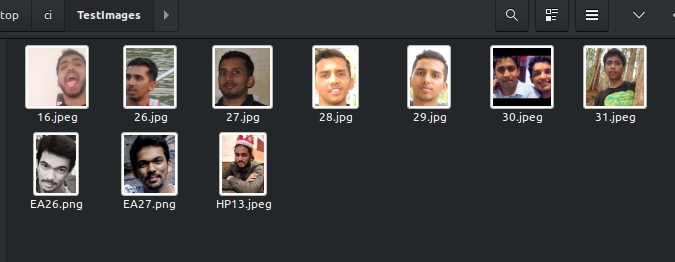
Create a Face Dataset.

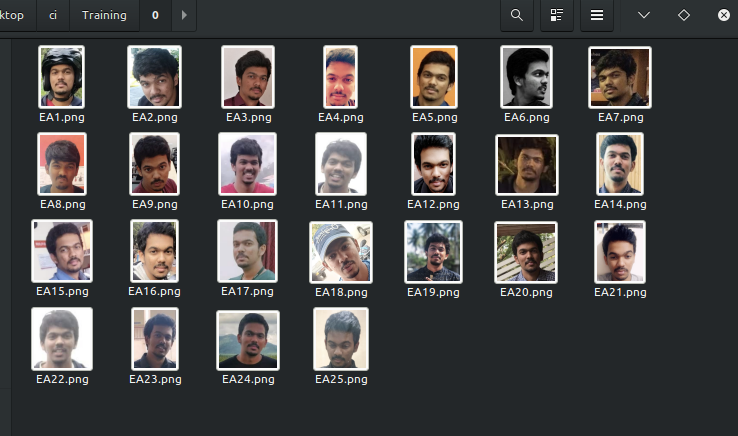
ML depends heavily on data, without data, it is impossible for an “AI” to learn. Dataset determines how well the model learns, and hence it’s import to train the model good quality, pre-processed data.

Dataset collection and creation involved the following steps:

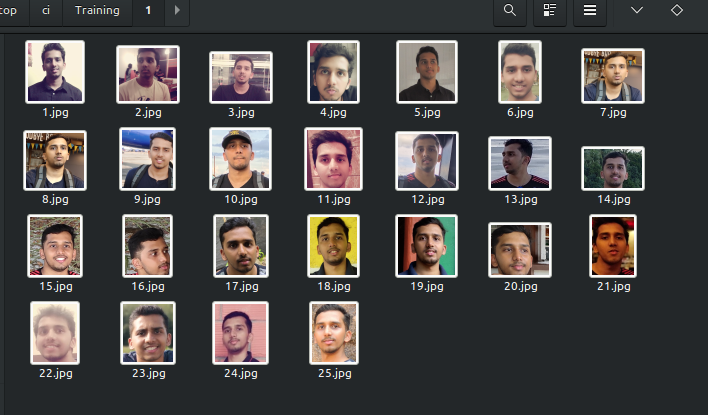
* Collection of 30 images each of 4 individuals.
* Preprocess - Crop the faces of those individuals in 1:1 ratio
* Choosing 25 images with regularities for training.
* Keeping the extra images for inferencing



