

**PROJECT REPORT**

**E-MAIL SPAM DETECTION**

**(AI&ML)**

**By:** -

Jaineet Kaur (20CSU046)

Pankil Kamboj (20CSU077)

**INTRODUCTION**

We all are using gmail in our day to day life and spam mails have always been a problem for us.We’ve all been the recipient of spam E-mails before.

Spam mail or Junk mail is a type of E-mail sent to a massive number of users at one time, frequently containing cryptic messages, scams, or phishing content.

In order to deal with spam mails, we need to build a real-time email spam detector that can efficiently and correctly flag incoming spam mail.

Google and other email services also provide these features for flagging email spam but they all are still developing and need feedback often. There is a great scope in building email spam classifiers, as the private companies run their email servers and want them to be more secure because of the confidential data, in such cases email spam classifier solutions can be provided to such companies.

**LITERATURE SURVEY**

According to a report from Kaspersky lab, in 2015, the volume of spam emails being sent was reduced to a 12-year low. Spam email volume fell below 50% for the first time since 2003. In June 2015, the volume of spam emails went down to 49.7% and in July 2015 the figures were further reduced to 46.4% according to anti-virus software developer Symantec.

To solve the problem of spam mails big companies like Gmail, Yahoo, etc have used a combination of different machine learning (ML) techniques such as [Neural Networks](https://www.sciencedirect.com/topics/social-sciences/neural-network)**, Natural Language Processing** in their spam filters. These ML techniques can learn and identify spam mail and phishing messages by analyzing loads of such messages throughout a vast collection of computers.

The spam filtering methods that are currently being used by every other company are listed here:

**1)Standard Spam Filtering Method**

Email Spam filtering process works through a set of protocols to determine whether the message is spam or not.

**2)Client Side and Enterprise Level Spam Filtering Methods**

A client can send or receive an email by just one clicking through an ISP. Client-level spam filtering provides some frameworks for the individual client to secure mail transmission. Enterprise-level spam filtering is a process where provided frameworks are installed on the mail server which interacts with the MTA for classifying the received messages or mail to categorize the spam message on the network.

**3) Case Base Spam Filtering Method**

The most frequently used NLP techniques are found to be **TF-IDF and word embeddings**. Furthermore, the most commonly used dataset for benchmarking phishing email detection methods is the Nazario phishing corpus. Also, Python is the most commonly used one for phishing email detection.

Social networking experts estimate that 40% of social network accounts are used for spam. Today, learning-based classiﬁers are commonly used for spam detection. In learning-based classification, the detection process assumes that spam emails have a specific set of features that differentiate them from legitimate emails. Many factors increase the complexity of the identification process of spam in learning-based models. These factors include spam subjectivity, idea drift, language problems, overhead processing, and text latency.

**Referenced Scholar articles:**

# 1. Machine learning for email spam filtering: review, approaches and open research problems.

Author links open overlay panel[Emmanuel GbengaDadaa](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)[Joseph StephenBassia](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)[HarunaChiromab](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)[Shafi'i MuhammadAbdulhamidc](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)[Adebayo OlusolaAdetunmbid](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)[Opeyemi EmmanuelAjibuwae](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)

**2. "Classification of Phishing Email Using Random Forest Machine Learning Technique",**

**3. Comparative Analysis of Classification Algorithms for Email Spam Detection**

# 4. Applicability of machine learning in spam and phishing email filtering: review and approaches

[Tushaar Gangavarapu](https://link.springer.com/article/10.1007/s10462-020-09814-9#auth-Tushaar-Gangavarapu), [C. D. Jaidhar](https://link.springer.com/article/10.1007/s10462-020-09814-9#auth-C__D_-Jaidhar) & [Bhabesh Chanduka](https://link.springer.com/article/10.1007/s10462-020-09814-9#auth-Bhabesh-Chanduka)

**5. Ham and Spam E-Mails Classification Using Machine Learning Techniques**

**M. Bassiouni,M. Ali &E. A. El-Dahshan**

**6. A systematic literature review on phishing email detection using natural language processing techniques**

S Salloum, T Gaber, S Vadera, K Sharan

7. **Spam Mail Scanning Using Machine Learning Algorithm**

Asma Bibi1, Rasia Latif1, Samina Khalid1, Waqas Ahmed2, Raja Ahtsham Shabir1, Tehmina Shahryar

### 8. [A survey of existing e-mail spam filtering methods considering machine learning techniques](https://www.researchgate.net/profile/Hanif-Bhuiyan-2/publication/332865507_A_Survey_of_Existing_E-Mail_Spam_Filtering_Methods_Considering_Machine_Learning_Techniques/links/5f12c1704585151299a4b7fb/A-Survey-of-Existing-E-Mail-Spam-Filtering-Methods-Considering-Machine-Learning-Techniques.pdf)

[H Bhuiyan](https://scholar.google.co.in/citations?user=gUdTzkcAAAAJ&hl=en&oi=sra), [A Ashiquzzaman](https://scholar.google.co.in/citations?user=Jn1xaIMAAAAJ&hl=en&oi=sra),

### 9. [Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges](https://www.hindawi.com/journals/scn/2022/1862888/)

N Ahmed, [R Amin](https://scholar.google.co.in/citations?user=qq6WosgAAAAJ&hl=en&oi=sra), [H Aldabbas](https://scholar.google.co.in/citations?user=n_43xdgAAAAJ&hl=en&oi=sra), [D Koundal](https://scholar.google.co.in/citations?user=1mRY0ZkAAAAJ&hl=en&oi=sra)

### 10. [Phishing email detection using natural language processing techniques: a literature survey](https://www.sciencedirect.com/science/article/pii/S1877050921011741)

[S Salloum](https://scholar.google.co.in/citations?user=MsZan18AAAAJ&hl=en&oi=sra), [T Gaber](https://scholar.google.co.in/citations?user=a9glOYUAAAAJ&hl=en&oi=sra), [S Vadera](https://scholar.google.co.in/citations?user=BQSRVTQAAAAJ&hl=en&oi=sra)

**Proposed Technique**

We have incorporated different techniques to design our email spam classifier.

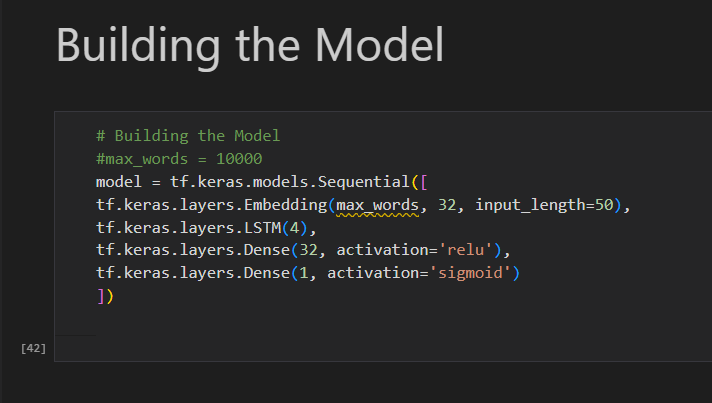
**NLP**

* This NLP tutorial will use Python NLTK library. NLTK is a popular Python library which is used for NLP. So what is NLP? And what are the benefits of learning NLP? What is NLP? What is NLP? Simply and in short, natural language processing (NLP) is about developing applications and services that can understand human languages.
* Natural language toolkit (NLTK) is the most popular library for natural language processing (NLP) which is written in Python and has a big community behind it. NLTK also is very easy to learn; it’s the easiest natural language processing (NLP) library that you’ll use. In this NLP Tutorial, we will use Python NLTK library.

**Recurrent neural networks (RNNs)**

RNNs are the state of the art algorithm for sequential data and are used by Apple’s Siri and Google’s voice search. It is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data. RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory.

