SENTIMENT ANALYSIS

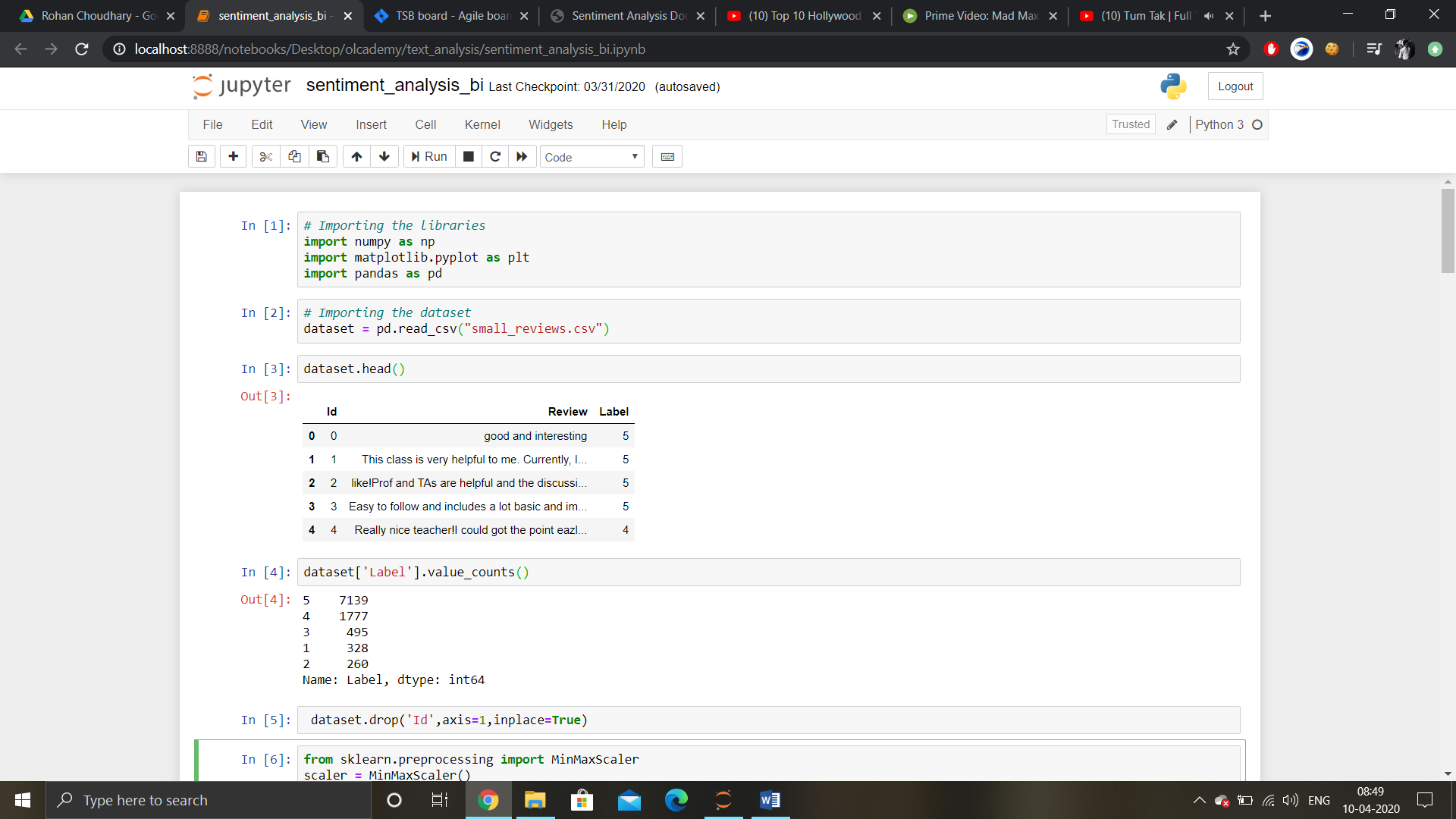
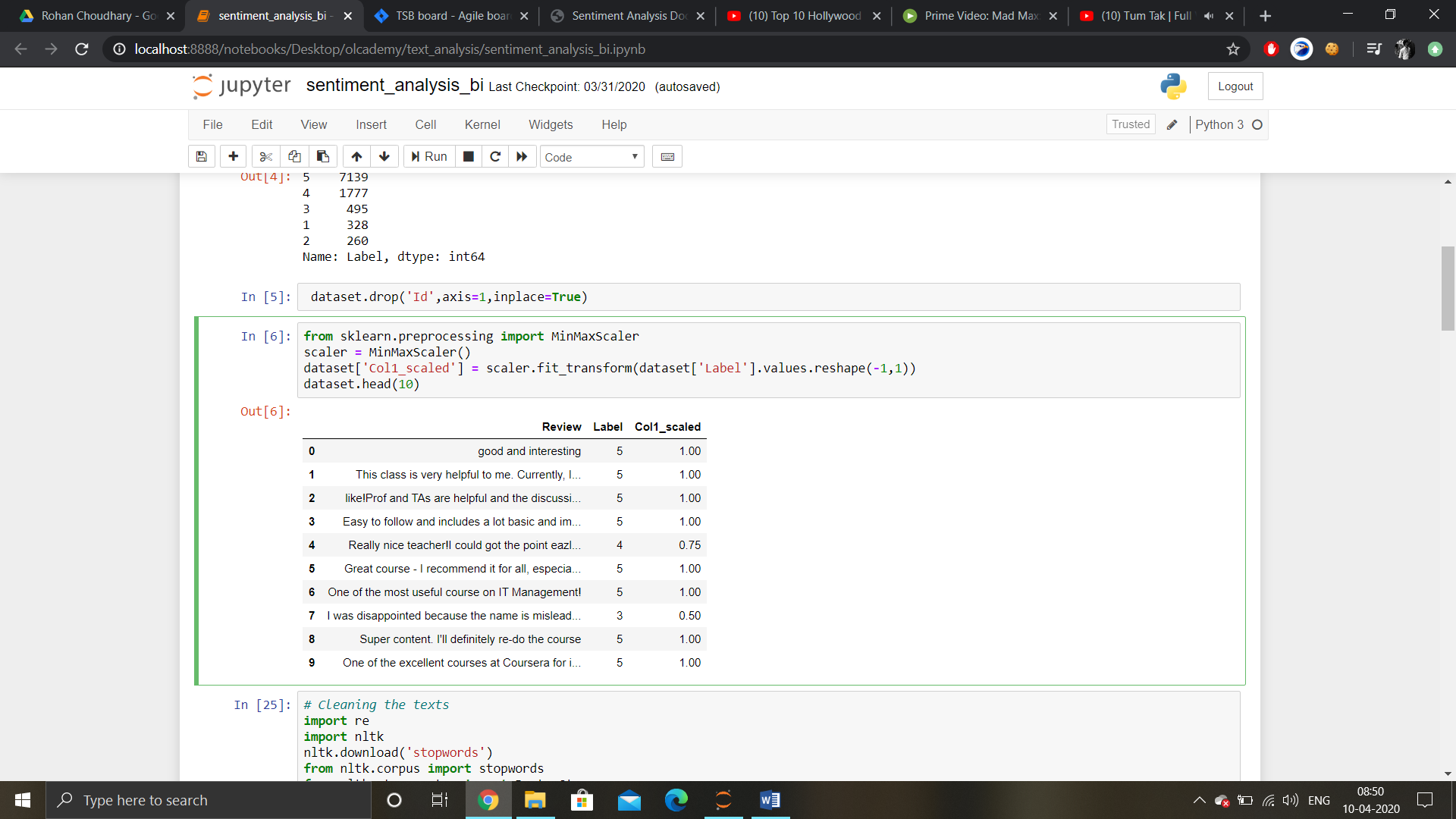
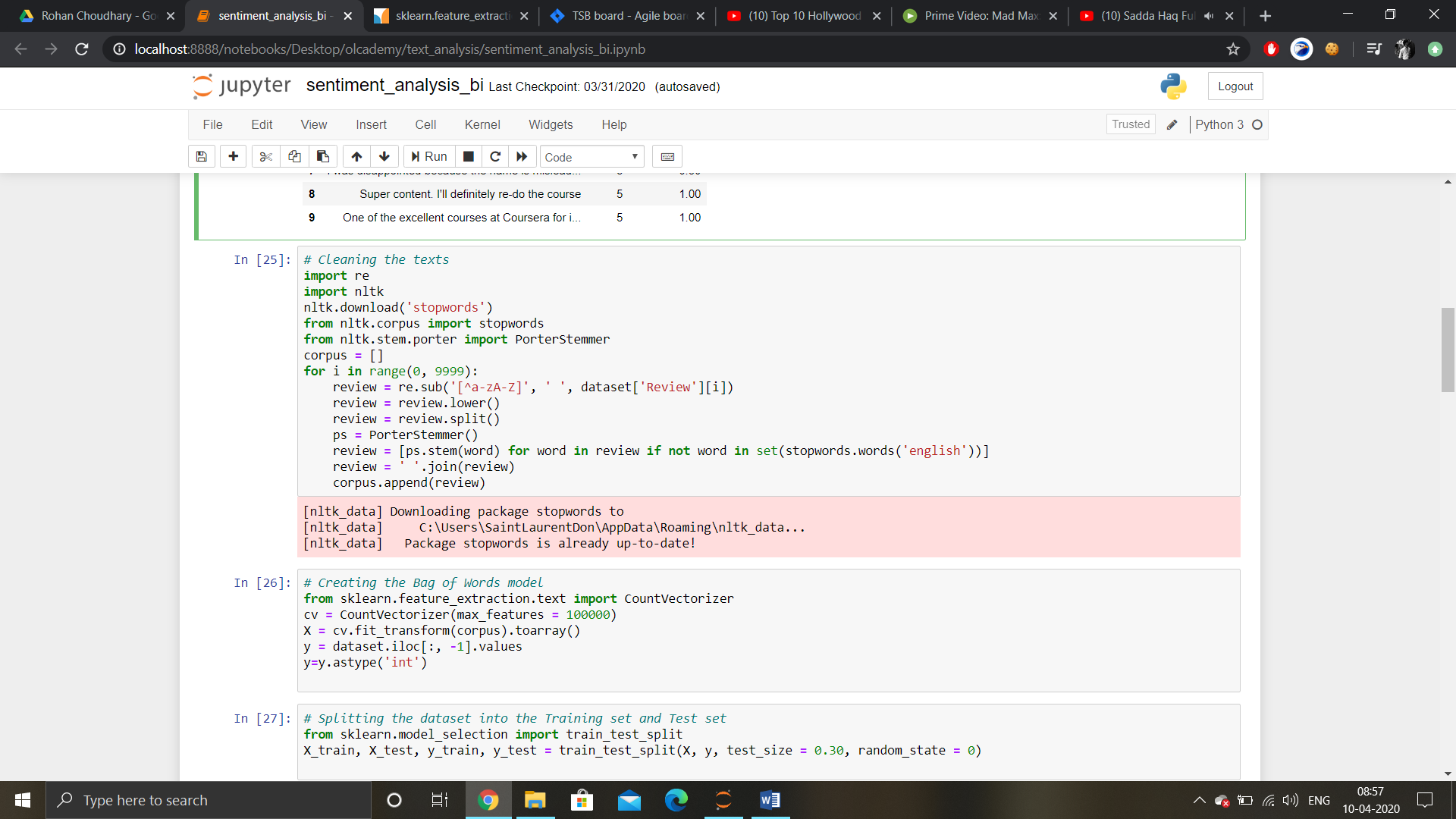
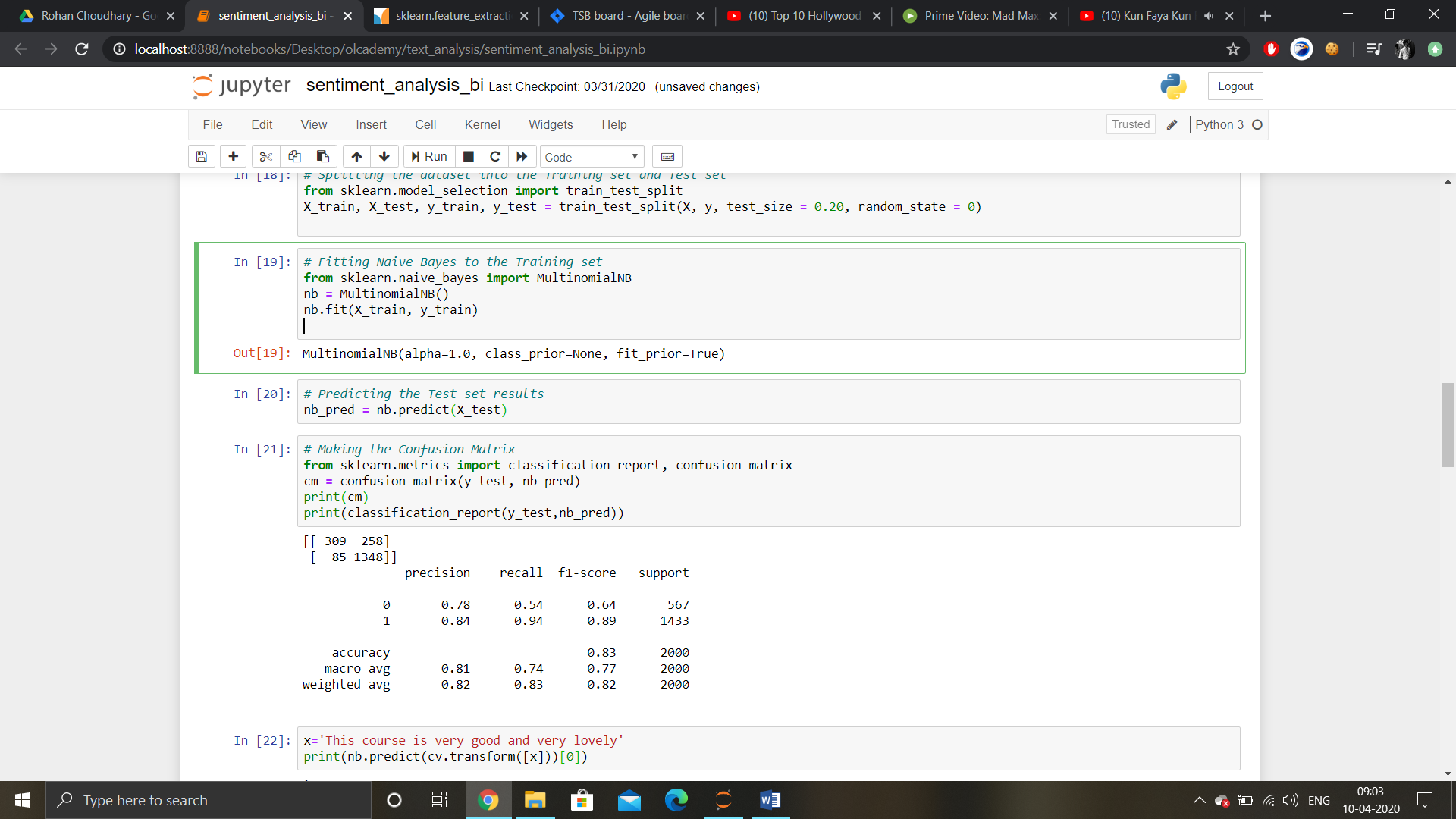
Authored by: Rohan Choudhary

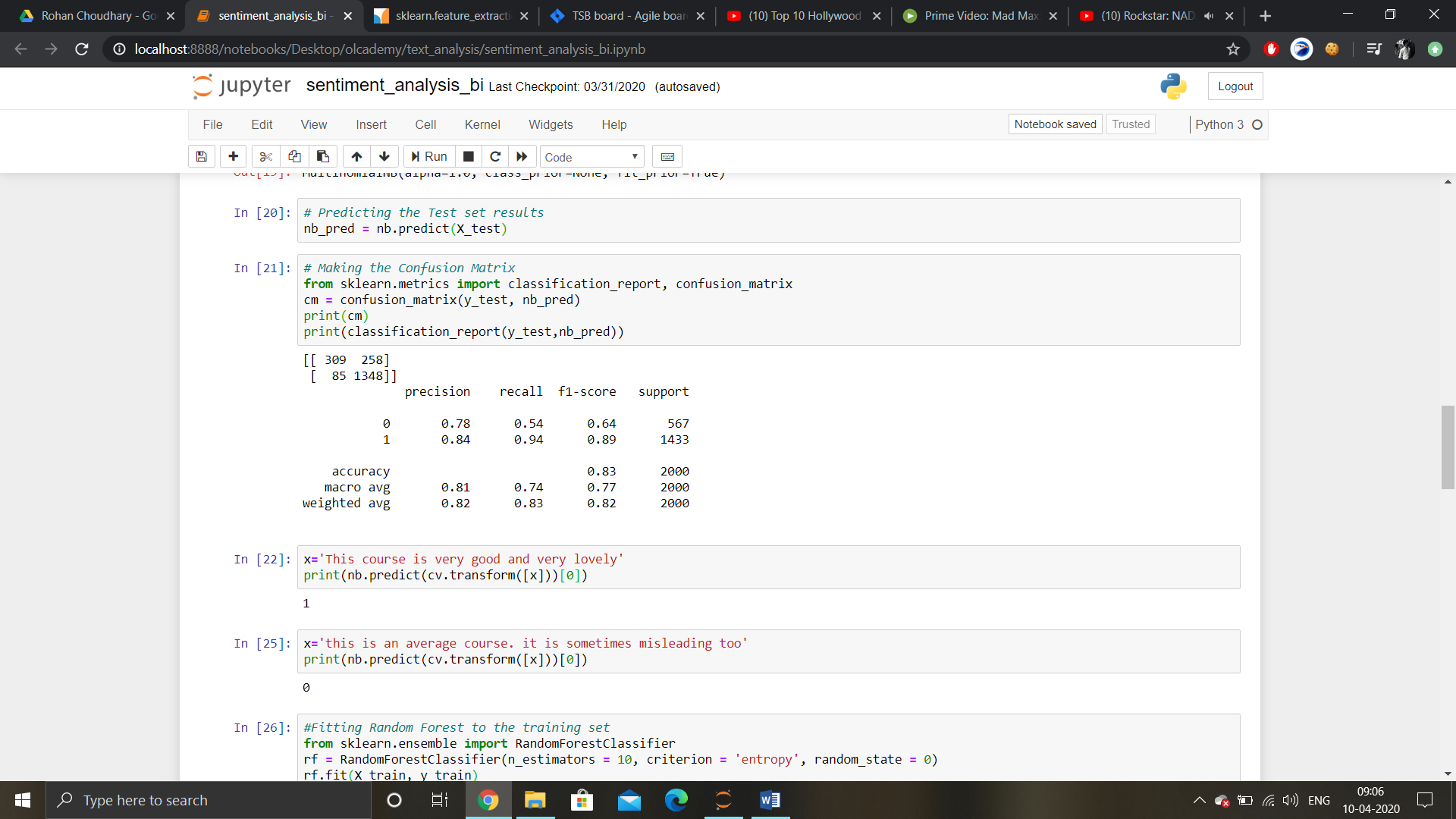
Objective: Try to predict the sentiment (Positive (1) or negative (0)) behind a comment.

