Mini Project Week 10

Analisis Data Media Sosial

Kelas B

Dennis Imanuel

NPM: 6161901066

UNIVERSITAS KATOLIK PARAHYANGAN

BANDUNG

2023

1. Data Preparation Review Motor

Dataset ‘Review_Motor.csv’ adalah dataset yang telah saya kumpulkan terkait hasil ulasan-ulasan pembeli sepeda motor melalui *marketplace* online seperti Tokopedia, Shopee, dan beberapa situs belanja online lainnya. Dataset tersebut disimpan dalam bentuk file *comma separated value* (csv) dan dapat diakses melalui halam *respository* Github saya menggunakan link: https://github.com/aiightvert/Data-UTS-Andat-Medsos/blob/0371e5926cf11d7f09e30e71053fe70da8efcf4c/Review_Motor.csv

Dataset tersebut mengandung 5 kolom parameter yaitu:

- No : penomoran baris dataset
- Nama : nama atau deskripsi produk
- Rating : nilai ulasan pembeli terhadap produk dalam skala 1 sampai 5
- Comments : komentar pembeli terhadap produk
- Links : tautan menuju laman penjualan produk

Dataset tersebut memiliki 202 baris data yang telah berhasil saya kumpulkan. Namun setelah melakukan penghilangan kolom kosong dengan perintah ‘df.dropna()’, maka baris dalam dataset berubah menjadi 201. Berikut adalah tampilan 5 baris teratas dataset ‘Review_Motor.csv’ beserta deskripsi dan tipe data pada masing-masing kolomnya.

```
df.head()
```

	No	Nama	Rating	Comments	Links
0	1	Honda BeAT CBS	5	seller fast respon, berintegritas dan helpful ...	https://www.tokopedia.com/ajaya-179/sepeda-mot...
1	2	Honda BeAT CBS	5	Semua layanan dan service toko nya bagus, komu...	https://www.tokopedia.com/ajaya-179/sepeda-mot...
2	3	Honda BeAT CBS	5	Kendaraan sangat bagus, tepat waktu, toko nya ...	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...
3	4	Honda BeAT CBS	5	Pelayanan cepet dan respon cepet pokoknya mantap	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...
4	5	Honda BeAT CBS	5	Proses dan pengiriman cepat sampai sesuai deng...	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...

```
df.describe()
```

	No	Rating
count	201.000000	201.000000
mean	101.641791	4.895522
std	58.567662	0.504014
min	1.000000	1.000000
25%	51.000000	5.000000
50%	102.000000	5.000000
75%	152.000000	5.000000
max	202.000000	5.000000

```
[ ] df.columns
```

```
Index(['No', 'Nama', 'Rating', 'Comments', 'Links'], dtype='object')
```

2. Text Pre-processing

Pada tahap *pre-processing*, ada beberapa langkah yang perlu dilakukan terhadap kolom ‘Comments’. Langkah pertama menghilangkan karakter-karakter spesial yang tidak dibutuhkan dalam analisis teks seperti tanda baca, situs website, username, dan lain-lainnya. Selain itu, kita juga akan mengubah seluruh karakter alfabetis ke dalam huruf kecil. Kemudian, kita akan mendefinisikan *stopwords* yang telah kita tulis dalam file txt terpisah yang tujuannya akan mengubah kata-kata ke dalam bentuk baku dan sederhananya sesuai KBBI. Langkah terakhir adalah memisahkan kolom ‘Comments’ yang telah dibersihkan tersebut ke dalam kolom baru yang akan kita beri nama ‘clean_text’, sehingga kita dapat membandingkan perbedaan sebelum dan sesudahnya dilakukan *text pre-processing*.

```
df.head()
```

	No	Nama	Rating	Comments	Links	clean_text
0	1.0	Honda BeAT CBS	5.0	seller fast respon, berintegritas dan helpful ...	https://www.tokopedia.com/ajaya-179/sepeda-mot...	seller fast respon berintegritas helpful pembe...
1	2.0	Honda BeAT CBS	5.0	Semua layanan dan service toko nya bagus, komu...	https://www.tokopedia.com/ajaya-179/sepeda-mot...	layanan service toko nya bagus komunikatif pen...
2	3.0	Honda BeAT CBS	5.0	Kendaraan sangat bagus, tepat waktu, toko nya ...	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...	kendaraan bagus tepat waktu toko nya amanah ha...
3	4.0	Honda BeAT CBS	5.0	Pelayanan cepet dan respon cepet pokoknya mantap	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...	pelayanan cepet respon cepet pokoknya mantap
4	5.0	Honda BeAT CBS	5.0	Proses dan pengiriman cepat sampai sesuai deng...	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...	proses pengiriman cepat sesuai pesanan

