**LAPORAN AKHIR SEMESTER MACHINE LEARNING**

**DETEKSI EMOSI PENGGUNA TWEETER**

Diajukan Untuk Memenuhi Tugas

Mata Kuliah Machine Learning

Yang diampu oleh:

**Ibu Adevian Fairuz Pratama, S.S.T, M.Eng.**

Semester Genap Tahun Akademik 2022/2023

****

**Disusun Oleh:**

**Iqri Mannisa’ Buchori (2041720066 / 12)**

**Wazir Qorni Abud (2041720124 / 21)**

**PROGRAM STUDI D-IV TEKNIK INFORMATIKA**

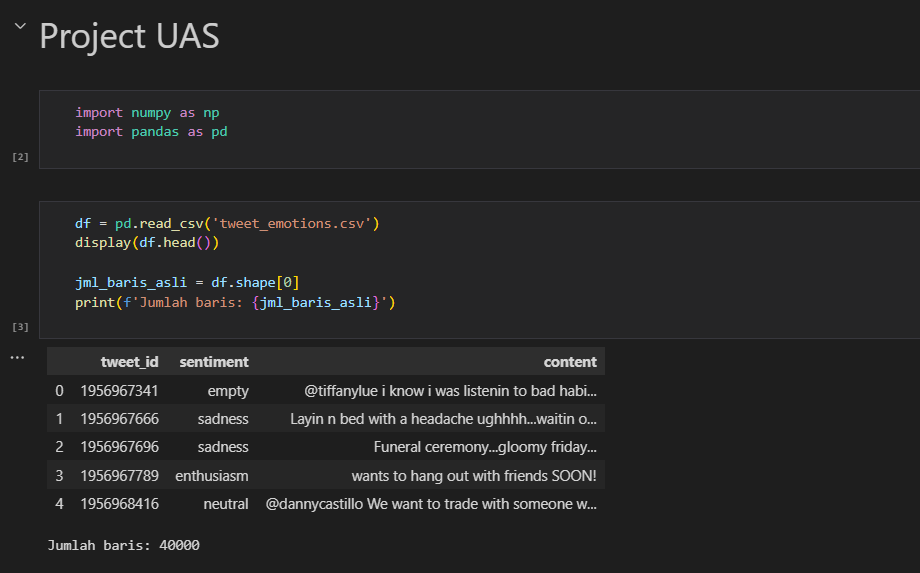
**JURUSAN TEKNOLOGI INFORMASI**

**POLITEKNIK NEGERI MALANG**

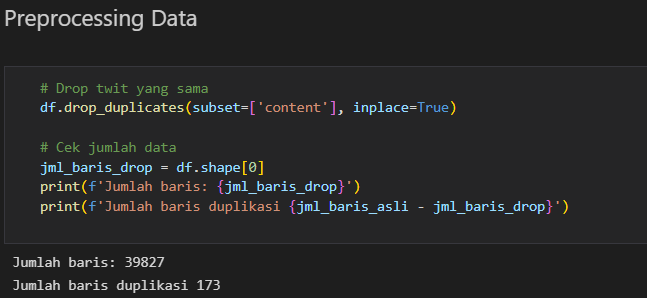
**2022**

1. **Prepocessing Data**

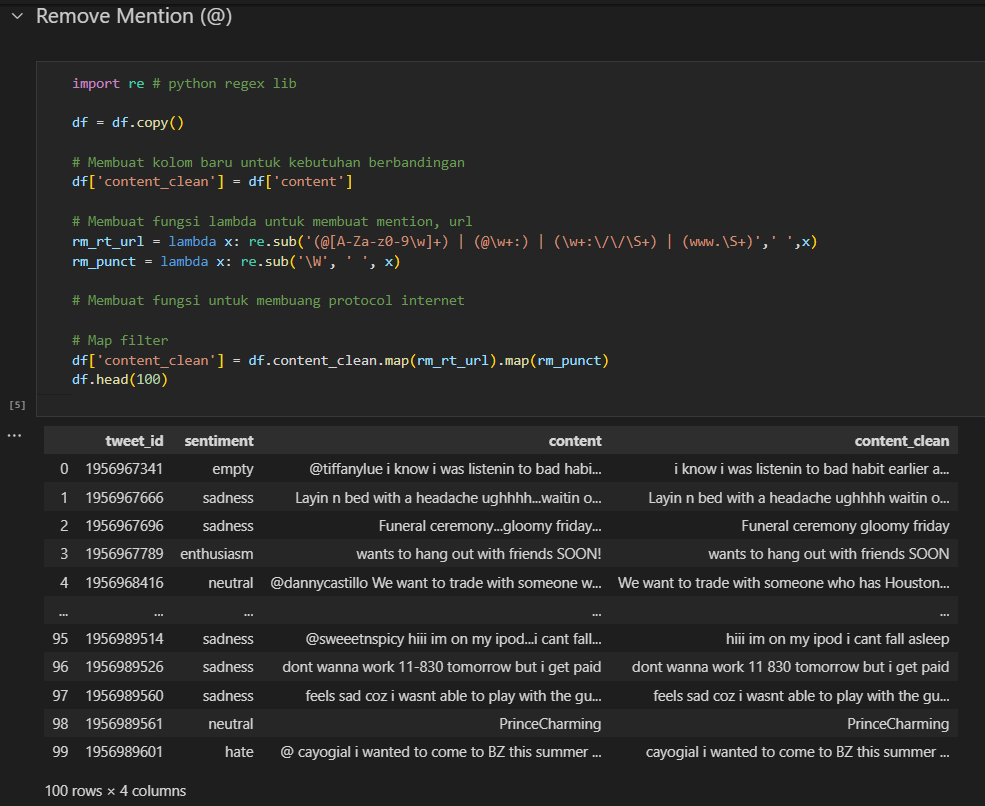
Data yang akan diolah untuk project UAS kali ini adalah data *tweet\_emotions.csv.* Tahap pertama yang dapat dilakukan adalah load dataset kedalam Dataframe menggunakan Pandas.