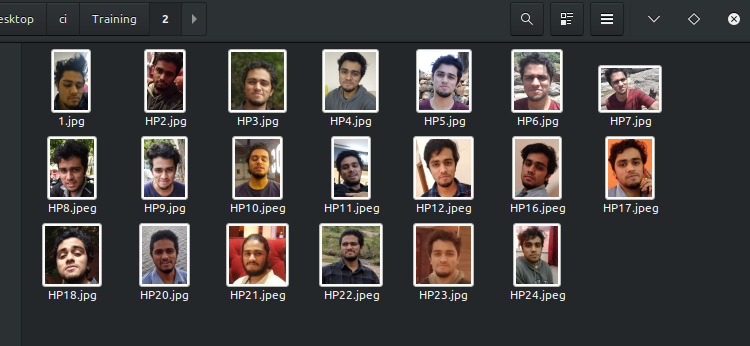
Test images



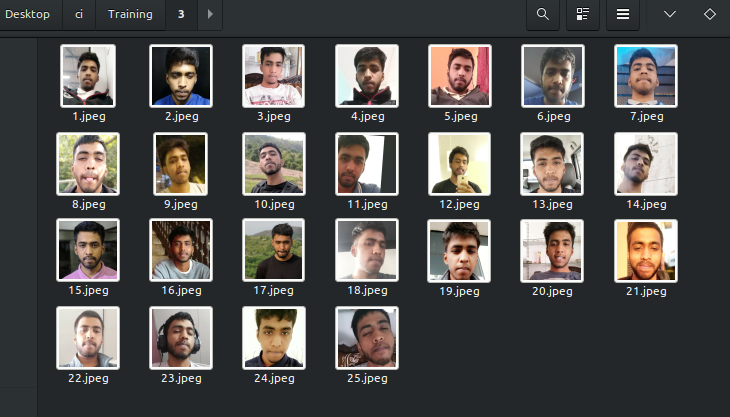
Training dataset 1



Training dataset 2



Training dataset 3



Training dataset 4

Choice of an appropriate Face Detection algorithm and its justification

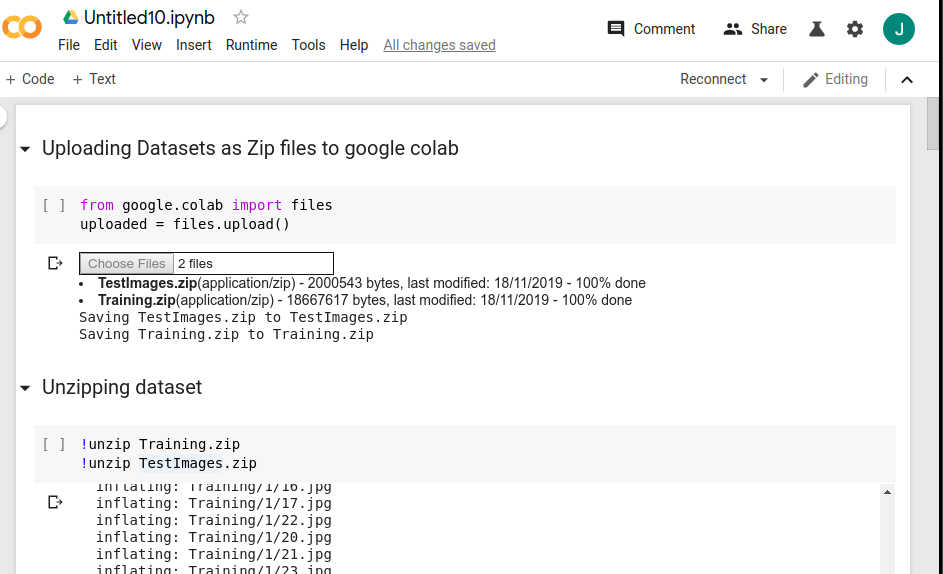
Appropriate face detection algorithm for this purpose is the Haar classifier

Justification:

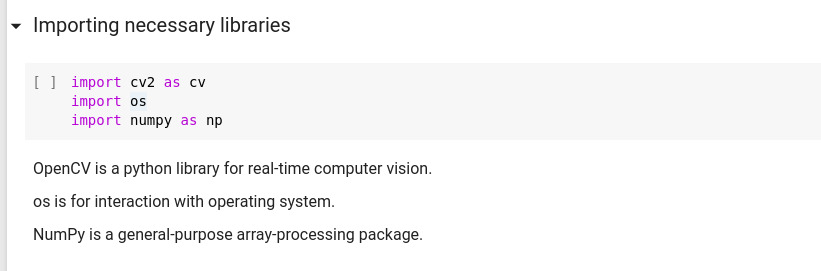
* One of the most important things in a face recognition system is actually detecting the faces in an image. Without the faces, you can’t really do any downstream tasks of person classification, gender classification, emotion classification, and so on. Within OpenCV, there’s a face detection module, which utilizes the technique called Histogram of Oriented Gradient (HOG). It trains a cascade function (boxes of shapes) that appears in images with faces and learns the general pattern of a face through the change in colours/shadows in the image.
* The other alternative is to use Deep learning i.e CNN( convolution neural network). MTCNN is one of the popular CNN architectures face rec. When it comes to computer vision, Deep Learning algorithms blow away all other models in terms of accuracy. This is due to CNN’s ability to learn features from an input image, as opposed to manual feature engineering done in traditional machine learning. In Multi-Task Cascaded Convolutional Neural Network (MTCNN), face detection and face alignment are done jointly, in a multi-task training fashion. This allows the model to better detect faces that are initially not aligned.
* Here, we can’t utilize CNN as the dataset available/created is very small. With such small dataset, CNN’s can’t pick the features and understand the regularities. Thus, it can’t fit well enough. Also, computational power available is limited.Hence, HOG is used.

