Problem Statement

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged as either ham (legitimate) or spam.
We have to train a model which may classify any given message as a ham or spam. This whole dataset can be used for the training of the classifier.

Source:

Tiago A. Almeida (talmeida ufscar.br)

Department of Computer Science

Federal University of Sao Carlos (UFSCar)

Sorocaba, Sao Paulo - Brazil

JosÃ© MarÃa GÃ³mez Hidalgo (jmgomezh yahoo.es)

R&D Department Optenet

Las Rozas, Madrid - Spain

Data Set

The dataset can be found at UCI Machine Learning repository

Attribute Information:

The collection is composed by just one text file, where each line has the correct class followed by raw message.

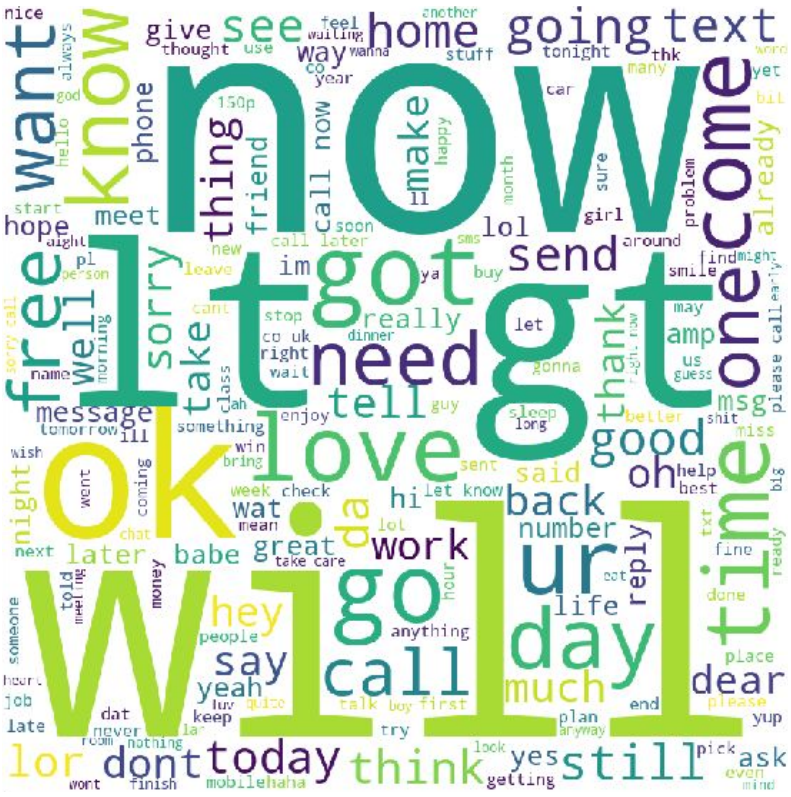
Data

Label	Either ham or spam
Message	The message or the sms which is classified.

HYPOTHESIS

Every message has a semantic and a grammatical meaning. By deciphering which type of semantic and grammatical meaning does a specific sentence carry, we may be able to figure out the difference between a spam and a ham message.

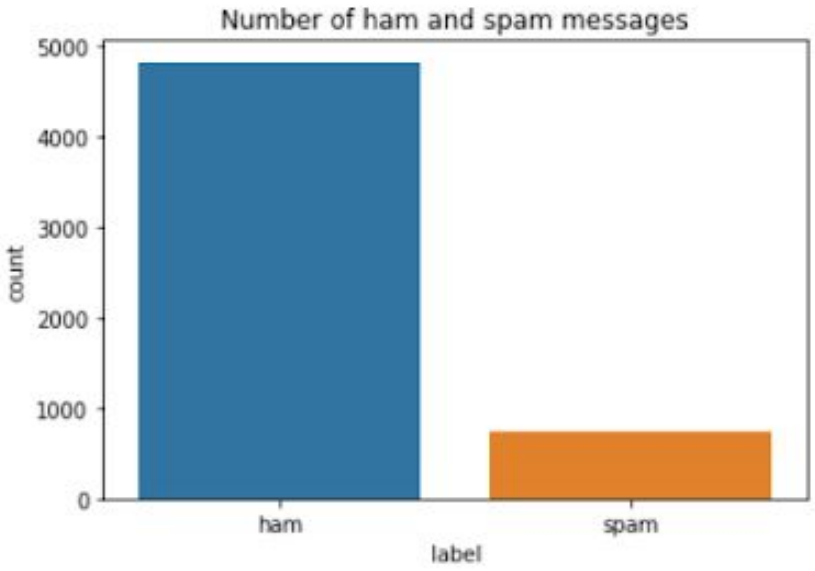
In the same way we can train models which can differentiate between the two categories.
We have been provided by just the Label, and the Message columns.