Setelah data terbaca oleh fungsi python, didapati jumlah baris yang terdapat pada data *tweet\_emotions.csv* berjumlah40.000. Kemudian masuk pada tahap Prepocessing Data dimana akan dilakukan inisialisasi terlebih dahulu apakah didapati data yang terduplikasi dari 40.000 data yang ada. Data yang terduplikasi akan dihapus dengan menggunakan method drop\_duplicates.

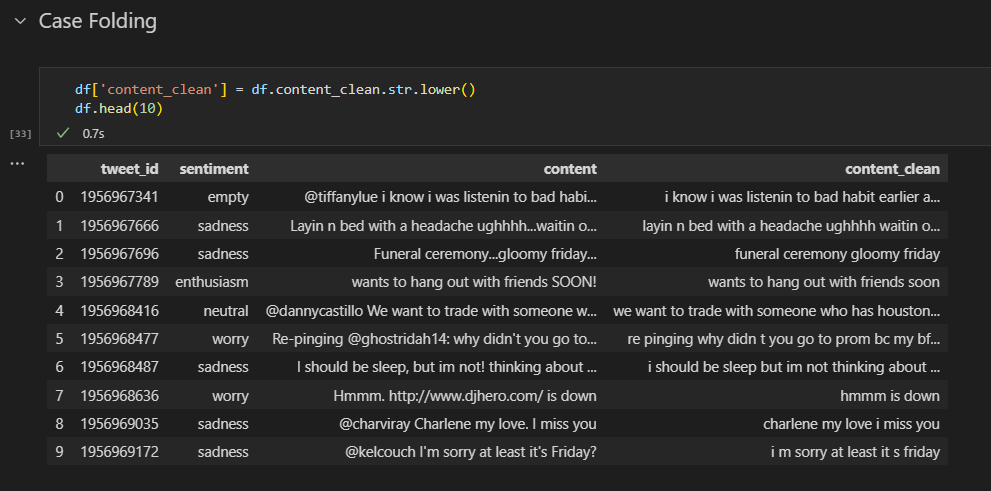


Setelah menghapus data yang terduplikasi masuk pada tahap operasi dasar yang digunakan pada tahap pra pengolahan data adalah Case Folding, Tokenizing, Filtering, dan Stemming. Namun sebelum itu hal yang perlu diperhatikan lagi adalah menghapus mention @.



