

**Fake News Classification Project Report**

Submitted by:

Parija Deshmukh

**INTRODUCTION**

* Business Problem Understanding

In our modern era where the internet is ubiquitous, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time.

The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits.

In this project, we aim to perform binary classification of various news articles with the help of concepts - Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real.

**Analytical Problem Framing**

* Data Sources and their formats

The dataset contains five columns. The total corpus of 44898 news. The descriptive feature consists of text and title. The target feature consists of two classes true and fake; the column name is target. The classes are labelled for each document in the data set and represent our target feature with a binary string-type alphabet of {true; fake}. Classes are further mapped to integer 0 (fake) and 1 (true).

* Data Preprocessing Done

The following steps we used for data preparation.

1. Identifying Missing values.
2. Converting all text to lower case.
3. Performing tokenization.
4. Removing Stop words.
5. Labelling classes: fake/true: {0;1}
6. Splitting Train and Test Data: 80% and 20%

* Hardware and Software Requirements and Tools Used

The preliminary step involved in devising a model is loading the required libraries. In this case, we mainly load four libraries namely pandas, numpy, scipy, sklearn and seaborn.

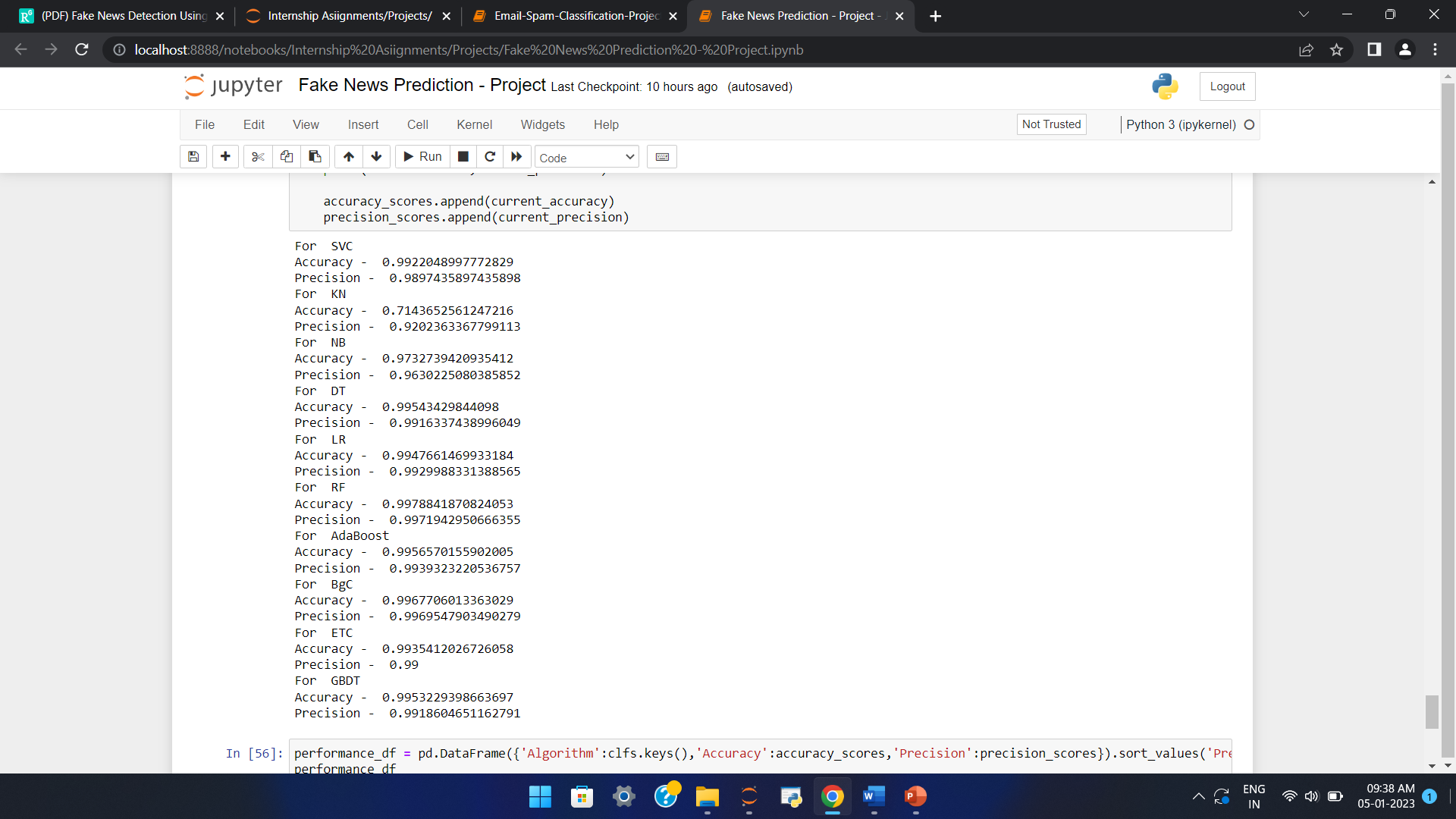
1. Pandas: Pandas is a python package which is quite quick, easy to use and structured in nature. Pandas data-frame is mainly used for data analysis purposes. Pandas also helps us to handle missing data, data to be reshaped and data transformation methodologies.
2. Numpy: Numpy is essentially used for creating very powerful and intuitive n-dimensional arrays. It offers various mathematical functions and also supports various kinds of computing hardware and software requirements. It is an open-source project and also contains various array-objects which are generally quick for usage.
3. Sci-kit: Sci-Kit is one of the most efficient and useful machine learning libraries in python. It provides various statistical and mathematical tools. It also has various techniques like regression, clustering etc. in it.
4. Seaborn: Seaborn helps us to create plots and live-interactive statistical plots and images. It also contains matplotlib which helps us to view our data in a much more intuitive manner.
5. Nltk: The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

