

**Fake-News**

**Classifiers Project**

Submitted by:

Vandana Jain

Internship 10

**ACKNOWLEDGMENT**

The internship opportunity I had with FlipRobo was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this project period.

I would like to thank our SME for suggesting this project and for his whole hearted cooperation and constant encouragement throughout the project.

**INTRODUCTION**

* Business Problem Framing

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

Our goal is to build a prototype of real and fake news classifier which can used to classify real to fake news so that it can be controlled and restricted from spreading the false information and cause any harmful impacts.

* Motivation for the Problem Undertaken

Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers”. Social media and news outlets publish fake news to increase readership or as part of psychological warfare. Ingeneral, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts.

* Research Problem

The project is concerned with identifying a solution that could be used to detect and filter out sites containing fake

news for purposes of helping users to avoid being lured by clickbaits. It is imperative that such solutions are identified

as they will prove to be useful to both readers and tech companies involved in the issue.

* OBJECTIVE

The objective of identification of comments are :

* To give knowledge to the user about the real and fake news.
* To classify that news is authentic or not.
* SCOPE OF THE PROJECT:
* It provides sensitivity to the client and adapts well to the future fake news detection techniques.
* It considers a complete message instead of single words with respect to its organization.
* It increases Security and Control.
* It reduces IT Administration Costs.
* It also reduce Network Resource Costs.

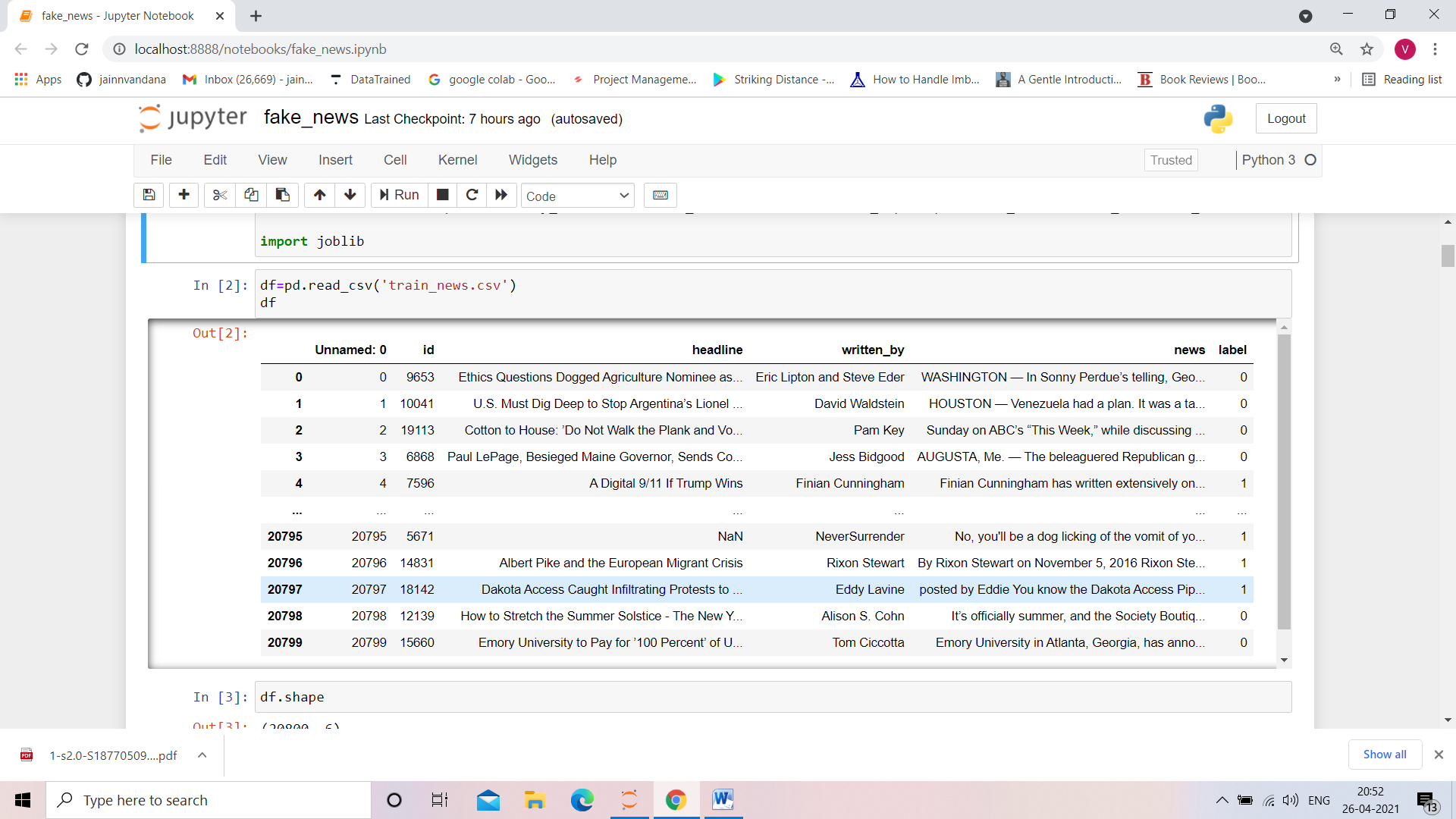
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Machine Learning is defined by Tom Mitchell in his book as “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. Supervised learning is when the output is known for the corresponding inputs, and is also provided for the machine to learn.

* EDA (Exploratory data analysis)
* Data Preprocessing
* Feature Extraction
* Scoring & Metrics
* Data Sources and their formats

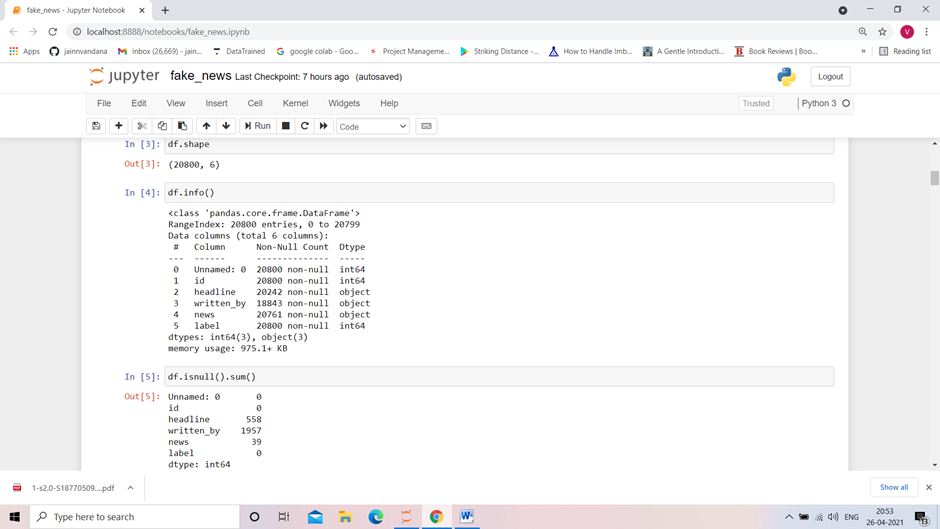
The data is provided to us from our client database. It is hereby given to us for the exercise to improve the selection of authentic news from fake one. It is given in the csv file format.



* Data Set Description

There are 6 columns in the dataset provided to you. The description of each of the column is given below:

* “id”: Unique id of each news article
* “headline”: It is the title of the news.
* “news”: It contains the full text of the news article
* “Unnamed:0”: It is a serial number
* “written\_by”: It represents the author of the news article
* “label”: It tells whether the news is fake (1) or not fake (0).

