STATEMENT OF WORK*- AI ALGO.*

**(SPAM SMS CLASSIFICATION)**

### 

### 

### **Issued By: Issued To:**

Sahil Kumar Grover Prof. Marcos Bittencourt

ID: 100799382

Contents

[Background Information 2](#_Toc381277520)

[Scope of Work 2](#_Toc381277523)

[Deliverables 2](#_Toc381277524)

[Milestones 2](#_Toc381277525)

[Gantt Chart 2](#_Toc381277523)

Data Source [3](#_Toc381277528)

EDA [3](#_Toc381277528)

Data Cleaning [4](#_Toc381277528)

[Testing 5](#_Toc381277529)

[Limitation 5](#_Toc381277530)

[Acceptance 5](#_Toc381277529)

2020-11-03

# Background Information

Short Message Service (SMS) has become a multi-million-dollar business industry with an incentive between 11.3 to 24.7 percent of the agricultural nations' Gross National Income (GNI) in the early year of 2013 [1]. In any case, the drawback of the expanding versatile clients and the modest SMS instant messages is that cell phones are pulling in more spontaneous mass messages particularly as promotions. Which brought about bothering the cell phone client. The dataset consists of 5572 rows and a label which classifies the message as “spam” or “ham”. It contains a collection of 747 SMS spam messages manually extracted from the Grumble text Web site.

# SCOPE of WORK

The objective of this task is to detect spam SMS. For the sake of simplicity, we say an SMS can contain information that is legit, or it can be spam. So, the task is to classify spam or ham SMS from other SMS. Spam SMS classification will follow classification Machine Learning Algorithm to determine whether the message is a spam SMS or a not.

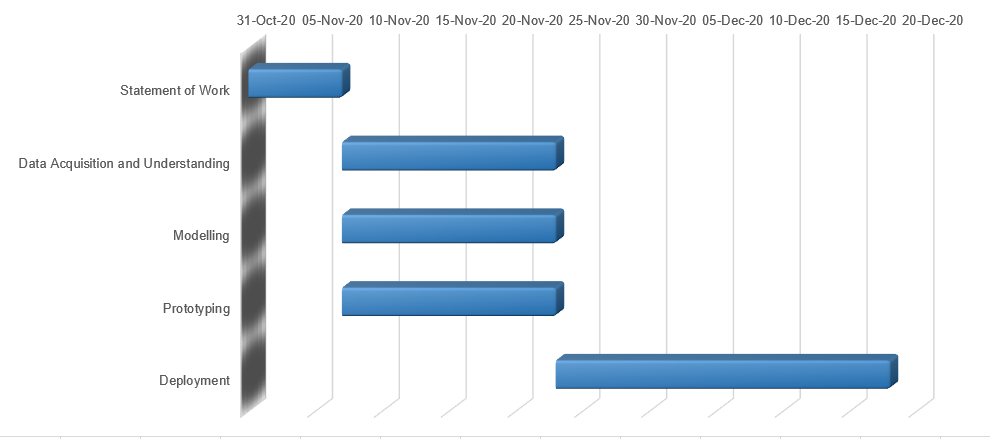
## Deliverables

The objective of this AI venture is to anticipate if a SMS is a SPAM or HAM given their segment variety. To accomplish this, different ML algorithms will be explored.

## Milestones

|  |  |
| --- | --- |
| Milestone | Estimated Delivery Date |
| Statement of Work | 06-Nov-2020 |
| Data Acquisition and Understanding | 23-Nov-2020 |
| Modelling | 23-Nov-2020 |
| Prototyping | 23-Nov-2020 |
| Deployment | 18-Dec-2020 |