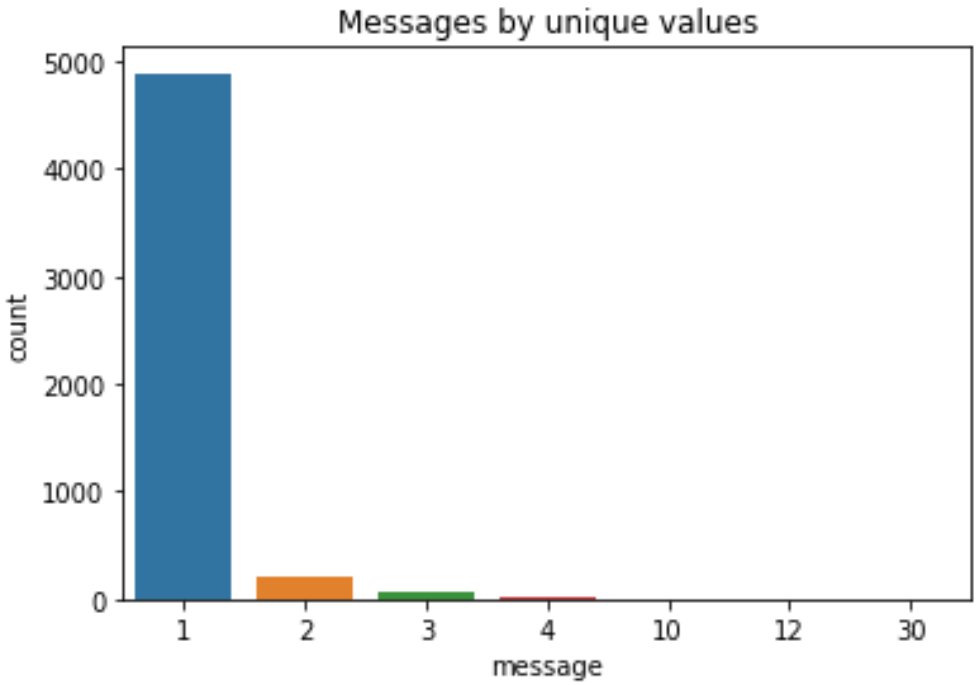
Data Wrangling & Exploratory Data Analysis

The dataset has two columns the label and the message.

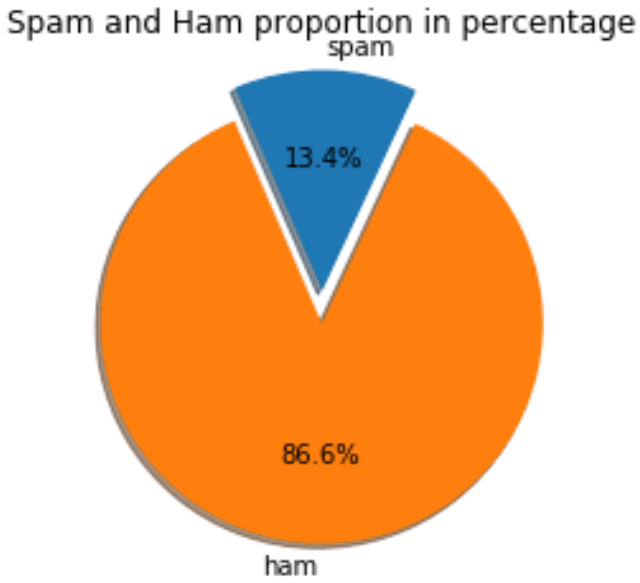
- We check the info() of the label and message columns. We find that there are 5572 non null values in both the columns.
- With the describe method we saw the most returned label is ham with 4825 entries and the most repeated message is Sorry, I'll call you later.
- 5169 messages are unique



We can see that the majority of the messages are ham.



