Athens University of Economics and Business

Department of Management Science &Technology - MSc in Business Analytics

**Big Data Content Analytics**

***‘Sentiment analysis on restaurant reviews from yelp website‘***

******

**Students:**

**Alexandros Ntzoufas**

**Romanos Oikonomidis**

**Sofia Zgkouri September 2019**

**Code repository:** <https://github.com/antzoufas/AI>

1. **Task description**

Sentiment analysis is the automated process of understanding an opinion about a given subject from written or spoken language. Currently, more than 2.5 quintillion bytes of data are generated in a daily basis worldwide, sentiment analysis has become a key tool for making sense of that data. This has allowed companies to understand the social sentiment of their brand, product or service while monitoring online conversations in order to get key insights and automate all kind of processes. With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research.

The main aim of this assignment is to identify whether reviews of the customers for hotels, restaurants, coffee shops, etc are negative, positive or neutral and if there is place for improvement. More specifically, it can be analyzed if the restaurants of a specific area gather positive or negative reviews, so another competitor can offer better services and gain higher share of the market. Another example is to identify if the restaurants of specific island in Greece (i.e. Santorini) that serve Mediterranean cuisine collect negative or possitive reviews by the visitors, so modifications in the sector should be implemented.

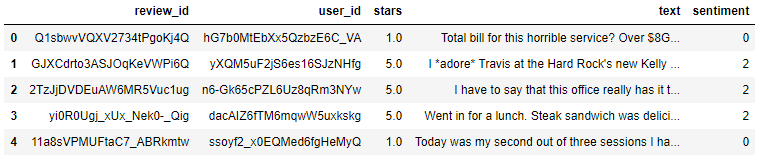
1. **Data**

The dataset under study is containing reviews from the Yelp website ([User Reviews and Recommendations of Best Restaurants, Shopping, Nightlife, Food, Entertainment, Things to Do, Services and More) was downloaded from Kaggle :](https://www.yelp.com/) [[https://www.kaggle.com/yelp-dataset/yelp-dataset](https://www.yelp.com/)](https://www.kaggle.com/yelp-dataset/yelp-dataset)

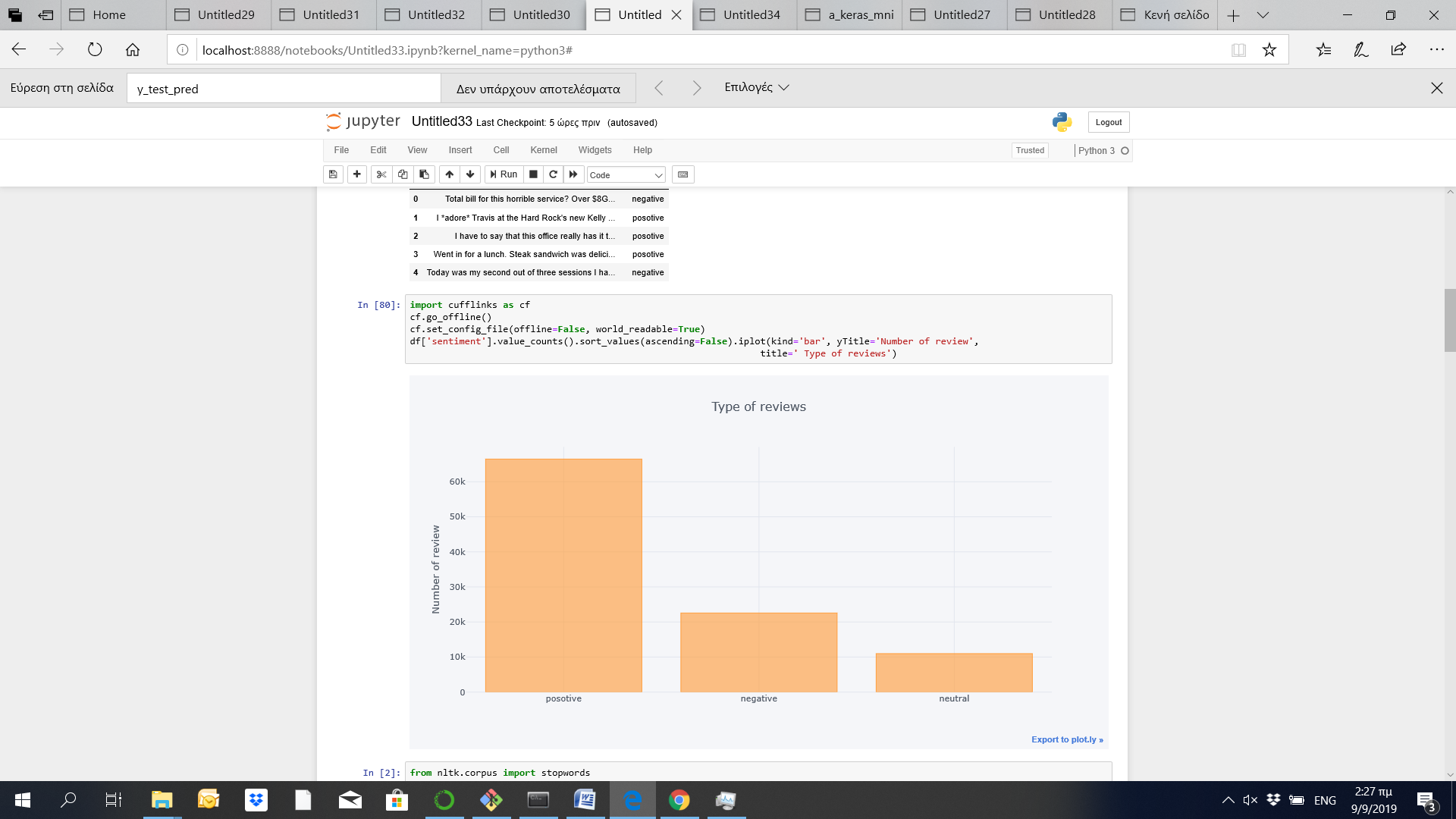
Sentiment analysis or in other words, opinion mining was performed on this data using Recurrent Neural Networks and Convolution Neural Networks.

From the initial dataset downloaded from Kaggle, of size ~5GB, a chunk dataset of ~800MB was handled in python. The initial file was split using Gitbash terminal commands (split <data.json> –n 6). From this cut file, 100.000 rows were used as an input to the model. The data were in json format.

The categoriasation of a review was done by taking into consideration the ratings of the users (stars column of the dataset). Reviews with 1 or 2 stars are considered as negative, reviews with 3 stars as neutral while reviews with 4 or 5 stars as positive. An additional column was created in the dataset with labels 0,1,2 for representing the 3 levels of characterization of a review (negative, neutral, positive for 0,1 and 2 respectively).



The columns review id, user id and stars were removed from the initial data. Then the dataset was split into train and test dataset, with an analogy of 90% - 10% .As we see from the following figure, the majority of the reviews are positive. Almost the one fourth of the reviews are negative while only the one tenth is neutral.



