**LAPORAN UJIAN TENGAH 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)**

**PROGRAM STUDI D-IV TEKNIK INFORMATIKA**

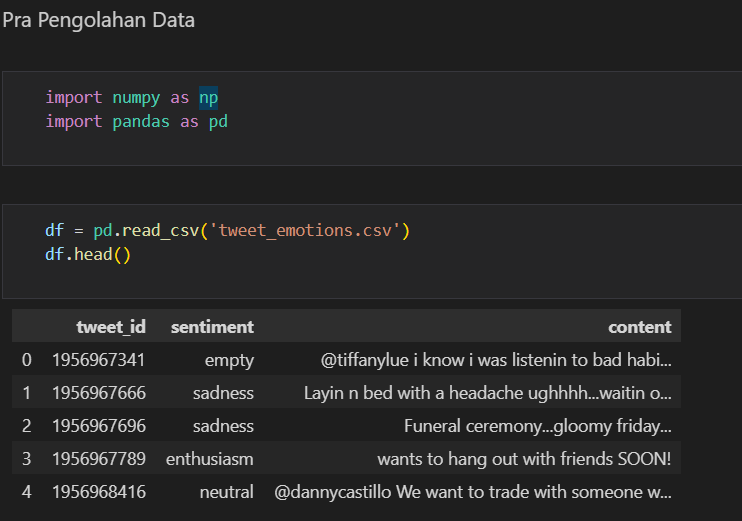
**JURUSAN TEKNOLOGI INFORMASI**

**POLITEKNIK NEGERI MALANG**

**2022**

**Pra Pengolahan Data**

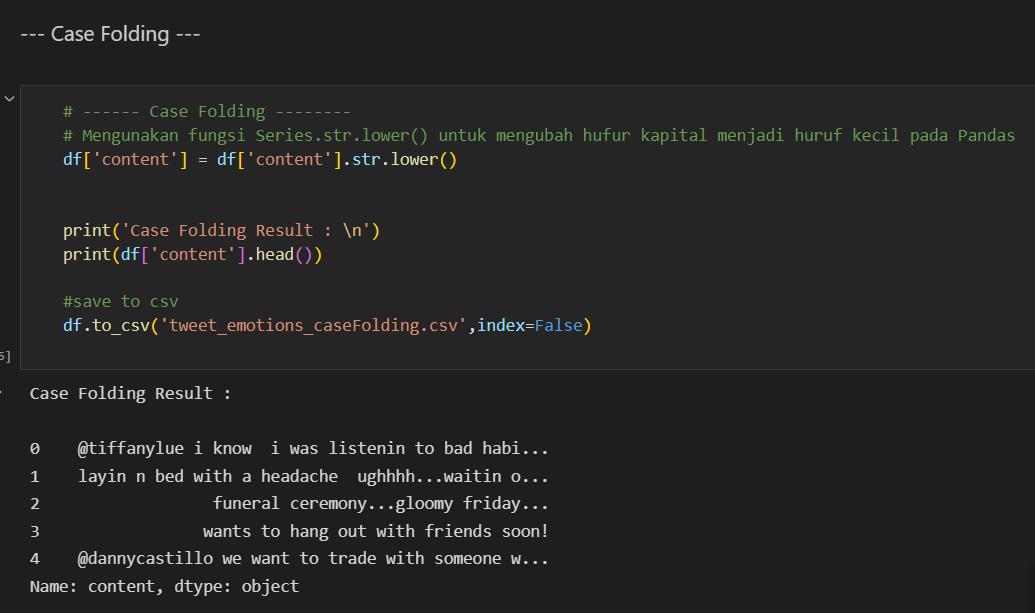
Data yang digunakan untuk pengolahan data kali ini adalah *tweet\_emotions.csv.* tahap pertama yang dapat dilakukan adalah load dataset kedalam Dataframe menggunakan Pandas.