**Models used**: **Random Forest Classifier**. Fake news classification done using traditional machine learning technique comprise Random Forest, due to not having sufficient hardware resources, takes less time to train. Also, not opting for neural algorithms due to less data and computing resources.

* Testing of Identified Approaches (Algorithms)



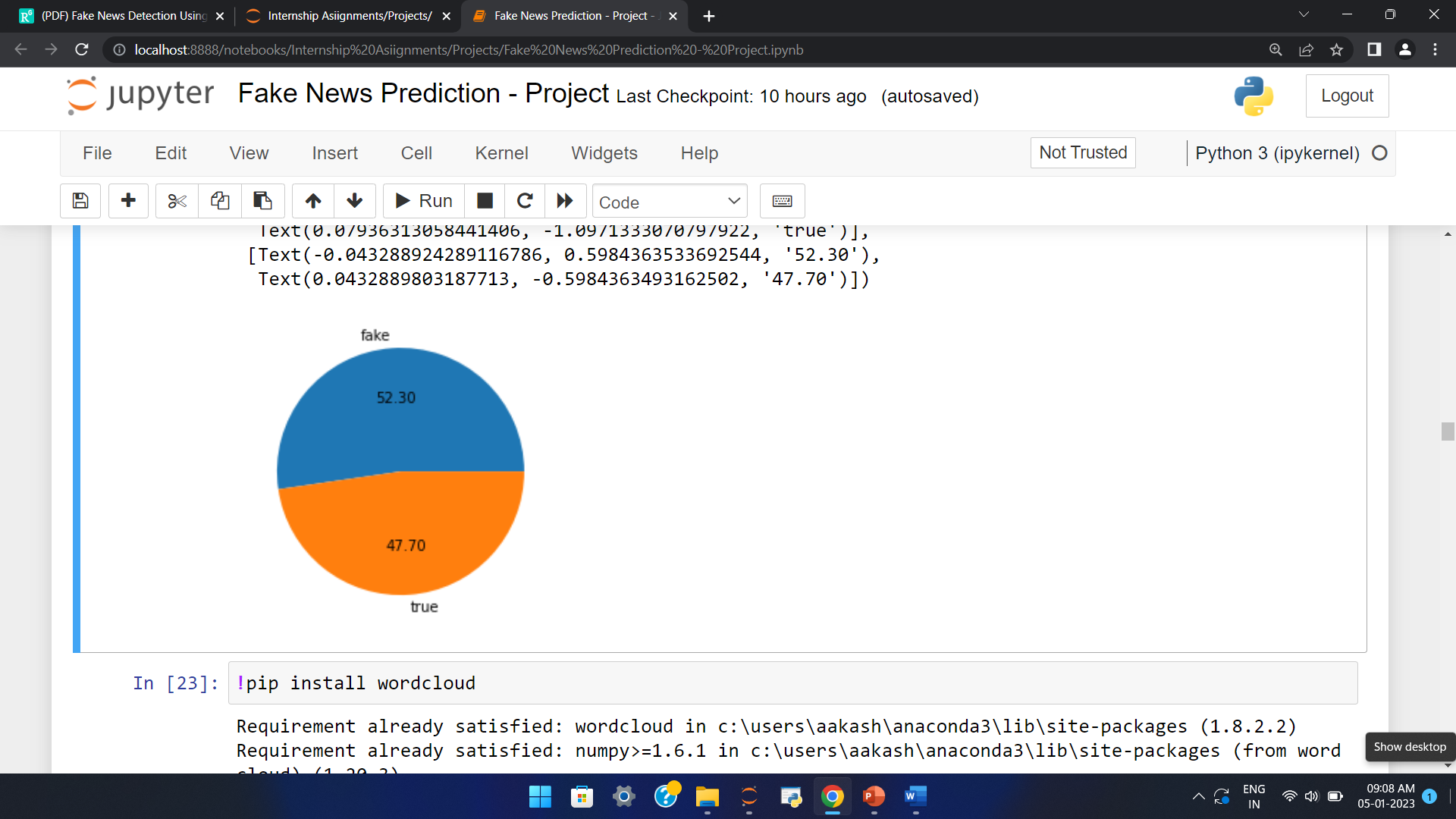
As we can observe RFC is performing well as compared to others with respect to precision and accuracy.

