MINI PROJECT WEEK 10

ANALISIS DATA MEDIA SOSIAL

Sentiment Analysis with Naïve Bayes

Amanda Gozali – 6162001169

Analisis sentimen merupakan metode atau proses yang digunakan untuk mengenali dan menilai sentimen, opini, dan emosi yang terkandung dalam sebuah teks. Fokusnya adalah untuk mengidentifikasi apakah suatu kalimat mengandung sentimen yang positif, negatif, atau netral. Sentimen ini berhubungan dengan perasaan atau penilaian yang terkait dengan topik tertentu. Dalam laporan ini, akan dilakukan analisis sentimen berdasarkan dataset komentar dan penilaian yang diberikan oleh pelanggan terhadap pembelian makanan cepat saji melalui situs-situs ojek *online* ataupun situs *review* makanan.

Tahap awal yang dilakukan adalah persiapan data (preparing data). Proses ini dimulai dengan membaca dataset, membuat DataFrame berdasarkan dataset yang ada, dan menghilangkan data kosong (null) dalam dataset. Hasil awal dari tahap ini adalah lima data pertama yang ditampilkan sebagai berikut:

	No	Brand	Rating	Review	Source
0	1	Pizza Hut Delivery	3	rose pizza, rotinya masih mentah yang bagian d...	https://gofood.co.id/en/bandung/restaurant/piz...
1	2	Mcdonald's	1	Saya pesan setau saya gambarnya pake nasi. Kal...	https://gofood.co.id/id/bandung/restaurant/mcd...
2	3	KFC	5	Ini mah nggak usah diragukan, udah langganan j...	https://gofood.co.id/en/bandung/restaurant/kfc...
3	4	Burger Bangor	3	Ukuran kecil untuk harga segitu. Tapi rasa ok.	https://gofood.co.id/en/bandung/restaurant/bur...
4	5	Burger King	5	enak bumbu madura nya, ayamnya pun juicy ga ke...	https://gofood.co.id/id/bandung/restaurant/bur...

Berdasarkan rating-rating tersebut, akan dilakukan pengelompokan kategori sentimen positif dan negatif. Pada dataset bentuk penilaian yang tertera adalah bintang 1 sampai 5 yang kemudian akan diubah bentuknya berdasarkan sentimennya yaitu ‘positif’ dan ‘negatif’ dengan bintang 1-3 diubah menjadi ‘negatif’ sementara bintang 4-5 diubah menjadi ‘positif’. Dengan mengacu pada hal tersebut, didapatkan 155 rating berkategori negatif dan 95 rating berkategori positif. Sehingga kolom Rating pada dataset akan berubah menjadi:

	No	Brand	Rating	Review	Source
0	1	Pizza Hut Delivery	negatif	rose pizza, rotinya masih mentah yang bagian d...	https://gofood.co.id/en/bandung/restaurant/piz...
1	2	Mcdonald's	negatif	Saya pesan setau saya gambarnya pake nasi. Kal...	https://gofood.co.id/id/bandung/restaurant/mcd...
2	3	KFC	positif	Ini mah nggak usah diragukan, udah langganan j...	https://gofood.co.id/en/bandung/restaurant/kfc...
3	4	Burger Bangor	negatif	Ukuran kecil untuk harga segitu. Tapi rasa ok.	https://gofood.co.id/en/bandung/restaurant/bur...
4	5	Burger King	positif	enak bumbu madura nya, ayamnya pun juicy ga ke...	https://gofood.co.id/id/bandung/restaurant/bur...

Langkah selanjutnya adalah melakukan *pre-processing* data. Pada langkah ini, data dalam kolom *Review* akan dibersihkan dari karakter yang tidak diinginkan. Kata-kata yang sering muncul namun tidak memiliki pengaruh signifikan dalam pemahaman arti kalimat akan dihilangkan berdasarkan daftar *stopwords* yang telah diunduh sebelumnya. Selain itu, seluruh huruf dalam kalimat akan diubah menjadi huruf kecil agar tidak menyebabkan *error* pada tahap selanjutnya. Setelah proses pembersihan ini selesai, komentar yang telah dibersihkan akan digabungkan kembali dengan DataFrame asli dengan nama kolom 'clean_text'. Berikut ini adalah hasil output yang dihasilkan:

	No	Brand	Rating	Review	Source	clean_text
0	1	Pizza Hut Delivery	negatif	rose pizza, rotinya masih mentah yang bagian d...	https://gofood.co.id/en/bandung/restaurant/piz...	rose pizza rotinya mentah bagian dalamnya
1	2	Mcdonald's	negatif	Saya pesan setau saya gambarnya pake nasi. Kal...	https://gofood.co.id/id/bandung/restaurant/mcd...	pesan setau gambarnya pake nasi kalo ga pake y...
2	3	KFC	positif	Ini mah nggak usah diragukan, udah langganan j...	https://gofood.co.id/en/bandung/restaurant/kfc...	mah nggak usah diragukan udah langganan wkwkwk
3	4	Burger Bangor	negatif	Ukuran kecil untuk harga segitu. Tapi rasa ok.	https://gofood.co.id/en/bandung/restaurant/bur...	ukuran untuk harga segitu rasa ok
4	5	Burger King	positif	enak bumbu madura nya, ayamnya pun juicy ga ke...	https://gofood.co.id/id/bandung/restaurant/bur...	enak bumbu madura nya ayamnya juicy ga kering