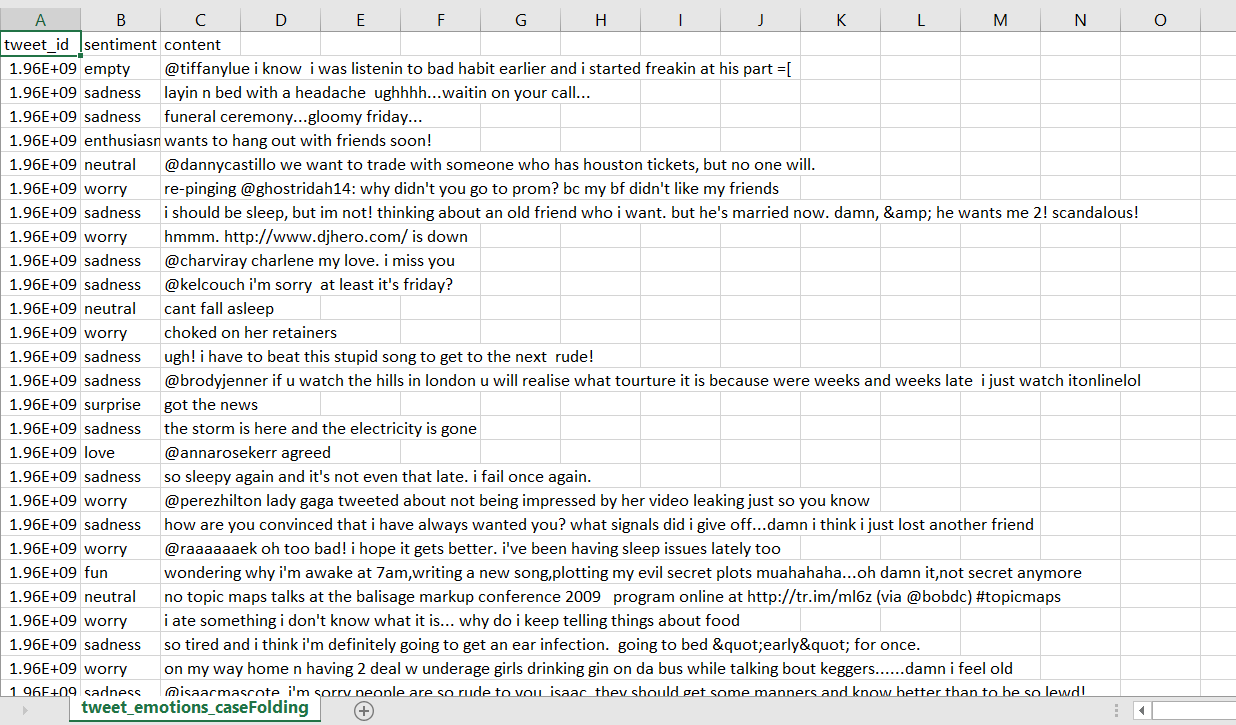
****

Operasi dasar yang digunakan pada tahap pra pengolahan data adalah Case Folding, Tokenizing, Filtering, dan Stemming.

1. **Case Folding**

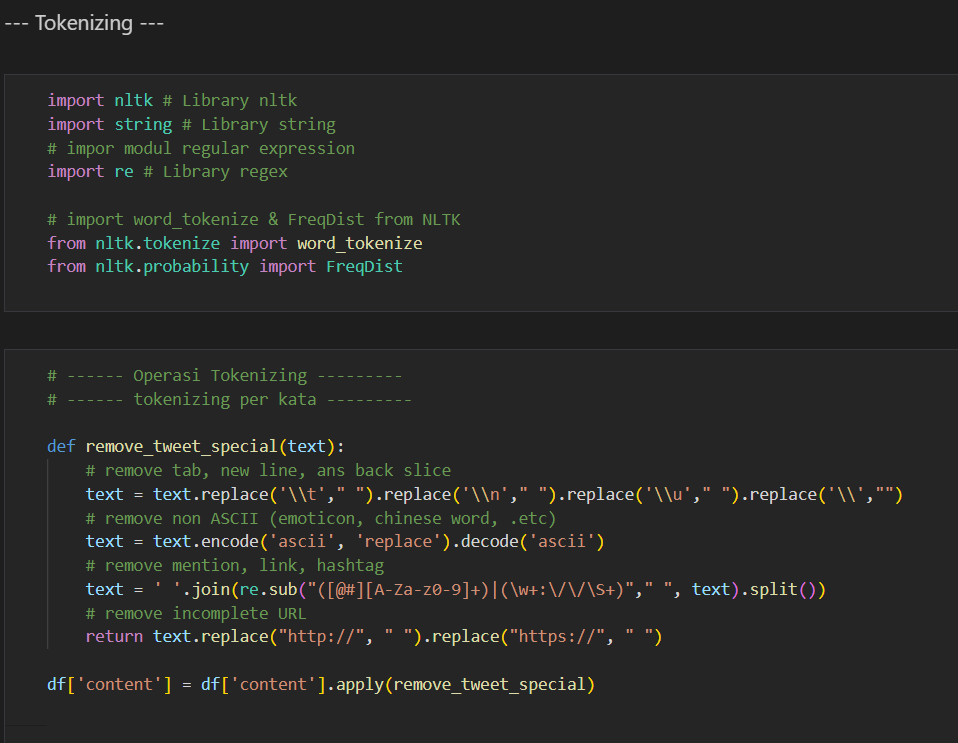
Case Folding digunakan untuk mengubah semua bentuk huruf dalam sebuah teks atau dokumen menjadi huruf kecil semua. Sementara itu, karakter lain yang bukan termasuk huruf dan angka, seperti tanda baca dan spasi dianggap sebagai delimiter. Delimiter ini bisa juga dihapus atau diabaikan dengan menggunakan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

