MACHINE LEARNING Navneet Singh

AM.EN.U4CSE19138

**SPAM MAIL DETECTION**

(PHASE-1)

Original Dataset :- [old](https://amritauniv-my.sharepoint.com/:x:/g/personal/navneetkumarsingh_am_students_amrita_edu/ER6bUu8DrpJKiyHMrbqZn3wBJM3SIcitAOtp5SrUibcIjQ?e=mpR51a)

Processed Dataset:-[new](https://amritauniv-my.sharepoint.com/:x:/g/personal/navneetkumarsingh_am_students_amrita_edu/EZ-APhY5zbZMpBgsp4zI40sB1XNzMKQwy50kLaw4ycZ5Gg?e=dn81vq)

horizontal line

# 

# 

## ABOUT:-

Most of us consider spam emails as one which is annoying and repetitively used for the purpose of advertisement and brand promotion. We keep on blocking such email-ids but it is of no use as spam emails are still prevalent. Some major categories of spam emails that are causing great risk to security, such as fraudulent emails, identity theft, hacking, viruses, and malware. In order to deal with spam emails, we need to build a robust real-time email spam classifier that can efficiently and correctly flag the incoming mail spam, if it is a spam message or looks like a spam message. The latter will further help to build an Anti-Spam Filter.

Google and other email services are providing utility for flagging email spam but are still in the infancy stage and need regular feedback from the end-user. Also, popular email services such as Gmail, Yandex, yahoo mail, etc provide basic services as free to the end-user and that of course comes with EULA. There is a great scope in building email spam classifiers, as the private companies run their own 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.