A1.3 Design and simulate the Face Detection algorithm using Python or MATLAB

The following snippets if code from a python notebook along with explaination is mention below.

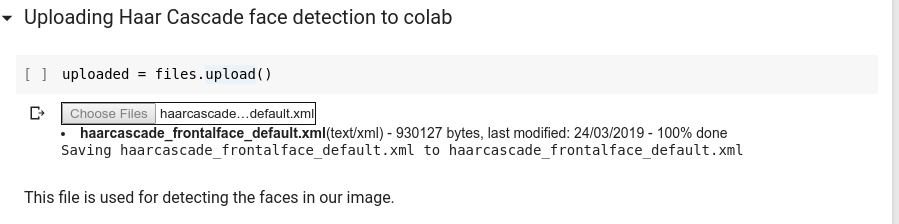


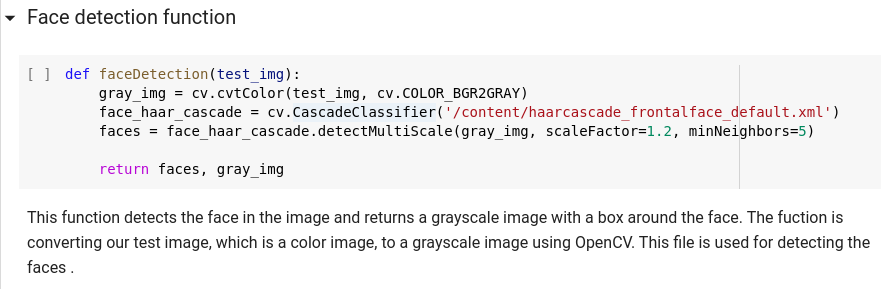
In this snippet, dataset is uploaded to colab and unzipped.