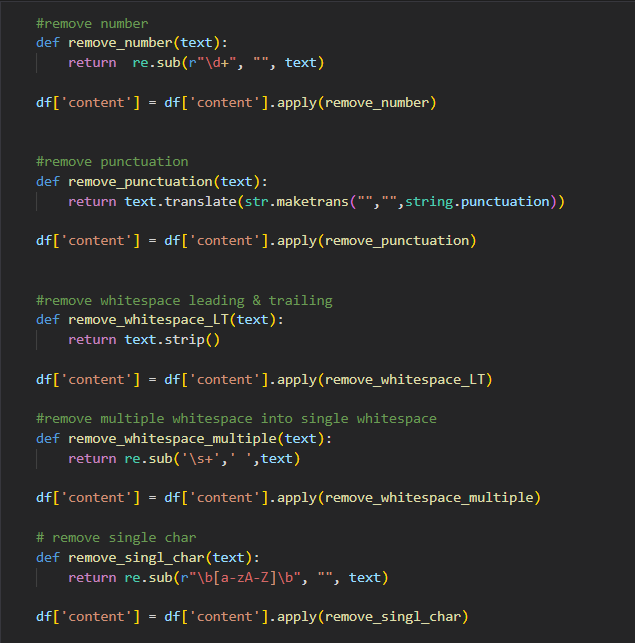


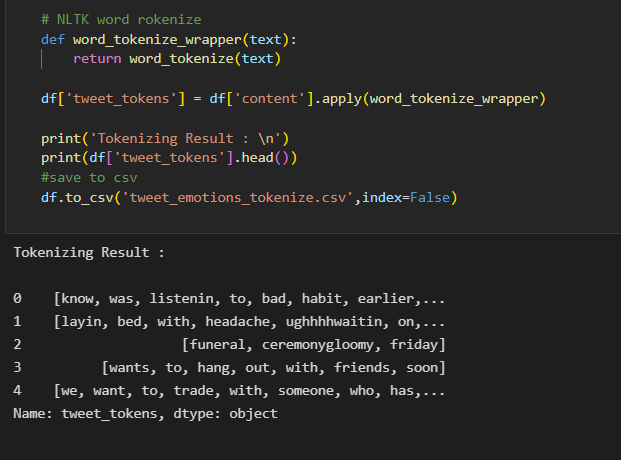


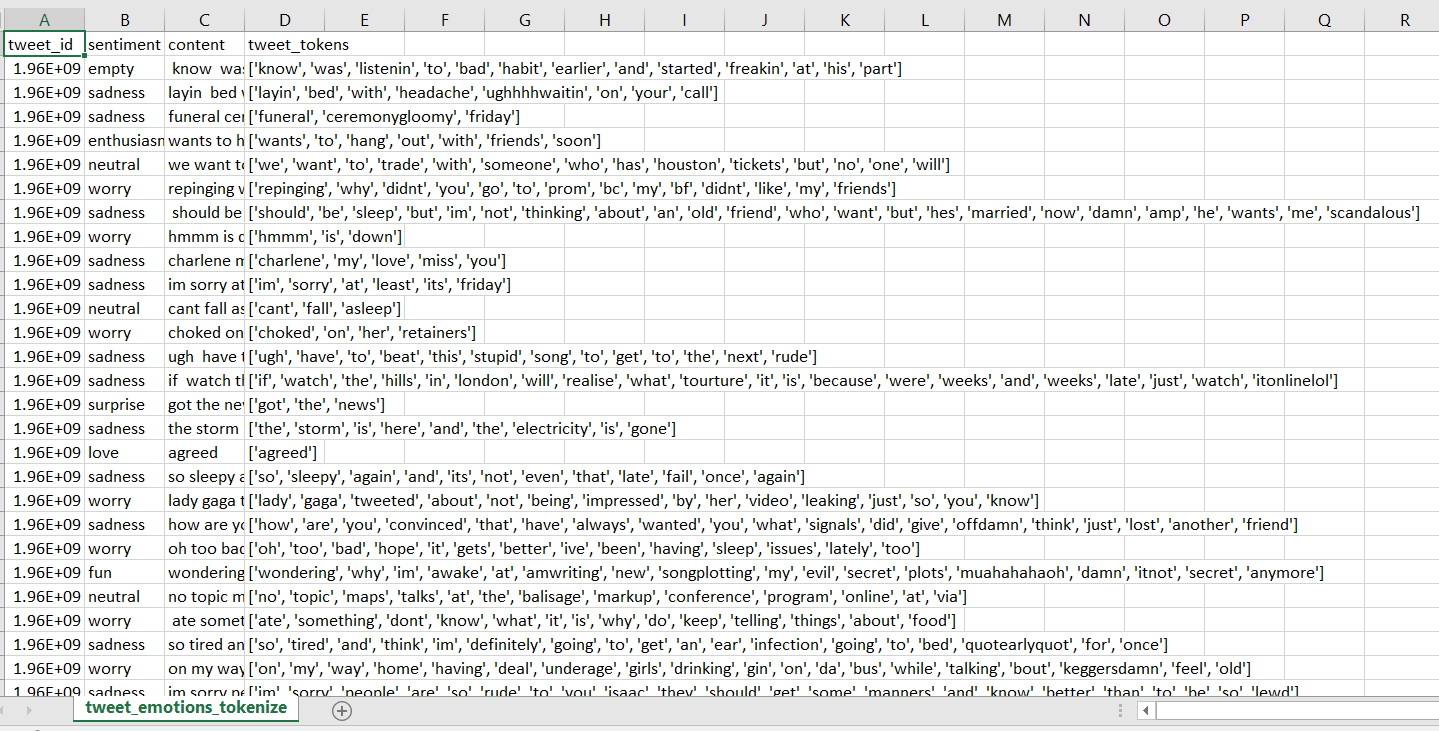
1. **Tokenizing**

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*().