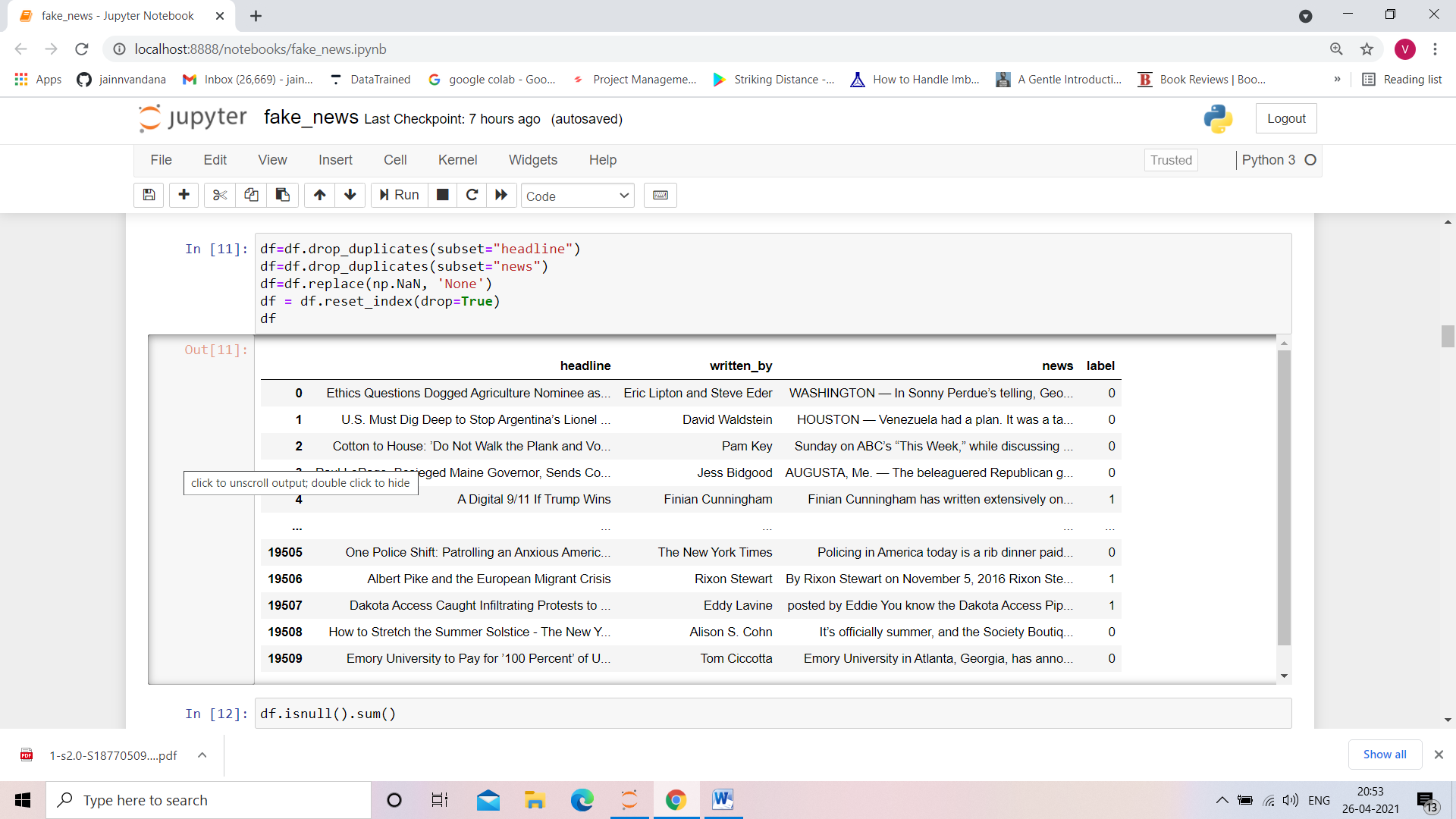


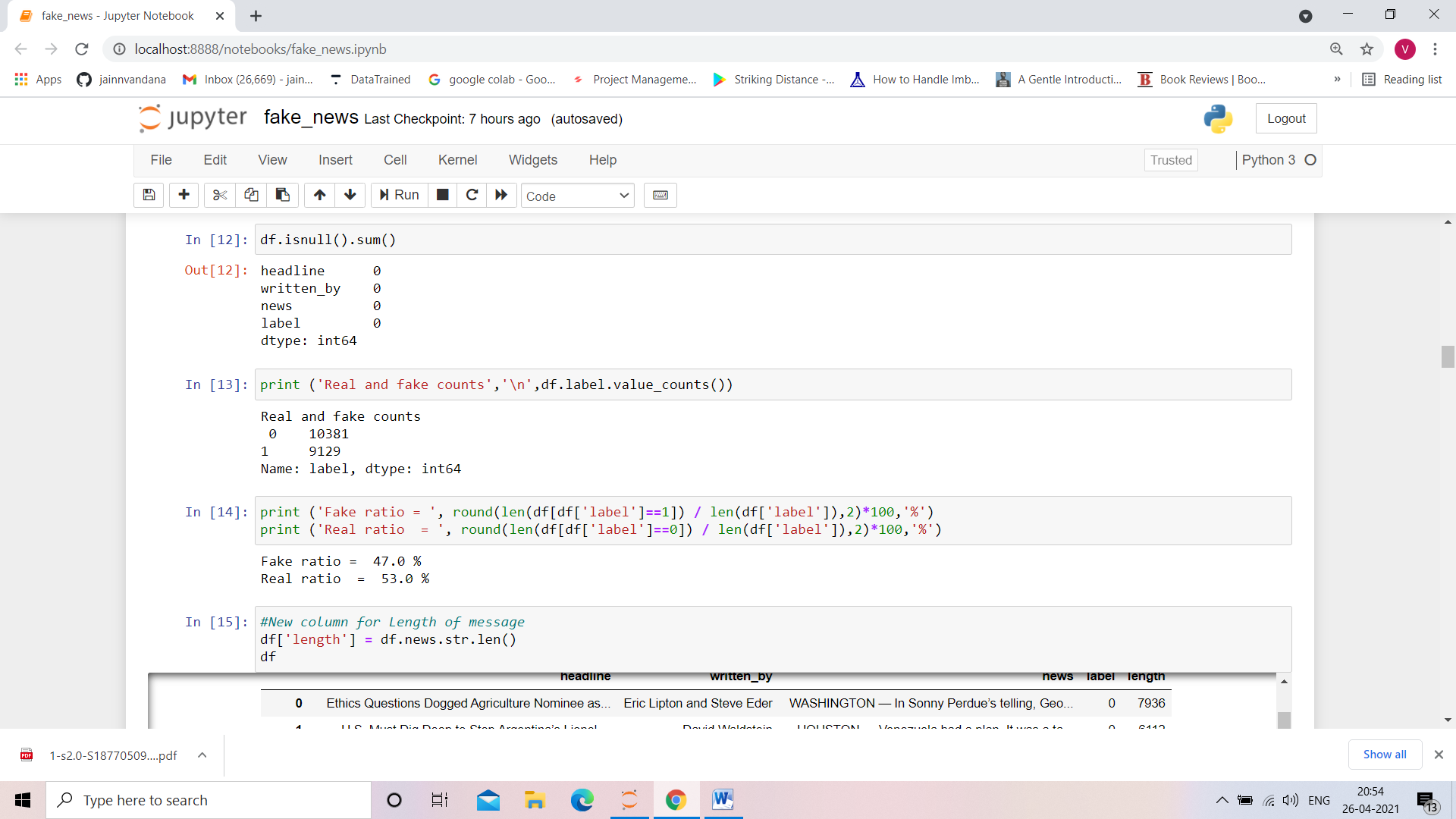
* Data Preprocessing Done

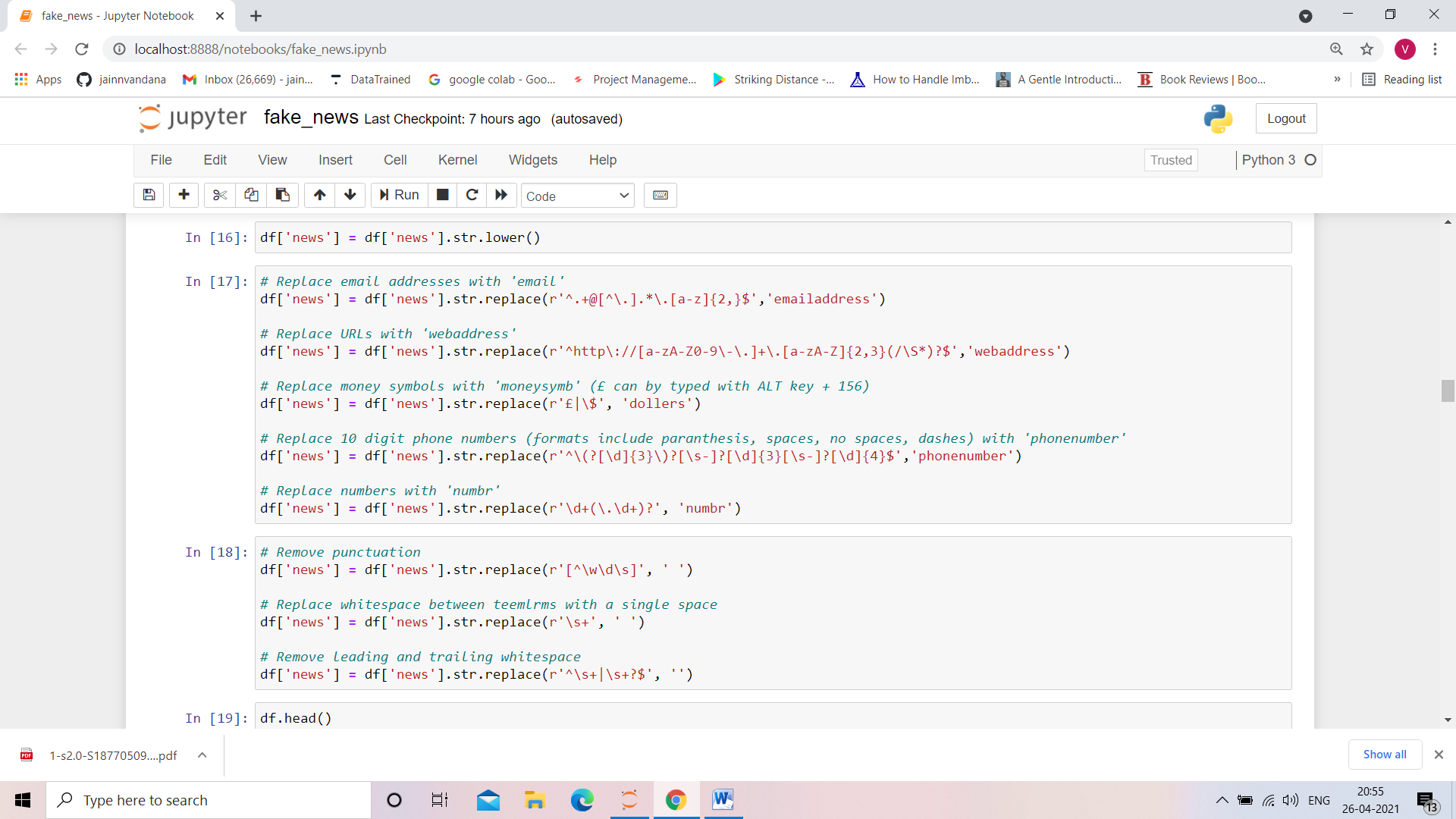
The dataset that will be used to train the model has some challenges. Text Cleaning is a very important step in machine learning because your data may contains a lot of noise and unwanted character such as punctuation, white space, numbers, hyperlink and etc.

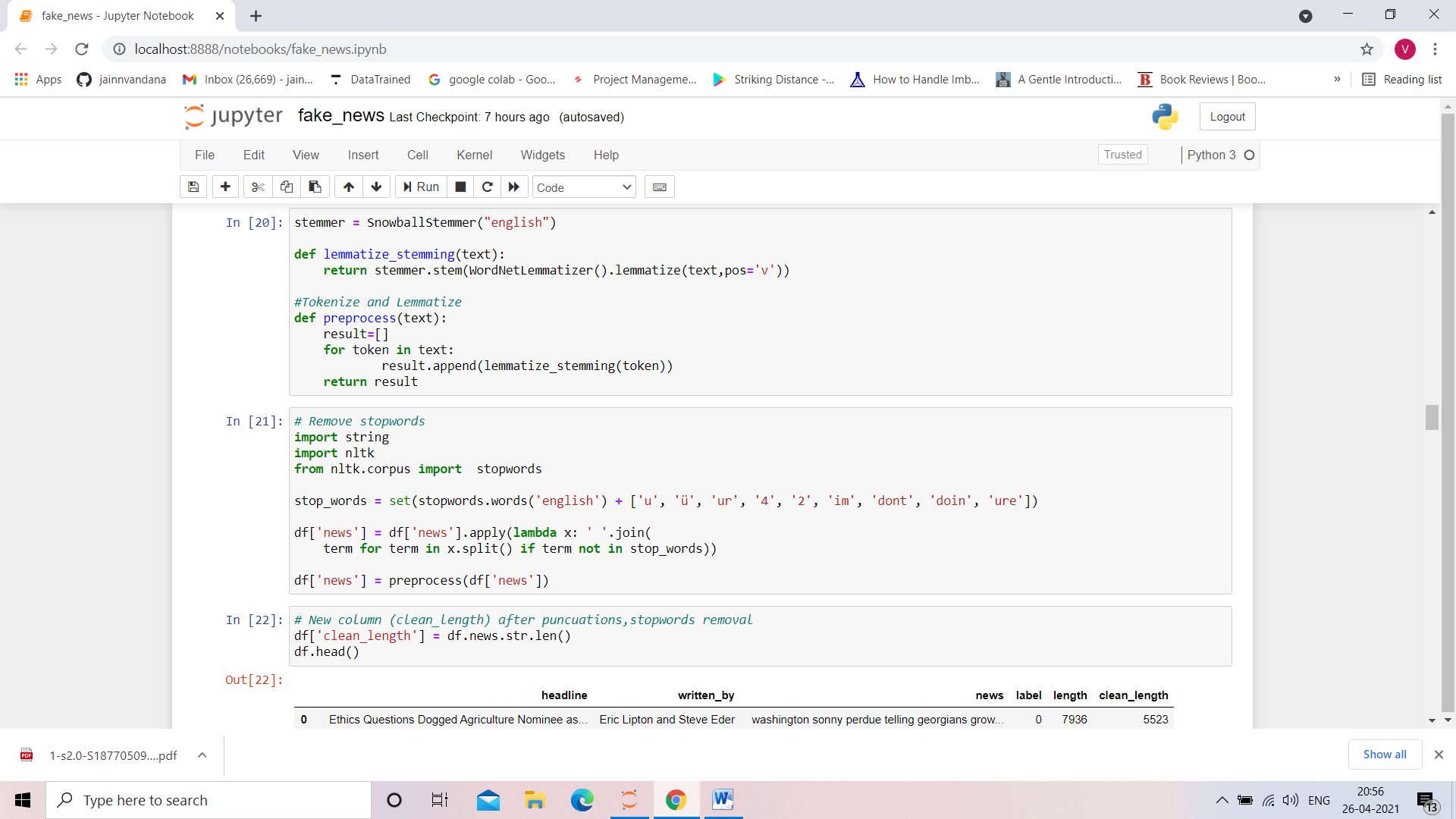
Some standard procedures are:

* convert all letters to lower/upper case
* removing numbers
* removing punctuation
* removing white spaces
* removing hyperlink
* removing stop words such as a, about, above, down, doing and the list goes on… Sometimes, the extremely common word which would appear to be of very little value in helping select documents matching user need are excluded from the vocabulary entirely.
* **Word Stemming:** Stemming algorithms work by removing the end or the beginning of the words, using a list of common prefixes and suffixes that can be found in that language.
* **Word lemmatization**: Lemmatization is utilizing the dictionary of a particular language and tried to convert the words back to its base form. It will try to take into account of the meaning of the verbs and convert it back to the most suitable base form.