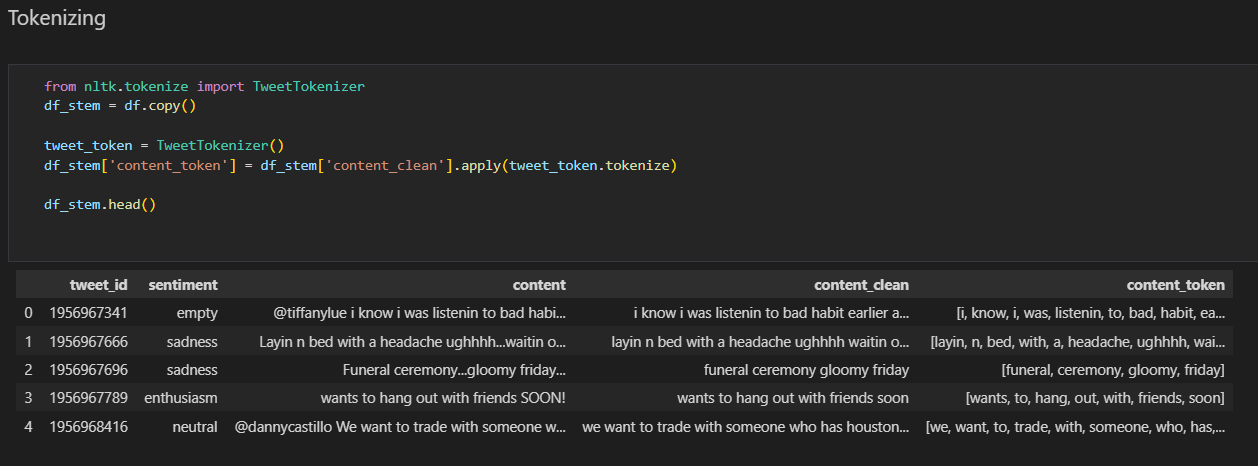
1. Case Folding

Case Folding digunakan untuk mengubah semua bentuk huruf dalam sebuah teks atau mengubah isi dokumen menjadi huruf kecil semua. Sementara itu, karakter lain yang bukan termasuk huruf dan angka, seperti tanda baca dan spasi dianggap sebagai delimiter. Delimiter ini bisa juga dihapus atau diabaikan dengan menggunakan perintah yang ada di Python.