In sequence prediction challenges, Long Short Term Memory (LSTM) networks are a type of Recurrent Neural Network that can learn order dependence. The output of the previous step is used as input in the current step in RNN. Hochreiter & Schmidhuber created the LSTM. It addressed the issue of RNN long-term dependency, in which the RNN is unable to predict words stored in long-term memory but can make more accurate predictions based on current data.



**LSTM(Long Short Term Memory)**

The LSTM Network model stands for Long Short Term Memory networks. These are a special kind of Neural Networks which are generally capable of understanding long term dependencies. LSTM model was generally designed to prevent the problems of long term dependencies which they generally do in a very good manner.

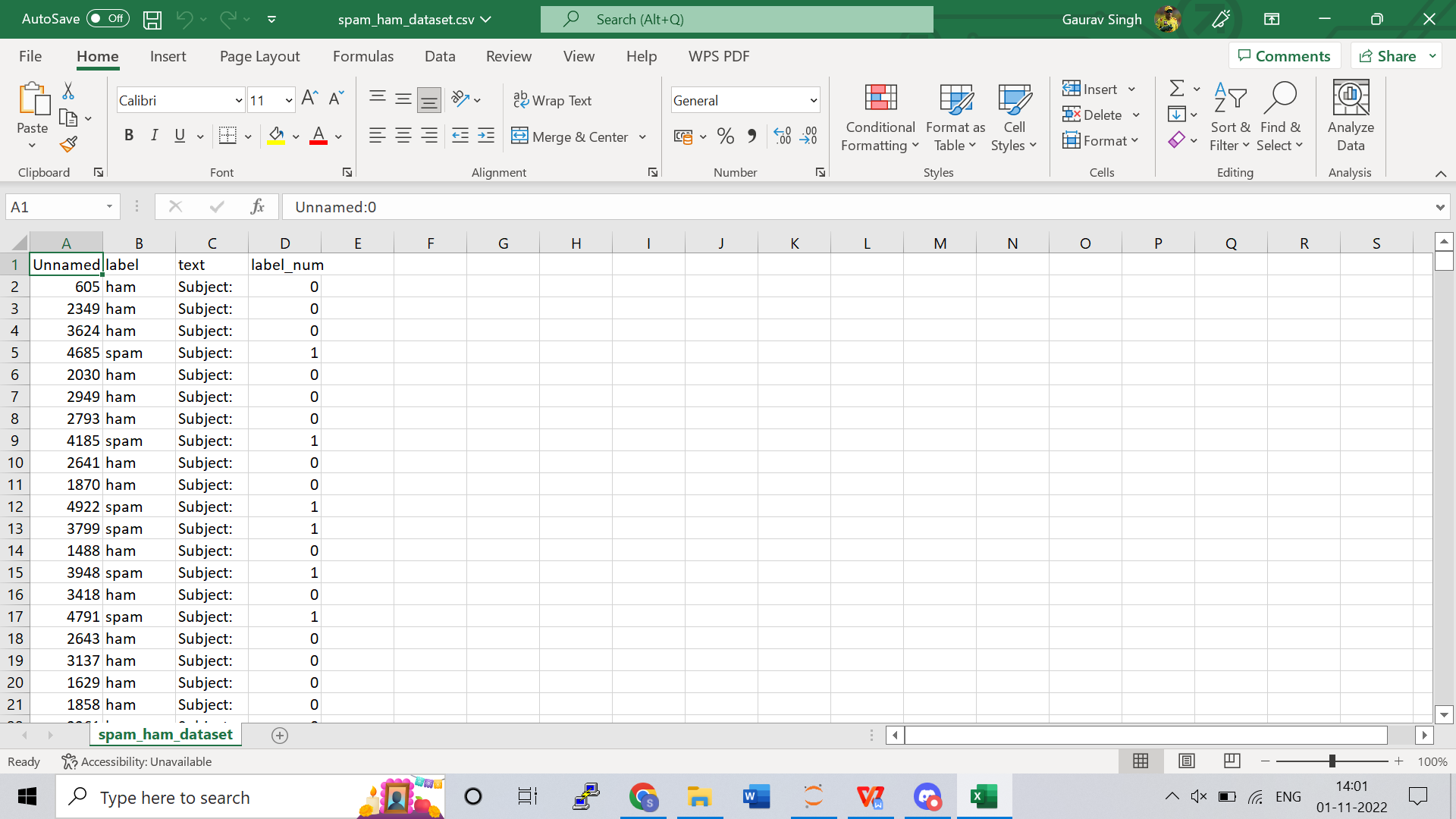
**Experimental Results**

**System configuration:**

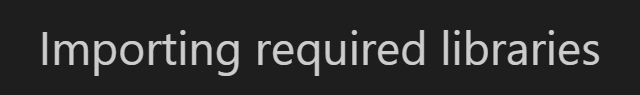
Visual Studio Code: Visual Studio Code is **a streamlined code editor with support for development operations like debugging, task running, and version control**. It aims to provide just the tools a developer needs for a quick code-build-debug cycle and leaves more complex workflows to fuller featured IDEs, such as Visual Studio IDEIt offers a simple, streamlined, document-centric experience.

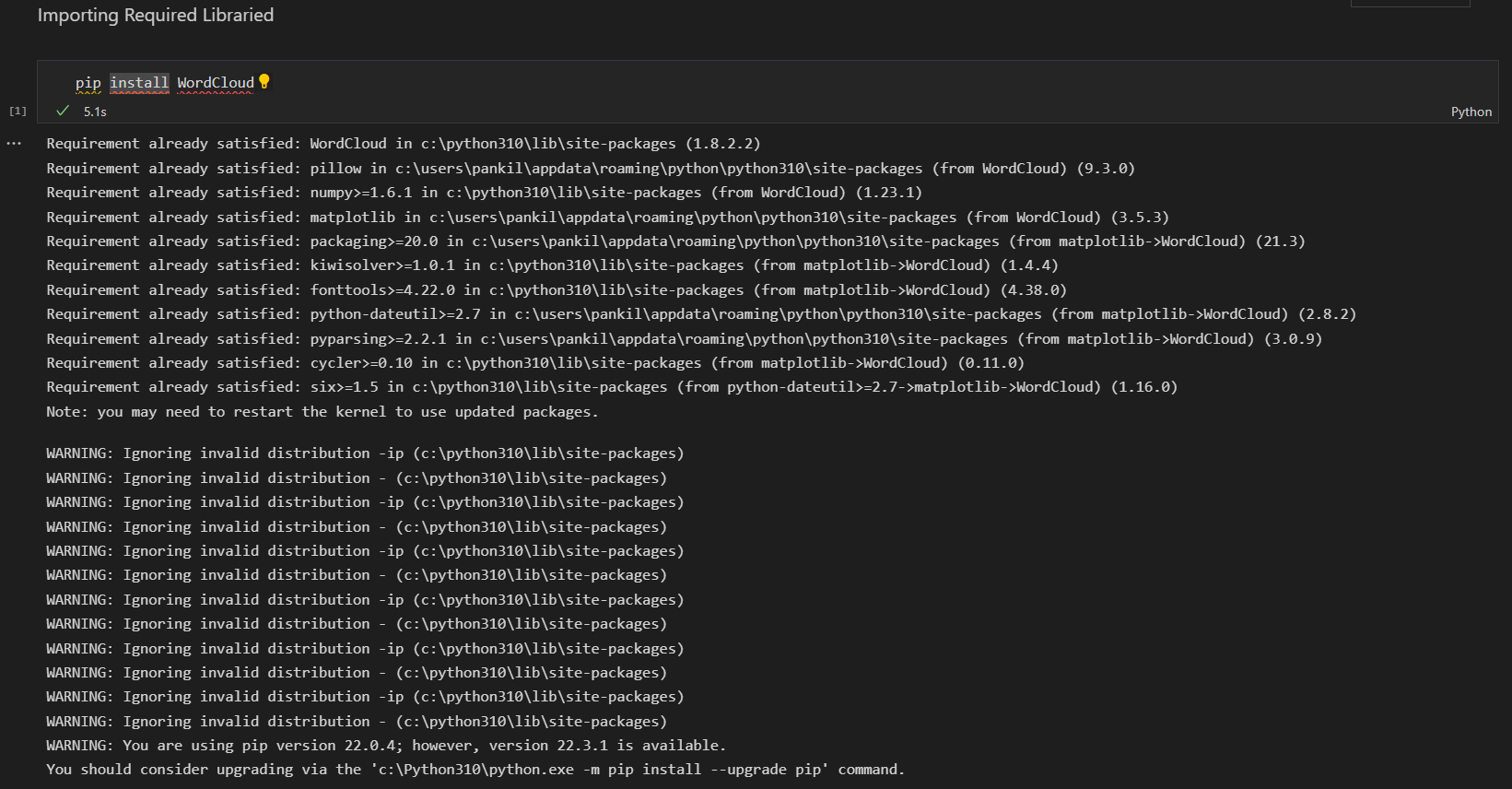
Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.