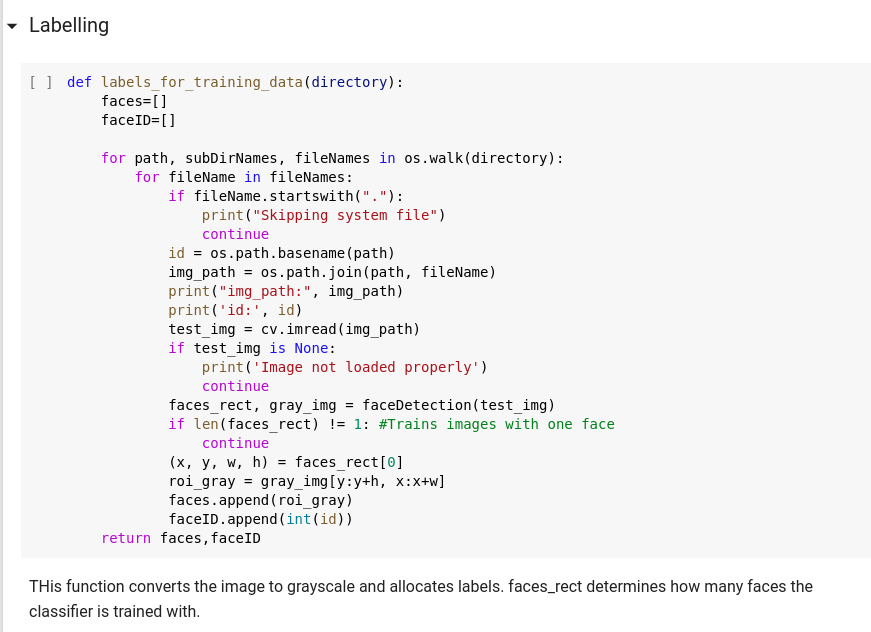




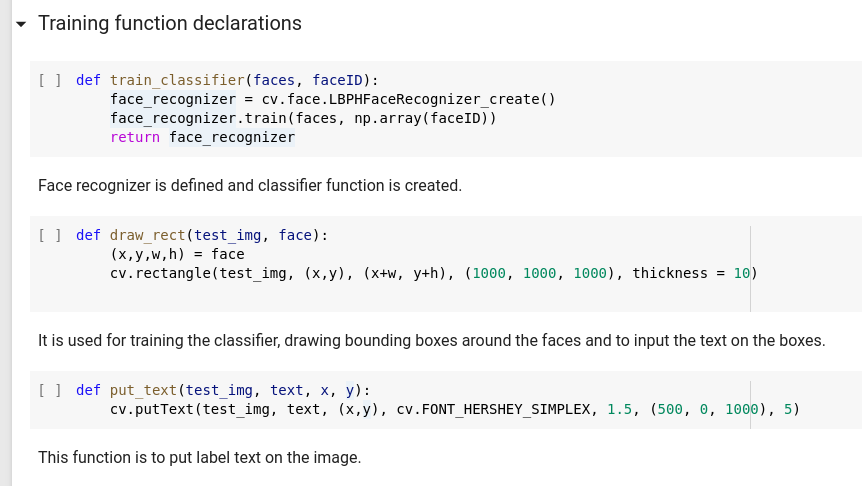
Resizing image from a given directory to 100\*100 pixels and writes all images to given directory

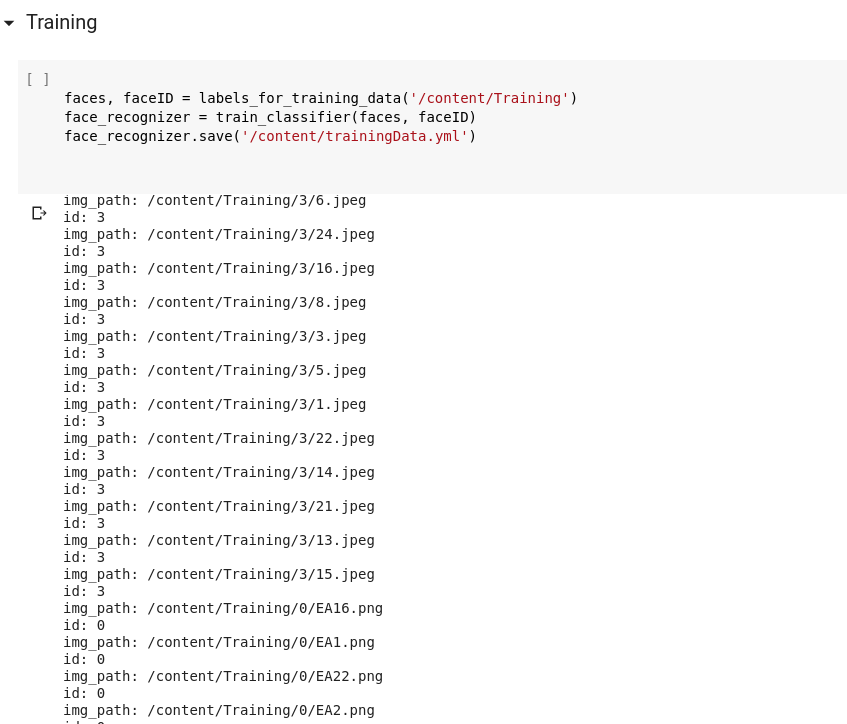






This function takes the directory as the input and returns the faces along with its labels. This function loads each image one by one in the subdirectories of our training data and detects the faces in each image and crops the faces in each image and returns the faces with their ID’s.





