# Sms spam and ham classification problem using Naïve Bayes algorithm

Problem statement: Classify the SMS text into Spam or Ham for the given unstructured data

The following is the data:



The above data contains a lot of strings, floats and ints. This is an unstructured data that requires some data processing

The following are few of the stop words that are eliminated from the above data using regular expressions:



After cleaning the data, we get the following:



Once the data is cleaned, we split this data into train and test data as follows:

The sentences containing the words in the training and testing data will be split to form individual features. This is done in the form of sparse matrix containing very few non binary digits.

We pass the sparse matrix data through Multinomial Naive Bayes and the following are the results:

Train data accuracy = 98%

Test data accuracy = 97%

We pass the sparse matrix data through Gaussian Naive Bayes and the following are the results:

Train data accuracy = 92%

Test data accuracy = 85%

From the above results we can evaluate that the results obtained in Multinomial Naïve Bayes are more accurate than Gaussian Naïve Bayes algorithm