* Key Metrics for success in solving problem under consideration

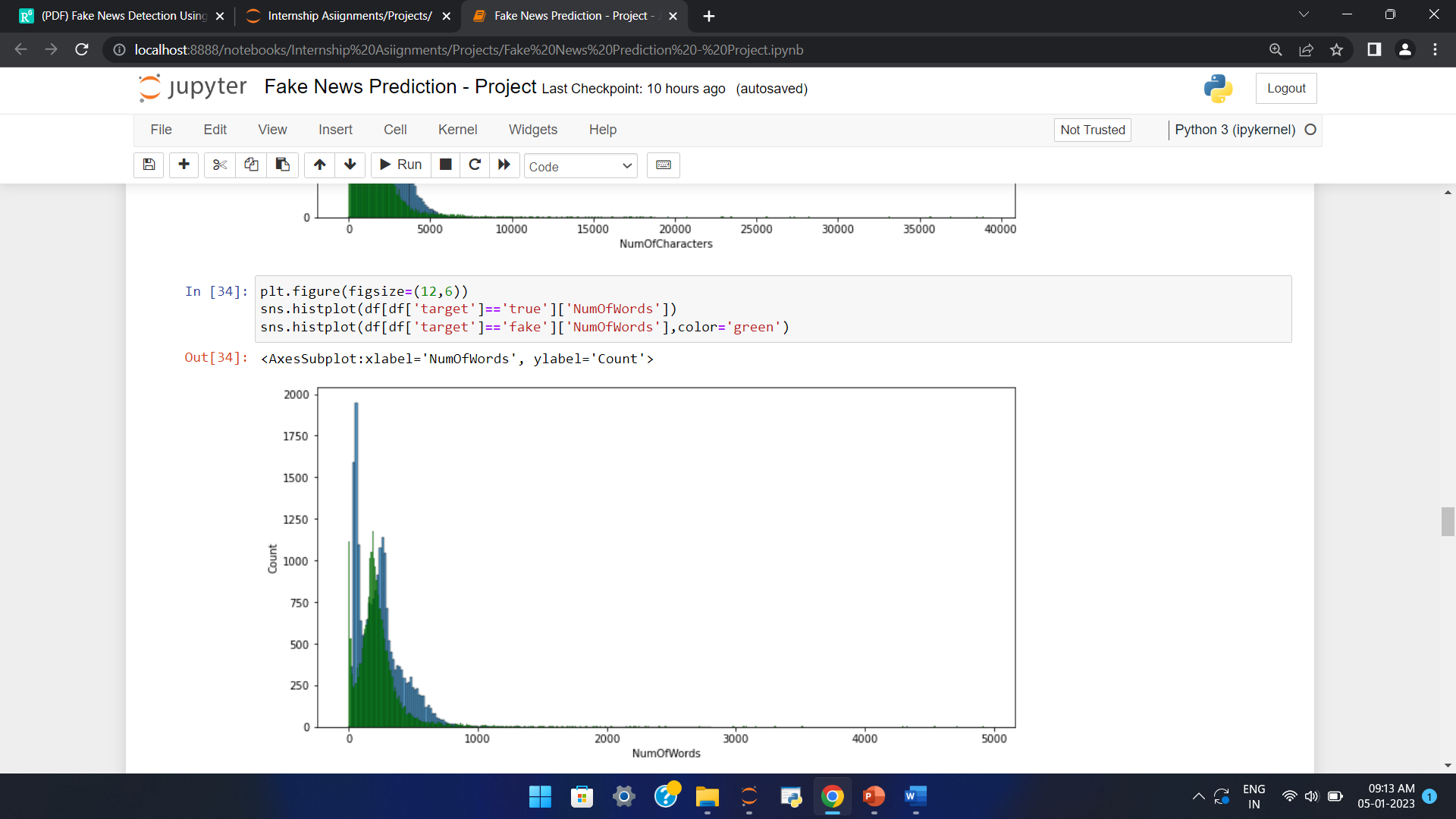
1. Confusion matrix has been used to identify recall/sensitivity, precision and F1 score.
2. Correlation matrix used to analyse the relationship between the columns.