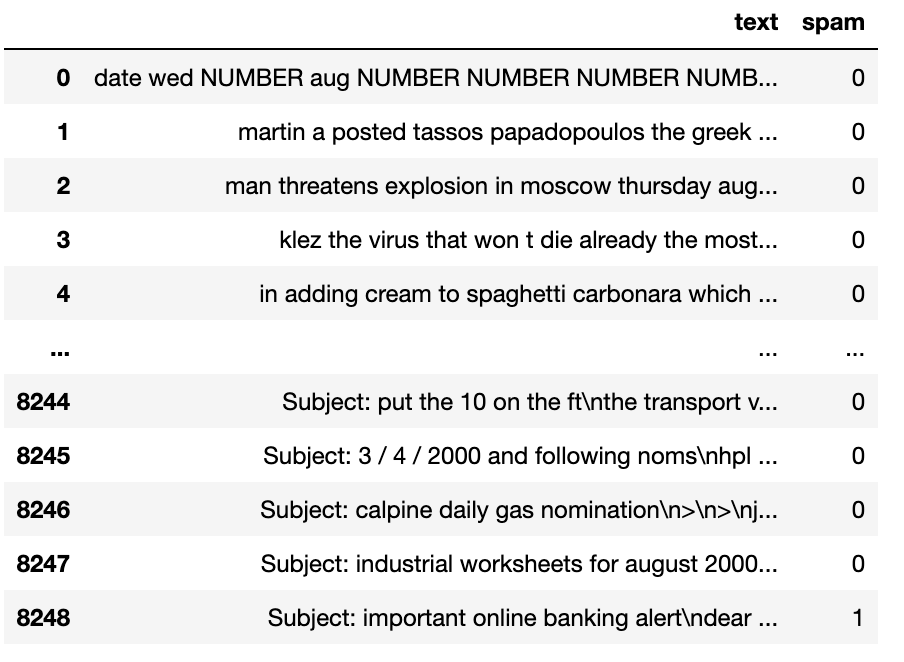
**Steps:-**

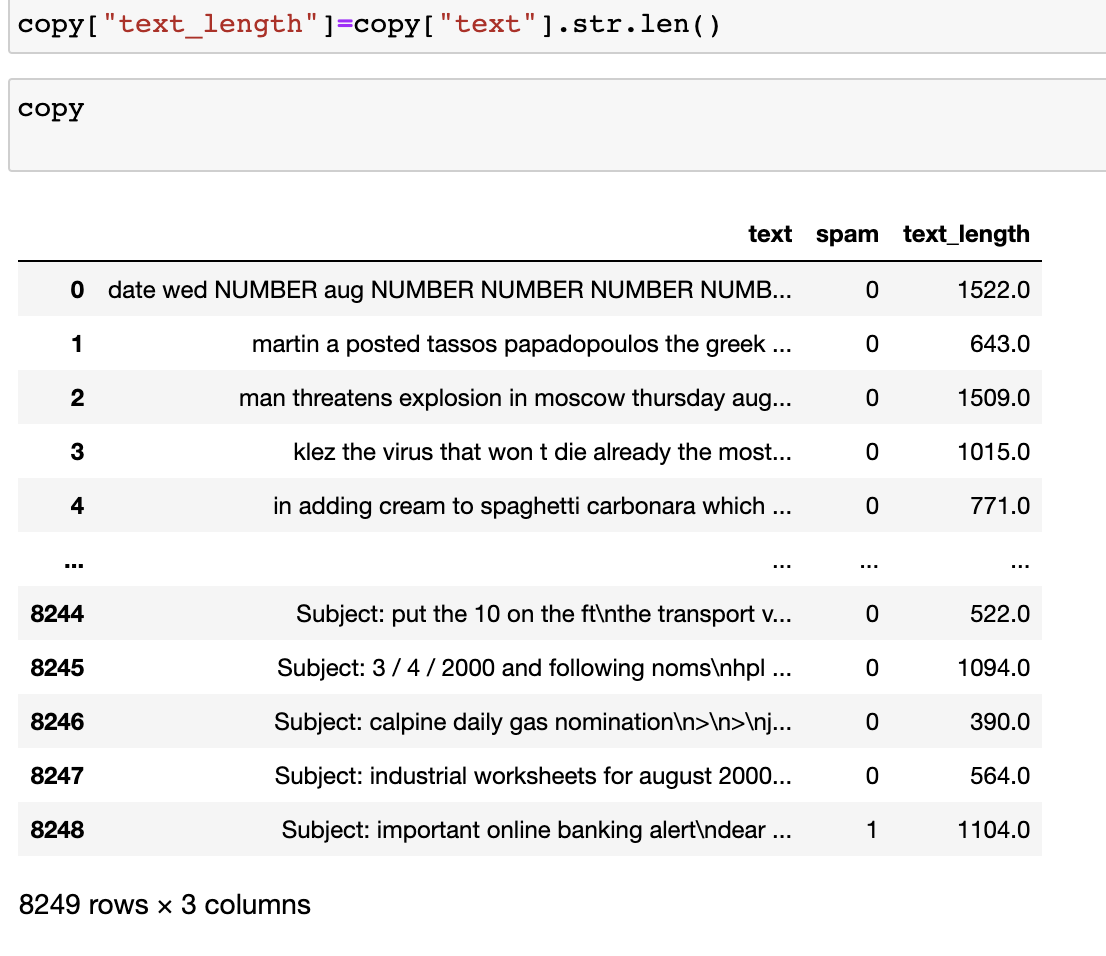
**1.)Text Preprocessing.