# GAntt chart

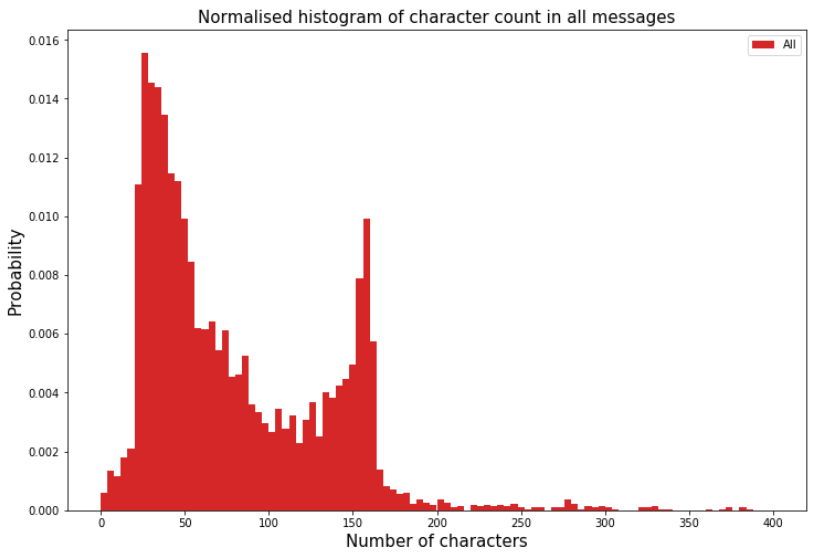


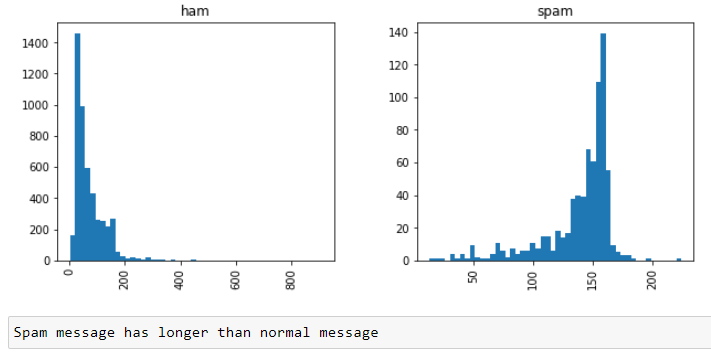
# DATA source

The dataset consists of 5572 rows. Having 747 spam SMS and 4825 Ham SMS. The dataset has two columns v1 and v2, where v1 consist of SMS label as spam or ham message and v2 consist of entire message. <https://www.kaggle.com/uciml/sms-spam-collection-dataset>

# EXploratory DATA Analysis

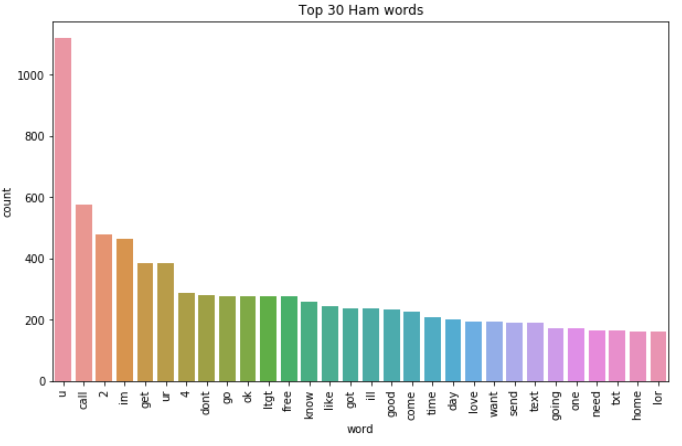
**EDA** is a thorough examination meant to uncover the underlying structure of a data set and is **important** for a company because it exposes trends, patterns, and relationships that are not readily apparent.

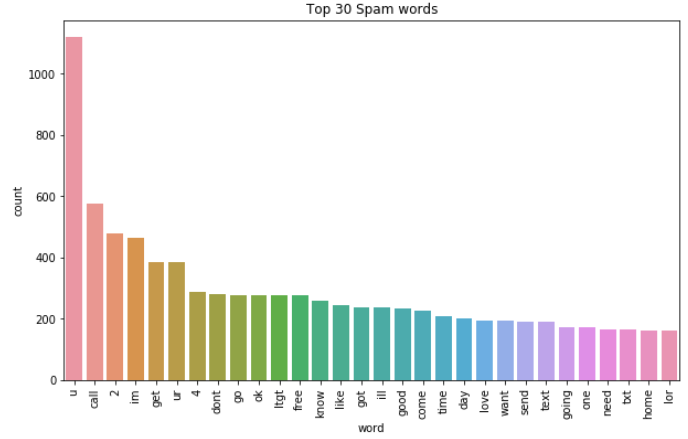




# DATA Cleaning

For the algorithm to give better results we need to be sure that we don’t have unnecessary data. In our case since the dataset has a lot of stop-words we need to be sure that we have removed them.





# TESTING

So, as to assess or decide the exactness of the versatile SMS spam separating procedures, certain presentation assessment measurements will be used:

Accuracy: will find the percentage spam SMS classified:

Accuracy = TP+TN/TP+TN+FP+FN

Precision: It calculates the fraction of cases for which the accurate outcome is returned:

Precision= TP/TP+FP

Recall: gives the quotient of accurate to inaccurate forecasts within real spam texts

Recall = TP/TP+FN

# Limitations

The dataset consists of two columns v1 and v2 where v1 is label and v2 is the text message. The column v1 has information in terms of “ham” and “spam” which needed to be converted in the form of 0 and 1 to make algorithm to understand the difference in “ham” message and “spam” message.

# Acceptance

Date:06-11-2020

By initialing each page and signing below, I Sahil Kumar Grover, in my capacity as student, of Durham College agree to and accept the terms set forth in this Statement of Work.

*(Durham College)*

By: Sahil Kumar Grover