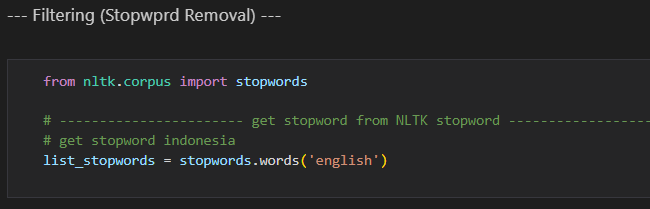




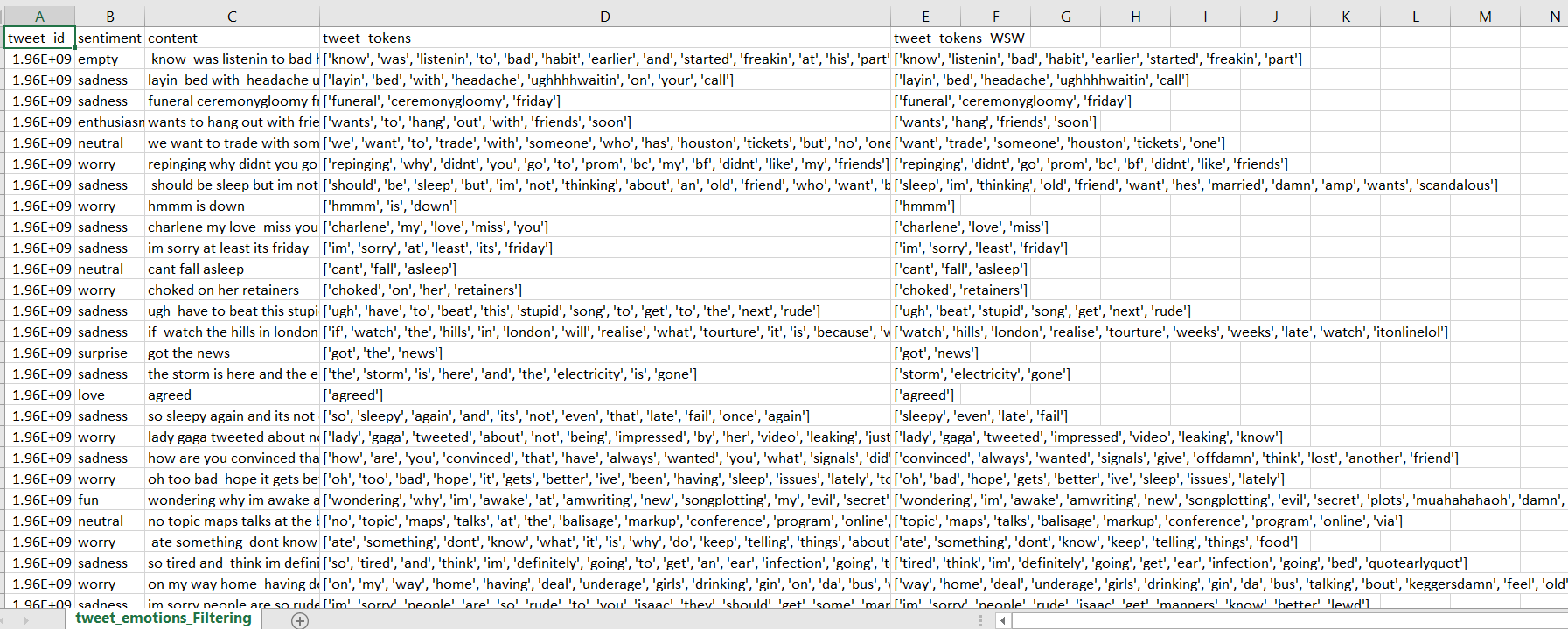


1. **Filtering**

Pada tahap ini kita akan menggunakan stopword bahasa English yang didapatkan dari library NLTK untuk filtering terhadap Dataframe. Juga dapat menambahkan list stopword dengan menggunakan fungsi *.extend()* terhadap list\_stopword , penggunaan fungsi *set()* bermanfaat untuk membuat iterable list menjadi sequence iterable element. Hasilnya kita dapat mempercepat proses pengecekan apakan sebuah token terdapat pada list\_stopword atau tidak.