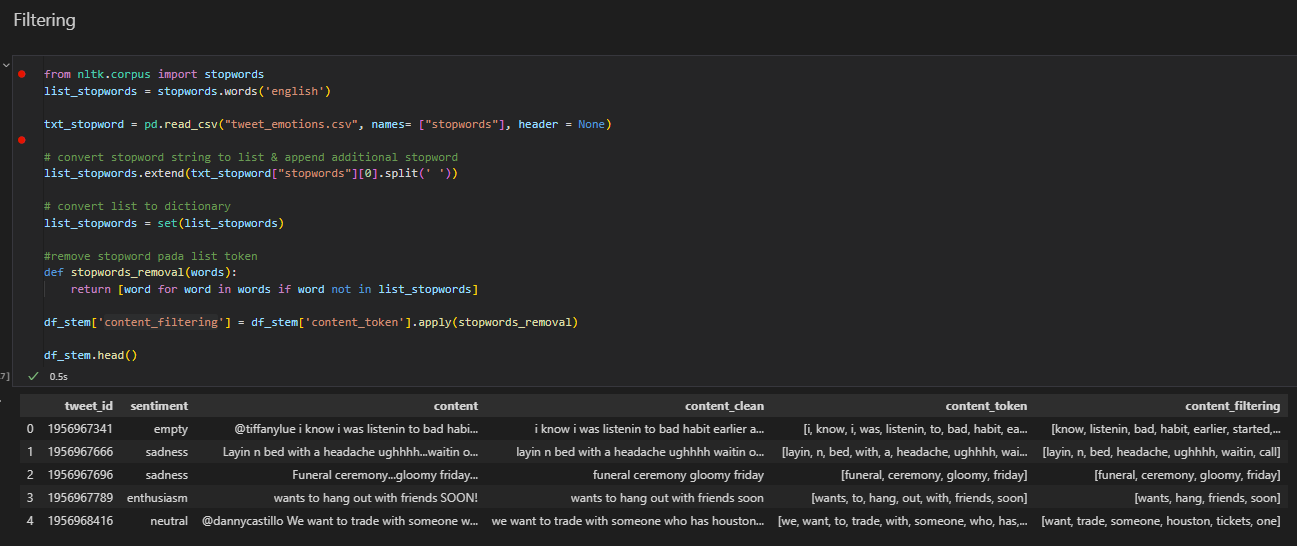
1. Tokenizing

Ditahap ini akan dilakukan proses number removal, whitecase removal, puntuation removal dan word\_tokenize() untuk memecah string kedalam tokens. Pandas Dataframe atau Series mampu menjalankan function external untuk di terapkan pada kolom atau baris dengan menggunakan fungsi .*apply*().



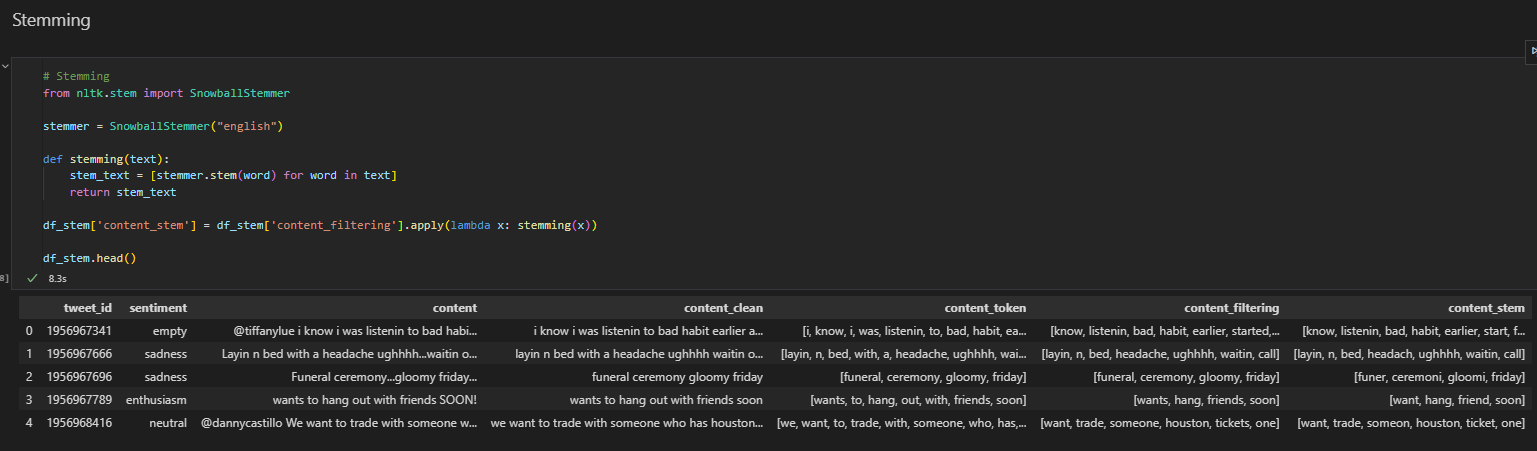
1. Filtering

Pada tahap ini kita akan menggunakan stopword bahasa English yang didapatkan dari library NLTK untuk filtering terhadap Dataframe. Tahapan filtering yang digunakan untuk mengambil kata-kata yang penting dari hasil token tadi. Kata umum yang biasanya muncul dan tidak memiliki makna disebut dengan stopword.