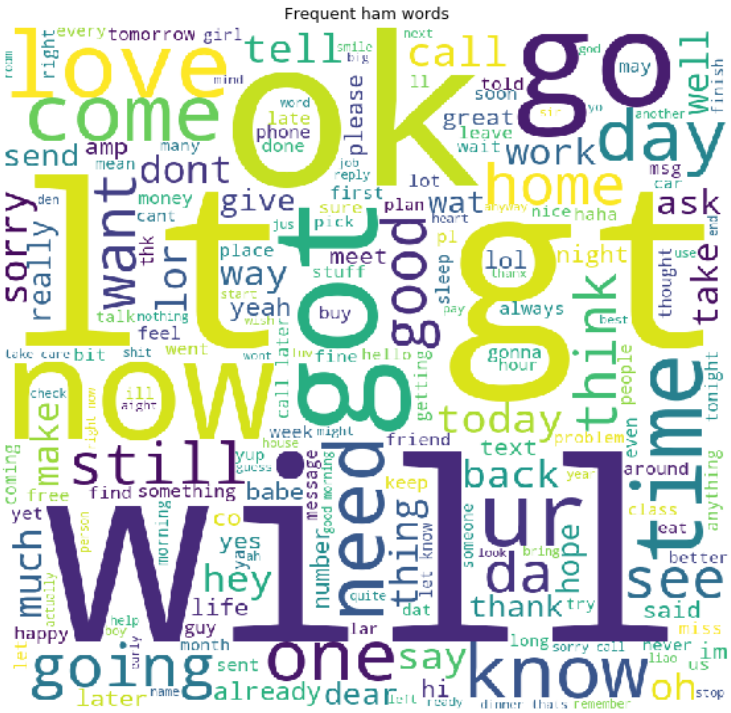
Most of the messages are unique as we can see the blue bar. The other frequencies recorded for messages are 2, 3, 4, 10, 12, and 30 times.



The proportion of ham messages is 86.6%, whereas that of spam messages stands at 13.4%.

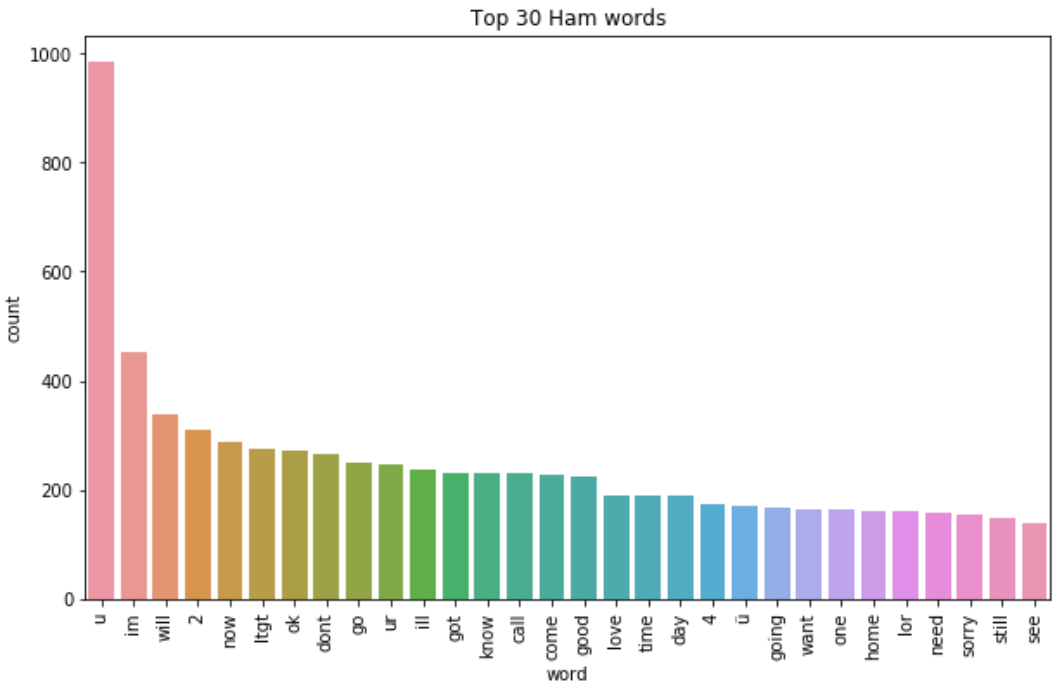
Frequent words used in the messages. Most frequent are will, now, want lt, gt. Most frequent words we can see are from the ham messages as these messages are more in number.