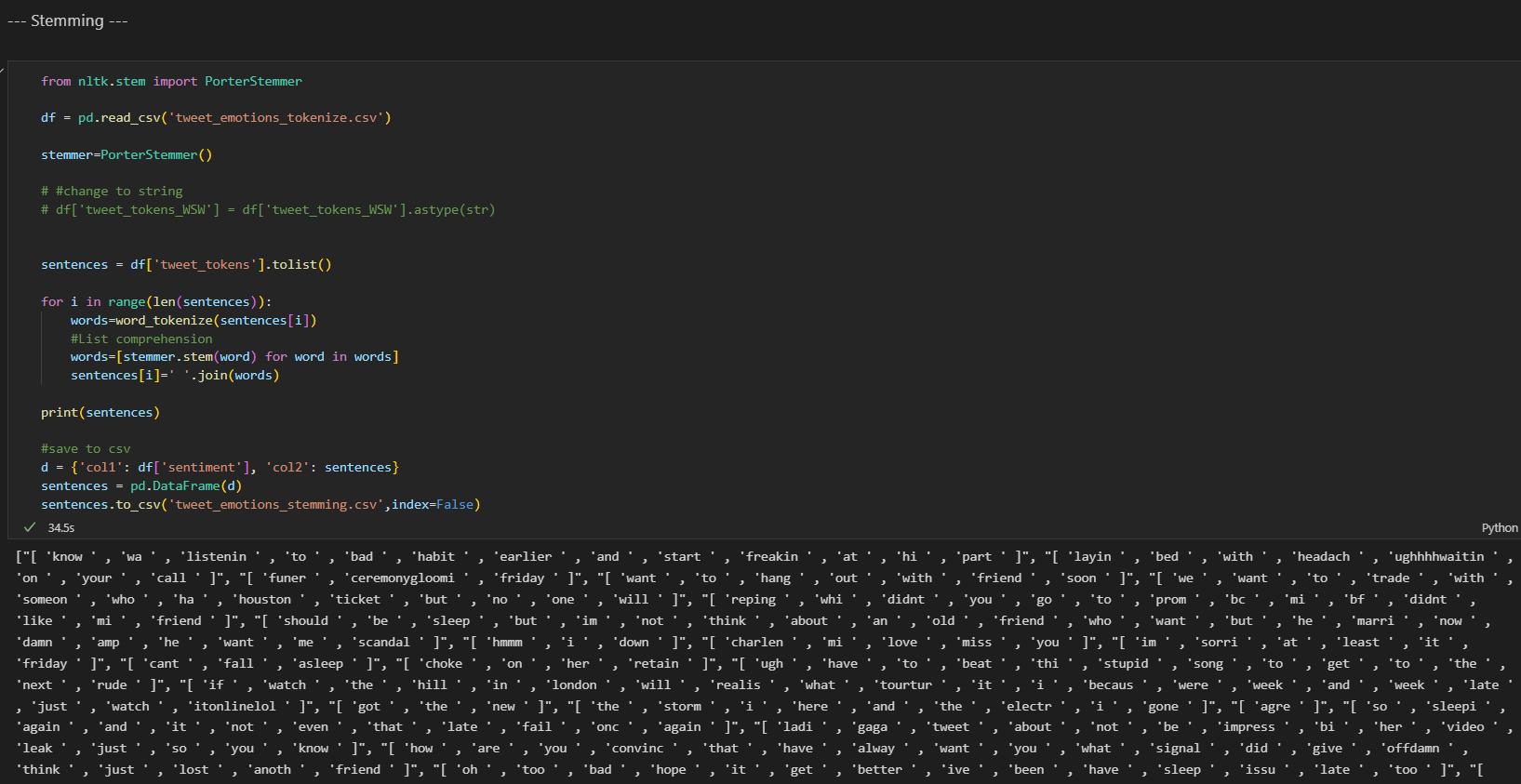


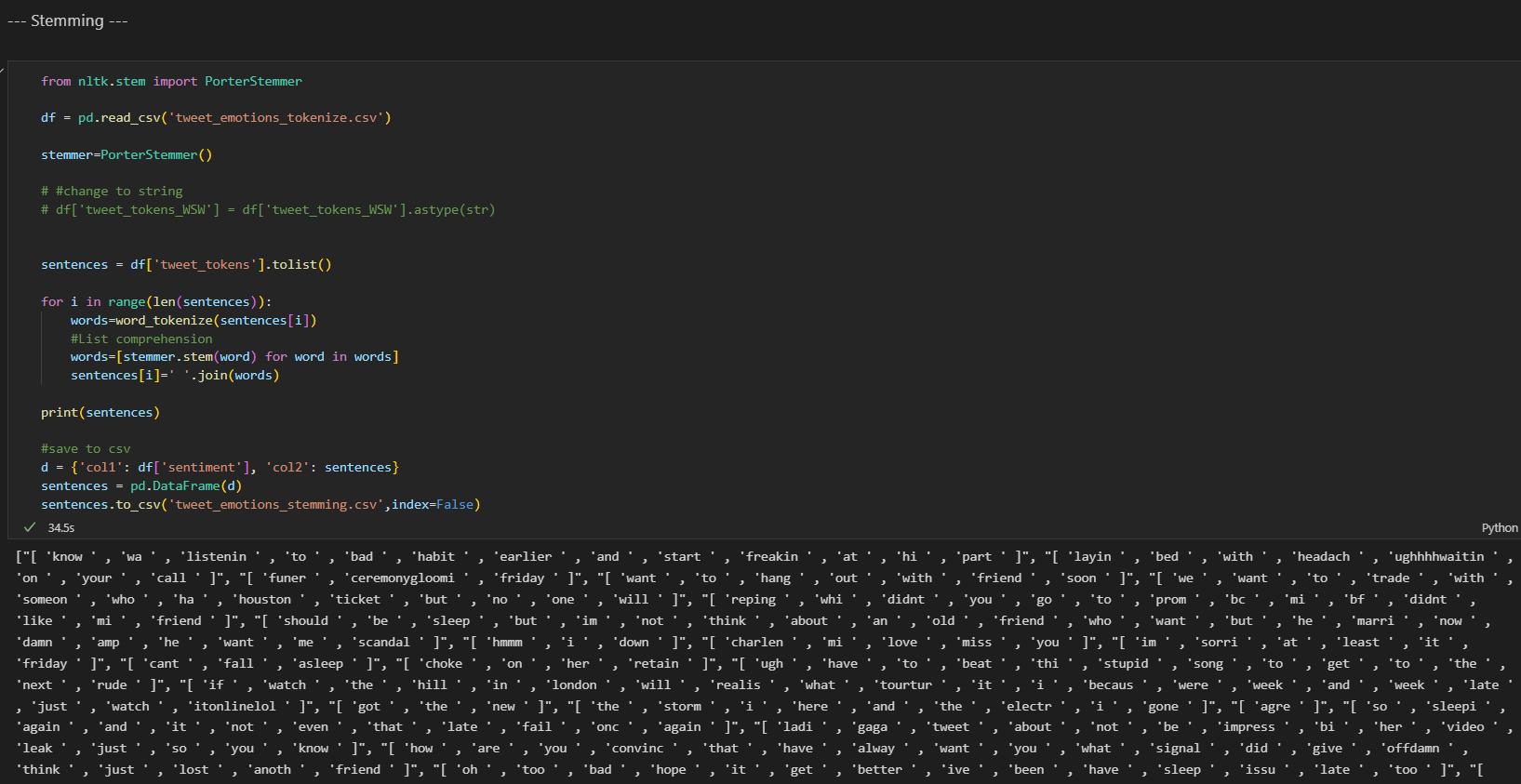


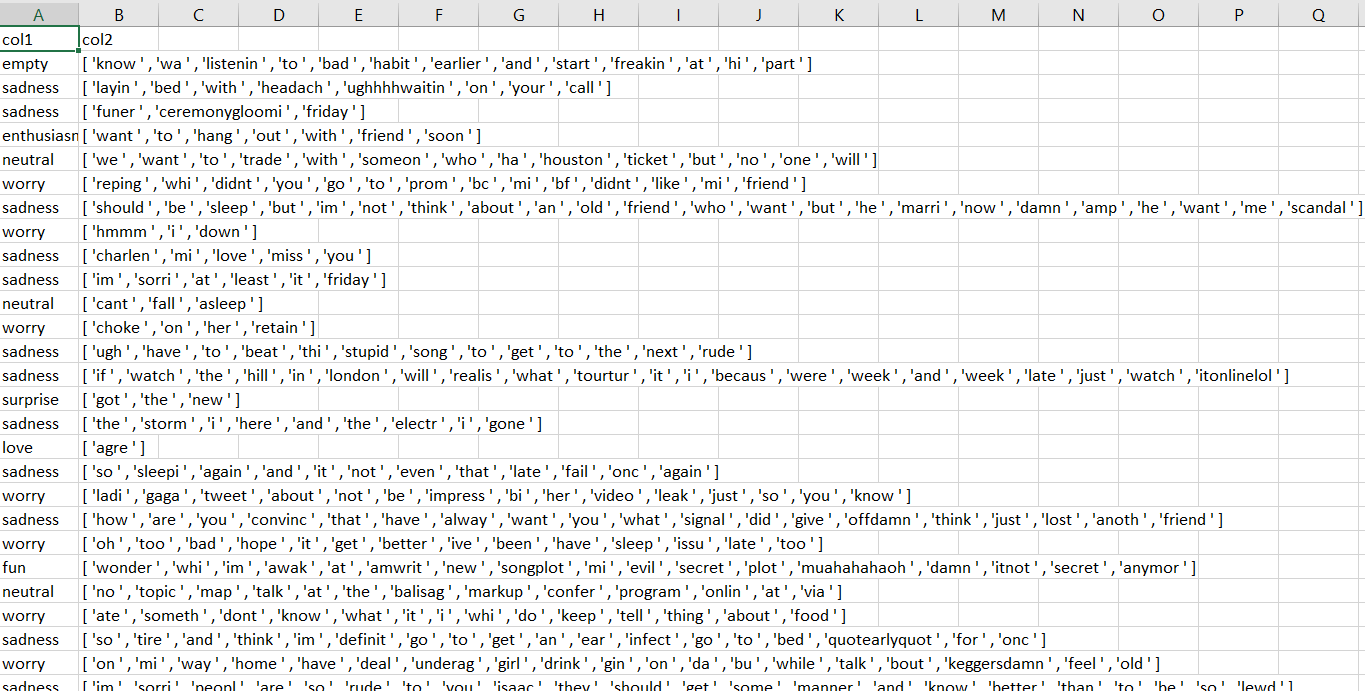


1. **Stemming**

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.