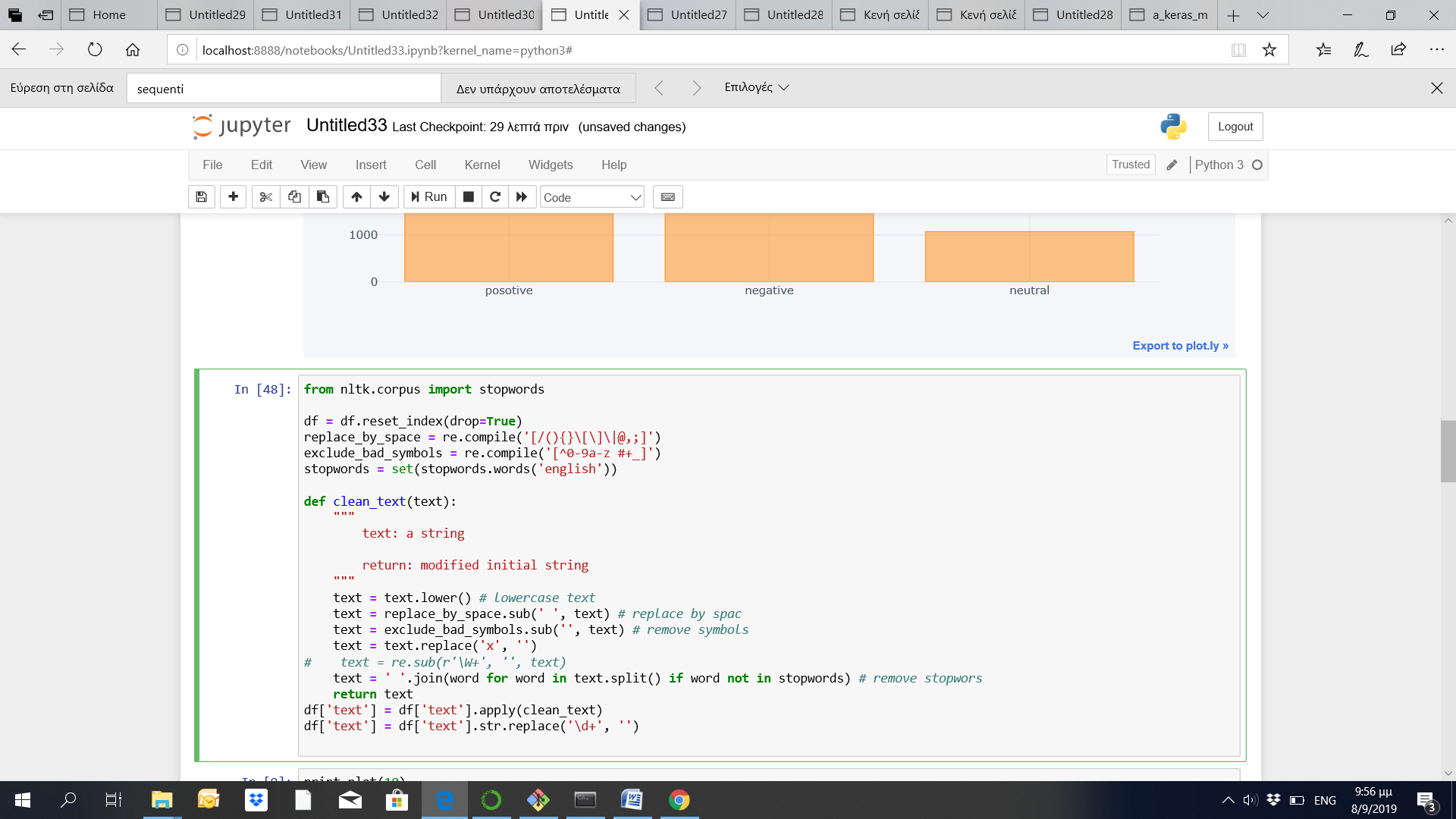
1. **Text Classification Using Convolutional Neural Network (CNN) :**

**Preprocessing**

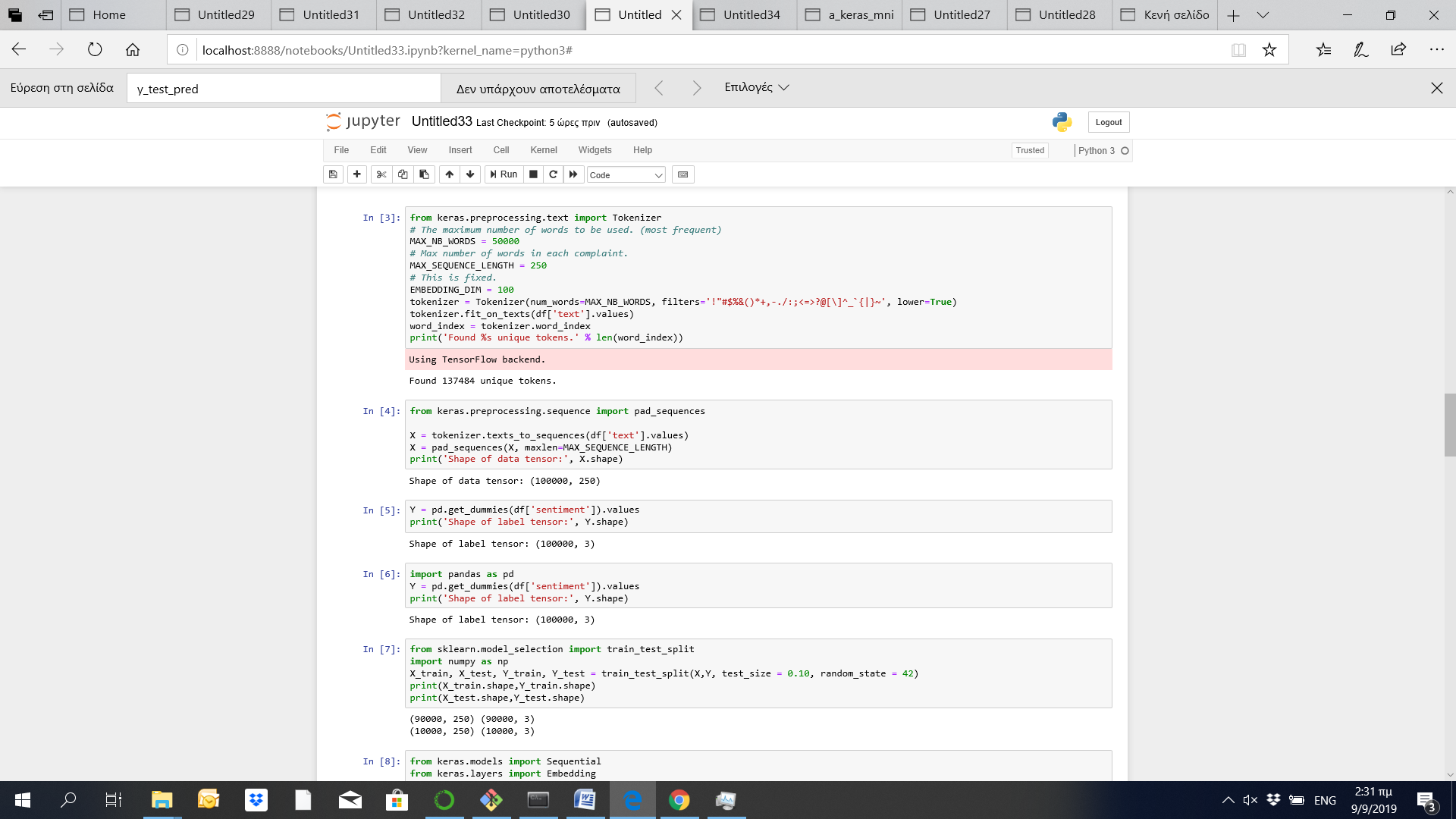
The text preprocessing that we followed included:

* Converting all text to lower case.
* Replace symbols by space in text.
* Delete symbols from text.
* Delete stop words.
* Delete digits in text.