**

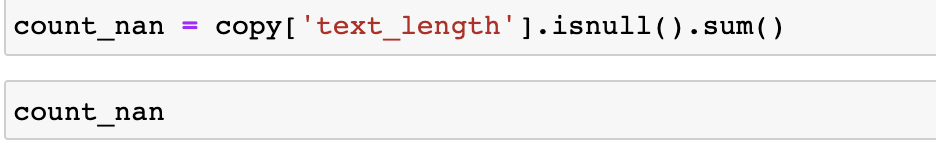
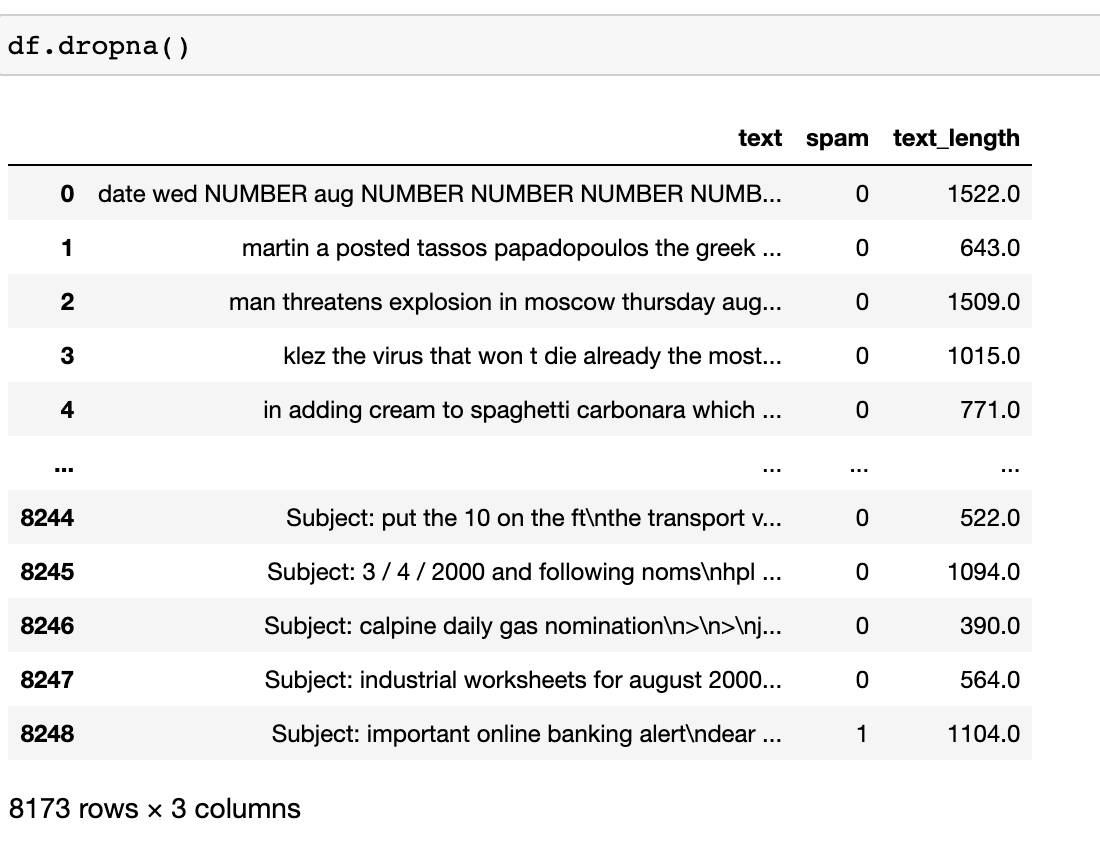
**2.)Text Sequencing.**