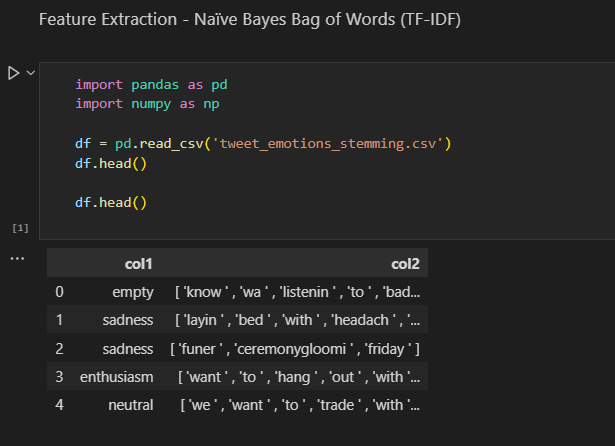
****

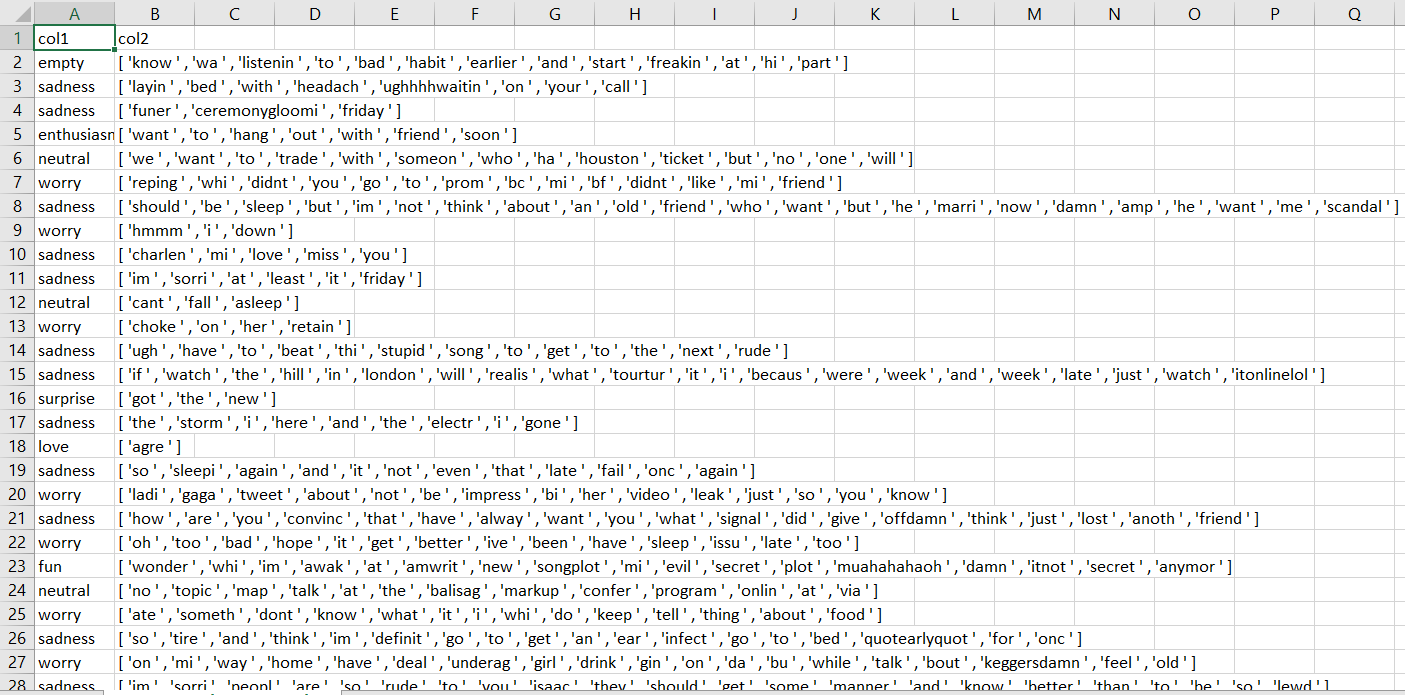
****

****

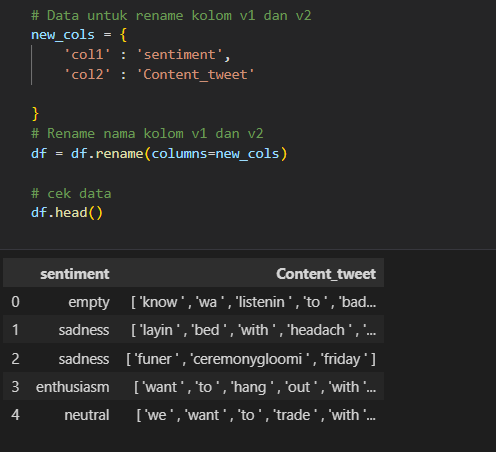
**Ekstraksi Fitur**

Ekstraksi fitur yang digunakan menggunakan metode Naïve Bayes Bag of Words (TF-IDF). TF-IDF merupakan nilai/bobot untuk mengukur seberapa penting sebuah kata. IDF akan menilai kata yang sering muncul sebagai kata yang kurang penting berdasarkan kemunculan kata tersebut pada seluruh dokumen. Semakin kecil nilai IDF maka akan dianggap semakin tidak penting kata tersebut, begitu pula sebaliknya.