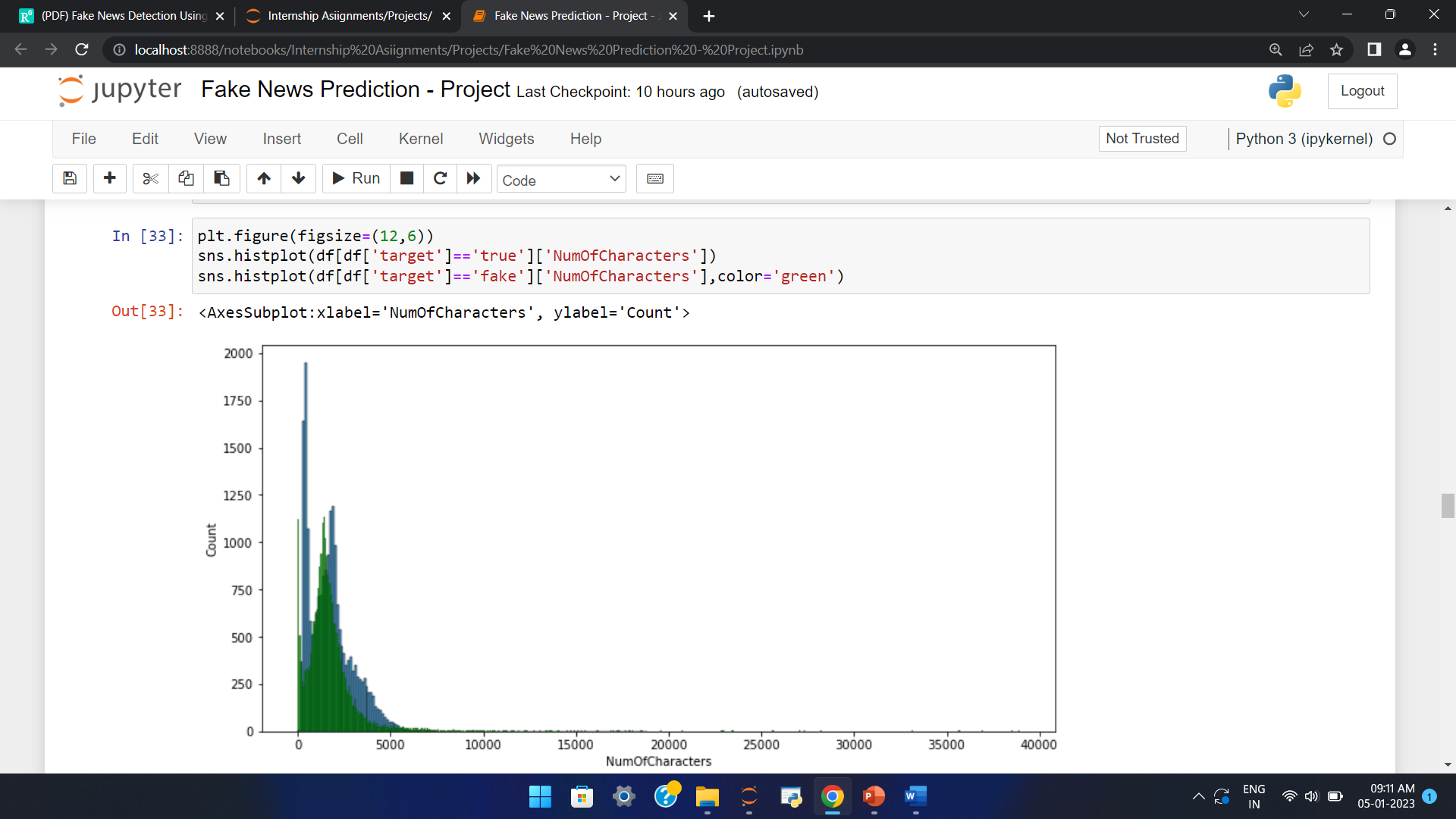
* Visualizations

1. Using Pie chart, we understand data is quite balanced as 52.30% data is of fake news and 47.70% data is of true news.