**3.) Visualization of the data.**

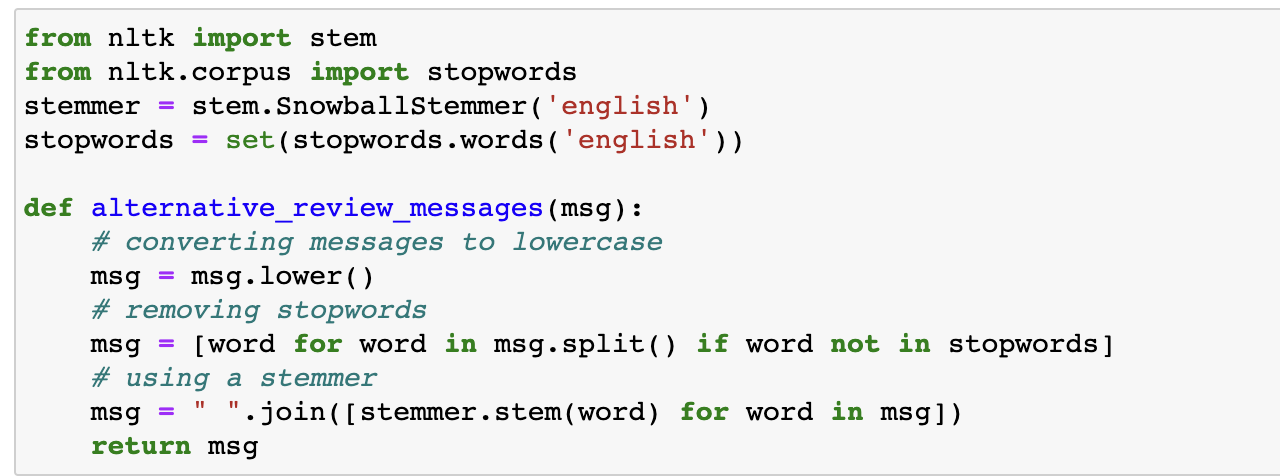
**Screenshots of the processes performed:-**