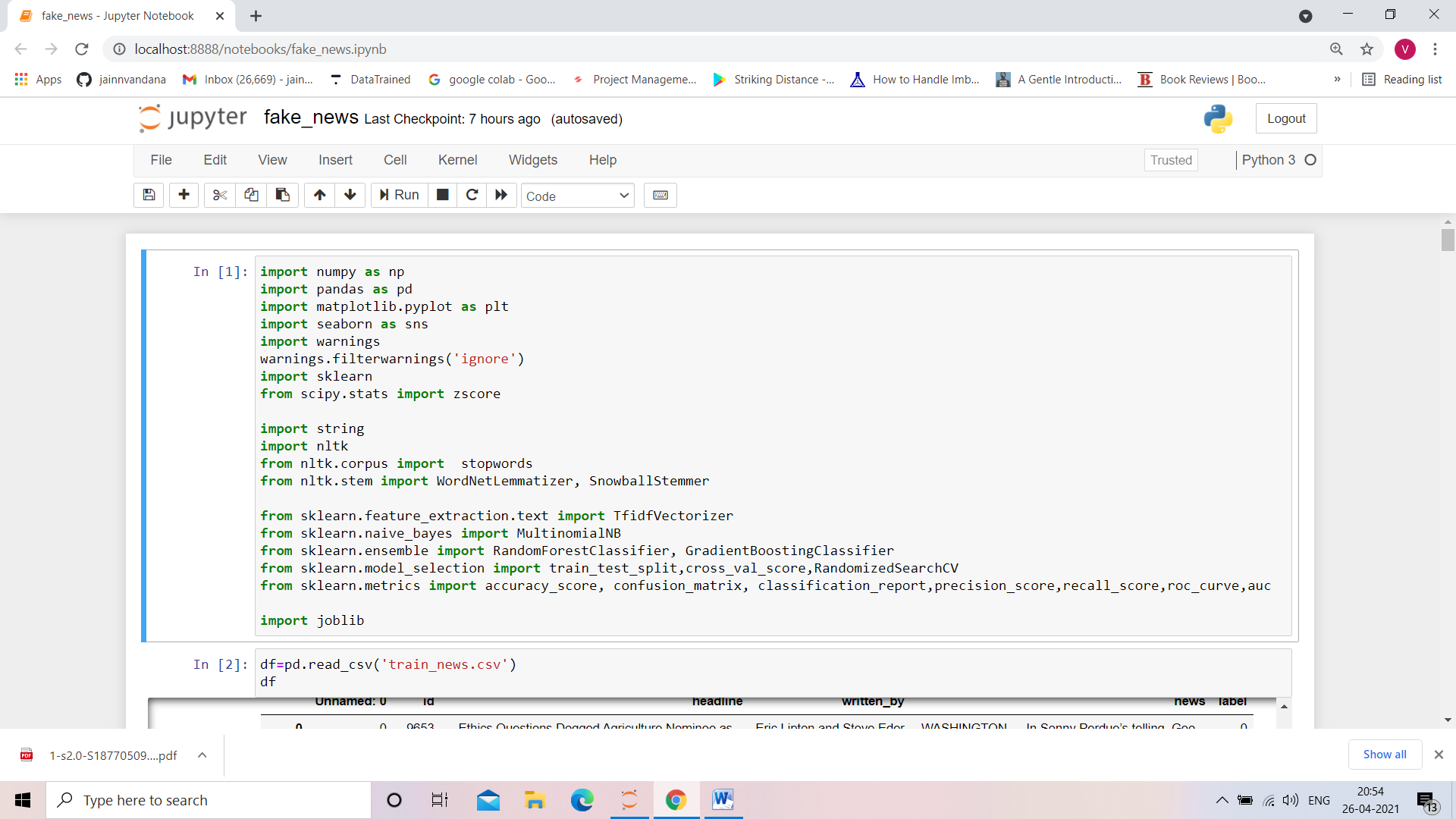


* Hardware and Software Requirements and Tools Used

Hardware : Since the computational aspect of the project is of importance to PANDA, it is important to know the hardware that was used in the evaluation process. The training and evaluation of the neural network model has been done on a Windows 10 computer using a quad-core CPU at i3.

Software : anaconda 3 , windows 10 ,Microsoft office.

Tools used : python , machine learning libraries, Nltk, Nlp libraries.



**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

1. NAÏVE BAYS CLASSIFIER

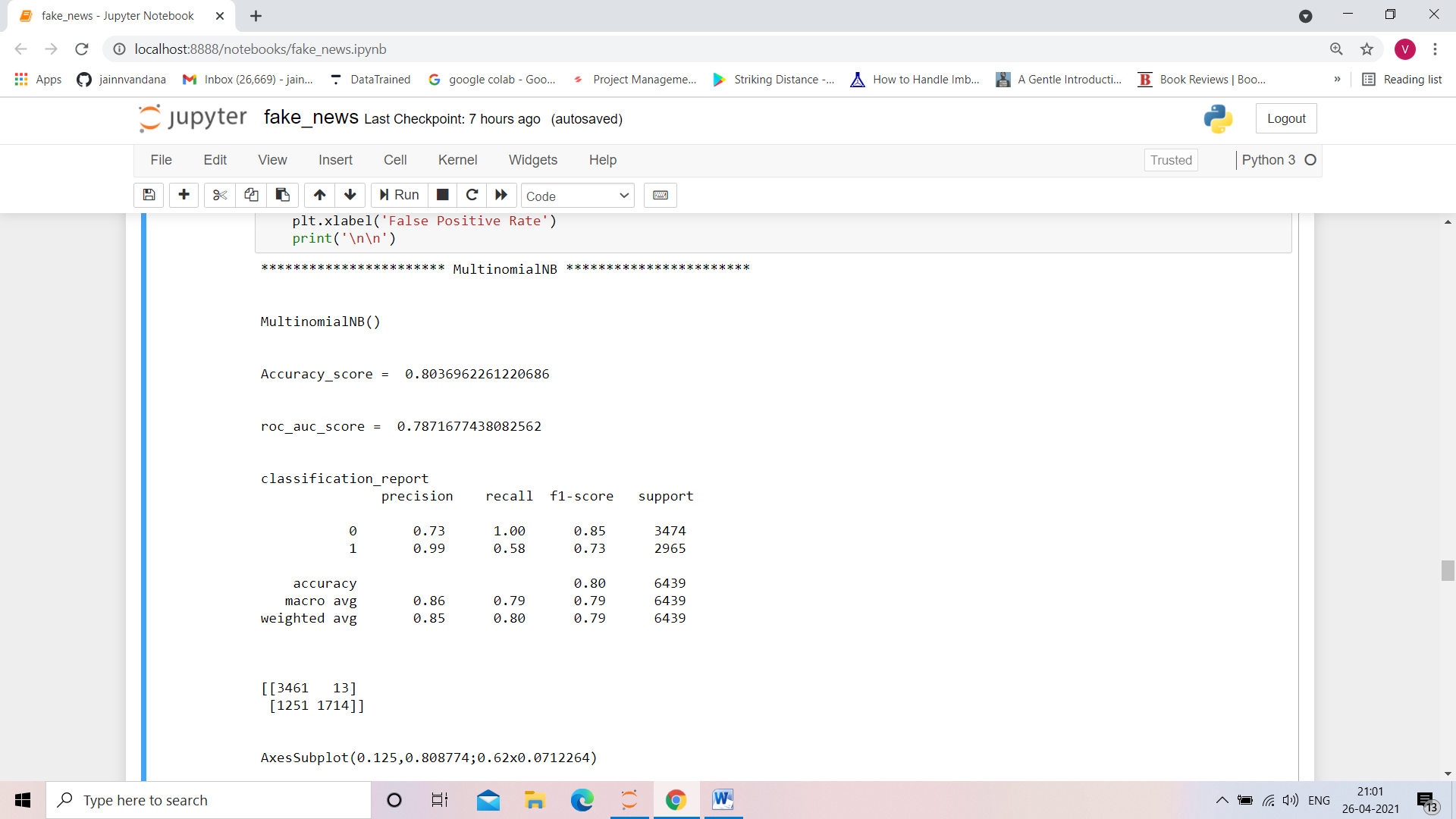
Simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values in a given dataset. Represent as a vector of feature values. It is very useful to classify the comments properly. The precision and recall of this method is known to be very effective.