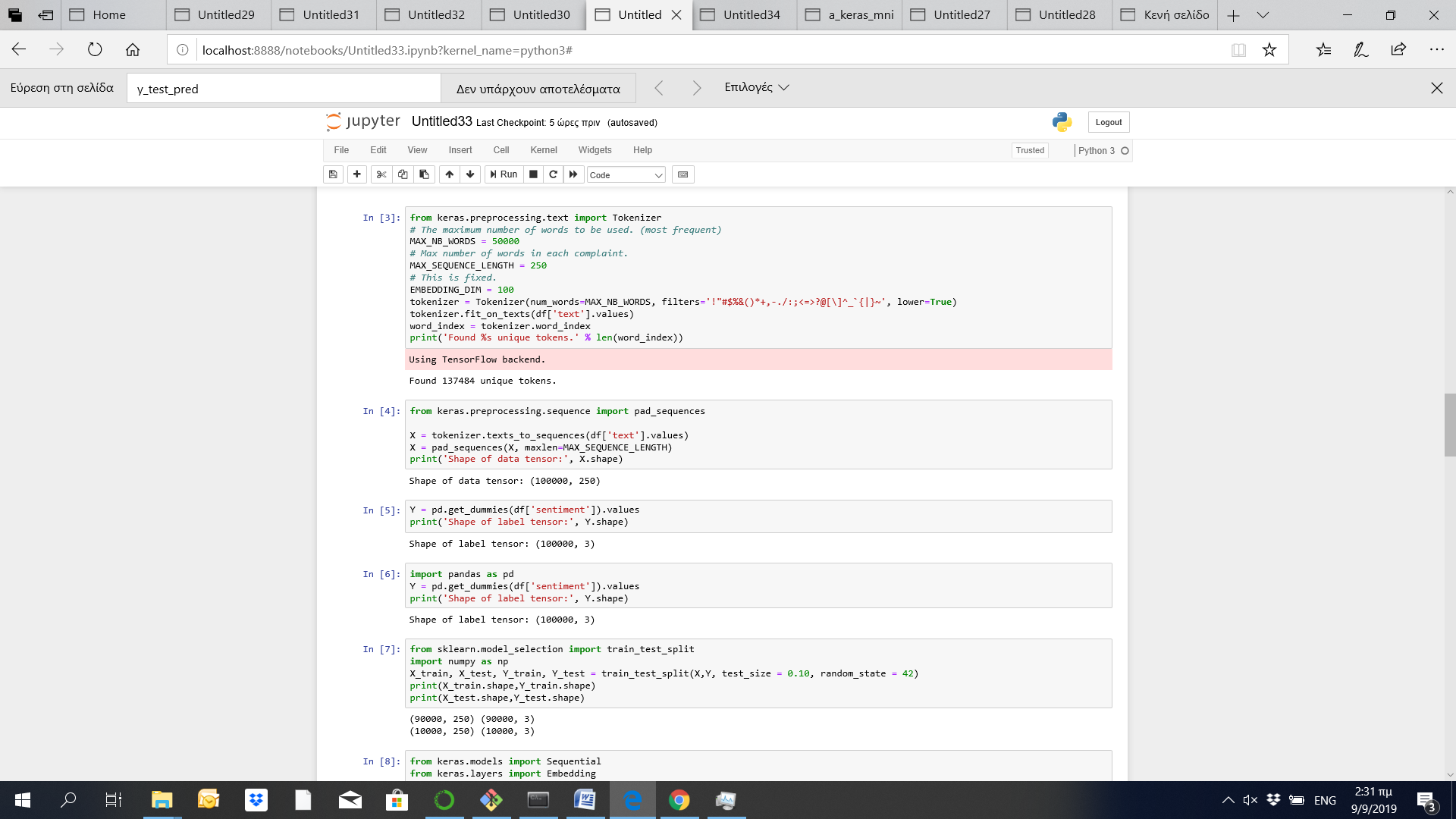
Then during the LSTM modeling, we vectorized consumers complaints text, by turning each text into either a sequence of integers or into a vector. Then we limited the data set (vocabulary) to the top 5,0000 words and we set the max number of words in each complaint to 250.



Then we used the pad\_sequences is used to ensure that all sequences in a list have the same length. Furthermore, since we have three different values in the categorical variable, negative positive and neutral, we converted the categorical labels to numbers.

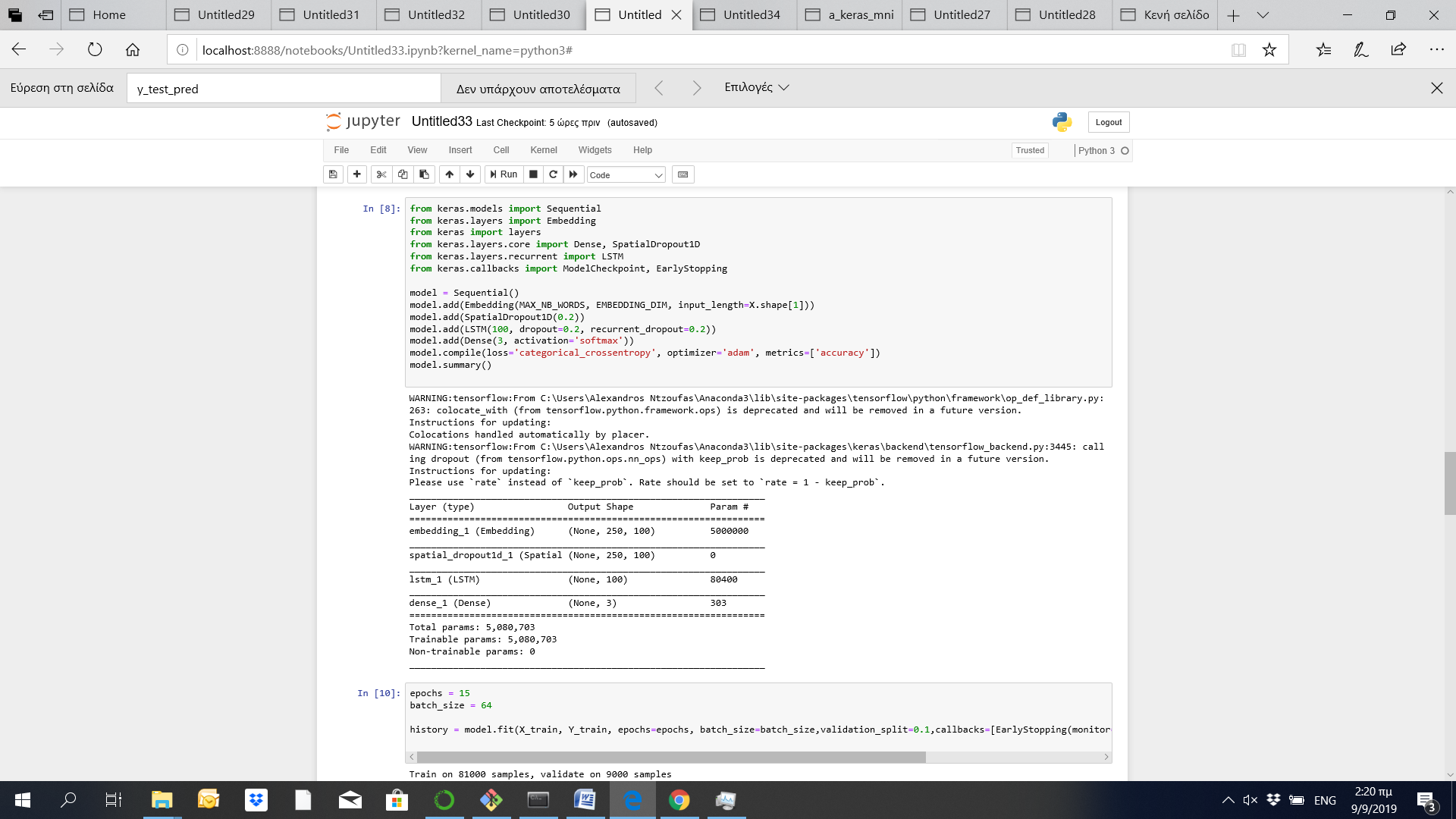


With the sklearn train\_test\_split function, we splitted the dataset to train and test, with proporsions, 90% and 10% respectively.