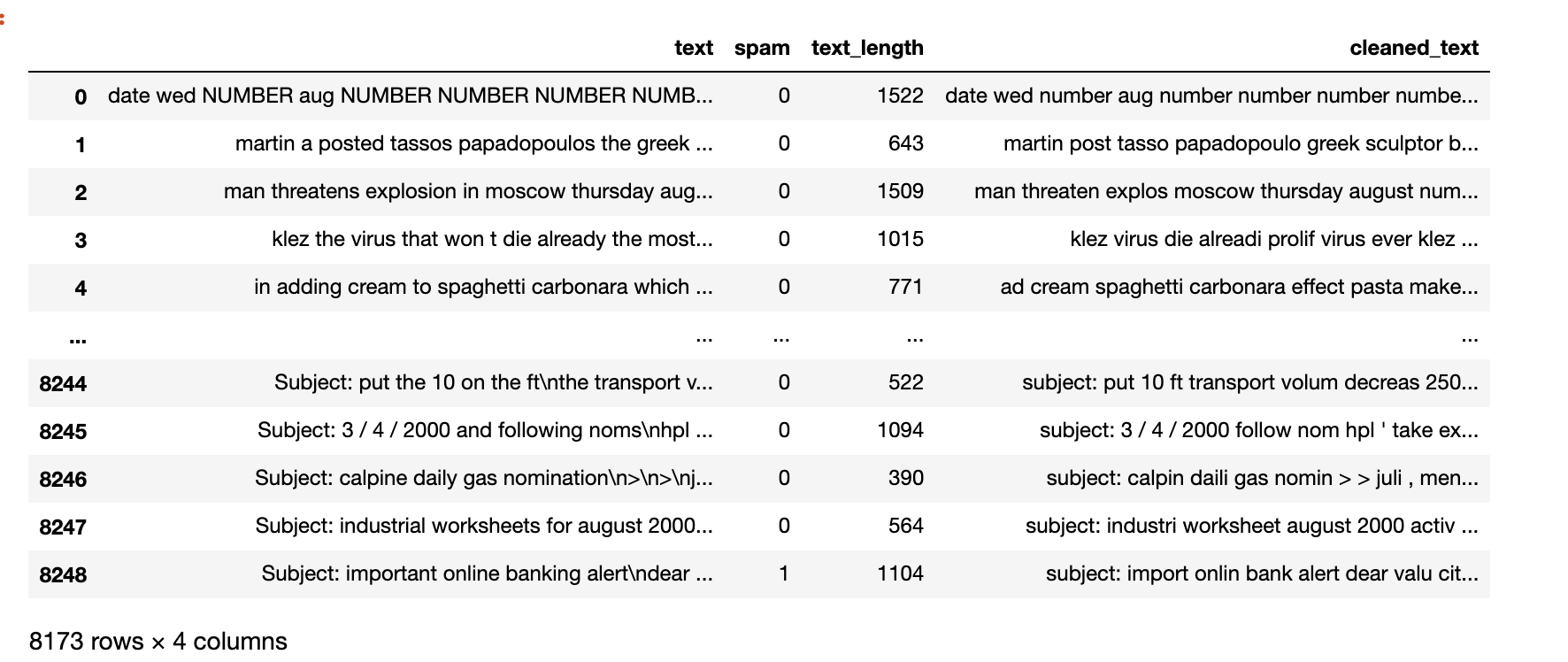
****

****

****

****

****

****

Processes that I have performed above are:-

**Processing the text:-**

Data usually comes from a variety of sources and often in different formats. For this reason, transforming your raw data is essential. However, this transformation is not a simple process, as text data often contain redundant and repetitive words. This means that processing the text data is the first step in our solution.

Some important steps include :-

1.)Cleaning the raw data.

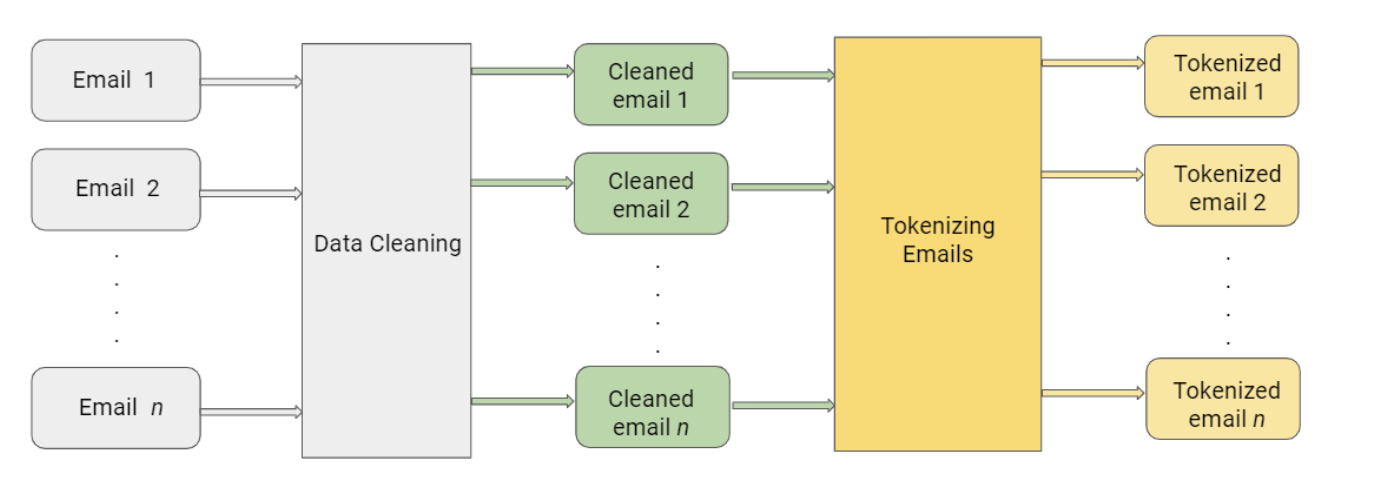
2.)Tokenizing the cleaned data.