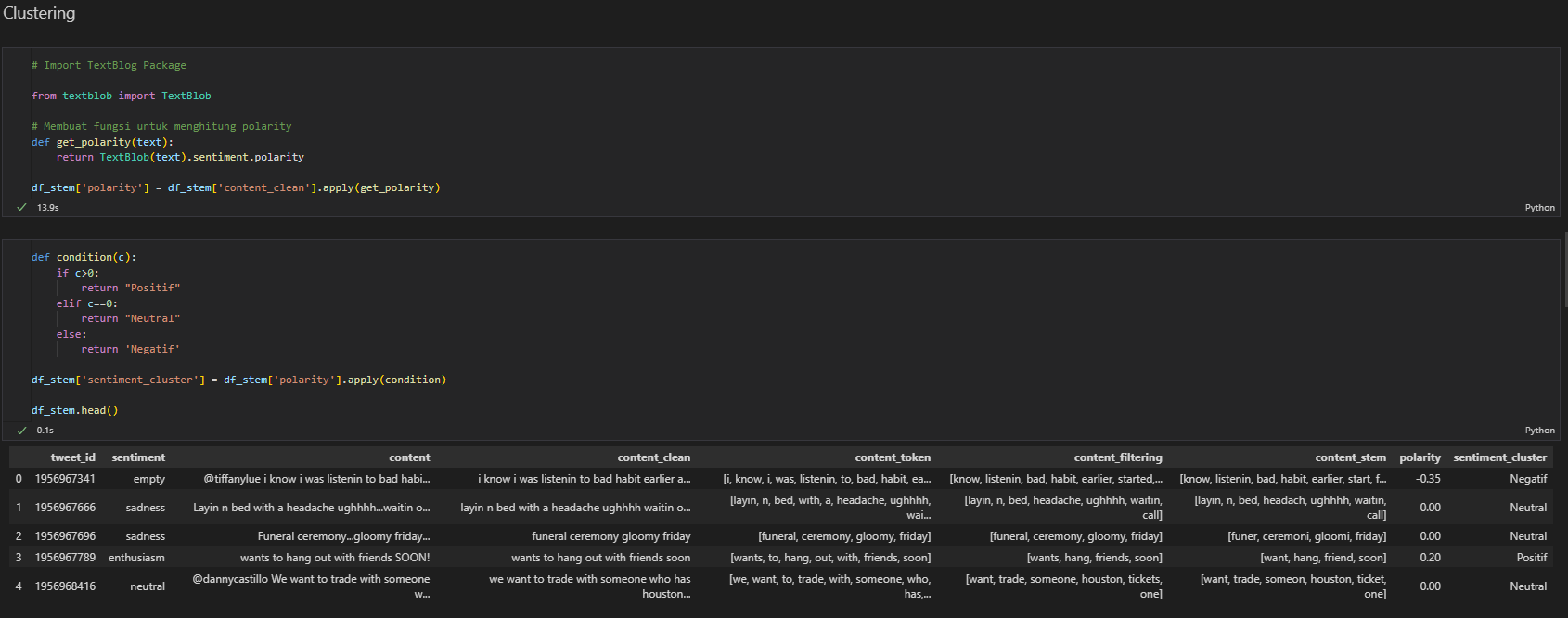
1. Stemming

Stemming adalah proses mengurangi infleksi kata-kata ke bentuk akarnya, seperti memetakan sekelompok kata ke batang yang sama, bahkan jika batang itu sendiri bukan kata yang valid dalam Bahasa.