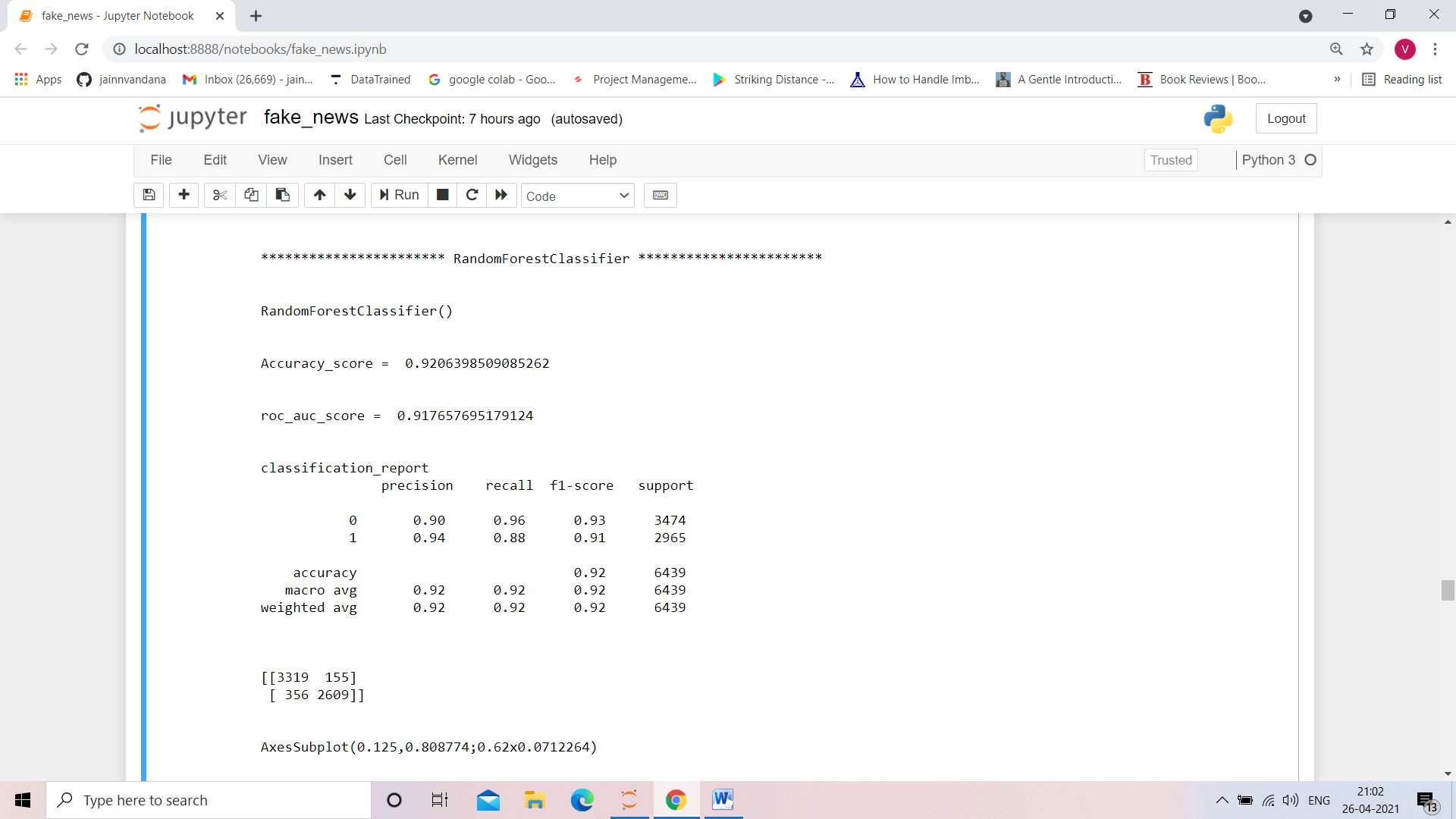
1. RANDOM FOREST CLASSIFIER

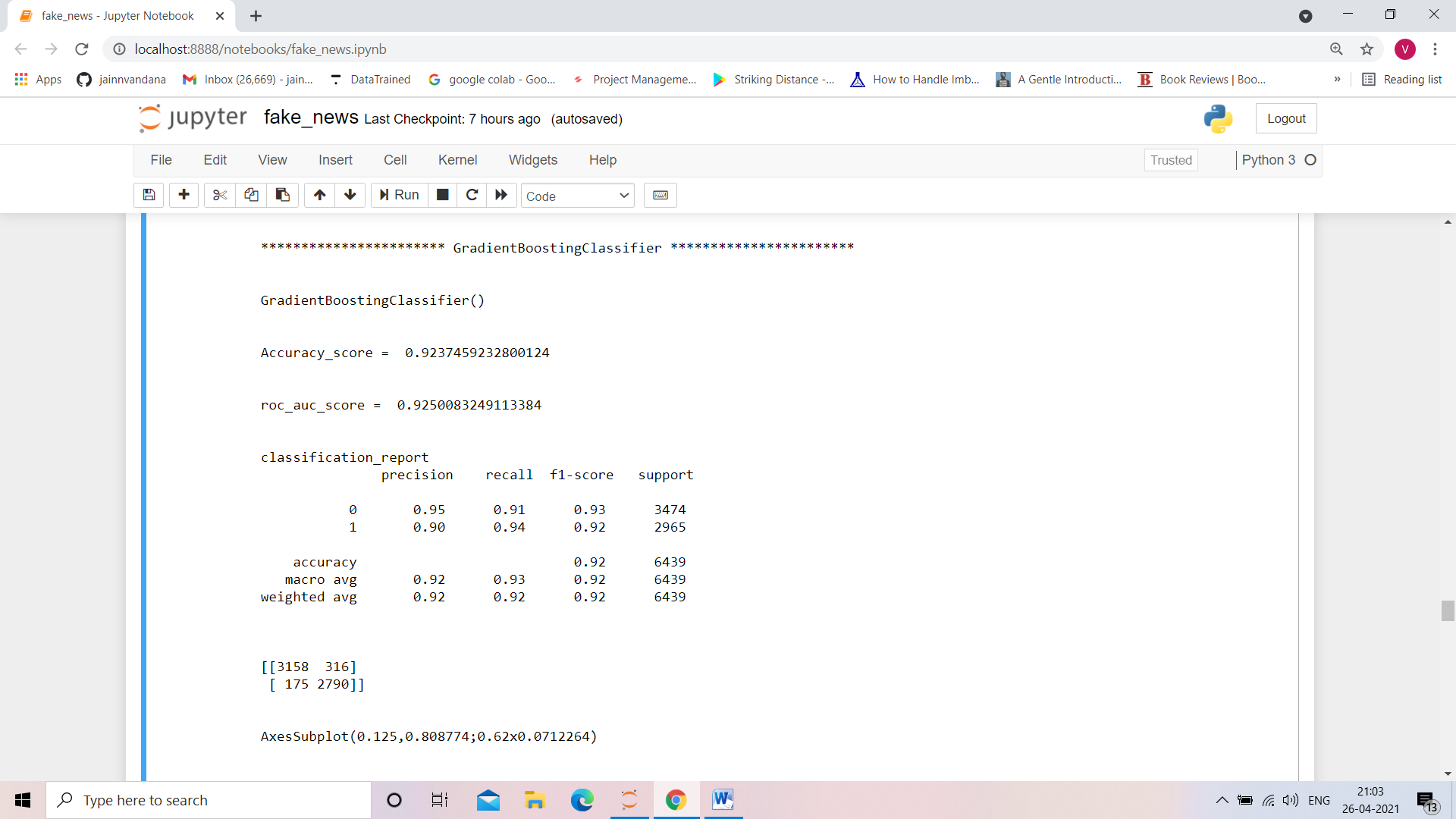
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