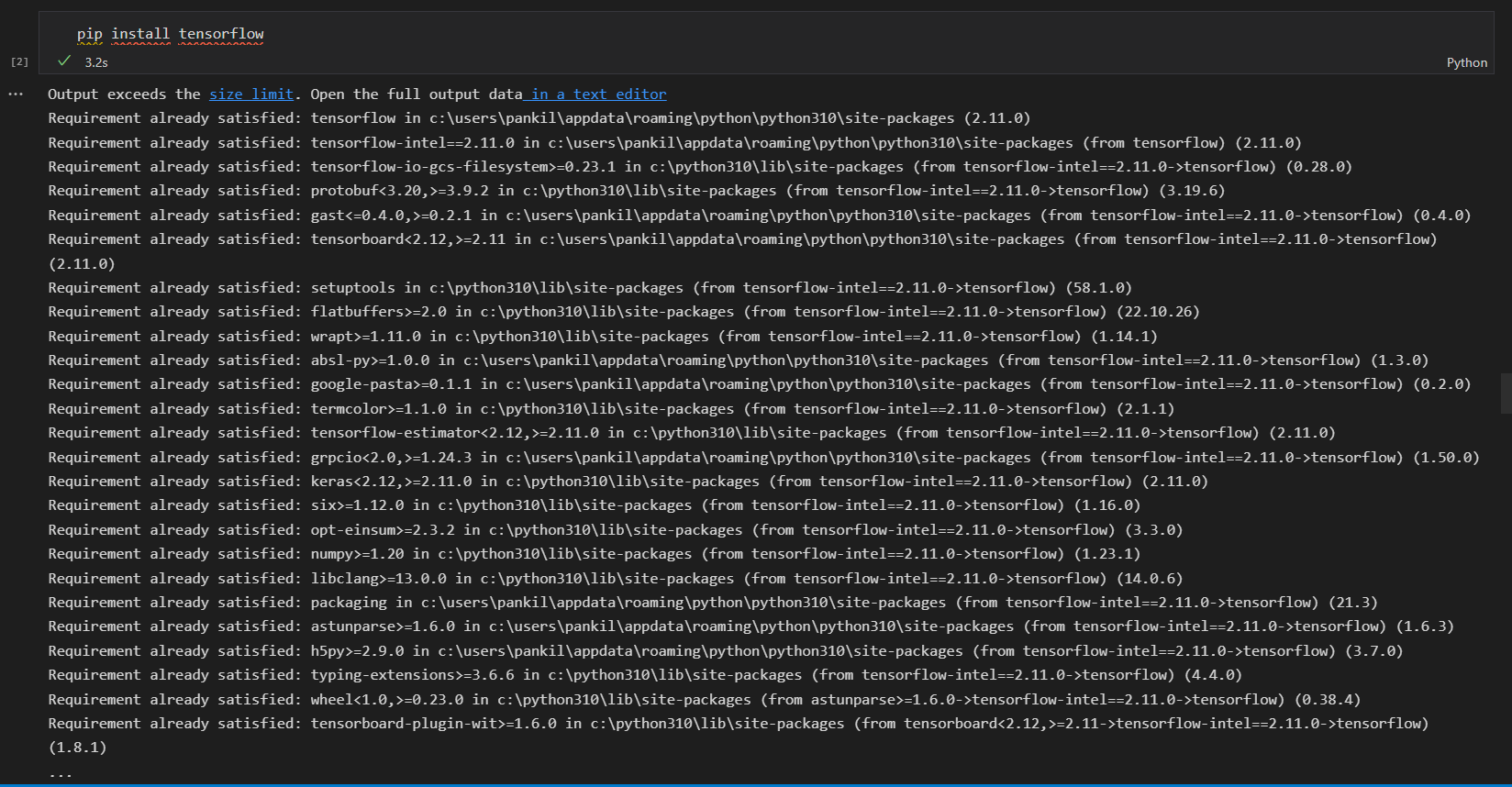
**Dataset Details:**

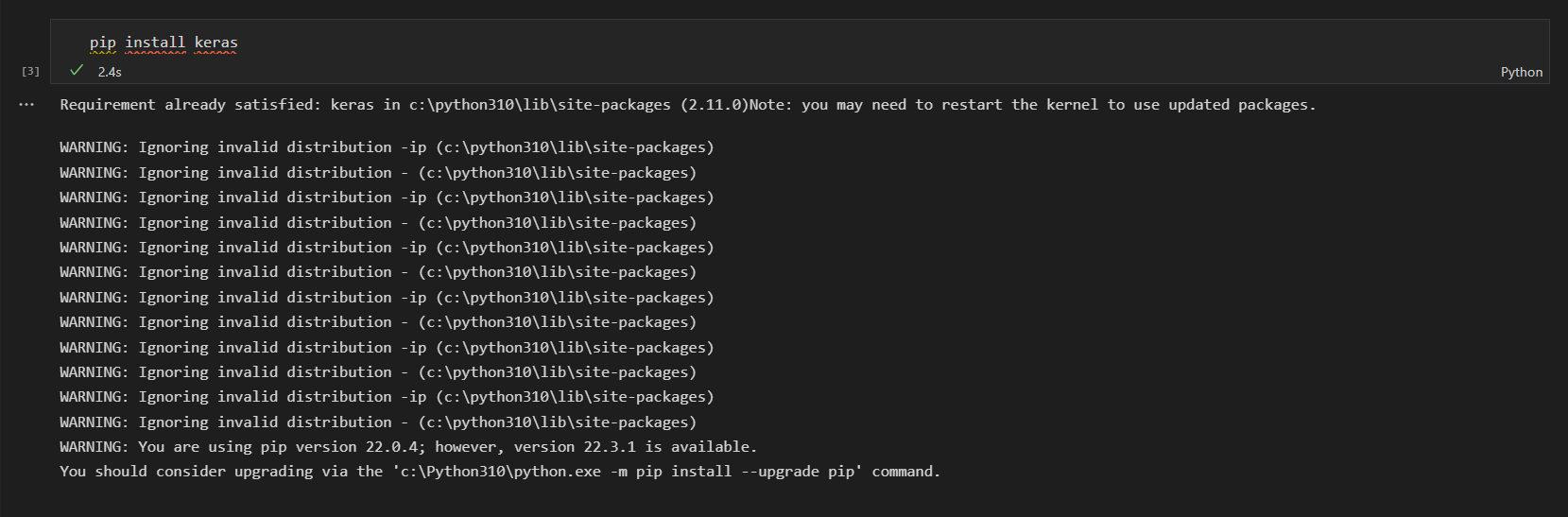


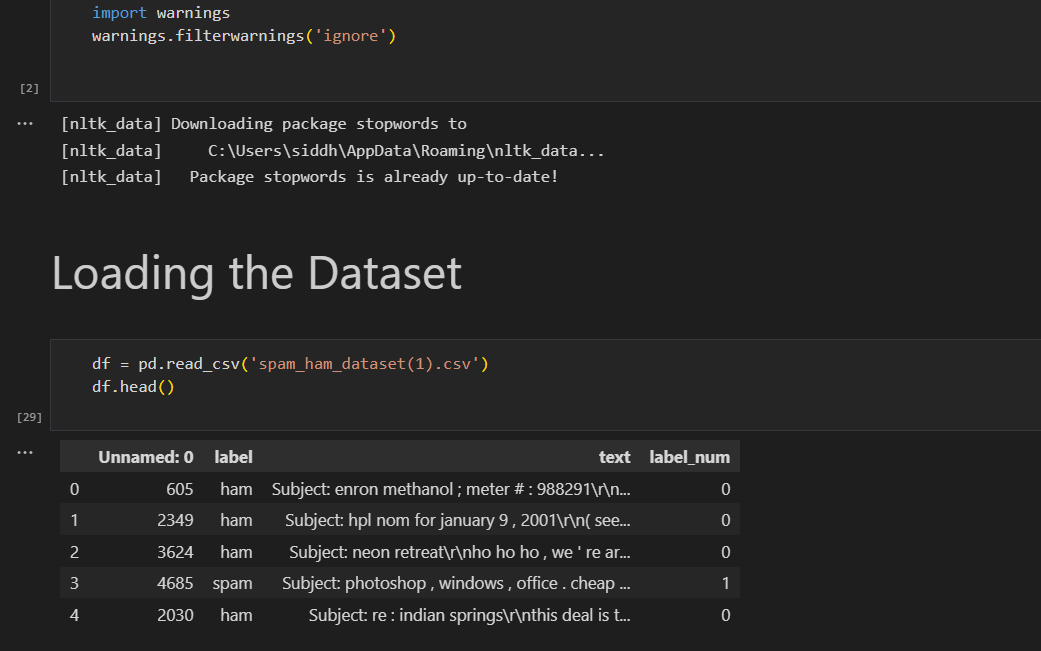
**Screenshots:**

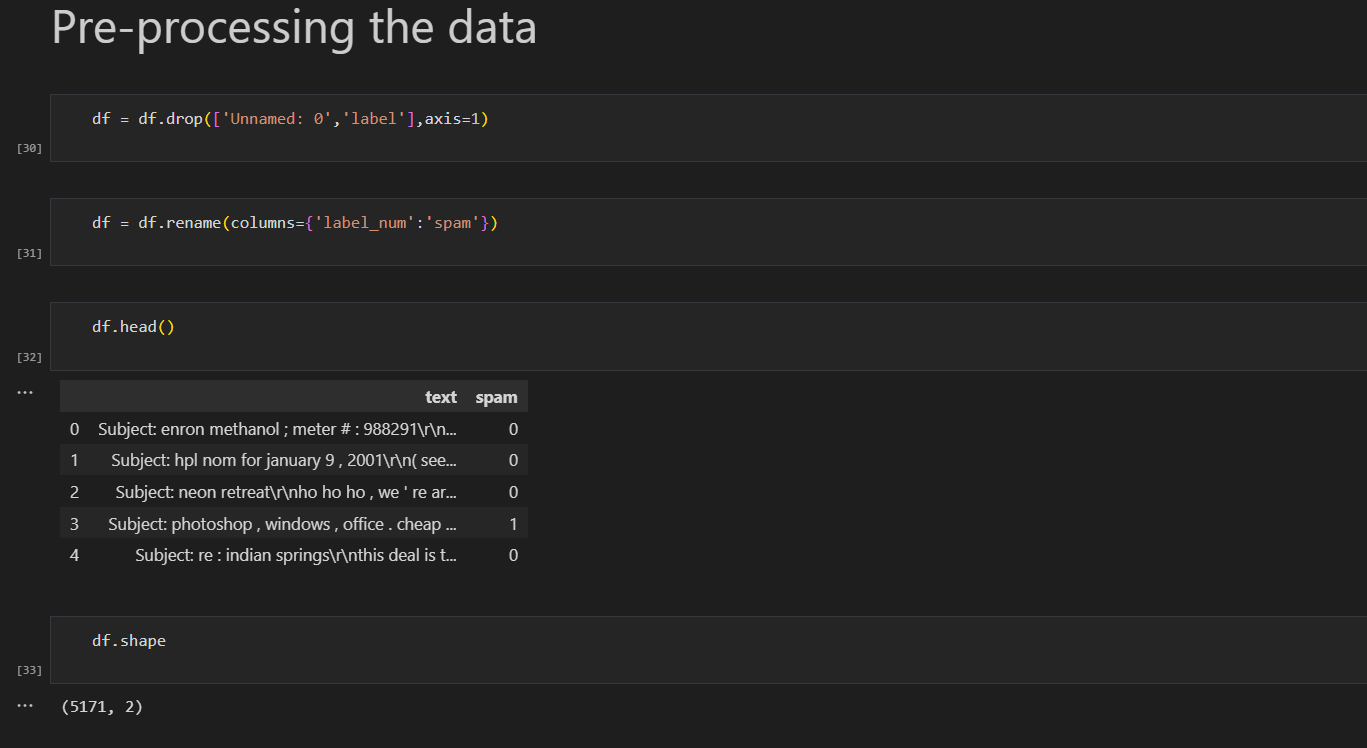
****

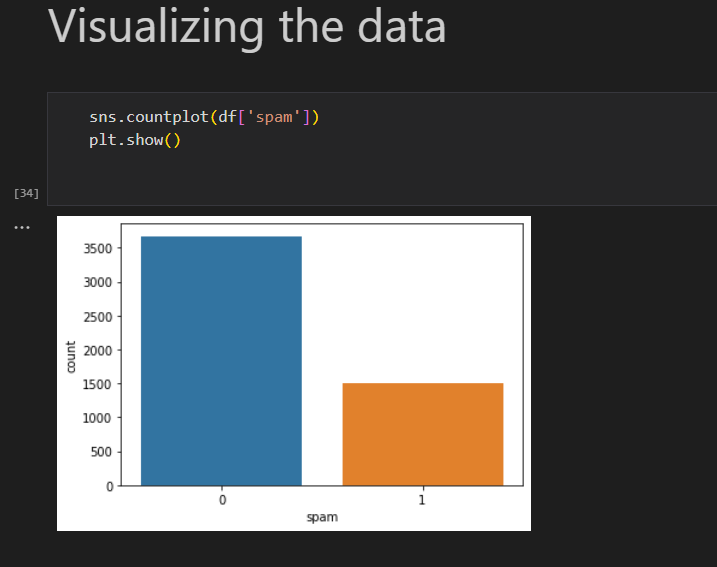


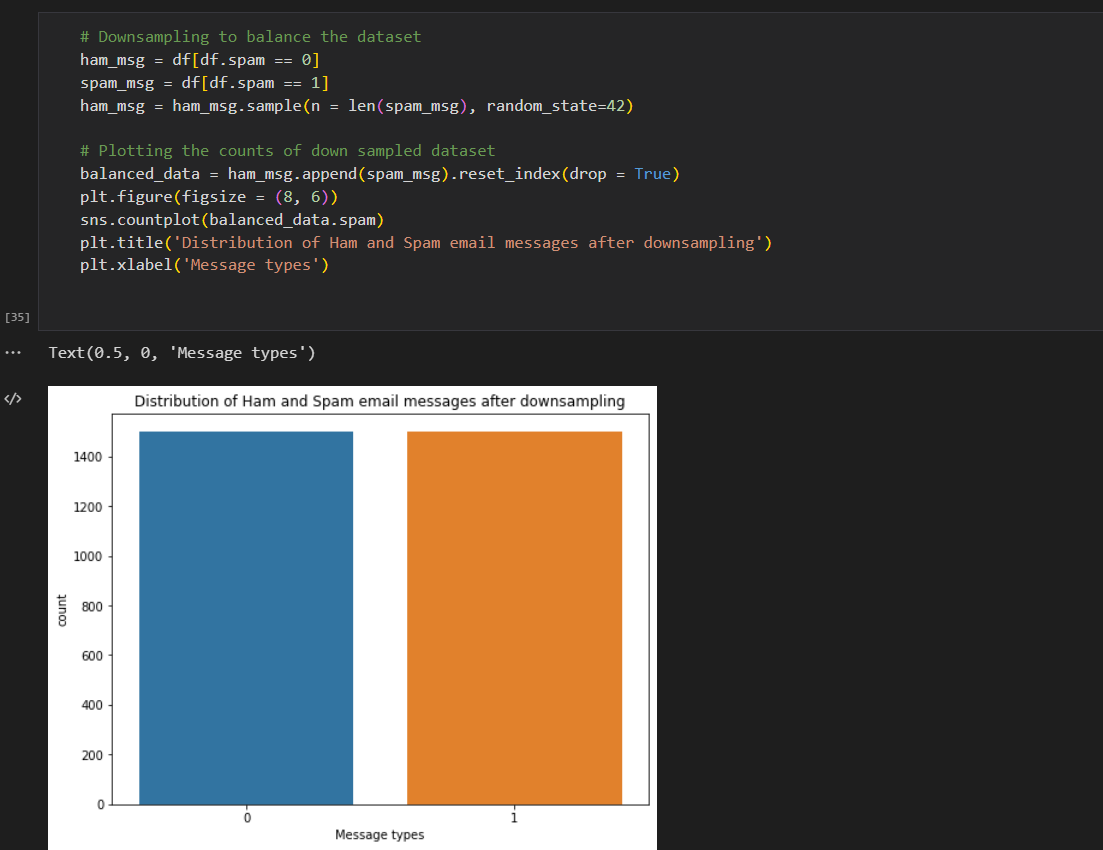


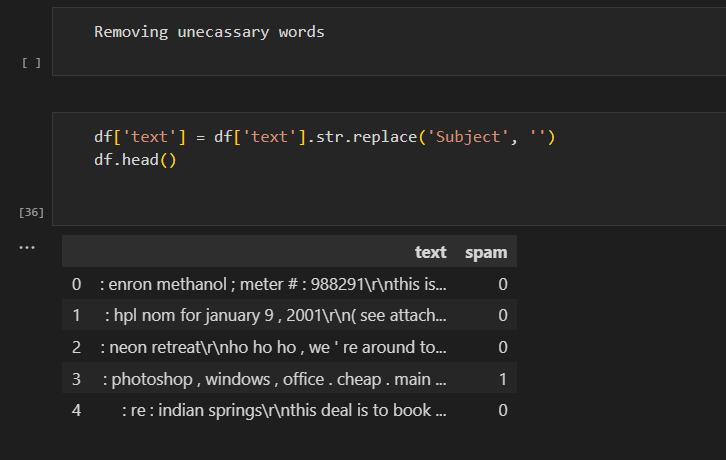


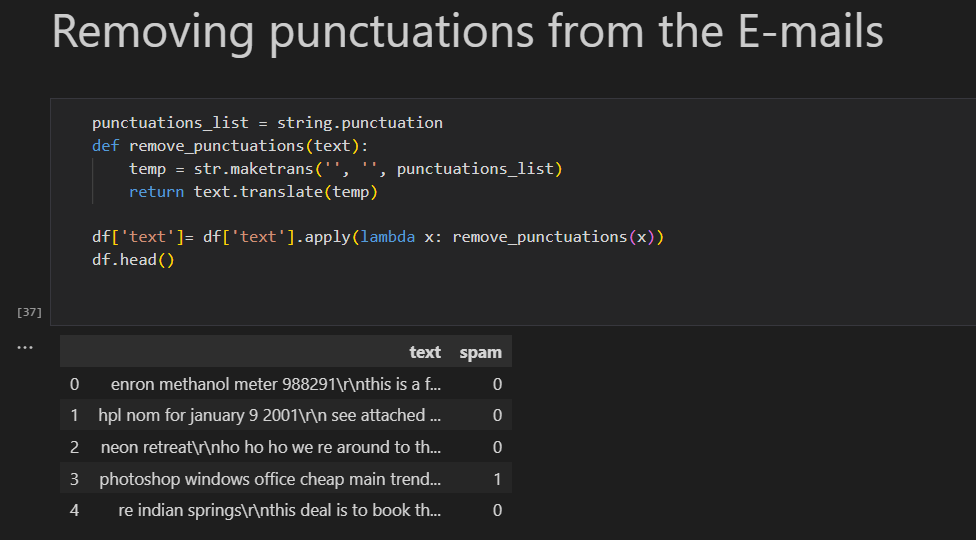