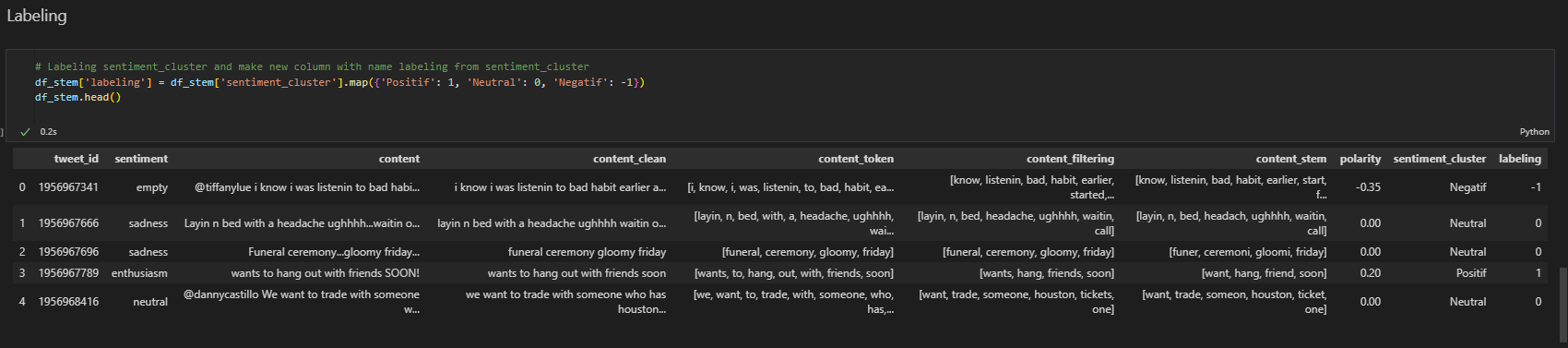
1. **Clustering**

Pengelompokan data ke dalam beberapa kategori atau cluster, yaitu komentar positif, netral, dan negatif. Clustering adalah sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum.

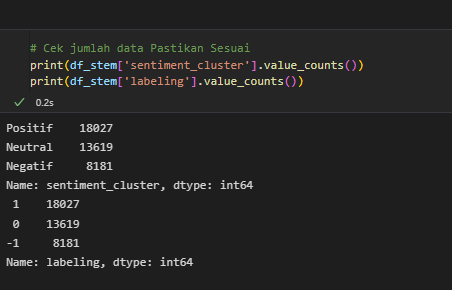


1. **Labeling**

Tahap selanjutnya adalah Labeling dimana hasil dari pengelompokan atau cluster diatas diberikan label hasil dari pengelompokan example melalui clustering. Seperti hasil kode program berikut yang memberikan label dengan keterangan Positif =1, Neutral = 0, Negatif = -1.

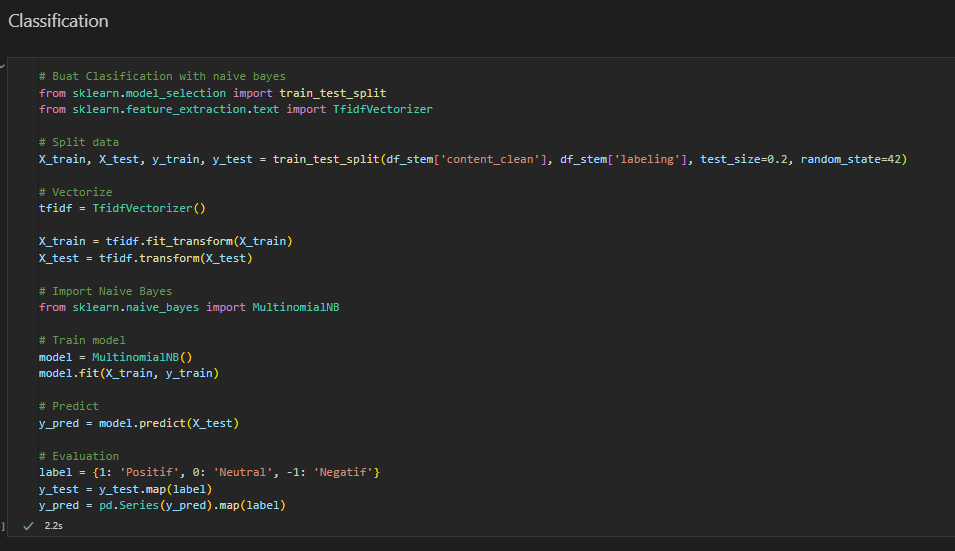


Setelah melakukan labeling perlu dicek kembali jumlah data yang telah dibuatkan labeling sesuai dengan pengelompokan sebelumnya dengan memanfaatkan *method .value\_counts().*