Passing the path to our training data and training the haar classifier on our data and creating a dictionary that contains names of each lab

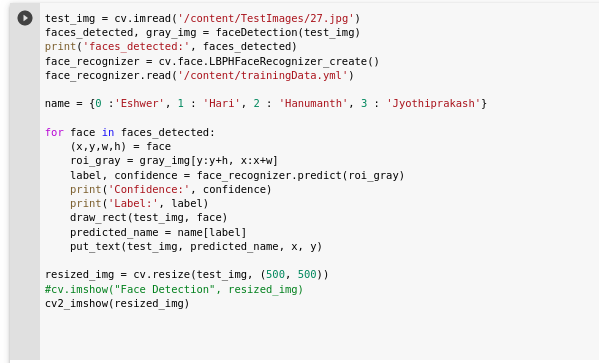
Testing and validation of the algorithm with the given test images

After training our model we shall pass our test image. In this final step, our model detects the face in the test image and draws a bounding box around the face and tries to predict the person in the image.

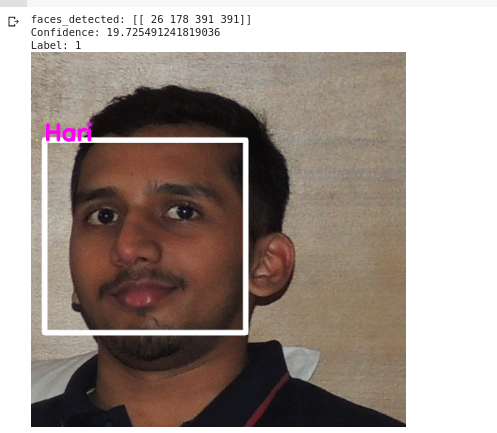


Inferencing on all the faces.

**1.**



In this snippet the grey test image is taken and the previous created functions and model is used to derive an inference as to who the person is.



Classifies as Hari with label 1 which is right.

**2***.*



**3.**



Correct inference though it’s an older image as compared to the ones used in training for that face

**4.**



Here, there’s an occurrence of something peculiar. Though the recognizer recognizes the face correctly, it identifies and misclassifies a wrinkle of a shirt as someone else’s face.

**5.**



Correct recognition with confidence of 51

Further improvements can be performed by

* Expanding dataset: Larger data helps the model learn better and allows the use of better algorithms.
* Data augmentation : Data augmentation can bring in more variane and improve generalization of the model.
* Third, better algorithms with higher computational and data can be used.

Introduction to NLP showing the steps to address the issue

"Natural Language" refers to human languages, and not programming/computer languages. Natural Language Processing (NLP) is a field in Artificial Intelligence and is also related to linguistics. On a high level, the goal of NLP is to program computers to automatically understand human languages, and also to automatically write/speak in human languages.

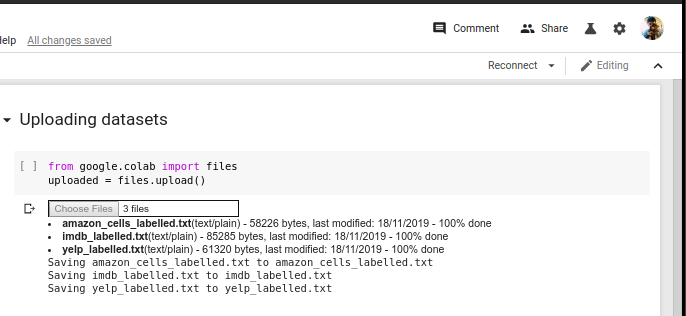
Comments passed on the internet contain lot of information that can be used to understand public behaviour. This understanding can provide insights which can be exploited to serve the customers well and gain higher profits. Also, it can be used to curb online toxicity and social media bullying by identifying negativeness and addressing them accordingly.

Here, we are classifying if a comment is positive or negative in nature. This is achieved by following three steps.