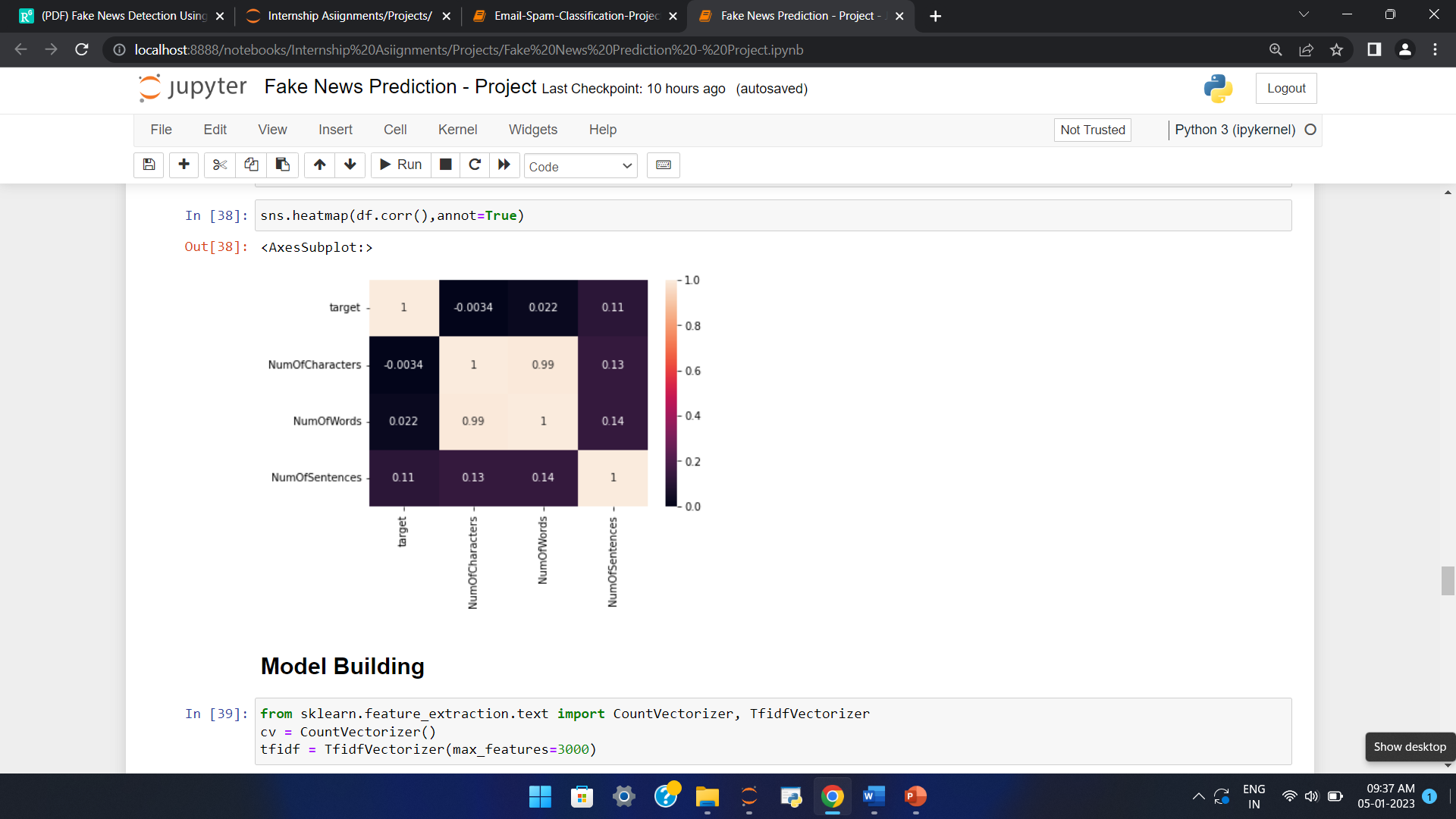


1. By using histplot we understand fake news contains a greater number of characters, words, sentences as compared to real news.