1. **Classification**

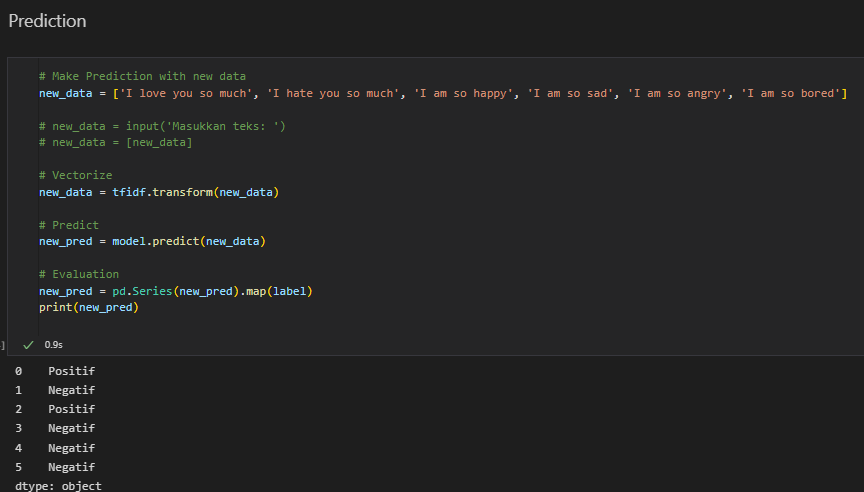
Pada tahapn classification ini menggunakan ekstraksi fitur TfIdfVEctorizer dan metode klasifikasi Naïve Bayes. Hal ini digunakan untuk klasifikasi teks yang melibatkan set data pelatihan dimensi tinggi.



Proses yang dilakukan mulai dari import library yang akan digunakan, kemudian melakukan split data, modeling dengan menggunakan TfidfVectorizer(), train model, melakukan prediksi atas model yang dibuat dan melakukan evaluasi.

1. **Predict**

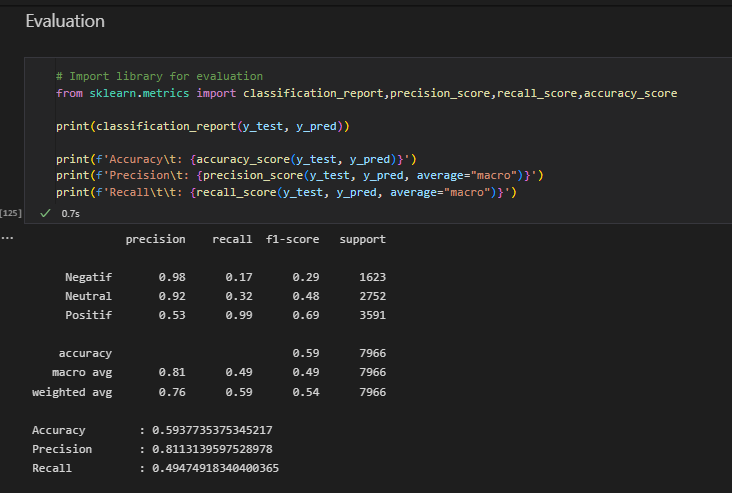
Tahapan ini adalah melakukan uji coba apakah machine learning yang dibuat bekerja dengan baik. Caranya dengan melihat hasil atau prediksi yang dihasilkan. Apakah sesuai dengan *input data*. Maka tahapan ini diputuskan untuk membuat sebuah prediksi yang didapatkan dari inputan kalimat baru seperti pada kode program berikut.



1. **Evaluasi**

Pada proses evaluasi, menggunakan metric akurasi dan juga menambahkan metric lain seperti Recall, Precision, F1-Score, detail Confussion Metric, ataupun Area Under Curve (AUC).

Pada pengevaluasian menggunakan fungsi accuracy\_score, classification\_report, precision\_score dari library sklearn.metrics untuk mendapatkan nilai-nilai matriks dari data yang digunakan.

