	2023-08-04	model jelek kualitas buruk harga	negatif

Next steps: View recommended plots

```
# Tokenisasi

tokenized = data['text_cleaning'].apply(lambda x:x.split())
tokenized
```

```
0       [saran, sih, bikin, harga, ionic, sama, kayak,...
1       [problem, subsidi, kualitas, diturunin, harga,...
2       [baik, kualitas, kembang, dulu, baik, kualitas...
3       [model, jelek, kualitas, buruk, harga, mahal, ...
4       [syarat, kacau, oi, anak, muda, punya, rumah, ...
                              ...
1512    [apa, kabar, padahal, negeri, luar, biasa, neg...
1513    [antar, anak, sekolah, antar, bantu, pasar, ka...
1514                               [esemka, bangga, solo]
1515    [cerdas, orang, dan, pasar, jalan, x, dan, bag...
1516    [niat, beli, ev, murah, malah, ikut, dinaiki, ...
Name: text_cleaning, Length: 1514, dtype: object
```

```
# Stemming

from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

def stemming(text_cleaning):
  factory = StemmerFactory()
  stemmer = factory.create_stemmer()
  do = []
  for w in text_cleaning:
    dt = stemmer.stem(w)
    do.append(dt)
  d_clean = []
  d_clean = " ".join(do)
  print(d_clean)
```

```
  return d_clean

tokenized = tokenized.apply(stemming)

tokenized.to_csv('mobil_listrik_clean.csv', index=False)

data_clean = pd.read_csv('mobil_listrik_clean.csv')
```

```
data_clean.head()
```

	text_cleaning
0	saran sih bikin harga ionic sama kayak brio in...
1	problem subsidi kualitas diturunin harga naik ...
2	baik kualitas kembang dulu baik kualitas motor...
3	model jelek kualitas buruk harga mahal crot
4	syarat kacau oi anak muda punya rumah jadi usa...

```
#kolom sentimen tidak muncul
#kita gabungkan data lama dengan data baru
```

```
at1 = pd.read_csv('mobil_listrik_clean.csv')
at2 = pd.read_csv('mobil_listrik.csv')
att2 = at2['sentimen']

data_clean = pd.concat([at1, att2], axis =1)
data_clean
```

	text_cleaning	sentimen
0	saran sih bikin harga ionic sama kayak brio in...	positif
1	problem subsidi kualitas diturunin harga naik ...	negatif
2	baik kualitas kembang dulu baik kualitas motor...	positif
3	model jelek kualitas buruk harga mahal crot	negatif
4	syarat kacau oi anak muda punya rumah jadi usa...	negatif
...	...	...
1512	cerdas orang dan pasar jalan x dan bagaimana k...	negatif
1513	niat beli ev murah malah ikut naik sama perint...	negatif
1514	NaN	positif
1515	NaN	netral
1516	NaN	negatif

1517 rows × 2 columns

Next steps: View recommended plots

```
#kita hapus sentimen NaN
data_clean = data_clean.dropna()
```

```
#hilangkan sentimen netral, karena kita butuh sentmen positif dan negatif saja

data_clean = data_clean[data_clean['sentimen'] != 'netral']
data_clean
```

	text_cleaning	sentimen
0	saran sih bikin harga ionic sama kayak brio in...	positif
1	problem subsidi kualitas diturunin harga naik ...	negatif
2	baik kualitas kembang dulu baik kualitas motor...	positif
3	model jelek kualitas buruk harga mahal crot	negatif
4	syarat kacau oi anak muda punya rumah jadi usa...	negatif
...	...	...
1509	apa kabar padahal negeri luar biasa negara	negatif
1510	antar anak sekolah antar bantu pasar kalau jau...	positif
1511	esemka bangga solo	positif
1512	cerdas orang dan pasar jalan x dan bagaimana k...	negatif
1513	niat beli ev murah malah ikut naik sama perint...	negatif

1371 rows × 2 columns

Next steps: View recommended plots

```
#ubah sentimen negatif dan postif menjadi numerik

data_clean = data_clean.replace({'positif' : 1, 'negatif':0})
data_clean.head()
```

	text_cleaning	sentimen
0	saran sih bikin harga ionic sama kayak brio in...	1
1	problem subsidi kualitas diturunin harga naik ...	0
2	baik kualitas kembang dulu baik kualitas motor...	1
3	model jelek kualitas buruk harga mahal crot	0
4	syarat kacau oi anak muda punya rumah jadi usa...	0