1. GRADIENT BOOSTING CLASSIFIER

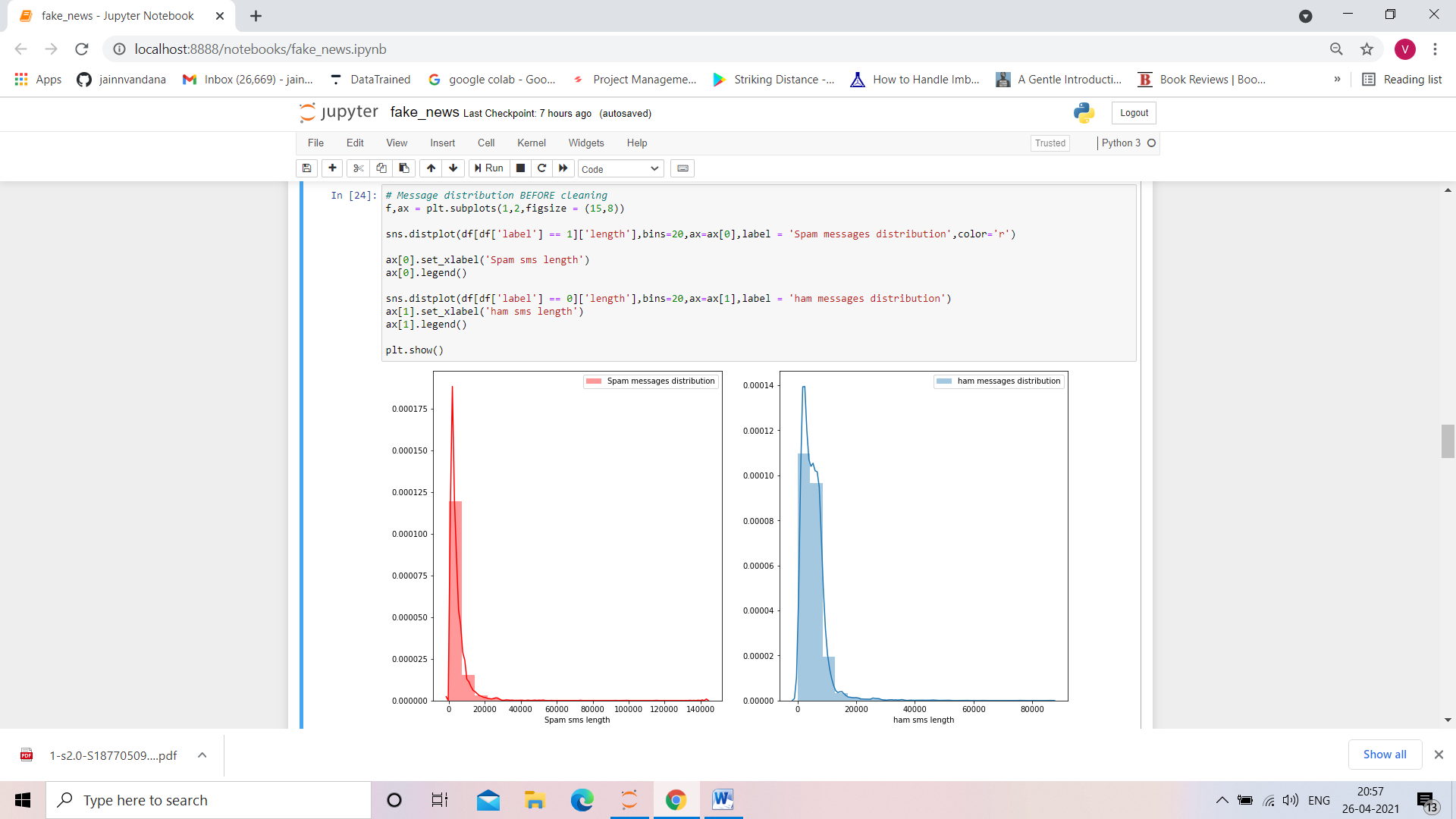
In Gradient Boosting, each predictor tries to improve on its predecessor by reducing the errors. But the fascinating idea behind Gradient Boosting is that instead of fitting a predictor on the data at each iteration, it actually fits a new predictor to the residual errors made by the previous predictor. Gradient Boosting has repeatedly proven to be one of the most powerful technique to build predictive models in both classification and regression. Because of the fact that Grading Boosting algorithms can easily overfit on a training data set, different constraints or regularization methods can be utilized to enhance the algorithm's performance and combat overfitting. Penalized learning, tree constraints, randomized sampling, and shrinkage can be utilized to combat overfitting.