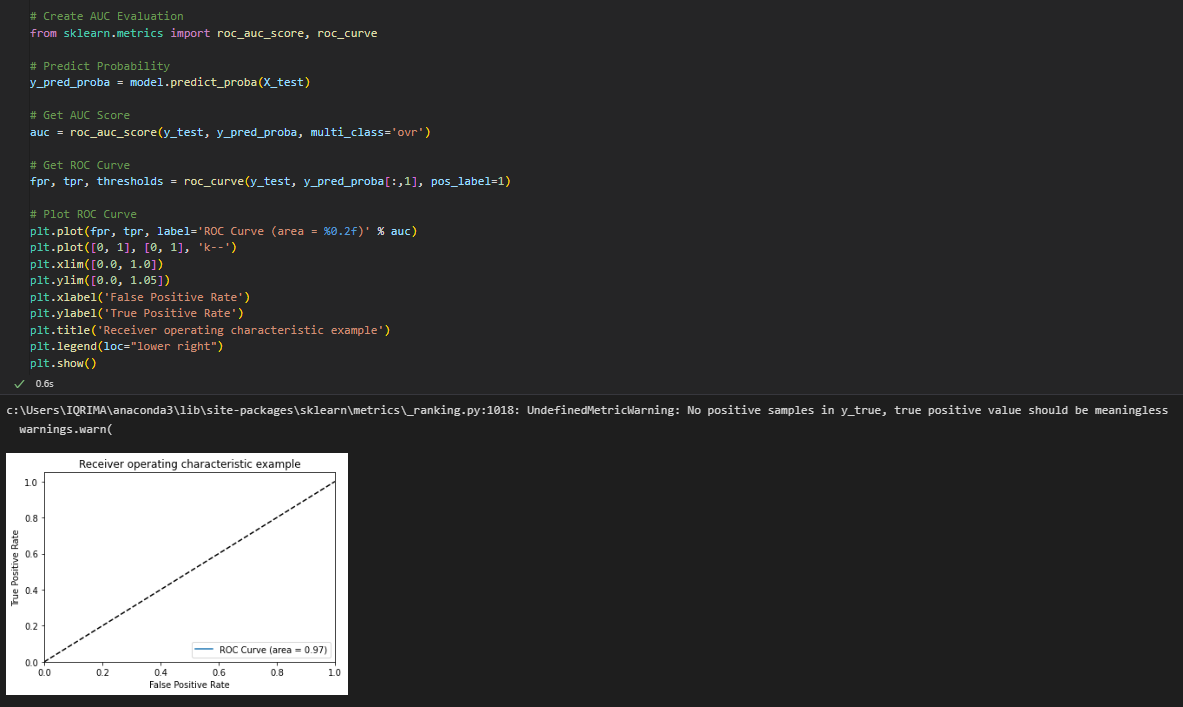


Dari perhitungan matrik diatas memunculkan nilai akurasi yang kurang baik yakni 0.59.

Kemudian membuat confusion Matrix Confusion Matrix merupakan metode evaluasi yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran dari proses klasifikasi seperti berikut



ROC Curve dibuat berdasarkan nilai telah didapatkan pada perhitungan dengan confusion matrix, yaitu antara False Positive Rate dengan True Positive Rate. Sehingga dihasilkan seperti pada gambar beriku ini.



Kesimpulan :

Ukuran besaran precision, recall, dan accuracy biasanya diberi nilai dalam bentuk presentase antara 1 sampai 100%. Sebuah sistem akan dianggap baik jika tingkat precision, recall, dan accuracy-nya tinggi.

Sedangkan dari hasil keluaran data tersebut menunjukkan bahwa keakuratan data training dan data testing rendah. hasil presisi juga menunjukkan rendah karena kurang dari satu. Bobot Akurasi atau tingkat kedekatan antara nilai yang didapat terhadap nilai sebenarnya dari data testing dan data training juga rendah. Presisi atau kecocokan antara bagian data yang diambil dengan informasi yang dibutuhkan rendah. Kemudian Recall atau tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi juga rendah. Intinya sistem ini dapat dianggap tidak baik karena memiliki nilai keakuratan yang rendah.

Link github : <https://github.com/iqrima4422/UAS_ProyekML>