3. Penambahan Kolom Sentimen (Positif/Negatif)

Pada tahap ini, kita akan menambahkan kolom baru yang menandakan apakah review lebih merujuk kepada review positif atau negatif, dan parameter argumen yang akan menentukan nilai itu sendiri akan diambil melalui kolom ‘Rating’. Akan dilakukan pengklasifikasian terhadap masing-masing baris dengan syarat kondisi apabila baris memiliki nilai rating lebih besar sama dengan tiga, maka akan disimpulkan bahwa review pada baris tersebut memiliki sentimen positif. Jika nilai rating lebih kecil daripada tiga, maka akan disimpulkan bahwa review pada baris tersebut memiliki sentimen negatif.

```
y = []
for i in df.Rating:
  if i >= 3:
    y.append("positif")
  else:
    y.append("negatif")
```

```
#Ubah ke bentuk tabel
sentimen = pd.DataFrame({'sentimen':y})
```

```
#Gabungkan ke tabel utama
df = pd.concat([df, sentimen], axis = 1)
```

```
df.head()
```

	No	Nama	Rating	Comments	Links	clean_text	sentimen
0	1.0	Honda BeAT CBS	5.0	seller fast respon, berintegritas dan helpful ...	https://www.tokopedia.com/ajaya-179/sepeda-mot...	seller fast respon berintegritas helpful pembe...	positif
1	2.0	Honda BeAT CBS	5.0	Semua layanan dan service toko nya bagus, komu...	https://www.tokopedia.com/ajaya-179/sepeda-mot...	layanan service toko nya bagus komunikatif pen...	positif
2	3.0	Honda BeAT CBS	5.0	Kendaraan sangat bagus, tepat waktu, toko nya ...	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...	kendaraan bagus tepat waktu toko nya amanah ha...	positif
3	4.0	Honda BeAT CBS	5.0	Pelayanan cepet dan respon cepet pokoknya mantap	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...	pelayanan cepet respon cepet pokoknya mantap	positif
4	5.0	Honda BeAT CBS	5.0	Proses dan pengiriman cepat sampai sesuai deng...	https://shopee.co.id/HONDA-BEAT-CBS-SEPEDA-MOT...	proses pengiriman cepat sesuai pesanan	positif

Tujuan dari penambahan kolom ini adalah untuk pelatihan data yang akan dilakukan dimana kolom ‘sentimen’ merupakan hasil keluaran yang diinginkan berdasarkan masukan teks yang terdapat pada kolom ‘Comments’.

4. Pelatihan Data

Pada tahap ini, kita akan menggunakan 2 kolom yaitu ‘Comments’ yang akan menjadi variabel independen atau masukan yang akan menentukan hasil keluaran variabel dependen ‘sentimen’. Kemudian, jumlah baris dalam kedua kolom tersebut akan dibagi ke dalam 2 kategori. Kategori data pertama akan digunakan untuk kebutuhan pelatihan yang artinya data tersebut akan menjadi bahan bagi pembelajaran mesin untuk mengidentifikasi keluaran kolom ‘Sentimen’ yang sudah disediakan berdasarkan masukan teks pada kolom ‘Comments’.

Kategori data kedua akan digunakan untuk kebutuhan uji coba, artinya setelah mesin mempelajari analisis teks pada kolom ‘Comments’, maka mesin akan berusaha memprediksi keluaran ‘sentimen’ yang diharapkan berdasarkan teks yang disediakan pada kolom ‘Comments’ uji coba kemudian membandingkan hasil dan akurasinya terhadap kolom ‘sentimen’ uji coba yang sebenarnya.

Pada keperluan, akan didefinisikan pembagian jumlah baris data uji coba sebanyak 20% dari total jumlah baris data yang telah disediakan.

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2, random_state=225)
```

```
[ ] print("Banyaknya data X_train : ",len(X_train))
    print("Banyaknya data X_test : ",len(X_test))
    print("Banyaknya data y_train : ",len(y_train))
    print("Banyaknya data y_test : ",len(y_test))
```

```
Banyaknya data X_train :  161
Banyaknya data X_test :  41
Banyaknya data y_train :  161
Banyaknya data y_test :  41
```

5. Model Bernoulli

Model Bernoulli merupakan bagian dari klasifikasi *Naive Bayes* yang akan digunakan sebagai model pembelajaran mesin terhadap analisis sentimen dalam percobaan data ini. Persamaan untuk menentukan *conditional probability* kelas c terhadap sebuah kata t_i menggunakan model Bernoulli multinomial adalah

$$P(t_i|c) = \frac{N_i + 1}{|V| + N'} \quad ,$$

dimana $|V|$ merupakan jumlah kata unik yang muncul pada data, N_i merupakan jumlah kemunculan kata t pada baris ke-i dari data dengan kategori kelas c, dan N' merupakan jumlah kata yang terdapat pada data dengan kategori kelas c [1].

Penerapan model Bernoulli pada uji coba ini akan dilakukan menggunakan *built-in function* yang telah disediakan oleh *package ‘sklearn.naive_bayes’*.

[1] Wardani, Nabila & Prahutama, Alan & Kartikasari, Puspita. (2020). ANALISIS SENTIMEN PEMINDAHAN IBU KOTA NEGARA DENGAN KLASIFIKASI NAÏVE BAYES UNTUK MODEL BERNOULLI DAN MULTINOMIAL. Jurnal Gaussian. 9. 237-246. 10.14710/j.gauss.v9i3.27963.