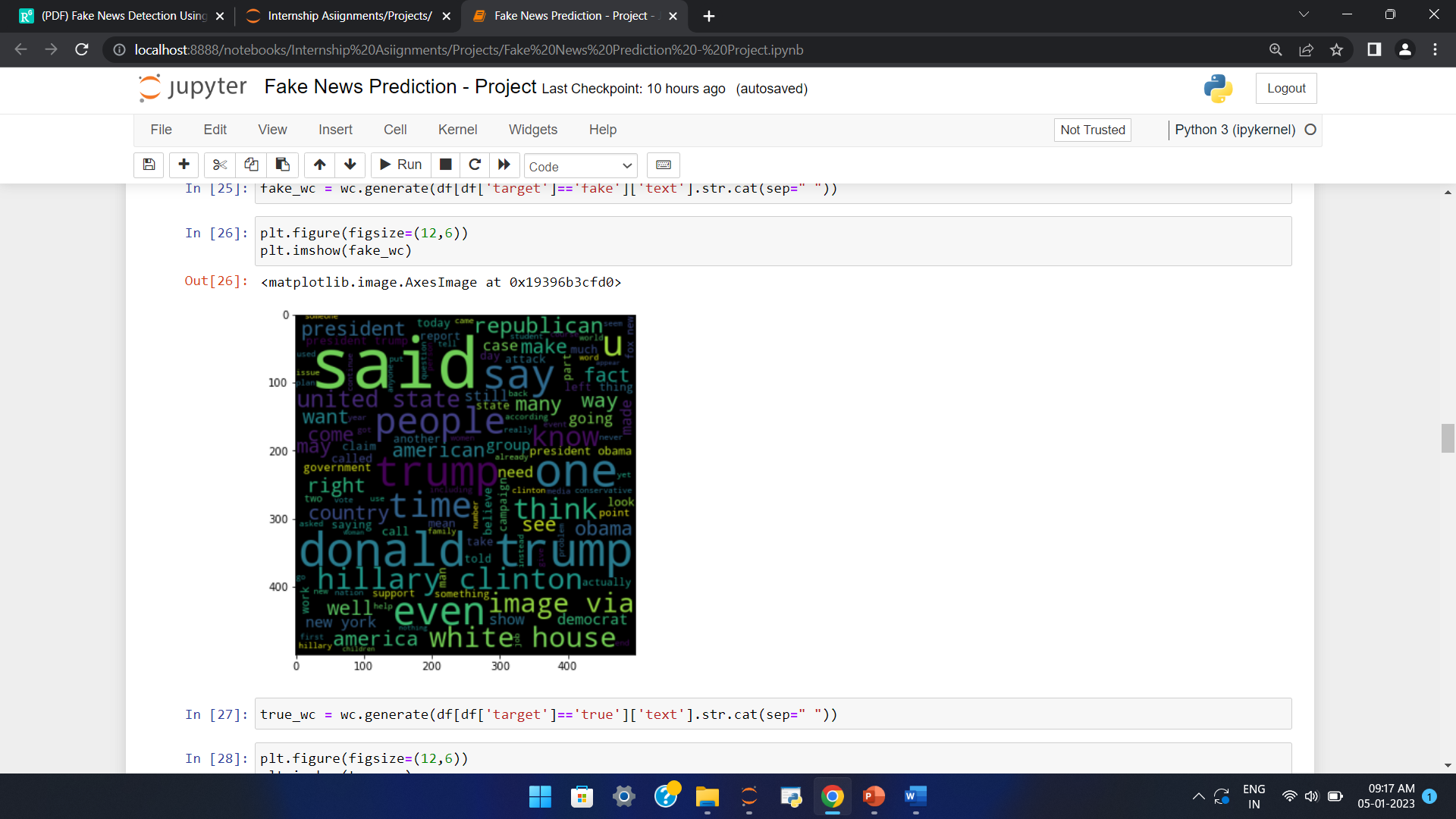


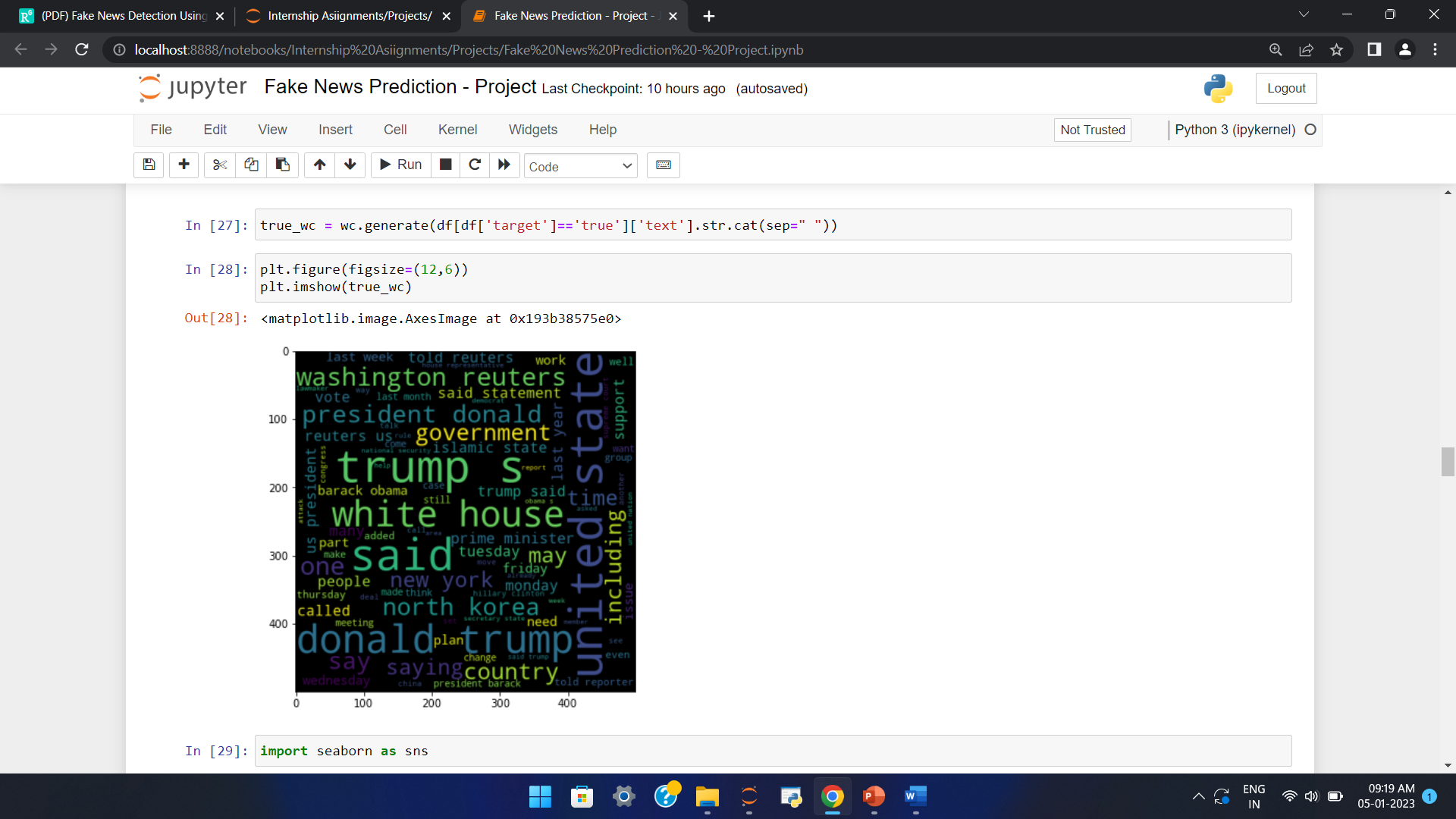


1. By using heatmap we have analysed correlation between the columns of dataset.



1. Used word cloud to observe most frequently used words in fake and real news.





* Interpretation of the Results

Random Forest Classifier is best fitted model with accuracy of 99.78% and with precision of 99.71%.