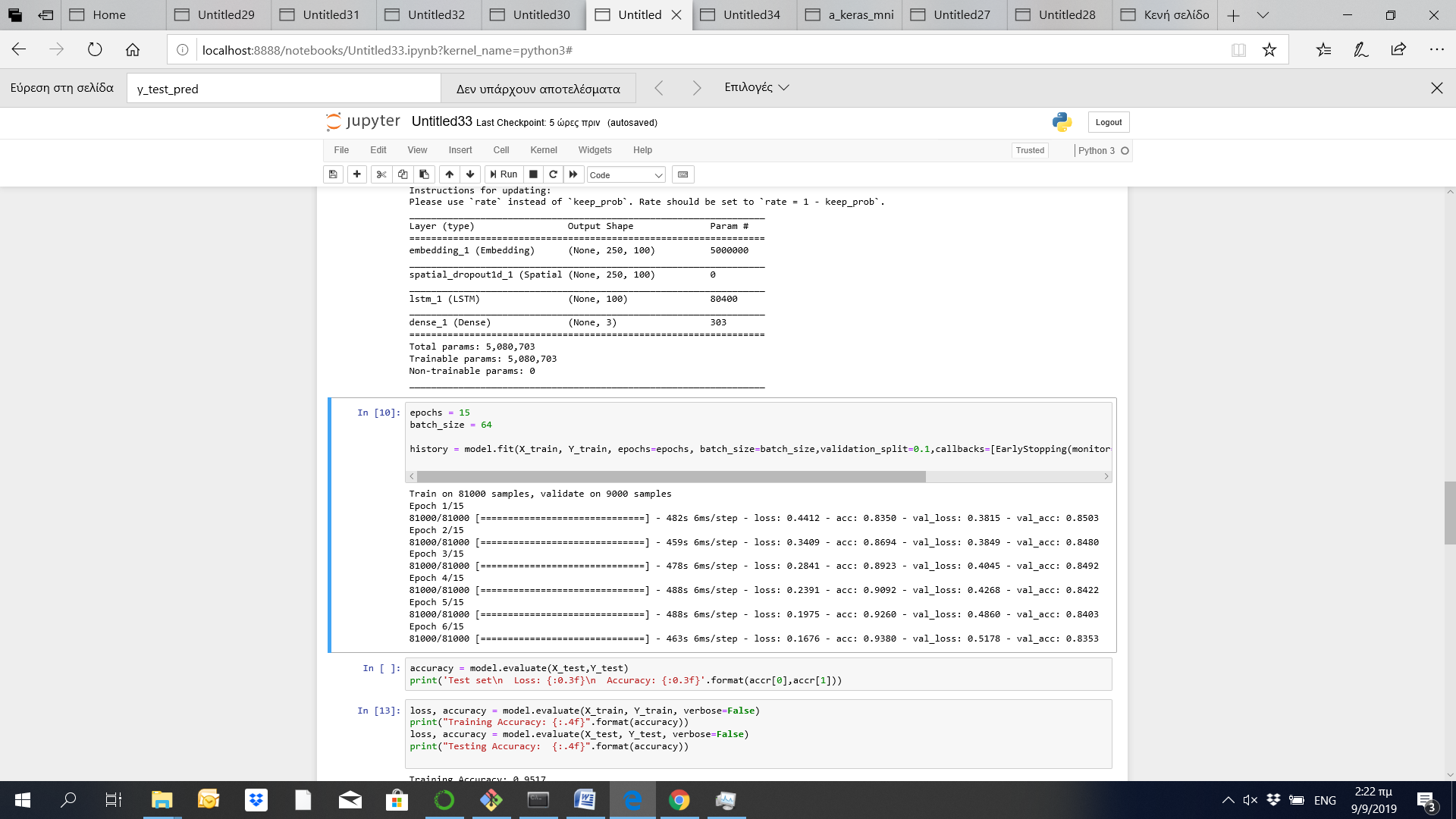


Long Short Term Memory networks , just called “LSTMs” are a special kind of RNN, capable of learning long-term dependencies. The model that we built model includes:

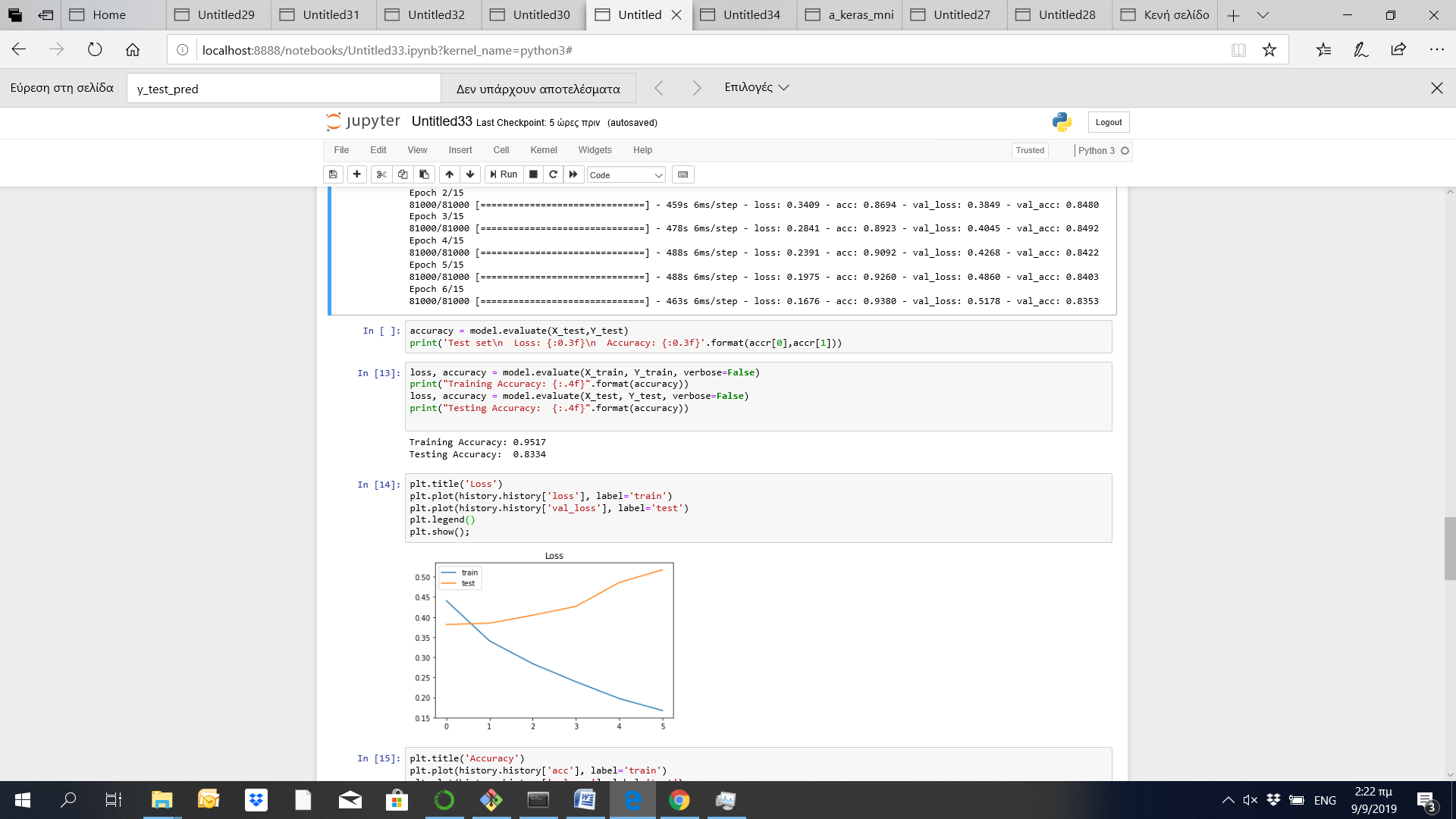
* Firstly we added the first layer which is a Sequential model which needs to receive information about its input shape.
* Then we added an embedded layer with 100 length vectors to represent each word.
* The we added a SpatialDropout1D performs variational dropout in NLP models. This version performs the same function as Dropout, however it drops entire 1D feature maps instead of individual elements
* Then we added the next layer is the LSTM layer with 100 memory units.
* The output layer must create 3 output values, one for each class for our labels.
* Since we have a multi class classification problem we added softmax as activation function.
* Finally, since we have a multi-class classification problem, categorical\_crossentropy is used as the loss function.