Second phase involves the deletion of words or characters that do not add value to the meaning of the text. Some of the standard cleaning steps are listed below:

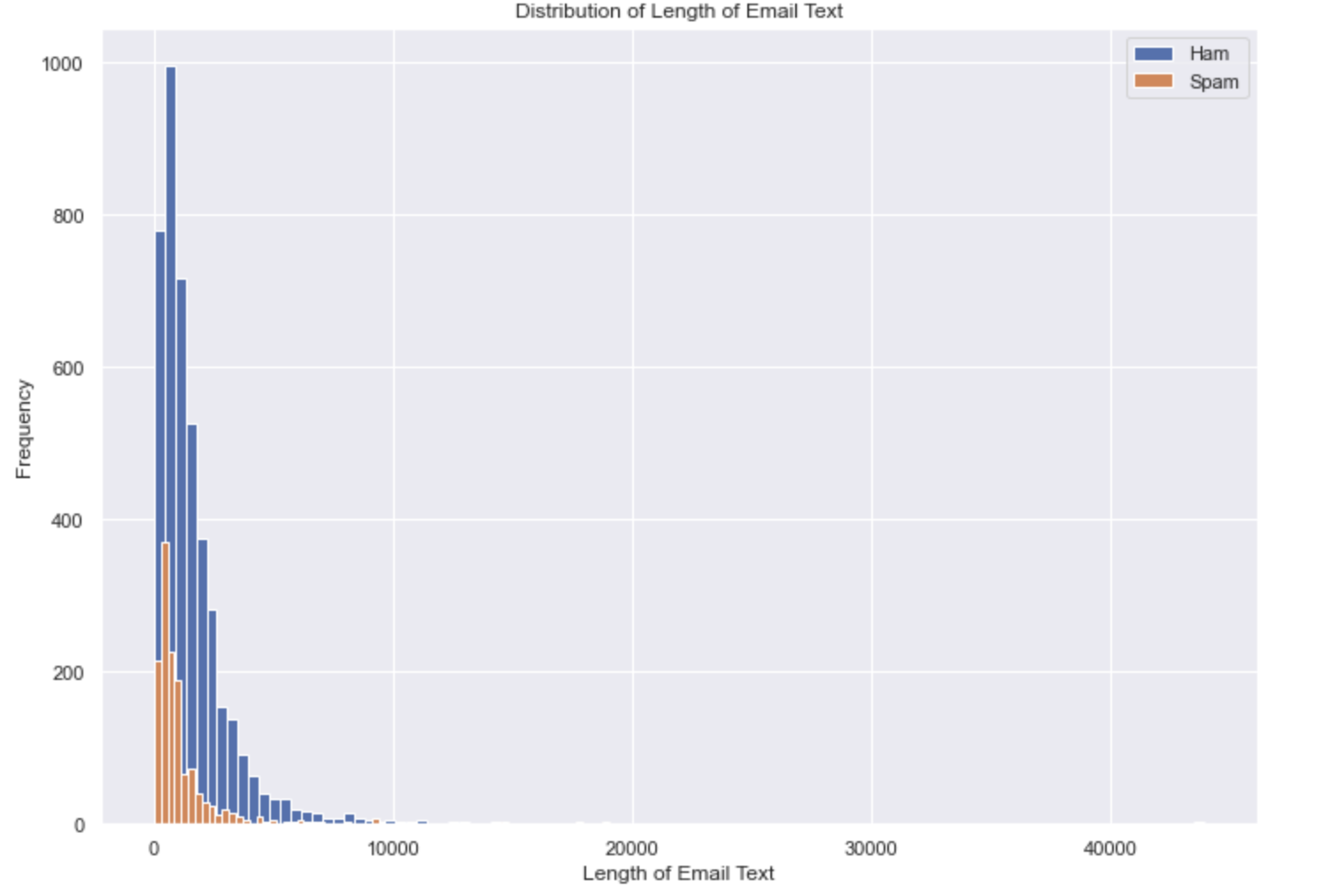
* Lowering case
* Removal of special characters
* Removal of stopwords
* Removal of hyperlinks
* Removal of numbers
* Removal of whitespaces