Next steps: View recommended plots

3. VISUALISASI KATA

```
data_negatif = data_clean[data_clean['sentimen'] == 0]
data_positif = data_clean[data_clean['sentimen'] == 1]
```

```
#visualisasi sentimen negatif

all_text_s0 = ' '.join(word for word in data_negatif["text_cleaning"])
wordcloud = WordCloud(colormap='Reds', width=1000, height=1000, mode='RGBA', background_color='white').generate(all_text_s0)
plt.figure(figsize=(9,6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title("Visualisasi Kata di Sentimen Negatif")
plt.margins(x=0, y=0)
plt.show()
```


```
#visualisasi sentimen positif

all_text_s0 = ' '.join(word for word in data_positif["text_cleaning"])
wordcloud = WordCloud(colormap='Blues', width=1000, height=1000, mode='RGBA', background_color='white').generate(all_text_s0)
plt.figure(figsize=(9,6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title("Visualisasi Kata di Sentimen Postif")
plt.margins(x=0, y=0)
plt.show()
```


```
data_clean['sentimen'].value_counts()
```

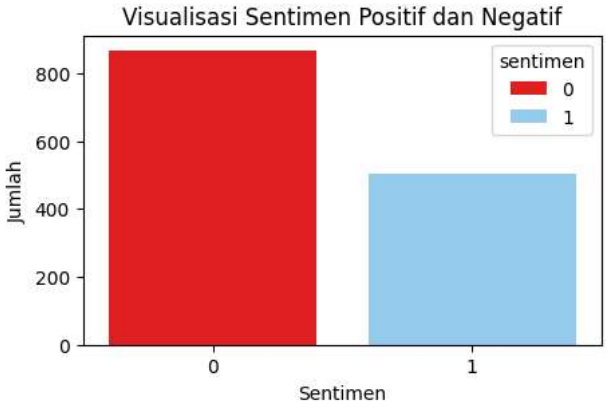
```
0    868
1    503
Name: sentimen, dtype: int64
```

```
data_clean.head()
```

	text_cleaning	sentimen
0	saran sih bikin harga ionic sama kayak brio in...	1
1	problem subsidi kualitas diturunin harga naik ...	0
2	baik kualitas kembang dulu baik kualitas motor...	1
3	model jelek kualitas buruk harga mahal crot	0
4	syarat kacau oi anak muda punya rumah jadi usa...	0

Next steps: View recommended plots

```
#visualisasi jumlah sentimen positif dan negatif
plt.figure(figsize=(5,3))
sns.countplot(data=data_clean, x='sentimen', hue='sentimen', palette={0:"red", 1: "lightskyblue"})
plt.title("Visualisasi Sentimen Positif dan Negatif")
plt.xlabel('Sentimen')
plt.ylabel('Jumlah')
plt.show()
```



4. DATA PREPARATION

```
# memisahkan data menjadi data latih dan data uji

X = data_clean['text_cleaning']
y = data_clean['sentimen']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, random_state=42)
```

```
# menggunakan CountVectorizer untuk mengubah teks menajdi fitur numerik

vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

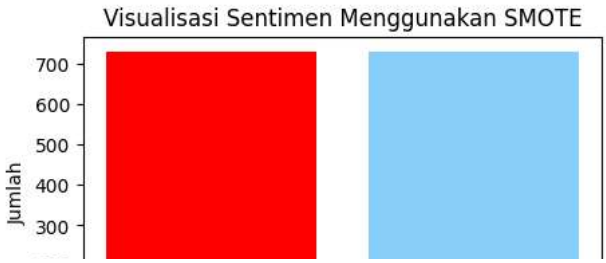
```
# melakukan oversampling mengunakan SMOTE pada data letih

smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train_vec, y_train)
```

```
# hasil data resampling

sentimen_counts = y_train_resampled.value_counts()

plt.figure(figsize=(5,3))
plt.bar(sentimen_counts.index, sentimen_counts.values, color=['red', 'lightskyblue'])
plt.xlabel('Sentimen')
plt.ylabel('Jumlah')
plt.title('Visualisasi Sentimen Menggunakan SMOTE')
plt.xticks(sentimen_counts.index, ['0', '1'])
plt.show()
```



. 5. MODELING

```
naive_bayes = MultinomialNB()
naive_bayes.fit(X_train_resampled, y_train_resampled)
```

```
MultinomialNB()
```

```
y_pred = naive_bayes.predict(X_test_vec)
```

```
# Evaluasi Model

accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred, target_names=['negatif','positif'])

print("Akurasi Model Naive Bayes : ", accuracy)
print("\nLaporan Klasifikasi :\n", classification_rep)
```

```
Akurasi Model Naive Bayes :  0.7815533980582524

Laporan Klasifikasi :
               precision    recall  f1-score   support

      negatif       0.82      0.86      0.84       137
      positif       0.69      0.62      0.66        69

     accuracy                           0.78       206
    macro avg       0.76      0.74      0.75       206
 weighted avg       0.78      0.78      0.78       206
```