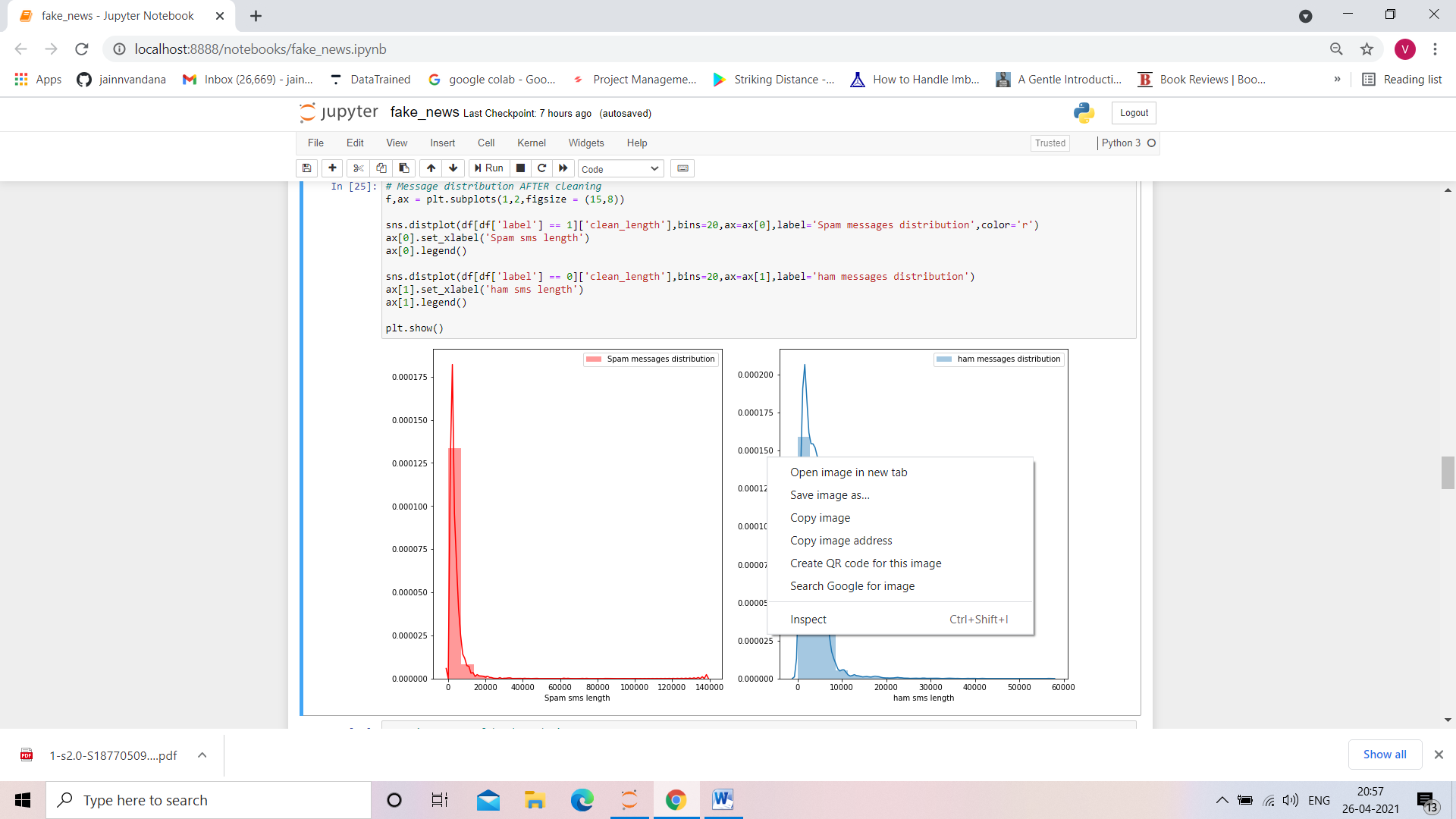
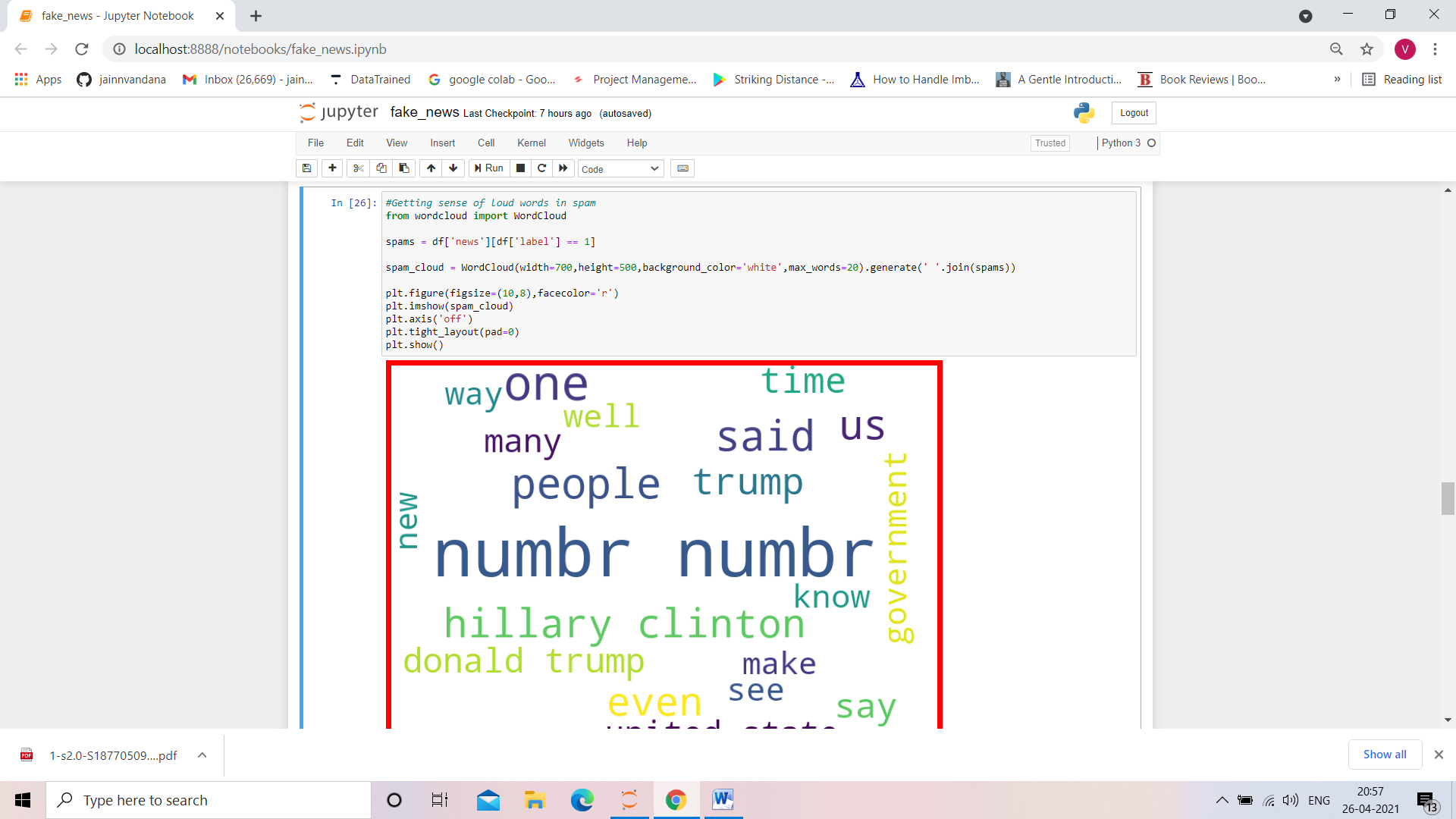
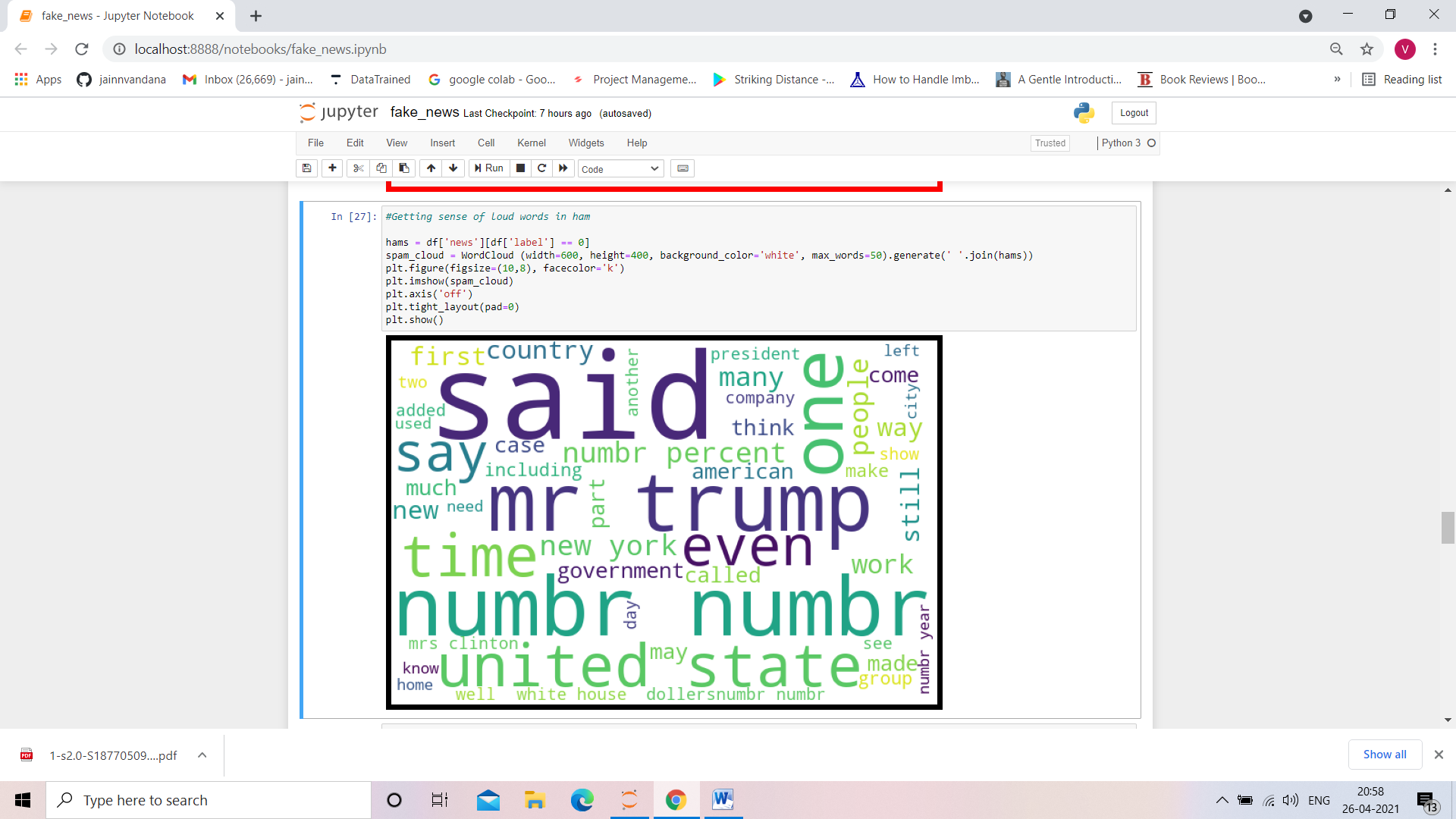






* Visualizations



* Key Metrics for success in solving problem under consideration

When it comes to evaluation of a data science model’s performance, sometimes accuracy may not be the best indicator.

Some problems that we are solving in real life might have a very imbalanced class and using accuracy might not give us enough confidence to understand the algorithm’s performance.

In the fake news problem that we are trying to solve, fake and real news data is approximately 47% of our data. Here we can see our dataset is balanced dataset so accuracy score nearly tells the right predictions.so the problem of overfitting in this problem is nearly not to occur. So here, we are using accuracy score as to find better model.