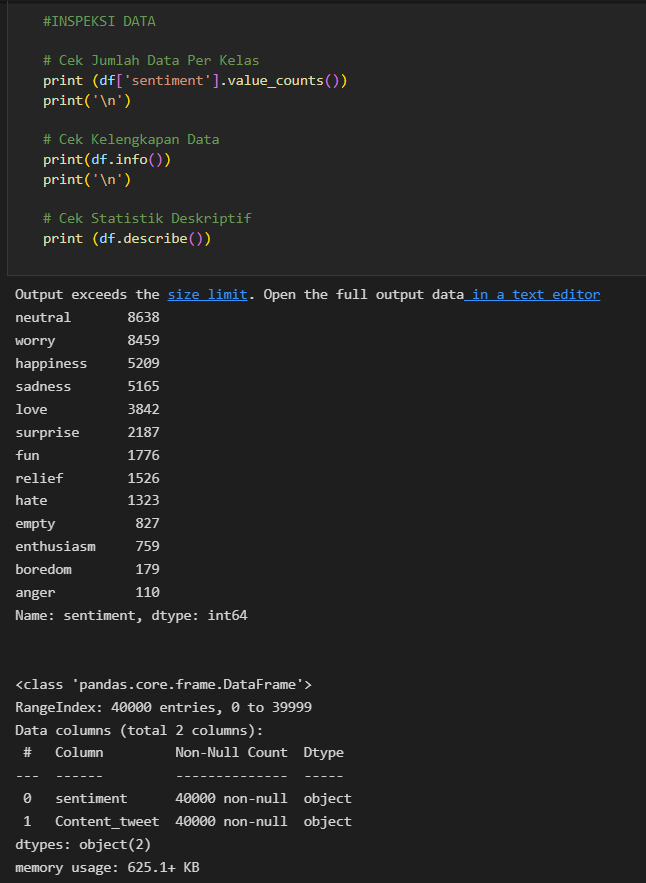


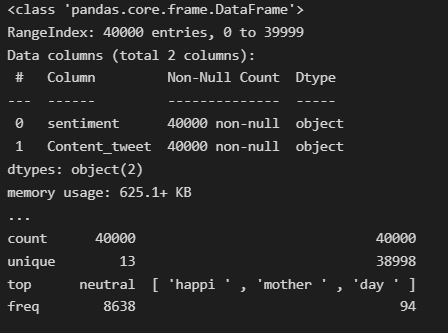


Melakuakn perubahan nama pada kolom 1 dan 2 menjadi sentiment dan Content\_tweet.



Setelah itu melakukan inspeksi data dengan tujuan untuk mengetahui hal-hal penting dalam data seperti cek jumlah data per kelas (sentiment). Cek kelengkapan data dan juga cek ststistik deskriptif.





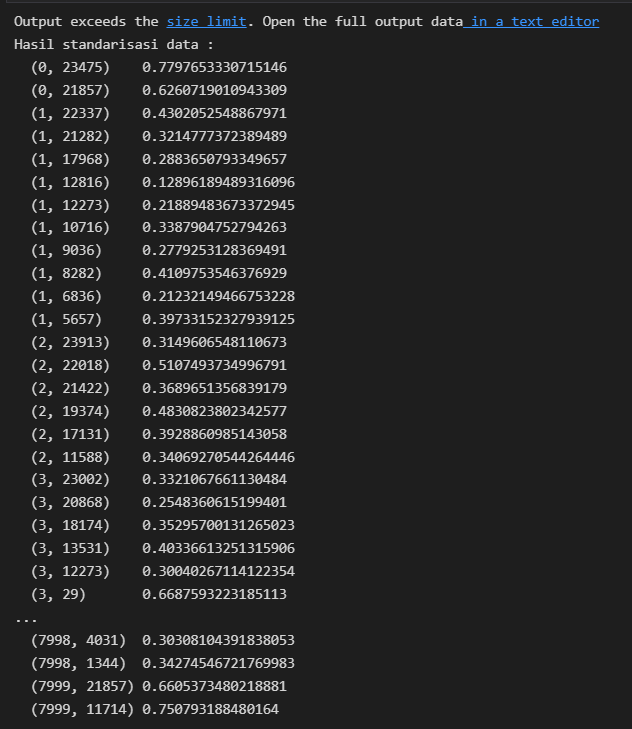
Kemudian, pada tahap ini dilakukan encode label pada kolom sentiment, dimana setiap kategorinya diubah menjadi angka seperti pada gamabar berikut.



Masuk pada tahap melakukan pemodelan ekstraksi fitur. Dataframe yang digunakan adalah kolom Content\_tweet sebagai sumbu X dan kolom sentiment sebagai sumbu y.