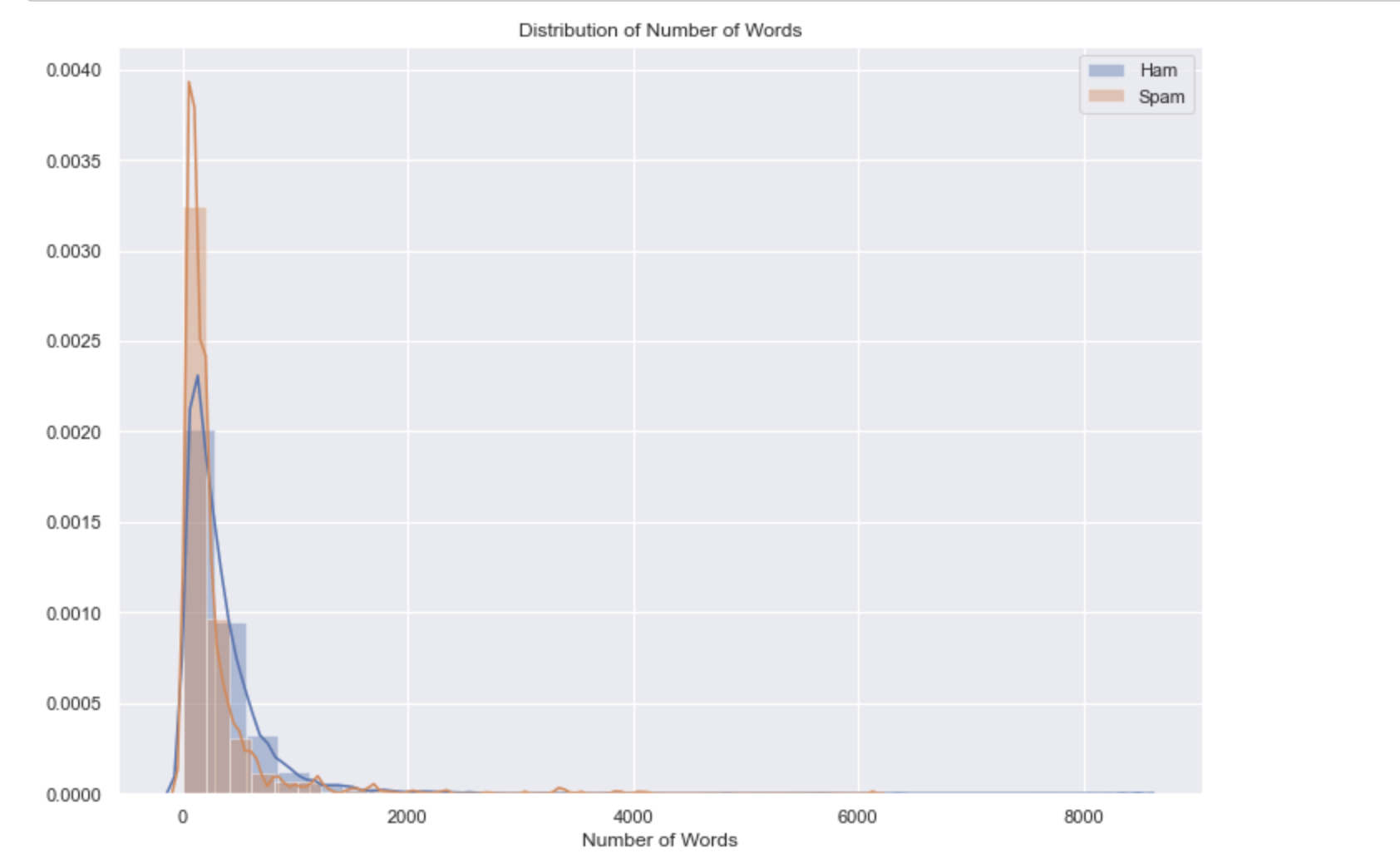
**Tokenizing the cleaned data:-**

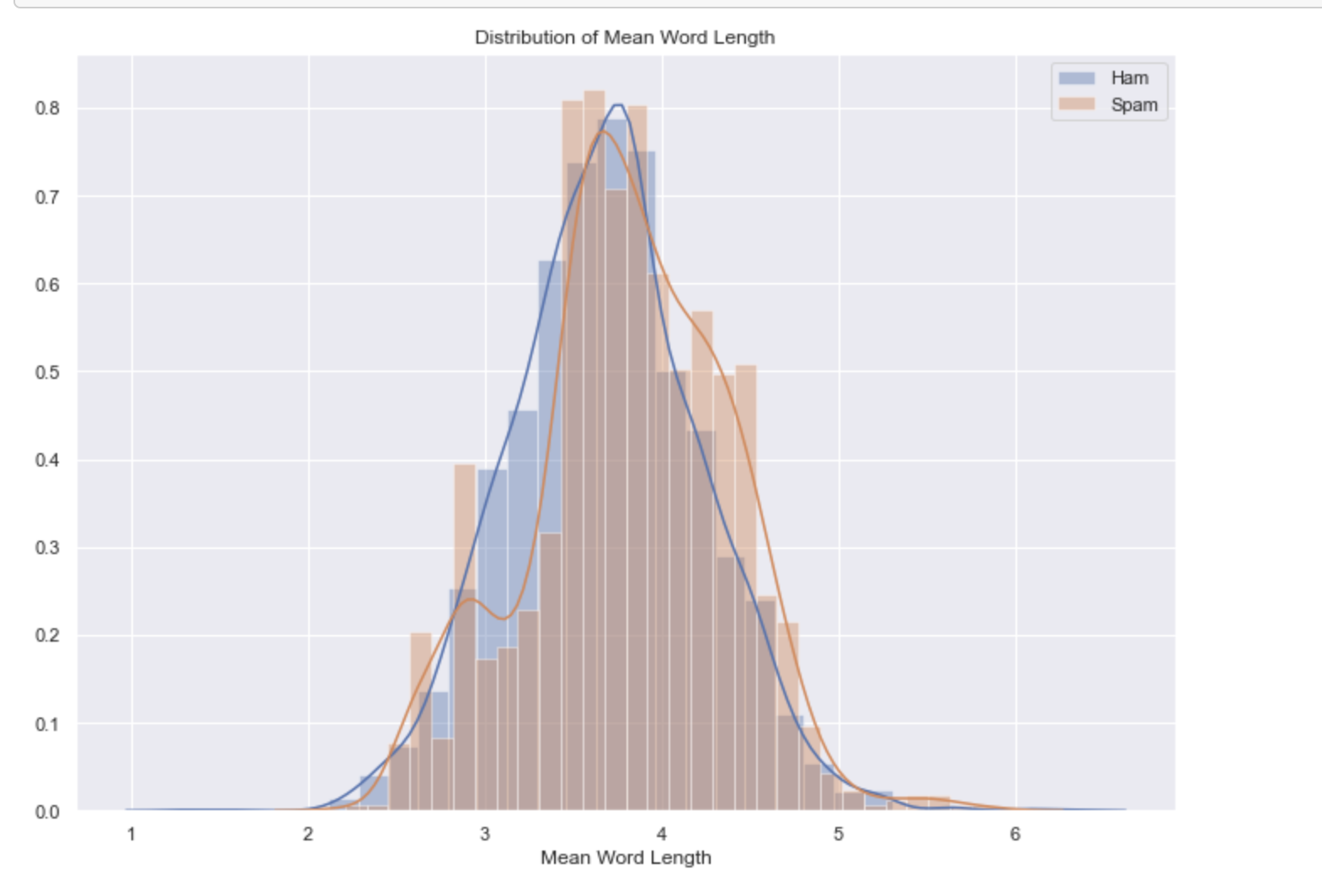
Tokenization is the process of splitting text into smaller chunks, called tokens. Each token is an input to the machine learning algorithm as a feature.

