**Fake News Project Report**

* Submitted by Smriti Mathur

**Acknowledgement**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project. With the help of Google, Kaggle, Github, and my previous project, I was able to complete this project.

**Introduction**

**Context:** Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

**Content:** What's inside is more than just rows and columns. Make it easy for others to get started by describing how you acquired the data and what time period it represents, too.

**What is fake news?**

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas. For media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So it is necessary to detect fake news.

**Analytical Problem Facing**

* **Mathematical/ Analytical Modeling of the Problem**

Naive Bayes, Logistic Regression, Decision tree, and Random Forest models are build for the analysis.

* **Data Sources**

Data source from Google, data in CSV format

* **Data Preprocessing Done**

As part of our text cleaning process, we steam, lemmatize, remove stopwords, special symbols, and numbers, among other things. As soon as we have cleaned the data, we should feed it into a vectorizer that will convert it into numerical features.

* **Hardware and Software Requirements and Tools Used**

1. Hardware - Laptop, Keyboard, Mouse
2. Software – Jupiter Notebook
3. Tools - Scikit-learn, scipy.stats
4. Library – Using python library – numpy , pandas, matplotlib, and sns
5. Machine algorithm – Naïve Bayes, Logistic Regression, Decision Tree Classifier, Random Forest Classifier

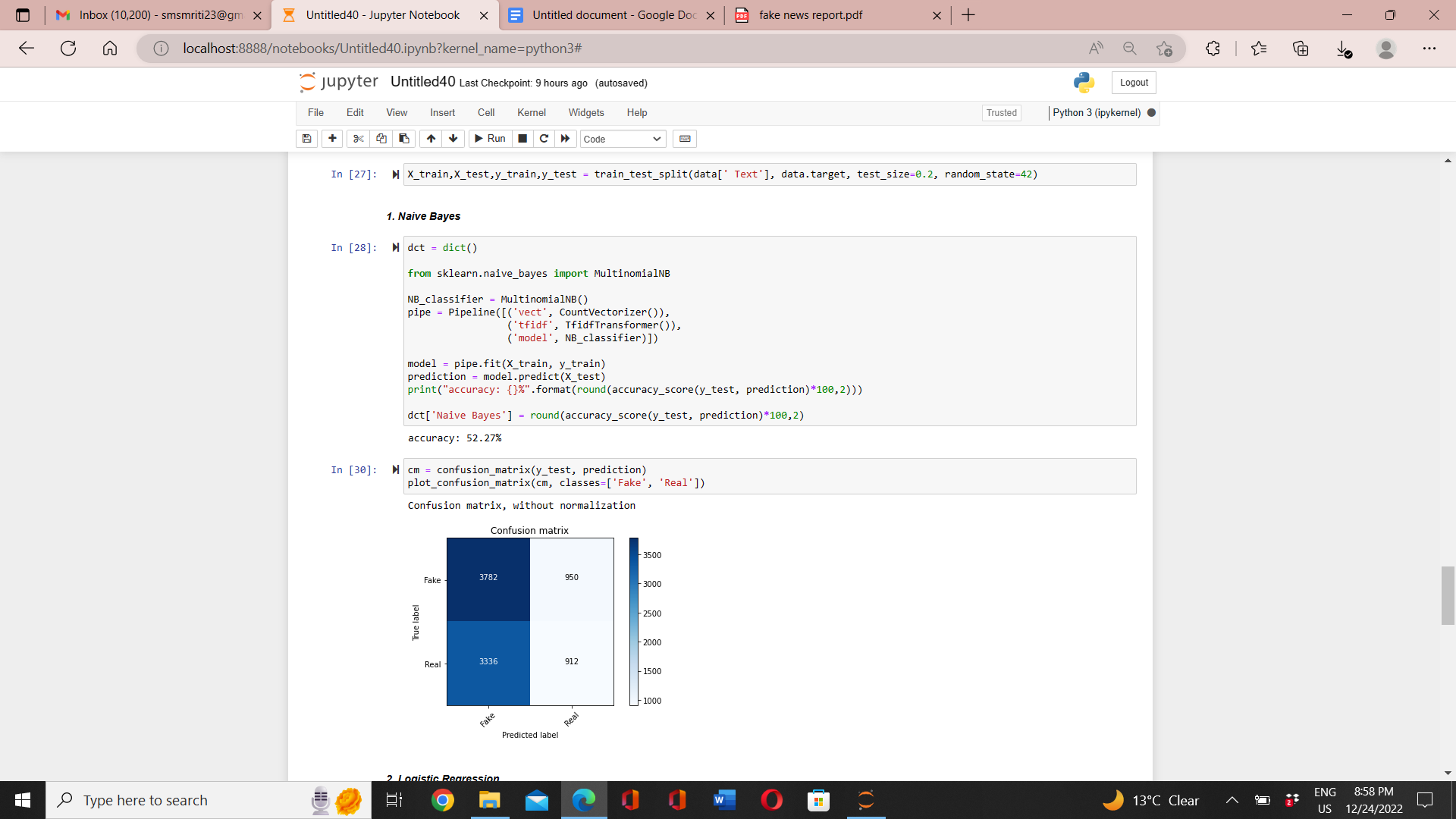
**Model/s Development and Evaluation**

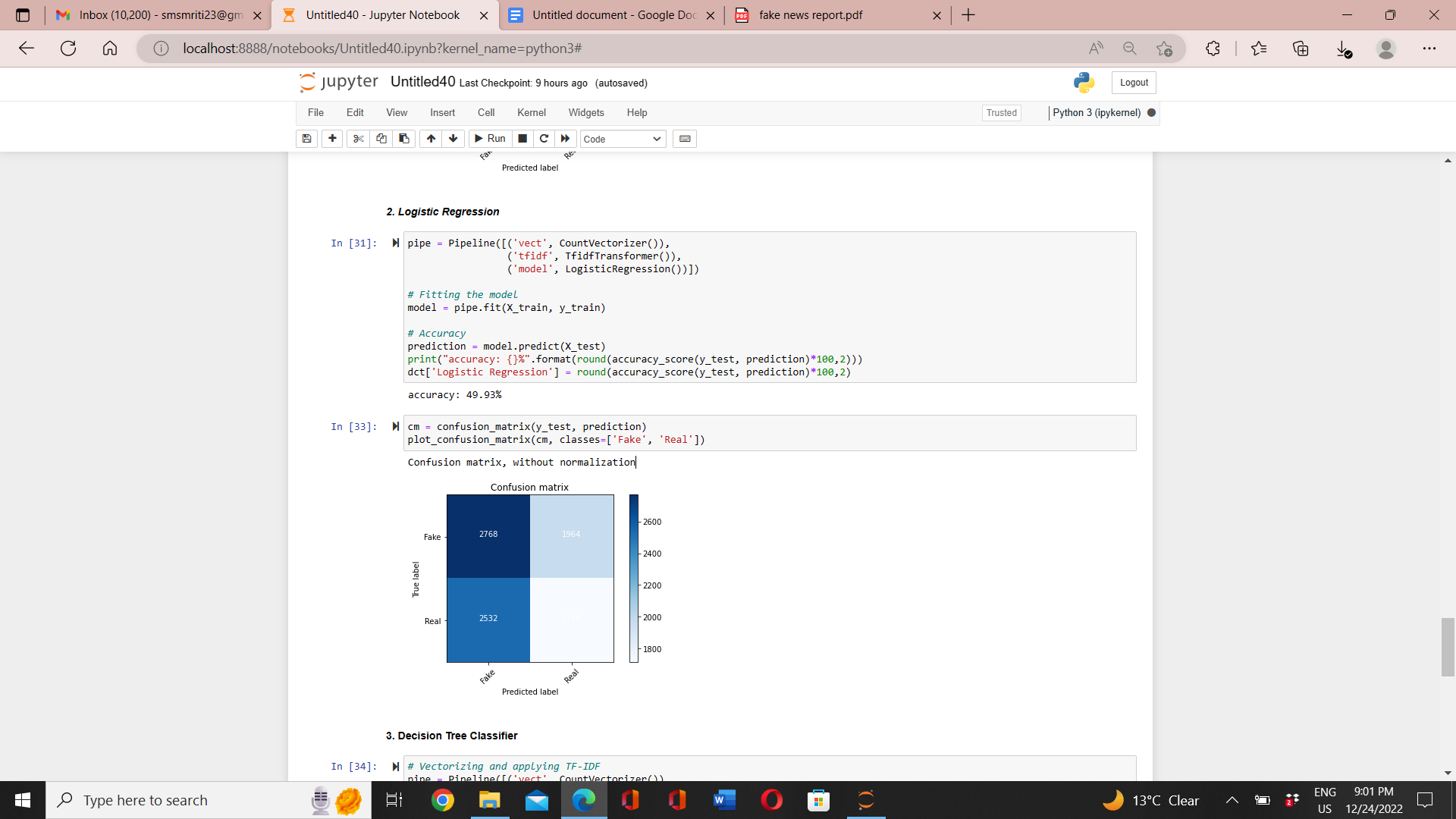
* **Testing of Identified Approaches (Algorithms)**