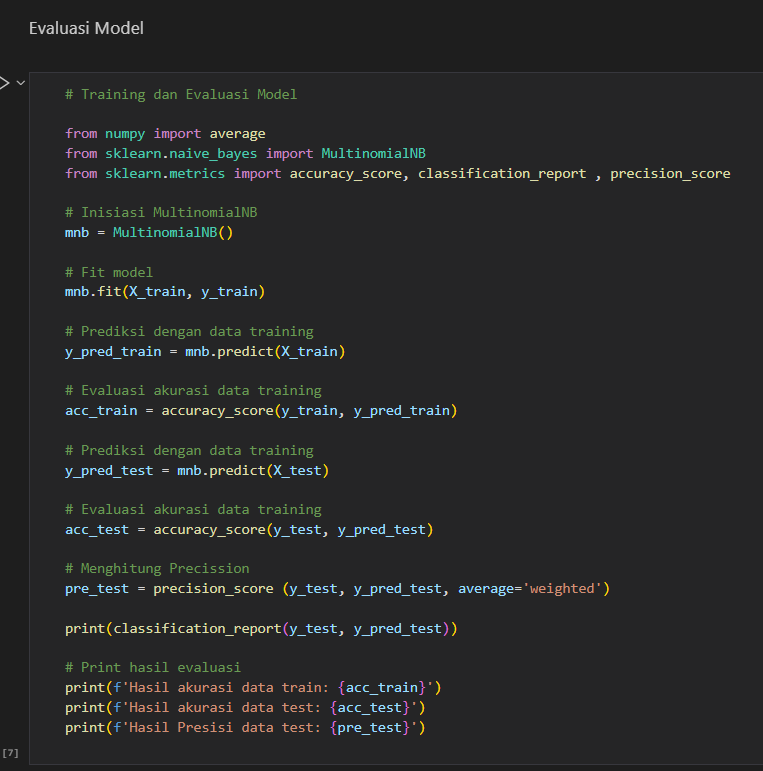
Dari gambar diatas data yang akan diolah akan displit dengan parameter X, y dengan size 0,2 dan mengambil data random sebanyak 50. Kemudian dari data yang telah di split dilakukan inisiasi pembobotan nilai menggunakan TfidfVectorizer dengan memanfaatkan stop\_word untuk mengurangi kemunculan jumlah kata dalam sebuah dokumen. Kemudian dilakukan fitting nilai dari data tersebut kemudian di transform. Hasil dari fitting tersebut akan dilakukan penskalaan atau standarisasi pembobotan antara data training dan data testing. Sehingga menghasilkan nilai seperti pada gambar dibawah ini.

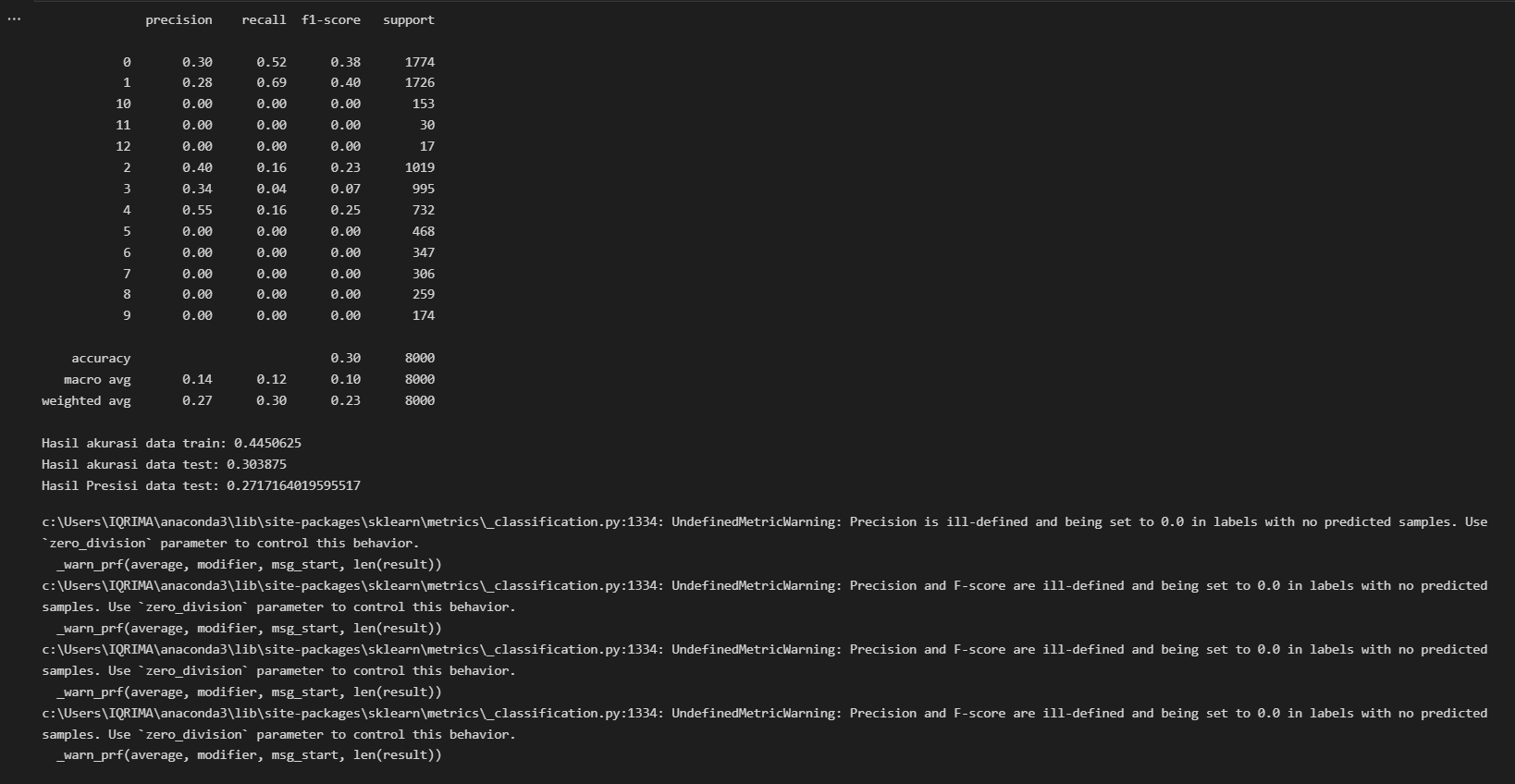


**Pemmbuatan Model dan Evaluasi**

Pembuatan model untuk klasifikasi algiritma menggunakan class MultinomialNB dari library sklearn.naive\_bayes. Pengklasifikasi multinomial Naive Bayes ini cocok untuk klasifikasi dengan fitur diskrit pada Deteksi Emosi Pengguna Tweeter (jumlah kata untuk klasifikasi teks).

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





**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 tidak memiliki keakuratan.