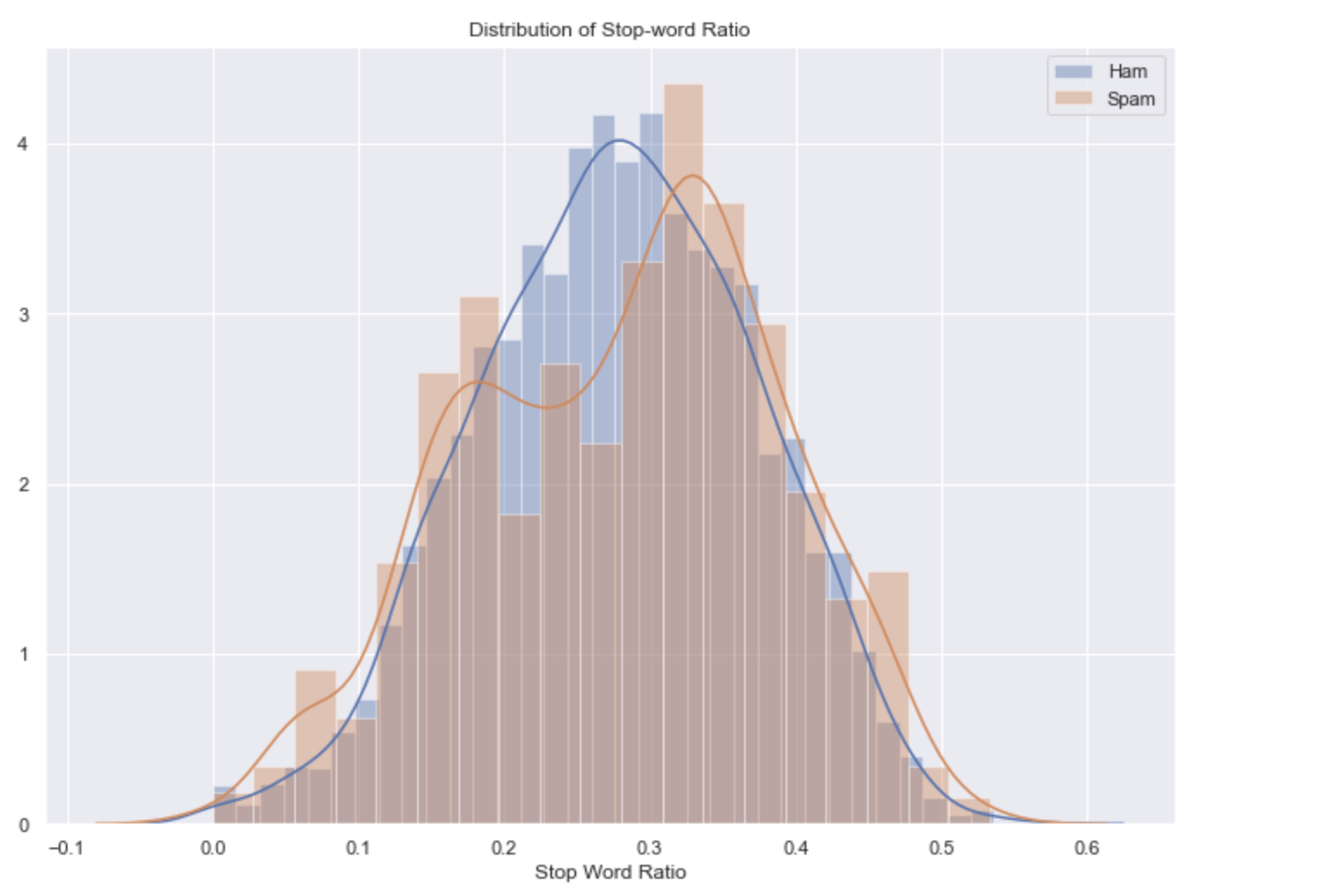


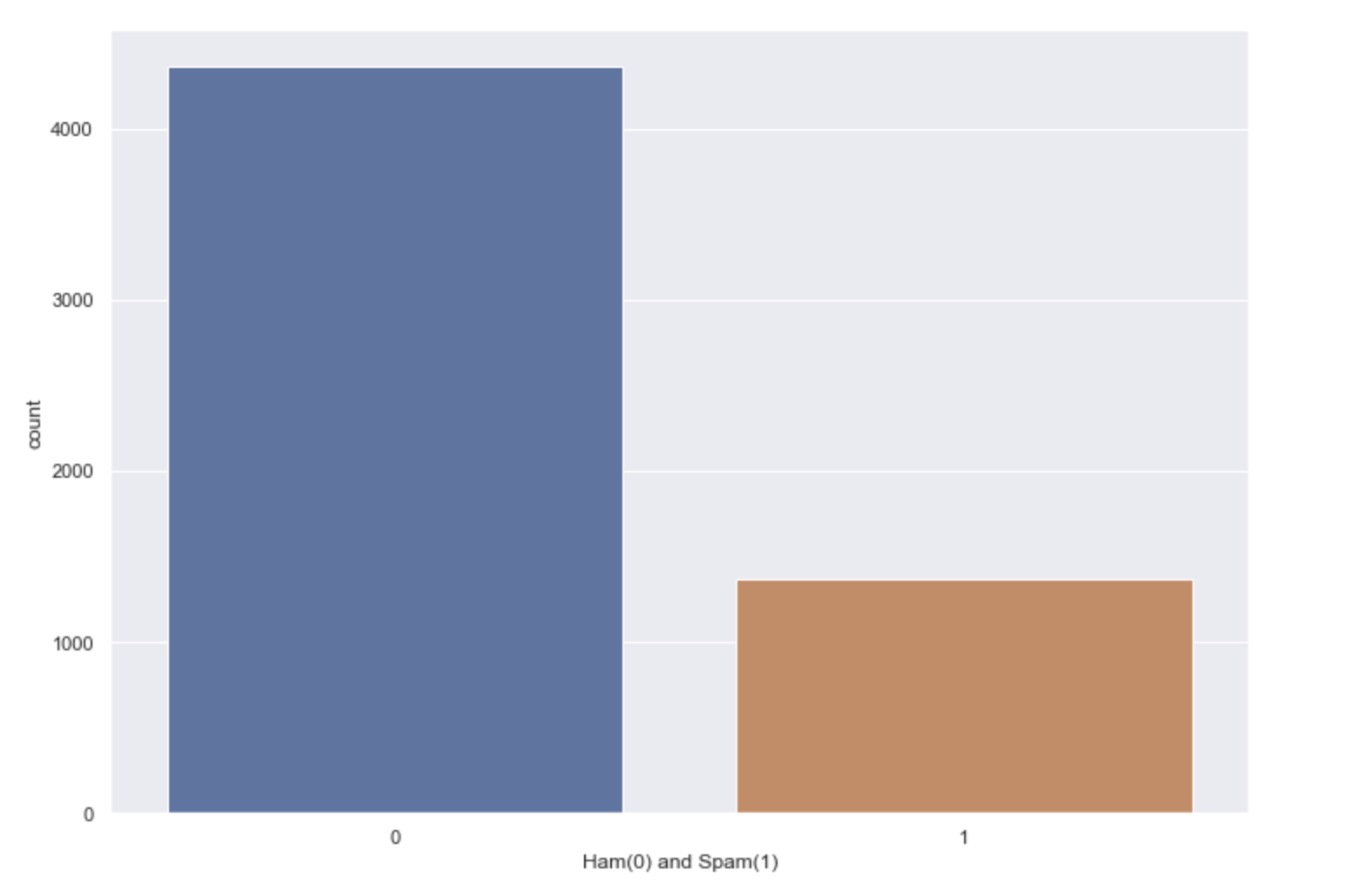
**Data Visualizations:-**

****

****

****

****

****



# 