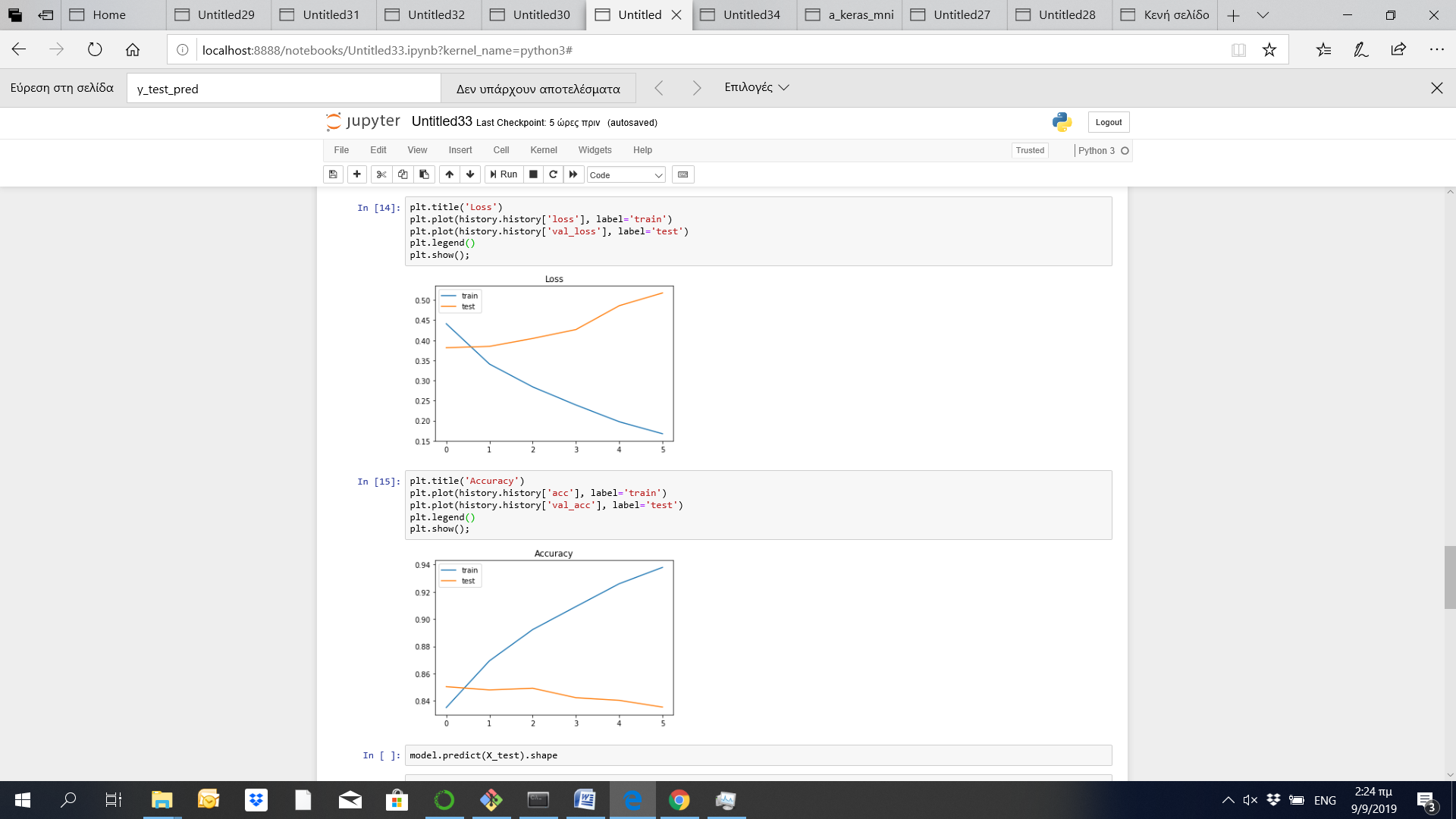
We train our model for 15 epochs, with an EarlyStopping, patience=5 and we yeld the following results.



The accuracy of the model in the training dataset is 95.17% while for the testing set is 83.34%.

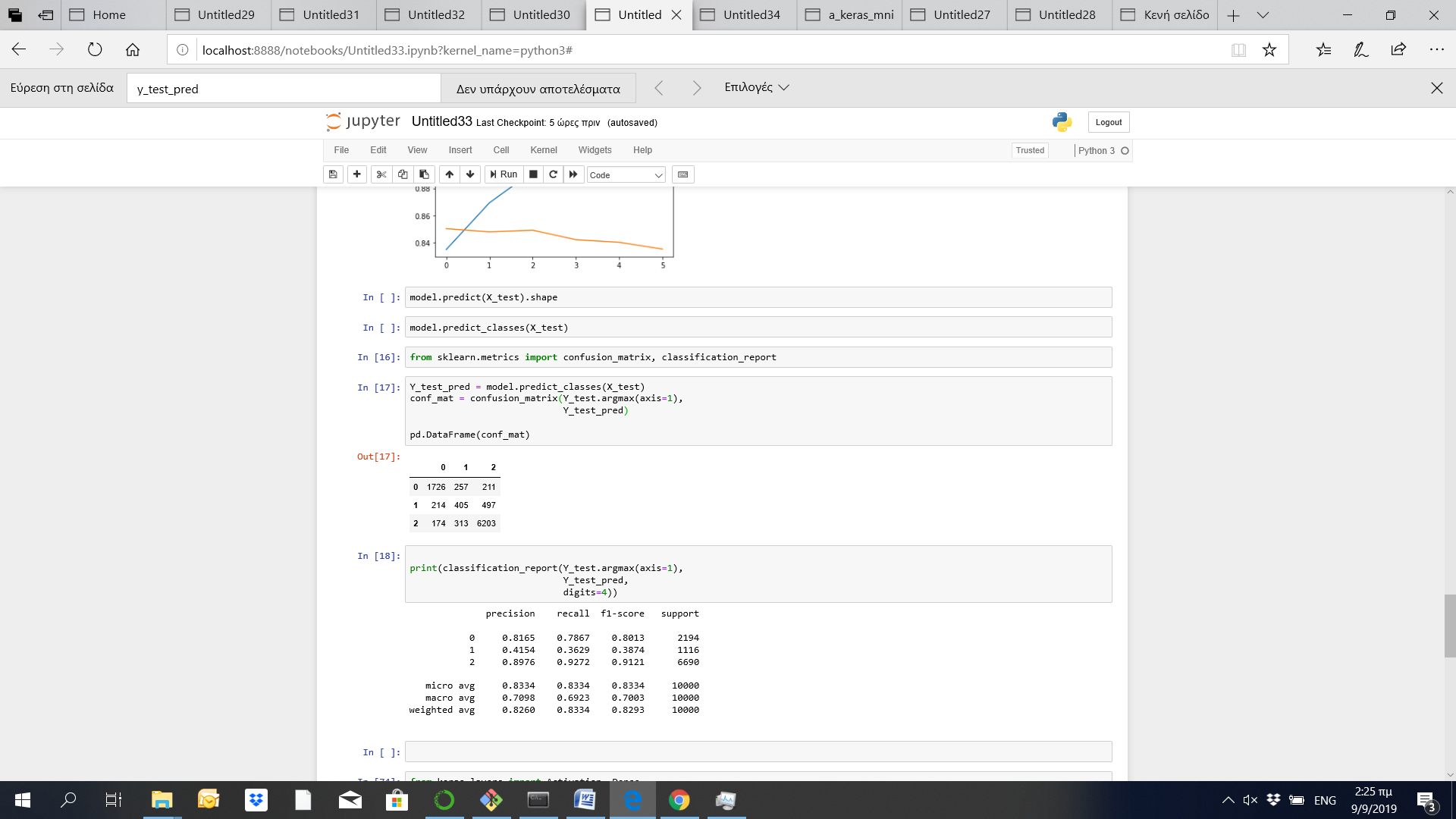


As it arises from the following figure, our model tends to overfit after the 6th epoch.