Frequent words used in the spam messages are free, text, call. These words show that the messages are trying to sell something and want the receiver to respond. They may even have malicious URLs.

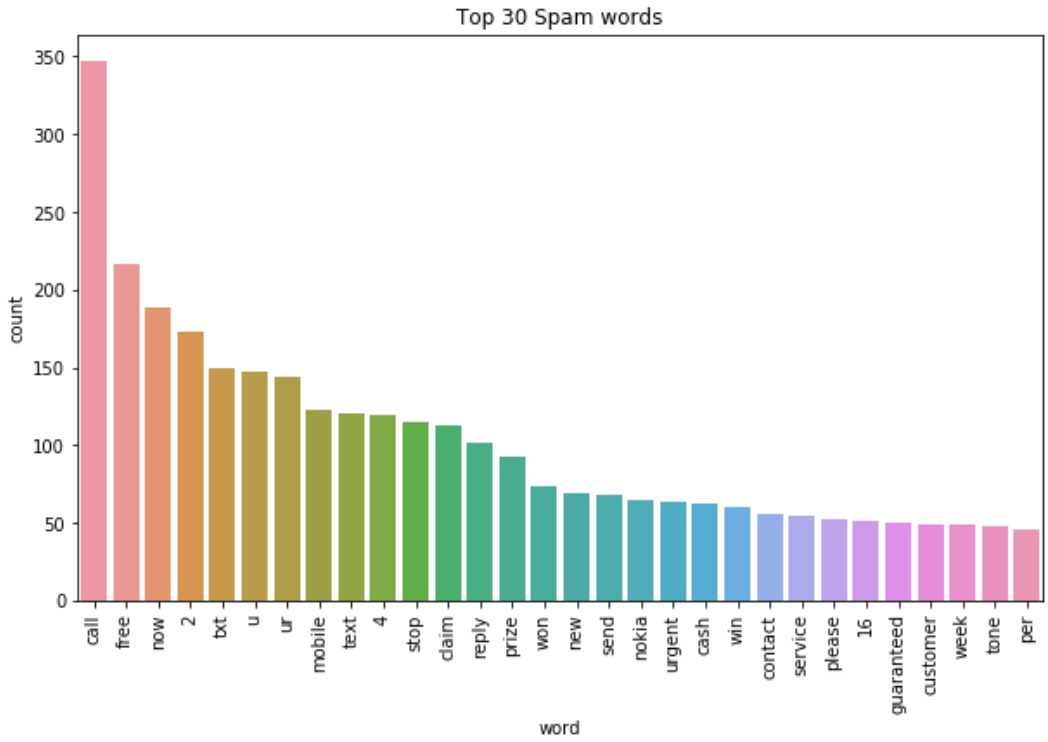


Frequent words used in ham messages are will, lt, ok, gt. They are talking about the real time conversation and talking about the near future.

The top 30 words used in all the messages.



The top 30 words used in the ham messages.



The top 30 words used in the ham messages.