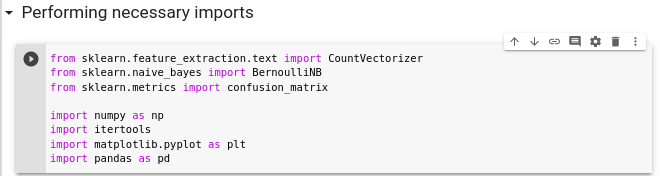
I.e Preprocessing, training and testing. Data required is collected from amazon, quora and yelp.The first step is about to collect, clean and organize your data. Fortunately, the selected data sets are pretty clean already.The second step is related to use your cleaned data sets to train your model through examples. By doing that, you’ll generate a classifier model.

This model contains a predict function which can receive any new commentary and classify it as a positive or negative one.The last step is when you validate if your model is good enough to solve the purposed problem. There’s no step by step validation, because it depents entirely on the purposed problem.

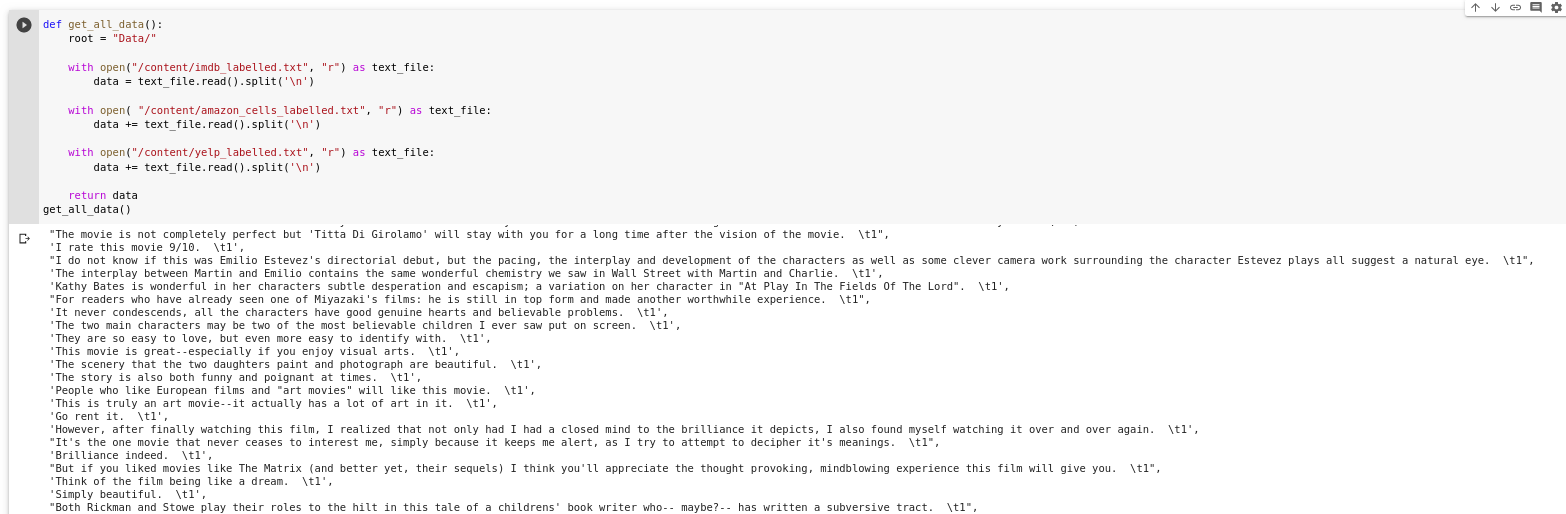
Below snippets are of a python notebook which consists of code with respective explainations.



Data files(text) are uploaded to colab



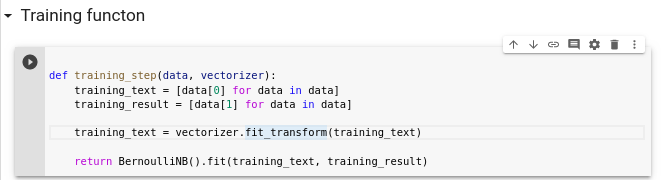
Necessary imports are performed. Scikit-learn is a machine learning library for the Python with various classification, regression and clustering algorithms including support vector machines,



The data is retrieved from the text files by accessing the paths to where they re uploaded and . the line breaker scape(\n ) to split our data into an array. Here ‘1’ means its positive and 0 means it’s negative.