The following matrices indicate the performance of the model. More specifically:

* Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. We see that our model has beter results for the negative and possitive labeles (0 and 2, respectively) than for neutral (1)
* Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Again, our model performs better for negative and positive reviews.
* F1 score - F1 Score is the weighted average of Precision and Recall. The same results applies for F1-Score as well.

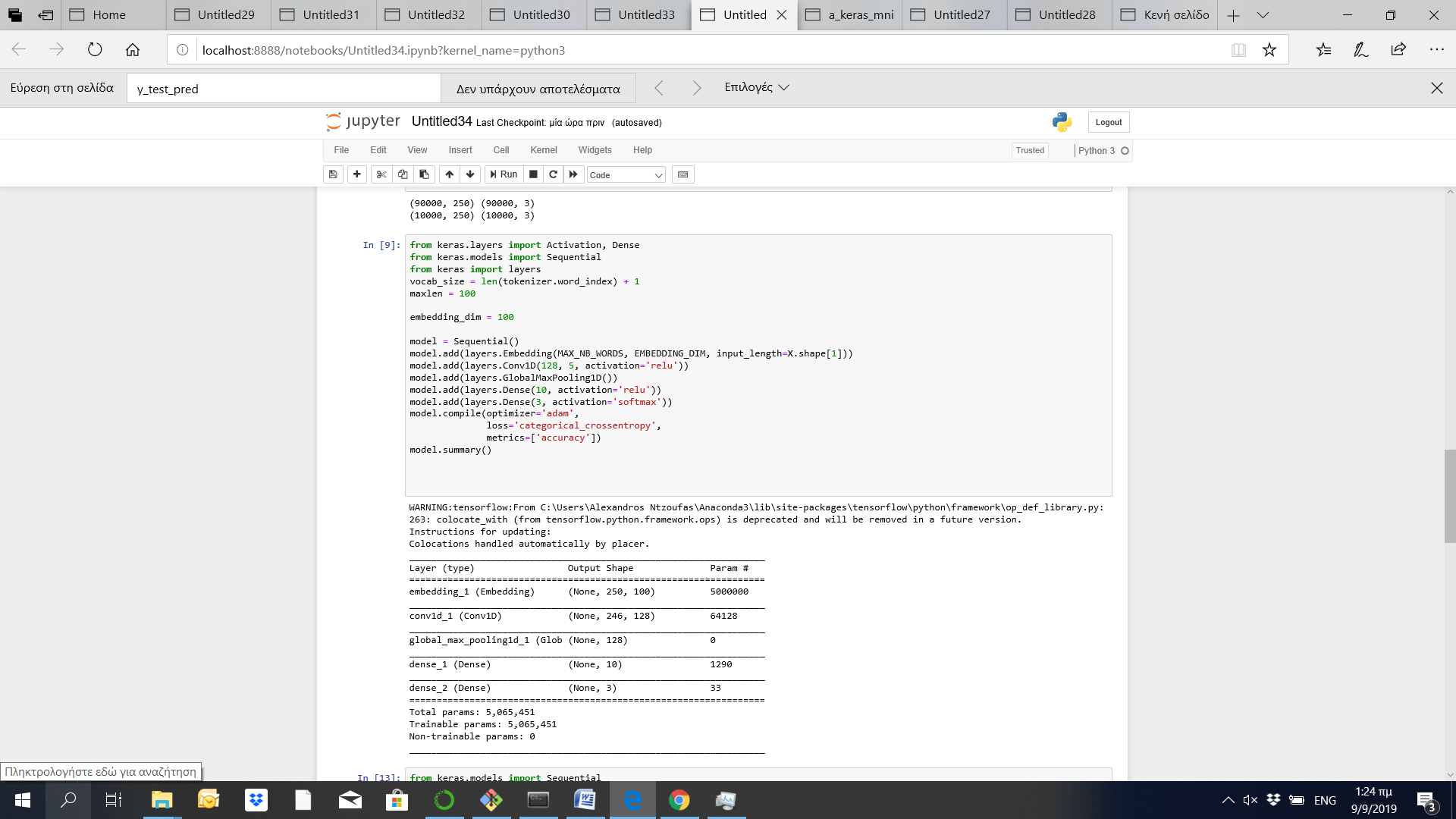


1. **Text Classification Using Convolutional Neural Network (CNN)**

[CNN](https://en.wikipedia.org/wiki/Convolutional_neural_network) is a class of deep, feed-forward artificial neural networks and use a variation of multilayer perceptrons designed to require minimal preprocessing. However CNNs are generally used in computer vision, they have been applied to various NLP tasks as well. In our case we applied the following CNN in our model in order to see the results and compare them with the other models.

More specifically, we applied the following model with a Sequential layer and an embedded layer with 100 length vectors. The we added a Conv1D layer which creates a convolution kernel that is convolved with the layer input over a single spatial (or temporal). Then we added a GlobalMaxPooling1D layer for temporal data which takes the max vector over the steps dimension

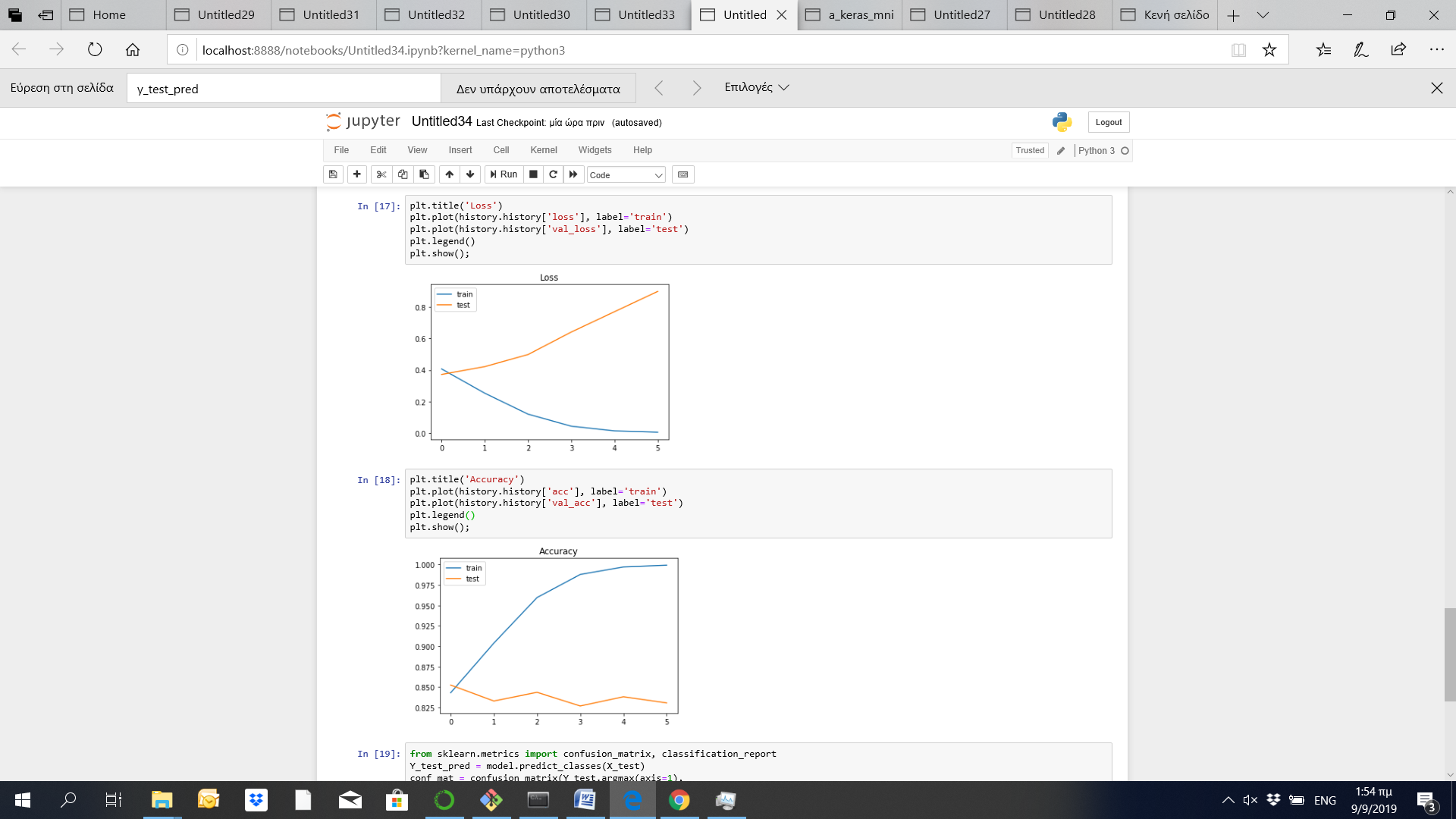
Then we added a a [ReLU activation function](https://en.wikipedia.org/wiki/Rectifier_(neural_networks)) to the output to introduce nonlinearities into the model.Since we have a multi class classification problem we added softmax as activation function. Finally, since we have a multi-class classification problem, categorical\_crossentropy is used as the loss function.