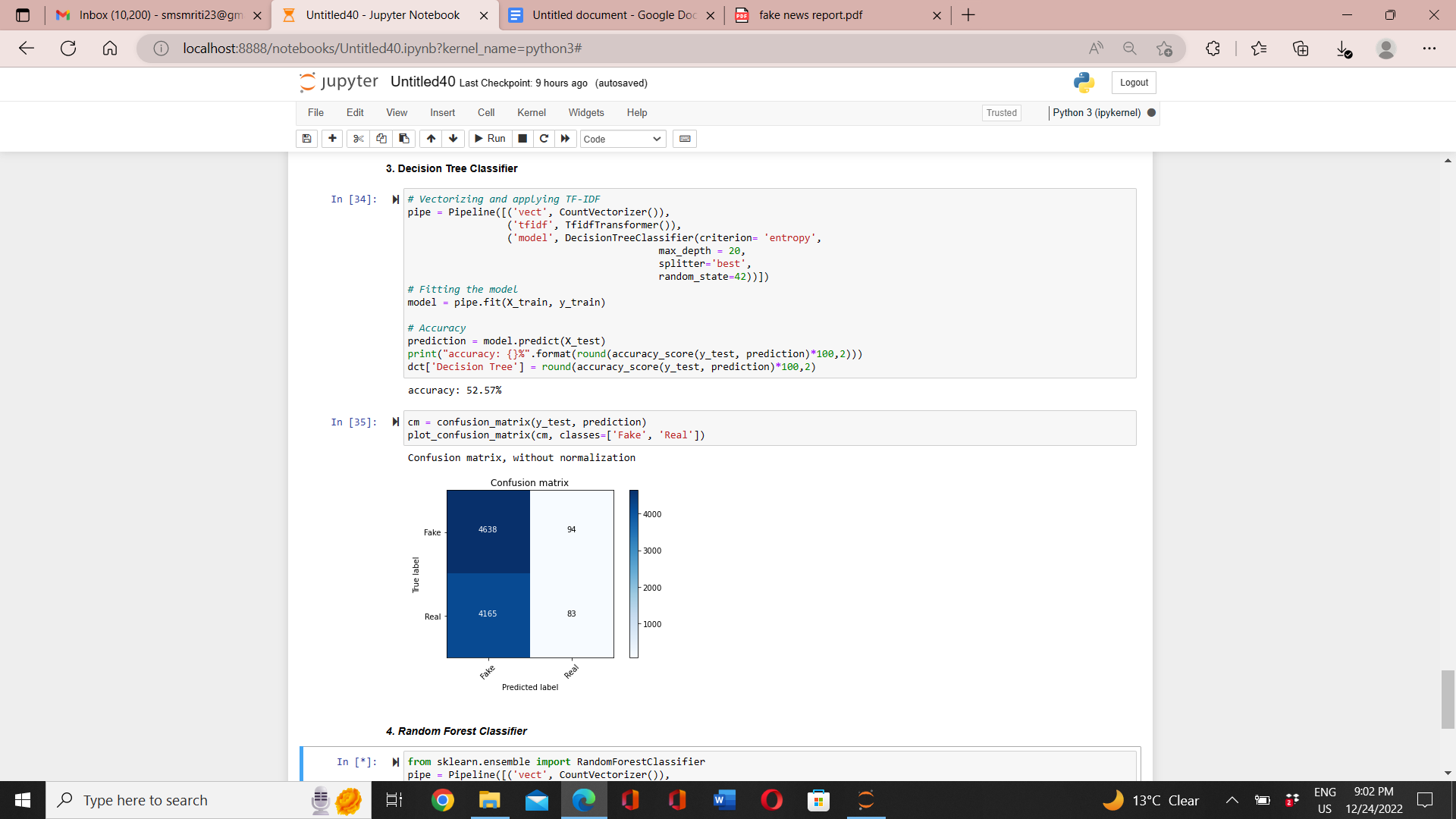
We used algorithms for training and testing:

1. Naïve bayes
2. Logistic regression
3. Decision tree
4. Random forest
5. Classification report
6. Confusion matrix
7. Decision tree has given the best accuracy score

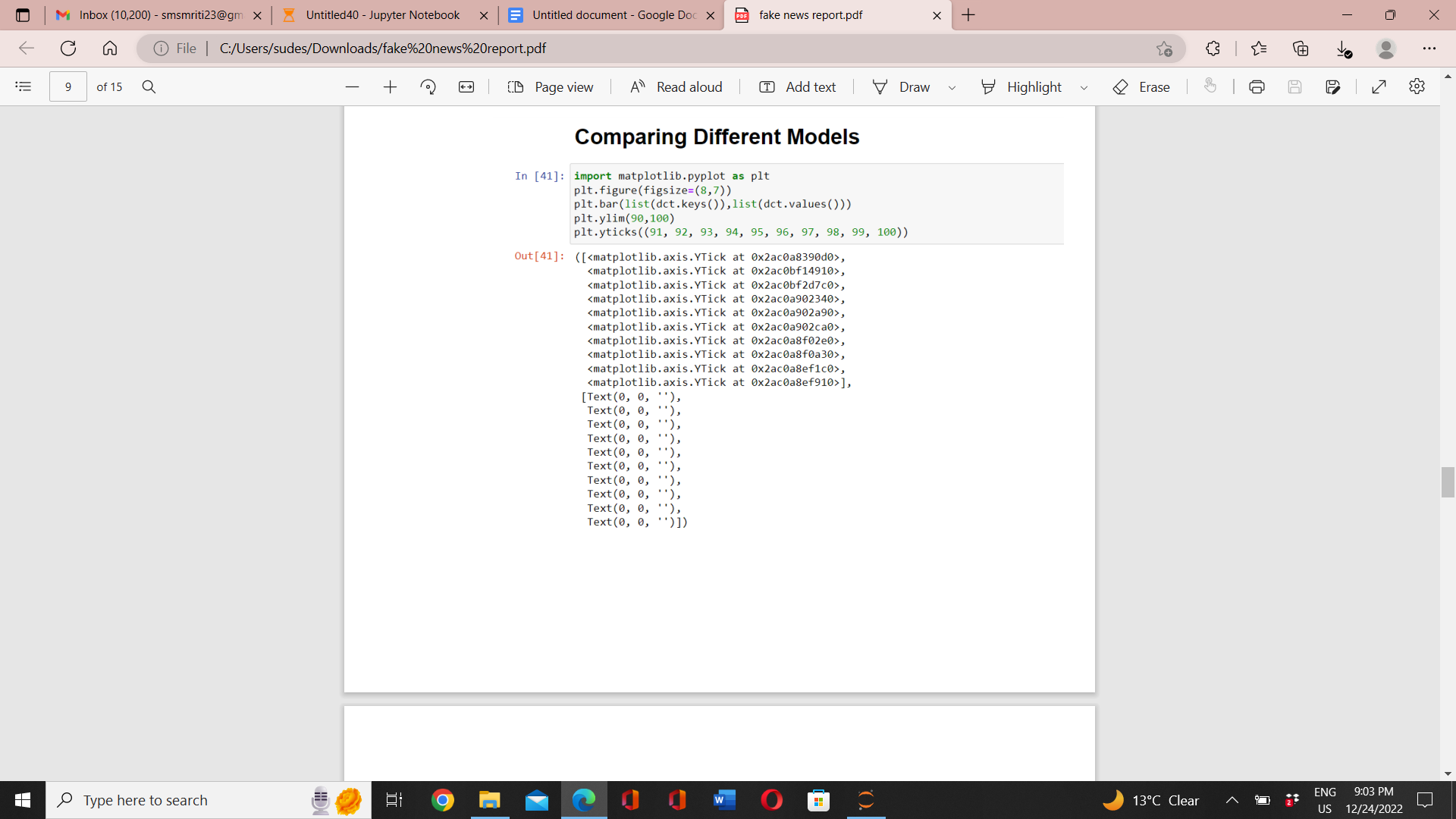
* **Run And Evaluate Selected Models**



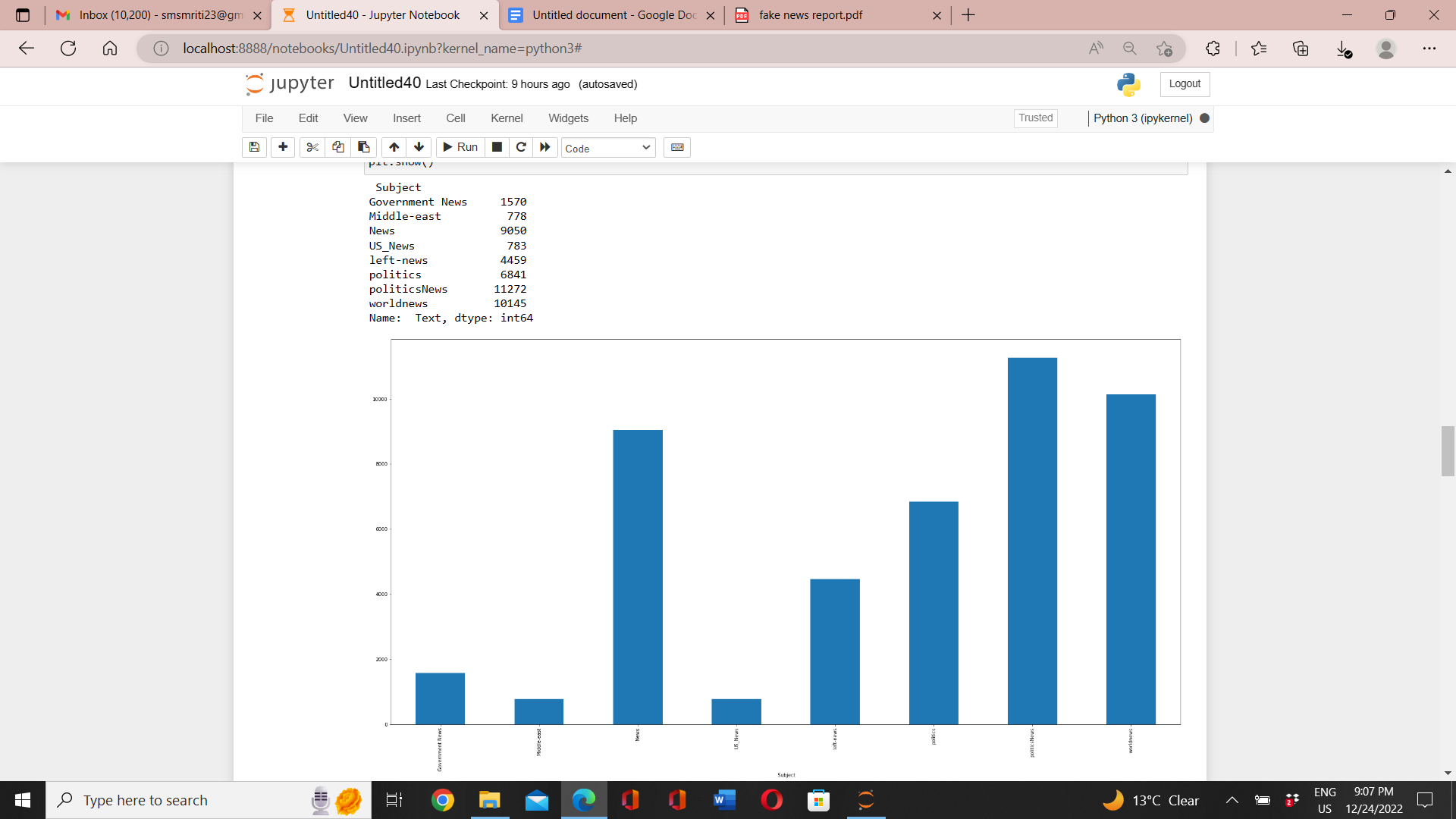


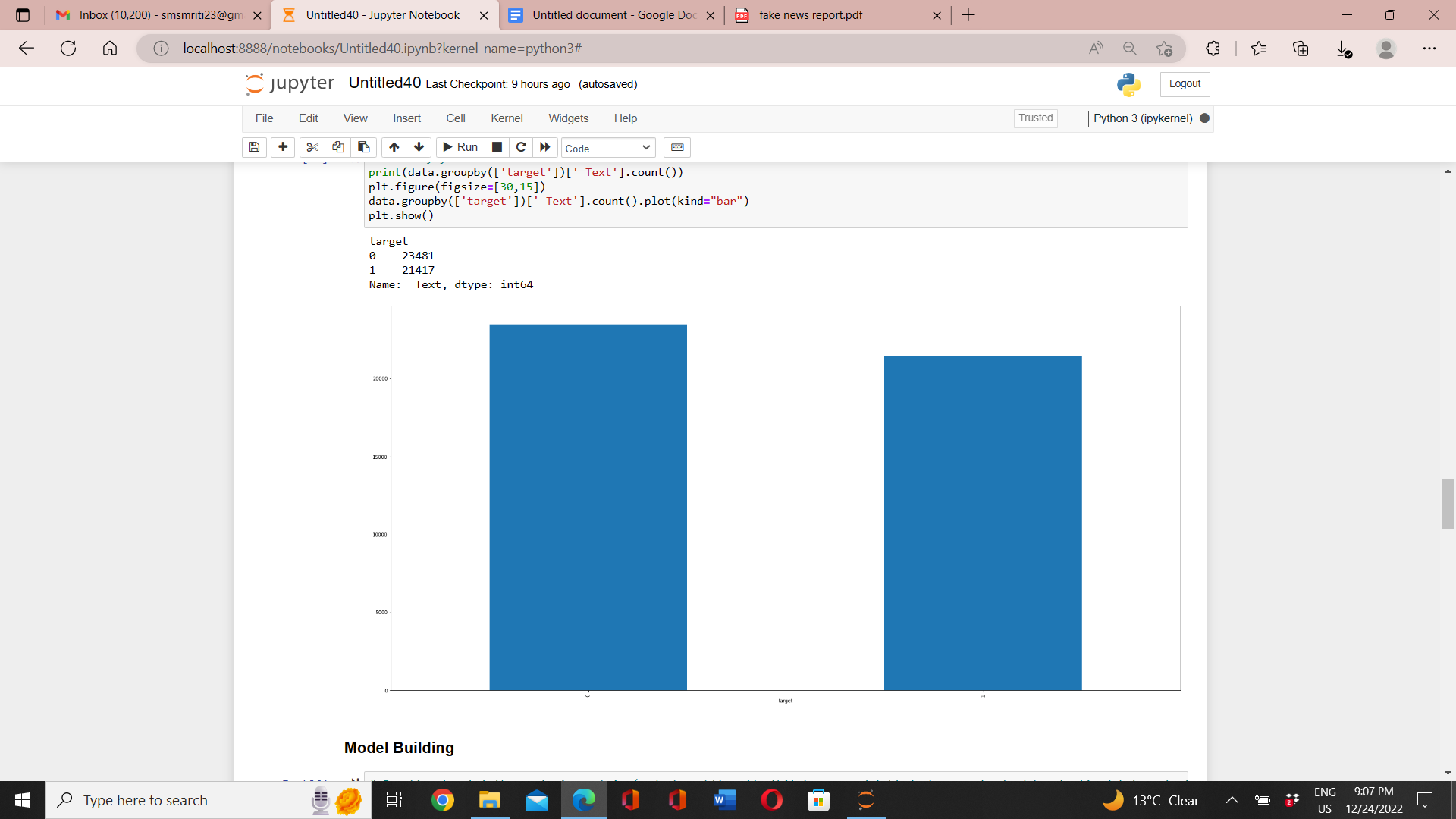
****

* **Comparing different models**



**Visualisation**





**Interpretation Of Results**

Describe how the visualisations, preprocessing, and modelling produced the results. As we noticed that the majority of articles on the politics news topic are fake, so we found the amount of fake news to be higher than that of real news. We see that 'trump' and 'said' words have been used several times in both fake and real news in this bar plot.

**Conclusion**

The model we developed does not belong specifically to a particular media, and in our dataset, all the data consists of news reports from various digital media, which means our model understands what constitutes fraud across any digital media.

Currently, we are cleaning up our text by removing stopwords, numbers, special symbols, etc.

A total of five machine learning algorithms have been used. It has been found that decision tree classifiers produce the best results around 99%.