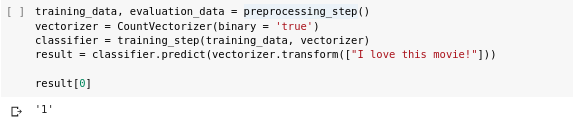
The above snippet consists of the preprocessing function which unifies other functions with same purpose.



The func splits training data set in two different lists, one for the sentences and other for the results.

and vectorizes the sentence by using fit\_transform function and finally returns the result of a BernoulliNB().fit function call.

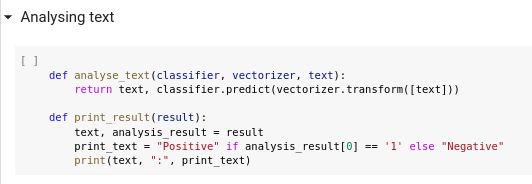
The BernoulliNB (NB for Naive Bayes) will generate classification model. The predict function of this model is used to predict the sentiment of any text



Above is an inference example.1 indicates that it is a good comment. And “I love a movie ” is indeed one.

To calculate the negative score one can simply use 1- Positive Score.

After these two scores are calculated, the Naive Bayes algorithm will use them to calculate the sentence score.

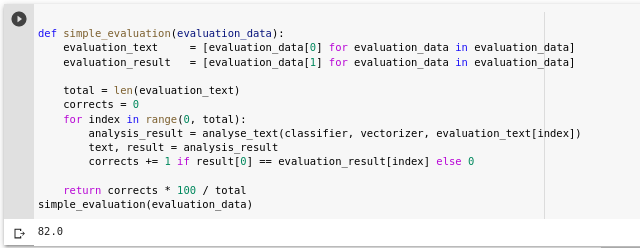


Vectorizer- The model will count, how many times the word good appear in a positive sentence and divide this number by how many times the word appears at all. It’ll create a positive score on this word.

This process is called vectorization.

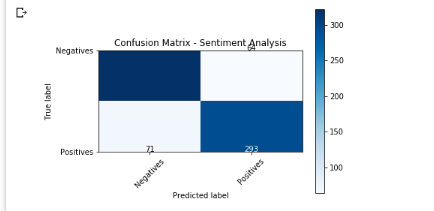
Evaluatiin

To check our predictions answer against the actual answer counting the right answers, the above code is run to iterate through all data by using our model to predict the sentiment analysis of each sentence, then, compare the model predicted result against the actual result in the data set.



To get information about prediction model the confusion matrix is utilized.

The confusion matrix is created using sklearn.metrics. False and true, positive and negatives are computed and and plotted diagramatically. This matrix is extremely useful because in several cases, there’s no big problem to create a false negative as creating a false positive or vice-versa. matplotlib.pyplot lib to create better visualization.



There are a total of 386 negatives sentences and a total of 364 of negatives, totalizing 750 sentences, which is 25% of the 3000 total sentences.

Of this 386 negatives sentences 322 were classified correctly while 64(about 16.5%) were classified as positive sentence, creating some false positives. On other hand there’s 293 positives sentences classified correctly and 71(about 24%) false negatives.

From this metrics like accuracy, recall, precision,FPR , TPR can be calculated to eveluate the model. As high accuracy doesn’t always mean that it’s a good model(cases where there is imbalaced data for example.)

Sentiment analysis is one of the metrics that will advance your marketing strategy as it has impact on many different aspects of your brand’s online presence. Analysing the sentiment around your brand helps you understand motivations behind your customers’ purchasing decisions and the intent behind their searches.

This problem gave an insight as to how nlp is used to derieve insights from textual data present in the web and throughout. But it’s hard to build programs to understand human language constructs ike sarcasm, jokes, exagerration etc. 'Disappointed' may be classified as a negative word for the purposes of sentiment analysis, but within the phrase “I wasn't disappointed", it should be classified as positive.