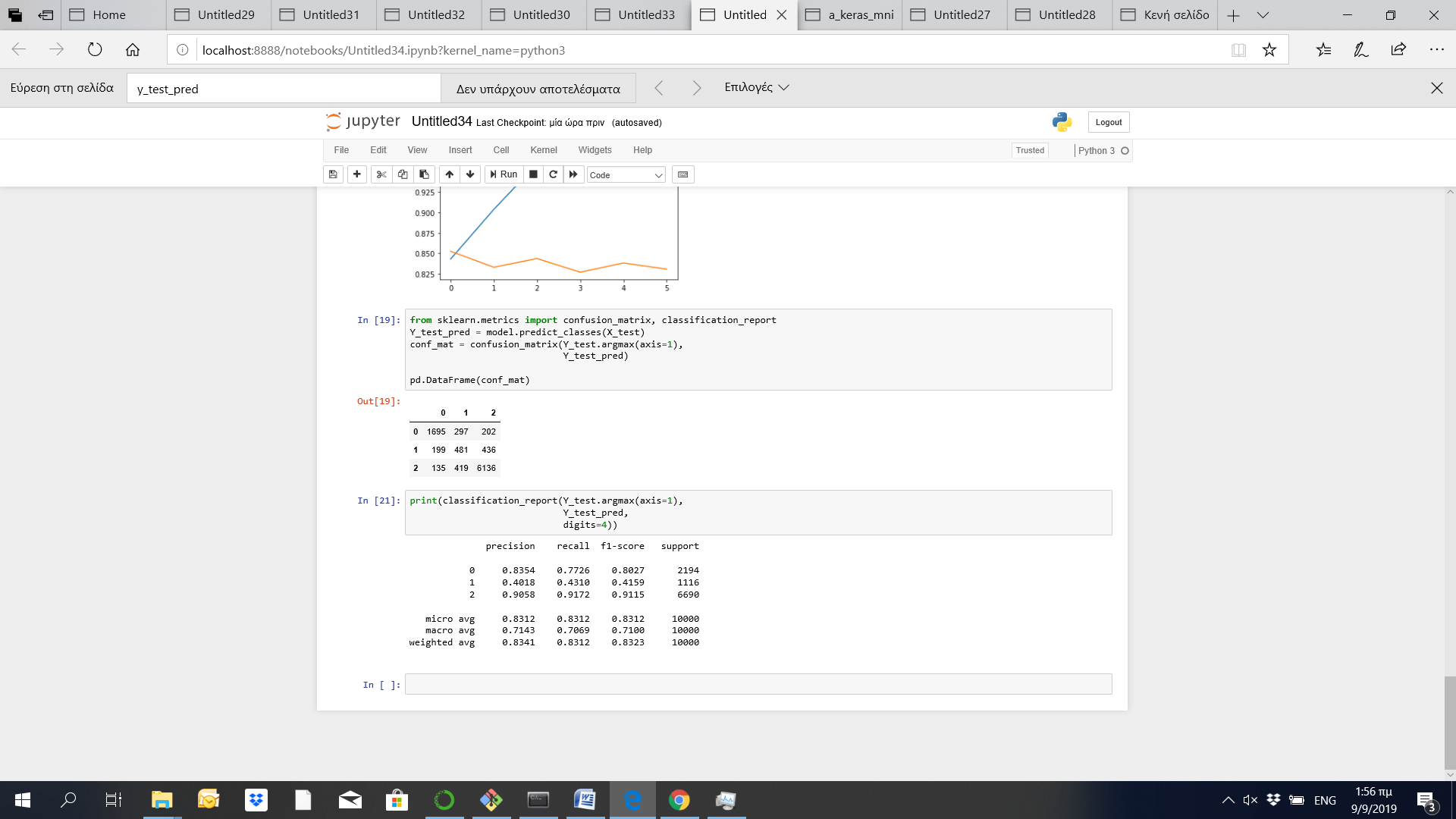
As we see from the following figure, we fited the model for 15 epochs, with an EarlyStopping with patience=5. The model trained fro 6 epochs, with an accuracy of 83.1% and loss of 0.919.

# 

As we see from the following figure, there is an over fitting in the training dataset.