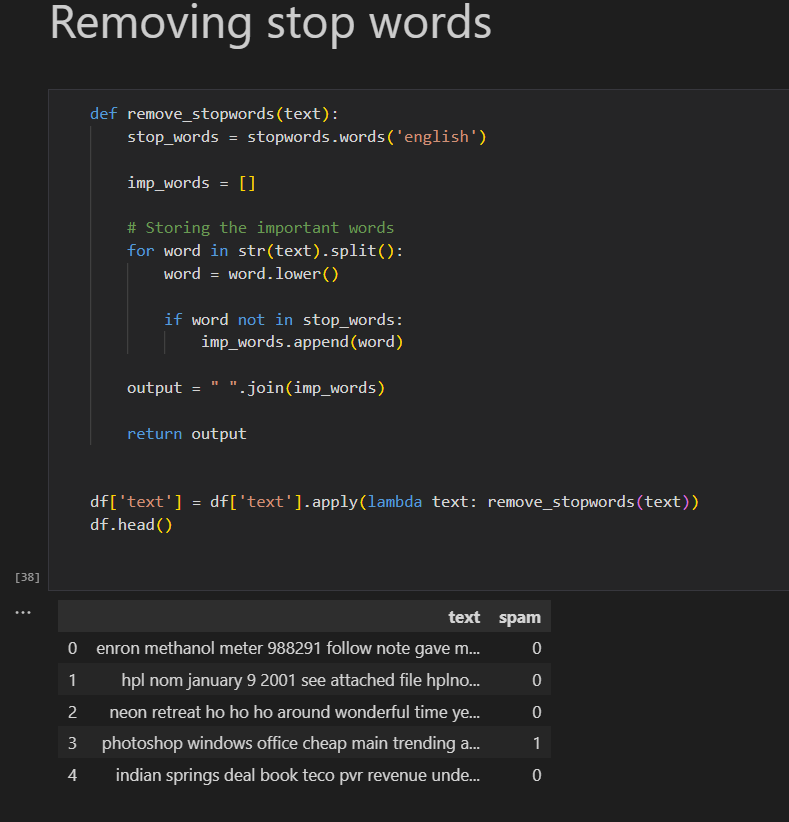


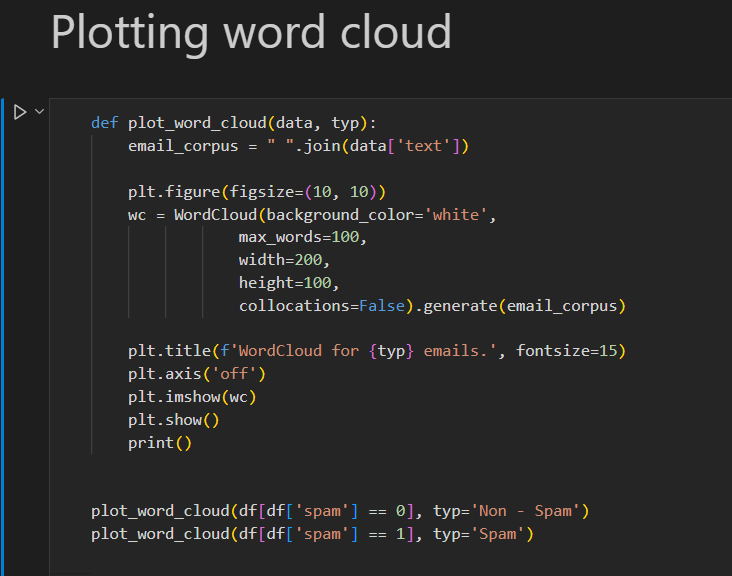


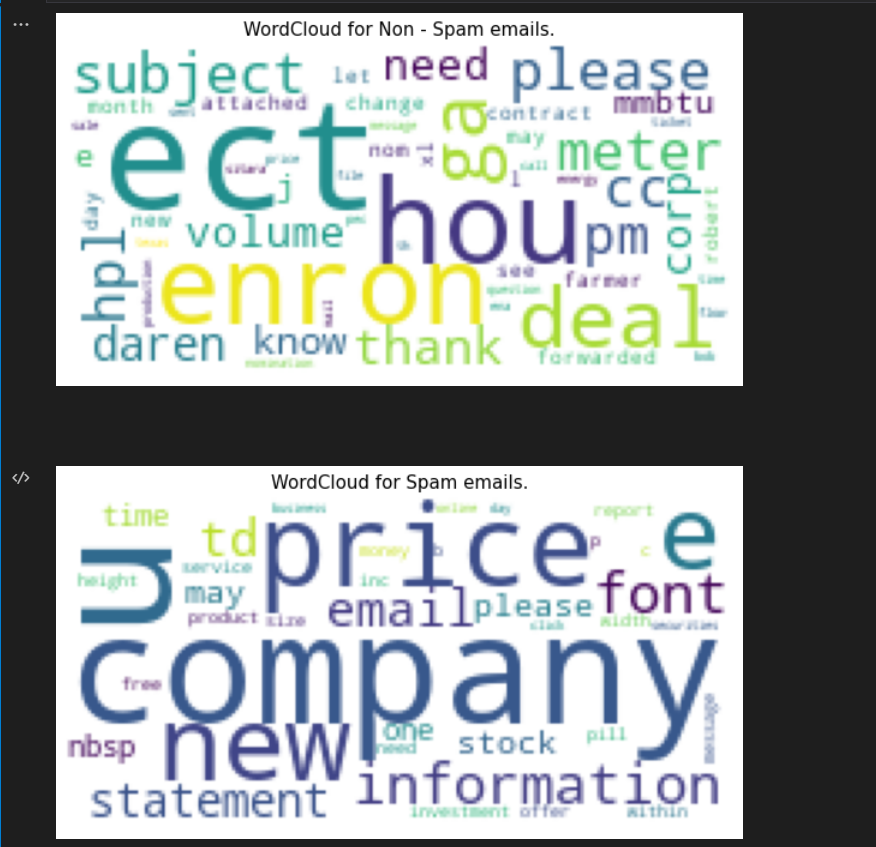


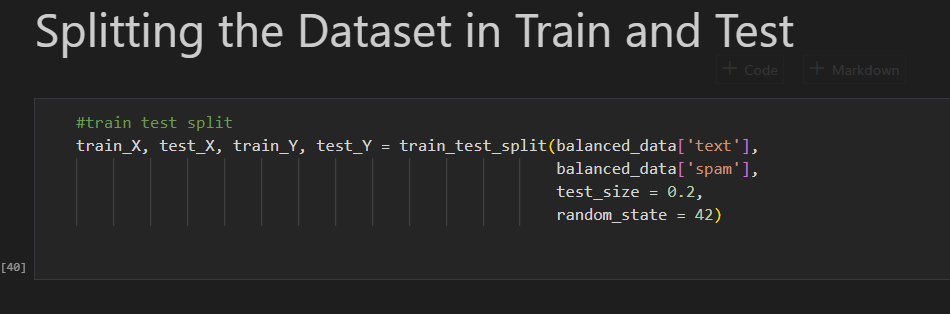


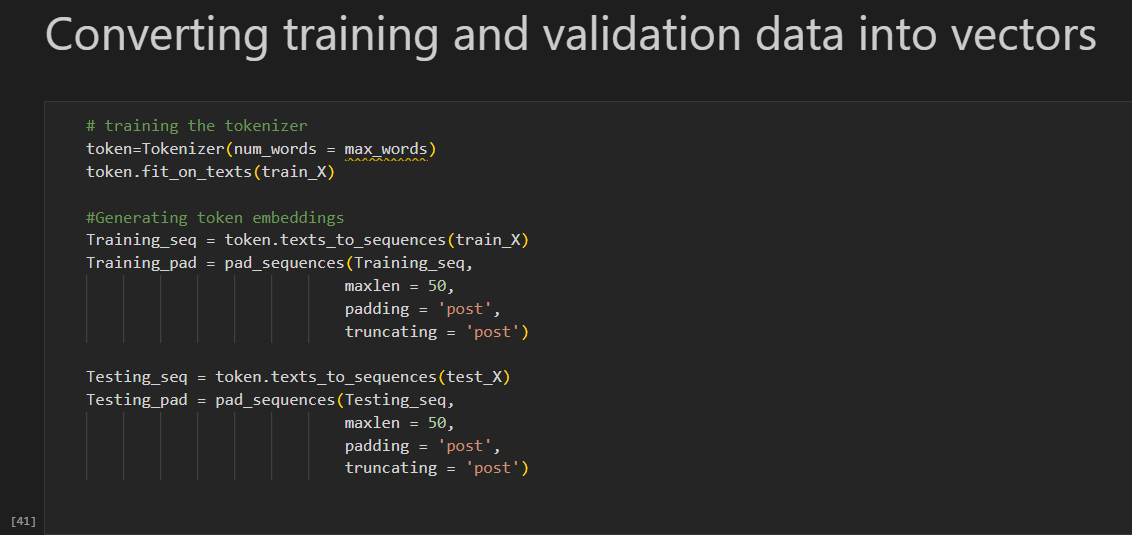


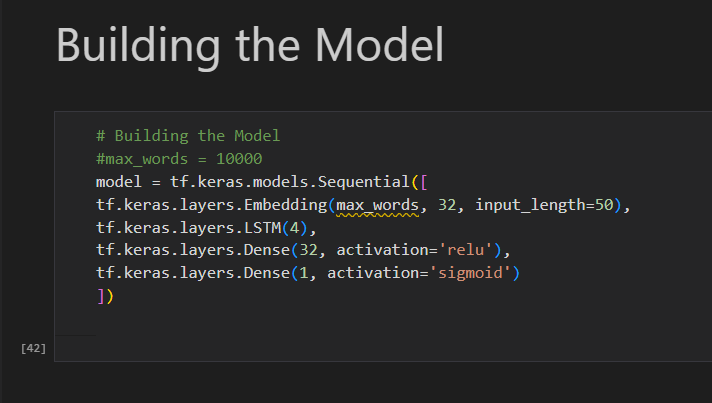


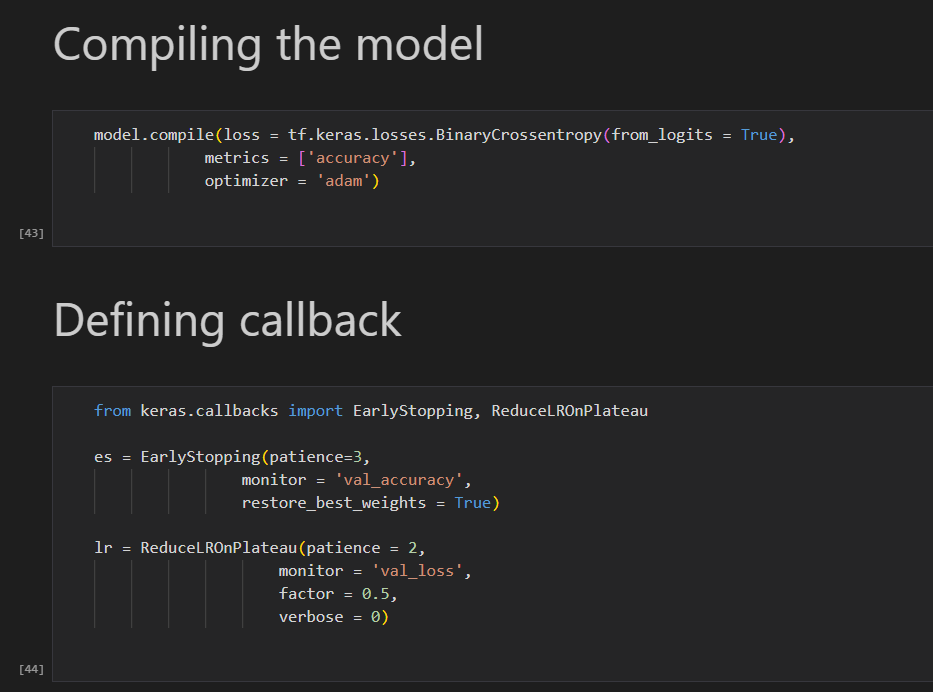


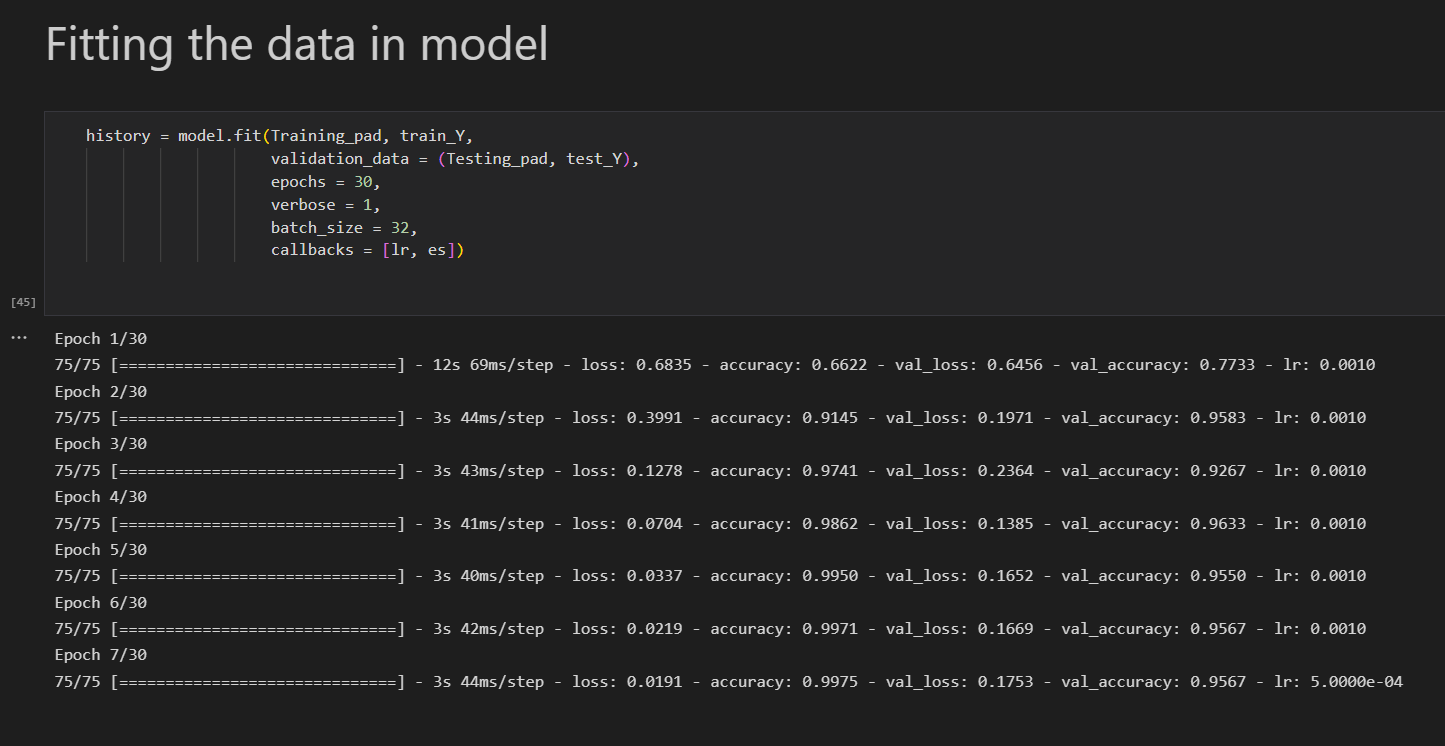


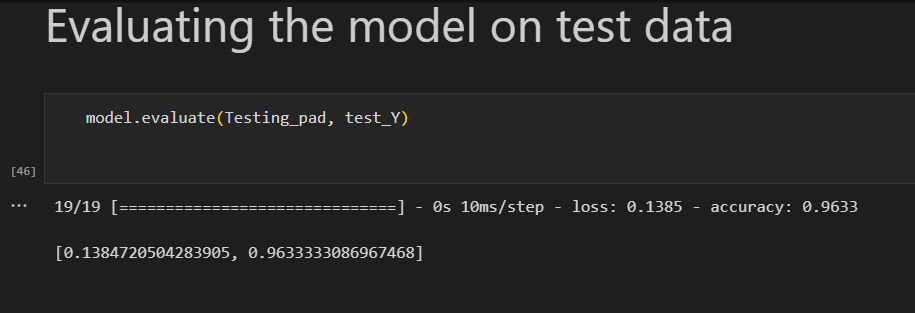