Skill Path: Python, Machine Learning algorithms, Natural Language Processing.

Note: Various ML models were used to predict sentiments, the one with the highest accuracy has been used in the end and only that algorithms is mentioned in this documentation. Naïve Bayes algorithms gave the best accuracy (83%). Other algorithms were also used: Random Forest (78% accuracy), Decision Tree (78% accuracy) and Logistic Regression (80% accuracy).

Filename: sentiment\_analysis\_bi.ipynb

1. The dataset small\_reviews.csv has been used in this script. The dataset contains almost 10000 entries for the sentiment classification task.
2. Exploratory Data Analysis on dataset is done and the id column of the dataset is removed. Also the label column is scaled between values 0 to 1 and a new column is created, namely col1\_scaled.
3. A for loop is created to remove everything except alpha characters and spaces from the dataset and perform data cleaning (removing stopwords, lowering all letters, stemming). [Bag of words](https://machinelearningmastery.com/gentle-introduction-bag-words-model/) model is used ([Count vectorizer](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html) is used)
4. Then the dataset is divided into training and test sets (80:20 ratio).
5. Multinomial Naïve Bayes is fitted to the model and the accuracy is checked using confusion matrix and f1 score, precision, recall and accuracy.

6) The model is checked with specific input from the user to test its prediction capabilities. 1 stands for positive and 0 for negative.