Langkah berikutnya melibatkan pembagian data menjadi dua kelompok, yaitu *training set* dan *testing set*, menggunakan pustaka sklearn dengan menggunakan fungsi train_test_split. Pada tahap ini, 80% dari data akan dimasukkan ke dalam training set, sementara sisanya sebesar 20% akan dimasukkan ke dalam testing set. Tujuan dari langkah ini adalah untuk menguji signifikansi atau kinerja model dalam memprediksi nilai yang sebenarnya dengan melatih model menggunakan *training set* dan mengujinya pada data yang belum pernah dilihat sebelumnya dalam *testing set*. Setelah data berhasil dibagi ke dalam dua kelompok, langkah selanjutnya adalah melakukan vektorisasi pada kedua kelompok tersebut, menghasilkan hasil sebagai berikut:

```
  (0, 101)      1
  (0, 114)      2
  (0, 123)      1
  (0, 299)      1
  (0, 349)      1
  (0, 356)      1
  (0, 464)      1
  (0, 478)      1
  (0, 522)      3
  (0, 824)      1
  (0, 904)      1
  (0, 987)      1
  (0, 1005)     1
  (0, 1065)     1
  (0, 1089)     1
  (0, 1105)     1
  (0, 1108)     2
  (0, 1112)     1
  (0, 1120)     1
  (0, 1148)     1
  (0, 1151)     1
  (0, 1182)     1
  (0, 1197)     3
  (0, 1199)     1
  (0, 1365)     1
  :     :
```

```
  :     :
  (45, 1384)    1
  (45, 1416)    1
  (45, 1571)    1
  (45, 1591)    1
  (45, 1623)    2
  (45, 1659)    1
  (45, 1696)    1
  (45, 1704)    1
  (46, 478)     1
  (46, 1153)    1
  (46, 1460)    1
  (46, 1621)    1
  (47, 478)     1
  (47, 691)     1
  (47, 767)     1
  (48, 134)     1
  (48, 158)     1
  (48, 229)     1
  (48, 306)     1
  (48, 346)     1
  (48, 735)     1
  (48, 1241)    1
  (48, 1339)    1
  (49, 197)     1
  (49, 1308)    1
```

Hasil yang diperoleh dapat dijelaskan bahwa pada baris pertama '(0, 101) 1', artinya terdapat kata dengan indeks 101 yang muncul satu kali dalam dokumen atau komentar pertama. Hal yang sama berlaku untuk baris-baris berikutnya.

Langkah kelima melibatkan pelatihan model *Naive Bayes* menggunakan algoritma *Bernoulli Naive Bayes* yang disebut sebagai *model definition*. Proses ini meliputi pembuatan model untuk melatih dan memprediksi data, menyesuaikan model dengan data pada *training set* agar dapat mempelajari pola data, menggunakan model yang telah dilatih untuk memprediksi data pada *testing set*, dan akhirnya

menghasilkan *output* berupa nilai prediksi dan nilai sebenarnya dari data pada *testing set*. Nilai prediksi yang diperoleh akan memiliki nilai berikut ini:

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

Sementara nilai sebenarnya,

47	positif
245	positif
71	negatif
49	positif
128	positif
182	negatif
237	negatif
99	negatif
229	negatif
113	negatif
226	positif
4	positif
58	negatif
31	positif
178	negatif
211	negatif
53	negatif
225	negatif
100	negatif
207	positif
50	negatif
155	positif
189	negatif
247	negatif
243	positif
86	positif

186	negatif
234	positif
217	negatif
94	negatif
11	negatif
56	positif
166	negatif
180	positif
107	positif
34	negatif
132	negatif
249	positif
171	negatif
184	negatif
136	negatif
165	negatif
123	positif
122	negatif
116	positif
151	negatif
218	negatif
22	positif
66	negatif
140	negatif

Berdasarkan hasil output tersebut, prediksi yang dihasilkan oleh model dapat dibandingkan dengan nilai sebenarnya pada data *testing set*. Dengan demikian, dapat disimpulkan seberapa akurat model dalam memprediksi nilai sebenarnya pada data *testing set*.

Tahap terakhir akan dicek keakuratan model melalui hasil yang diperoleh berikut ini

	Positif	Negatif
Positif	30	18
Negatif	1	1

Berdasarkan tabel diatas dapat diperoleh bahwa nilai *True Positive* adalah 30, *False Positive* adalah 18, *False Negative* adalah 1, dan *True Negative* adalah 1. Dapat diperoleh juga nilai akurasi, presisi, recall, f1-score, dan support dalam tabel berikut:

	precision	recall	f1-score	support
negatif	0.97	0.62	0.76	48
positif	0.05	0.50	0.10	2
accuracy			0.62	50
macro avg	0.51	0.56	0.43	50
weighted avg	0.93	0.62	0.73	50

Berdasarkan hasil di atas dapat disimpulkan bahwa akurasi model untuk sentimen negatif adalah 0.97, yang berarti model dapat mendekati nilai sebenarnya dari sentimen negatif dengan akurasi 97%. Dalam hal memprediksi sentimen positif, model memiliki akurasi yang lebih rendah yaitu 5%, sehingga tidak sebaik dalam memprediksi sentimen negatif. *Recall* adalah ukuran sejauh mana model dapat mengenali kalimat yang mengandung sentimen negatif dan positif. Dari hasil yang diperoleh, model dapat mengenali 62% komentar yang mengandung sentimen negatif, tetapi hanya dapat mengenali 50% komentar yang mengandung sentimen positif. Secara keseluruhan, model dapat mengklasifikasikan 62% dari seluruh data yang ada (dilihat dari akurasi pada hasil output yang diperoleh).