```
[ ] from sklearn.naive_bayes import BernoulliNB
```

```
[ ] Model_NB = BernoulliNB()
```

```
[ ] #Proses pembelajaran mesin
    Model_NB.fit(X_train,y_train)
```

```
▾ BernoulliNB
BernoulliNB()
```

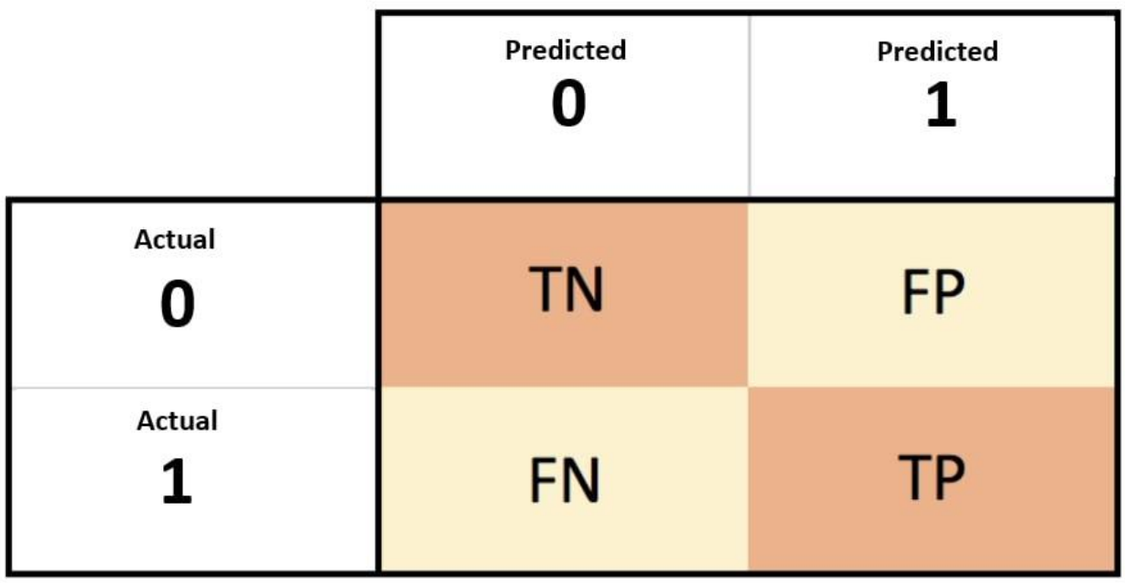
```
    #Uji hasil pembelajaran
    hasil = Model_NB.predict(X_test)
```

```
[ ] print(hasil)
```

```
['positif' 'positif' 'positif' 'positif' 'positif' 'positif' 'positif'
 'positif' 'positif' 'positif' 'positif' 'positif' 'positif' 'positif'
 'positif' 'positif' 'positif' 'positif' 'positif' 'positif' 'positif'
 'positif' 'positif' 'positif' 'positif' 'positif' 'positif' 'positif'
 'positif' 'positif' 'positif' 'positif' 'positif' 'positif' 'positif'
 'positif' 'positif' 'positif' 'positif' 'positif' 'positif']
```

6. Uji Akurasi

Uji akurasi terhadap hasil yang telah kita peroleh menggunakan model Bernoulli akan dilakukan menggunakan matriks *confusion*. Matriks *confusion* dapat digunakan untuk melihat seberapa banyak klasifikasi kelas hasil prediksi berada pada klasifikasi kelas yang sebenarnya. Misalkan C merepresentasikan matriks *confusion*, maka $C_{0,0}$ merepresentasikan hasil prediksi negatif yang benar (*True Negative*), $C_{1,0}$ merepresentasikan hasil prediksi negatif yang salah (*False Negative*), $C_{1,1}$ merepresentasikan hasil prediksi positif yang benar (*True Positive*), dan $C_{0,1}$ merepresentasikan hasil prediksi positif yang salah (*False Positive*).



Berikut adalah matriks *confusion* yang dihasilkan berdasarkan uji coba kita.

```
[ ]  from sklearn.metrics import accuracy_score,confusion_matrix
     from sklearn.metrics import classification_report
```

```
confusion_matrix(hasil, y_test)
```

```
array([[ 0,  0],
       [ 1, 40]])
```

Berdasarkan analisis terhadap 41 baris data uji coba, kita memiliki 40 keluaran prediksi sentimen positif yang benar dan 1 keluaran prediksi negatif yang salah. Total nilai akurasi dari hasil uji coba ini adalah 97.56%.

```
[ ]  accuracy_score(hasil,y_test)*100
```

```
97.5609756097561
```

```
eprint(classification_report(y_test, hasil))
```

```
              precision    recall  f1-score   support

     negatif       0.00      0.00      0.00         1
     positif       0.98      1.00      0.99        40

    accuracy                           0.98        41
   macro avg       0.49      0.50      0.49        41
weighted avg       0.95      0.98      0.96        41
```