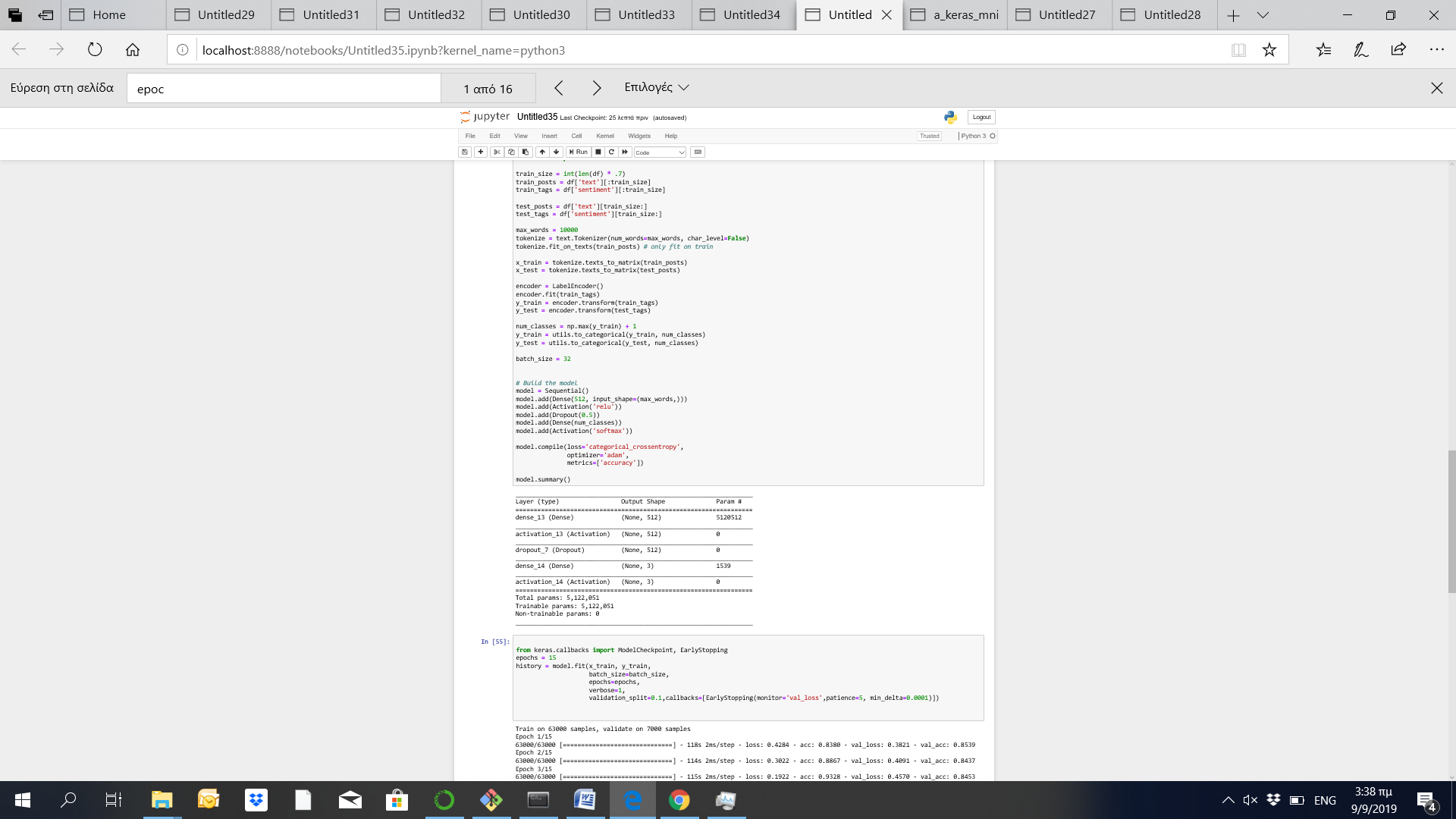


With regards to the performance in the training dataset, we can see that the model can predict with a high accuracy the positive and the negative reviews; however the probability to predict the neutral reviews is less than 50%. This may happens since the number of the neutral reviews were low in number.



1. **Text Classification BOW (Bag of Words) with Keras**

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). More specifically, we used tokenizer methods to count the unique words in our vocabulary. Then we feed a one-hot vector to our model and we built our text classification model as follows:



* Firstly we added the first layer which is a Sequential
* Then we added an Dense layer of dense 512-dimensional vectors
* Then we added a a [ReLU activation function](https://en.wikipedia.org/wiki/Rectifier_(neural_networks)) , a Dropout layer, and again a Dense layer
* Since we have a multi class classification problem we added softmax as activation function.
* Finally, since we have a multi-class classification problem, categorical\_crossentropy is used as the loss function.

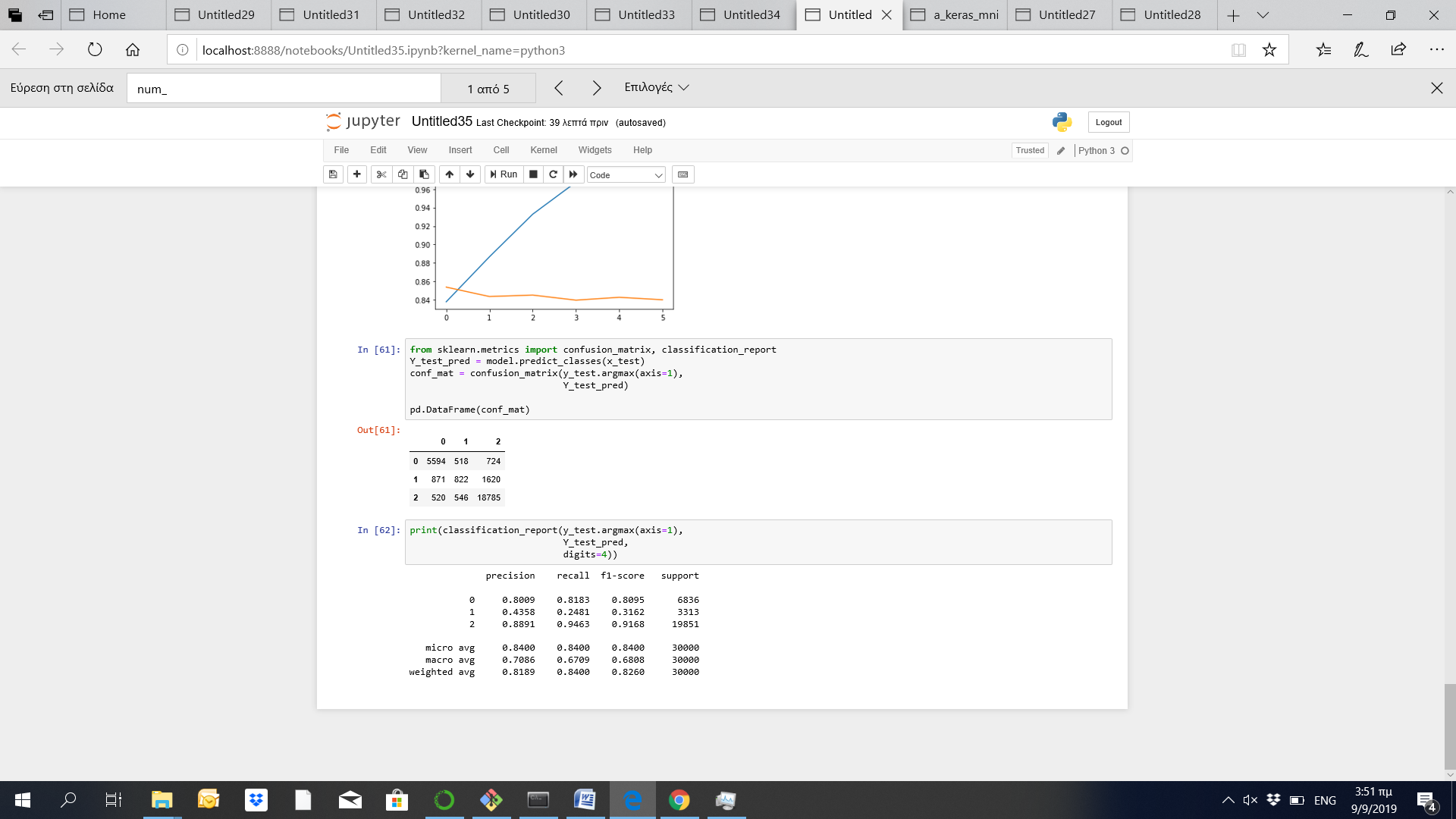
# 

As we see from the following figure, we fited the model for 15 epochs, with an EarlyStopping with patience=5. The model trained fro 6 epochs, with an accuracy of 98.3% in the training set and 84% in the test set.

# 

As we see from the following figure, there is an over fitting in the training dataset close to the 6th epoch.

# 

Considering the performance in the training dataset, we can see that the model predicts with a high accuracy the positive and the negative reviews; however the probability to predict the neutral reviews is less than 50%. This may happens since the number of the neutral reviews were low in number. 

**6. Conclusions**

In the context of this assignment, various models were tested in order to identify whether the sentiment polarity of the reviews of customers can be predicted. More specifically, the yelp dataset was processed and tested with LSTM RNN model, a CNN model and and a BOW with Keras model. The scope of this analysis is to identify the positive and negative reviews of consumers for specific sectors of interests (restaurants, hotels, coffee shops) and specific areas of interest. With this process, we can identify sectors or places where potential for improvement or investments can be initiated.

As it is presented in the following table, all models brought relevant results. The models brought good prediction with regards to the positive and negative reviews of the customers; however the neutral reviews were more difficult to be identified.

|  |  |  |
| --- | --- | --- |
| **Accuracy(%)** | **Train dataset** | **Test dataset** |
| **LSTM RNN** | **95.17%** | **83.34%** |
| **CNN** | **98.3%** | **83.6%** |
| **BOW with KERAS** | **98.3%** | **84%** |