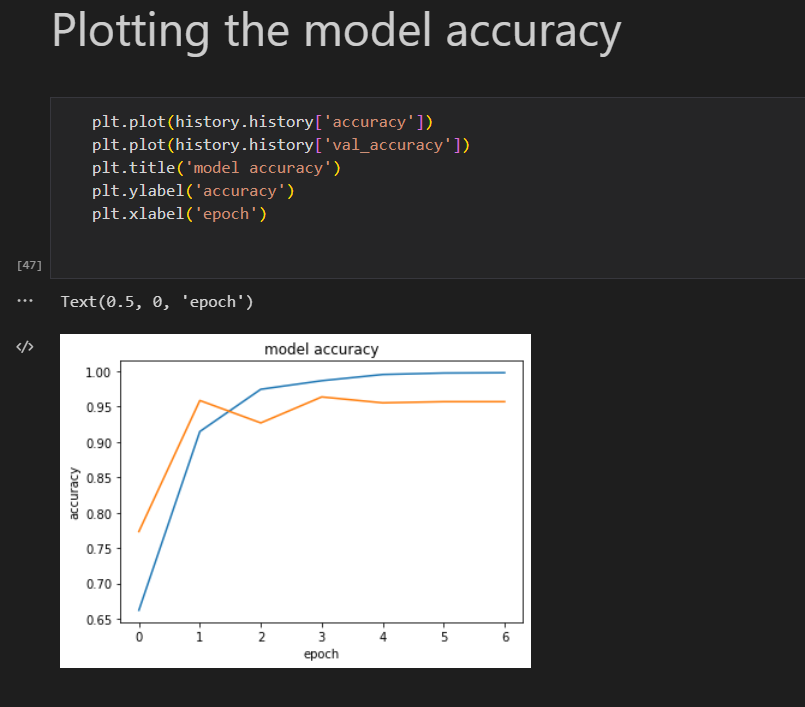


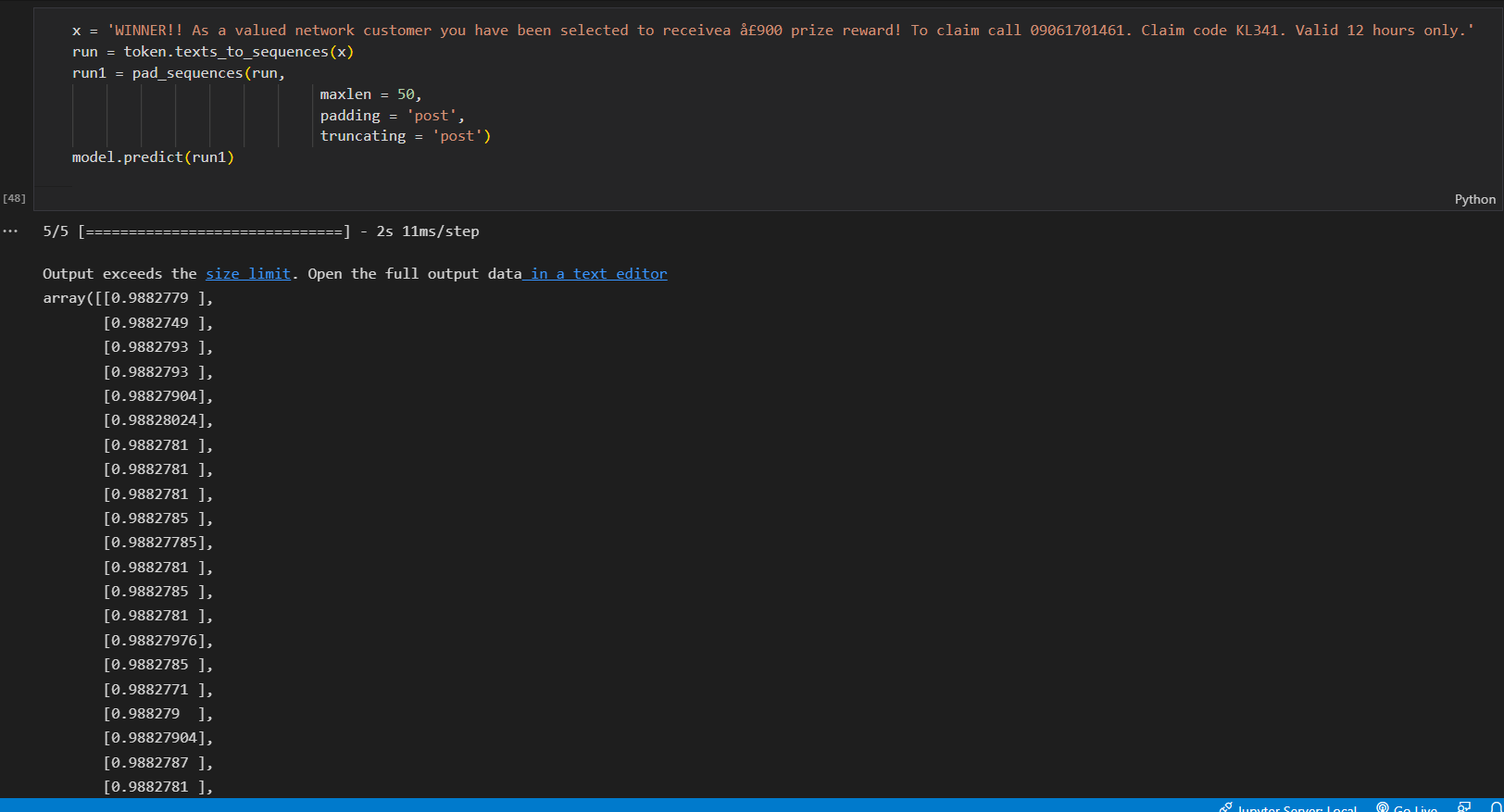


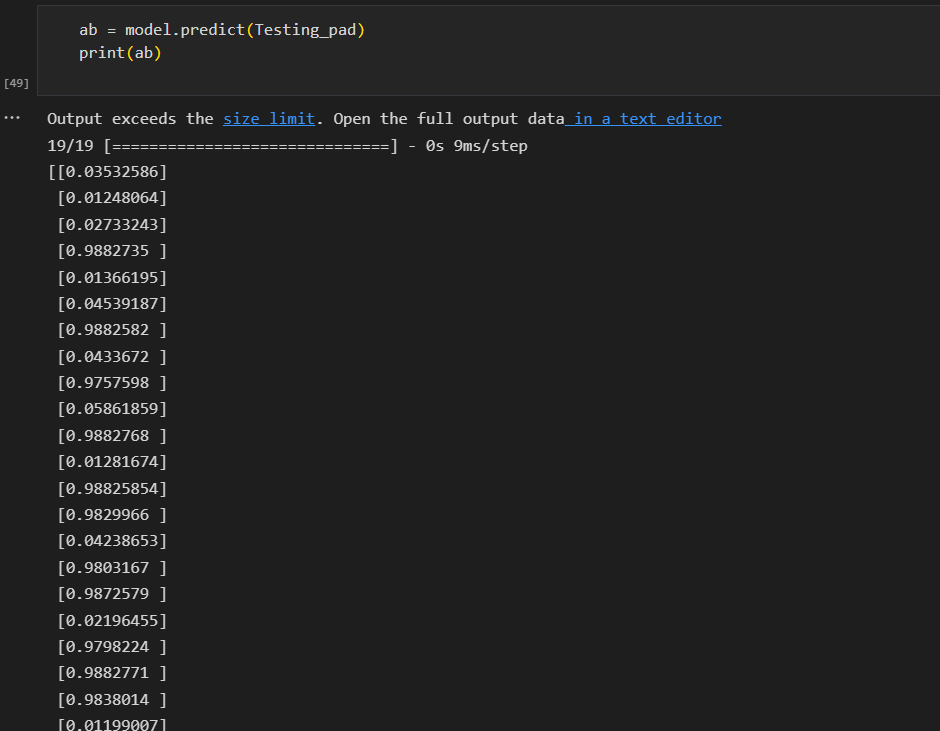












**Conclusion**

All of the known methods for email spam screening are effective. Some have successful outcomes, while others are attempting to establish a different procedure to improve their accuracy rate. We discussed the problems faced by current algorithms and models to classify spam mail.

We developed a machine learning model that will help us to classify spam mail by incorporating various python libraries. We must pre-process our data to resolve any discrepancies and errors from it. Then we split the data into train and test parts to train our model. In this process, we utilize various libraries like TensorFlow, and NLP to match the syntax and semantics of the mails so that it will be easy for the system to classify the text part, Matplotlib to visualize our dataset, etc. Then we developed our LSTM model using Tokenizers and validated our dataset into vectors. After properly fitting our dataset in the model and training the model to improve accuracy.

The final model that we have developed can classify spam mail and can achieve a 96.3% accuracy.

**Reference**

1. <https://pythonbaba.com/email-spam-classification-project-report/>
2. <https://www.sciencedirect.com/science/article/pii/S2405844018353404>
3. <https://www.kaggle.com/datasets/venky73/spam-mails-dataset>
4. <https://www.geeksforgeeks.org/detecting-spam-emails-using-tensorflow-in-python/>