Accuracy score

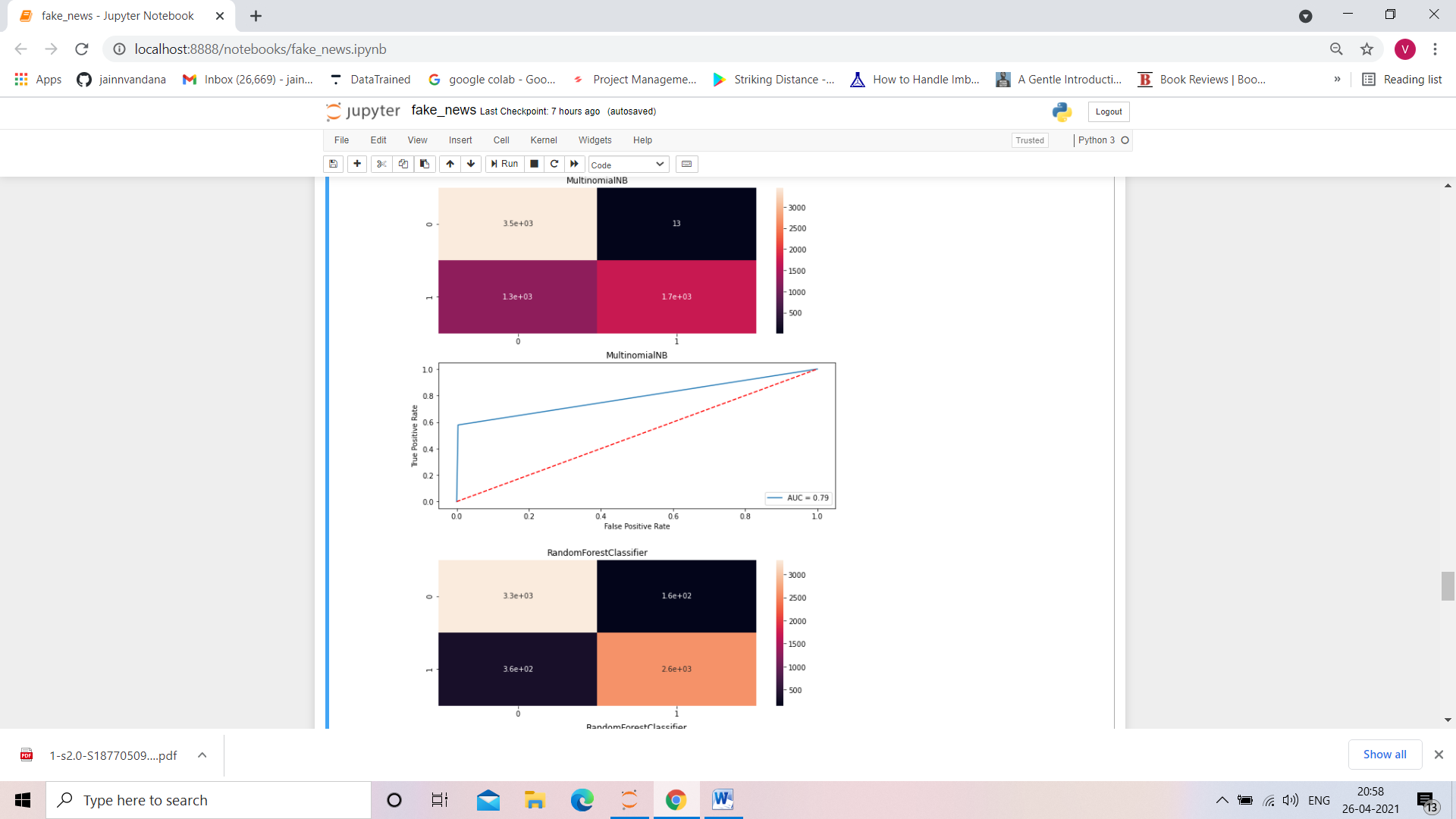
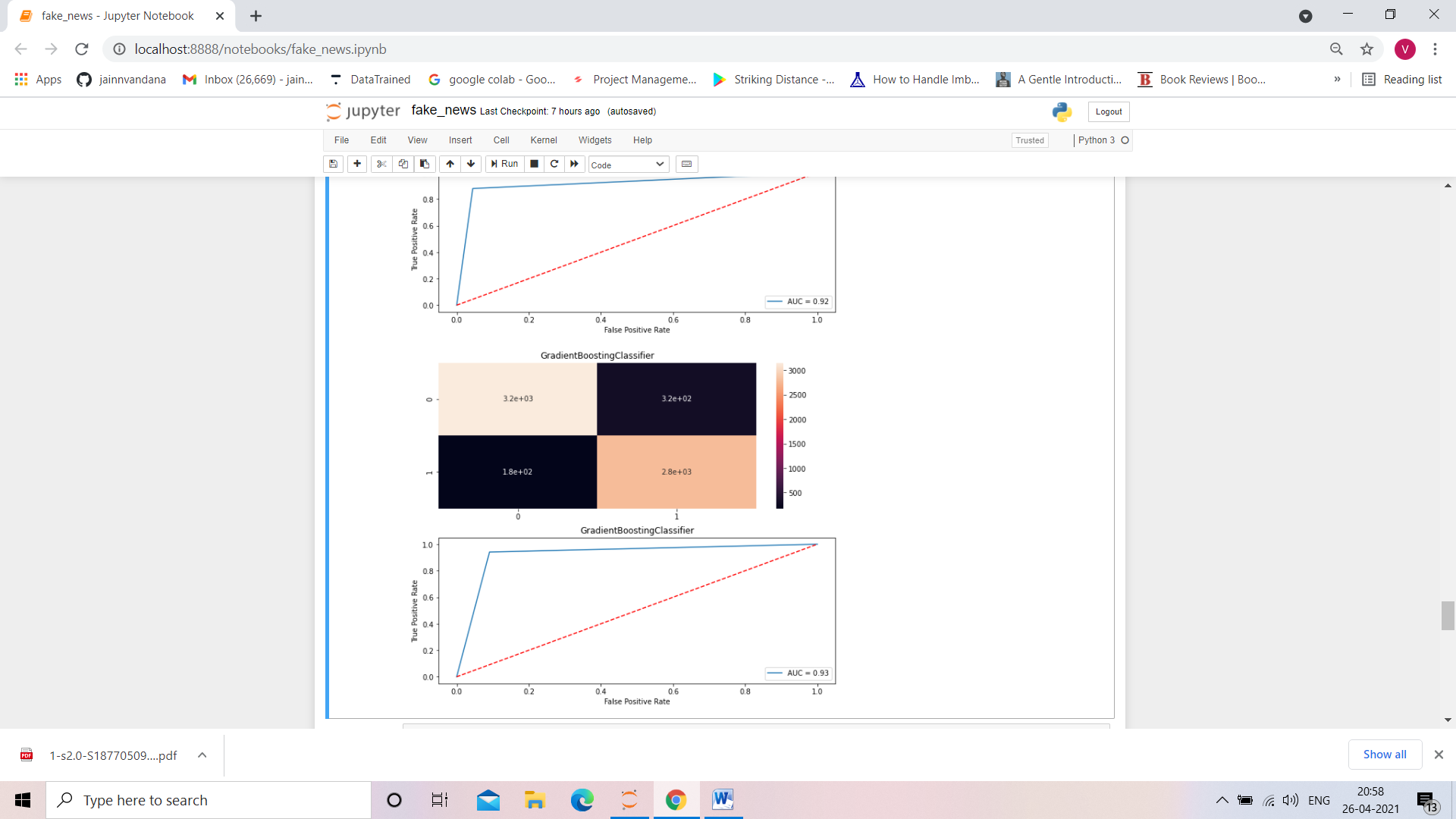
Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

Confusion Matrix

Confusion Matrix is a very good way to understand results like true positive, false positive, true negative and so on.

Sklearn documentation has provided a sample code of how to plot nice looking confusion matrix to visualize your result..

Confusion Matrix of the result

**CONCLUSION**

Now that we have implemented the algorithm.

Label ‘1’ has approximately 47% records, while, label ‘0’ has approximately 53% records. Let’s have the results. We have implemented the dataset using Naive Bayes, Random Forest, Gradient Boosting. We have the accuracy score of prediction of fake or real news is 92% given by Gradient Boosting Classifier.

