**AN ANALYSIS OF SPAM SMS FEATURES**

IFN 701 Project 1- Data Analysis and Research Project

Project Student Name: Harshita Jain

Student ID: n9539361

Academic Supervisor: Dr. Guido Zuccon

Project Coordinator: Dr. Charles Wang

# **ABSTRACT**

The increasing dependence on short text messages for communicating with other people has led to an increase in submission of Spam Short Message Service(SMS) which intends to extract confidential, personal and valuable information from the recipients. This growth is difficult to curb because of the availability of affordable and unlimited prepaid SMS packages and high response rates as customers are more comfortable with sharing their confidential and personal information via SMS. The previous solutions built in this context are simple, straightforward and do not consider all the core characteristics of Spam SMS. One of the major problems that has limited the research on this topic is scarce data available for research on Spam SMS.

This document details a data analysis and research project involving the investigation of SMS Spam. This project aims to create a data analysis to analyze the features that differentiate a Spam SMS from a Legitimate SMS and a predictive model that is enabled to understand and automatically identify whether an SMS is a Spam message or a Legitimate message. This project will be performed using a publicly available dataset acquired from Kaggle and by creating a supporting codebase in R Markdown.

The outcome of analysis is the most frequent words used in Spam SMS that differentiate it from a Legitimate SMS. Moreover, the result of building predictive models using 4 classifiers: *Naïve Bayes, Support Vector Machine, Logistic Regression and Decision Tree* indicated that Support Vector Machine performed better than the rest of the three classifiers and can be used in mobile devices to prevent arrival of Spam SMS to the subscriber.

The outcome of this project would benefit the mobile networks operators as, now, they would not have to spend more on maintaining the networks and operations which get hampered by Spam SMS. Customers would be benefitted as their confidential, personal and valuable information would remain protected.

# **1. INTRODUCTION**

This document aims at elaborating a data analysis and a research project. The project used a publicly available dataset from Kaggle to analyze the features that differentiate a Spam SMS from a Legitimate SMS and to build a predictive model capable of accurate prediction of a Spam SMS. The project was performed by creating a codebase in R Markdown.

## **1.1 Context of the Project**

SMS is an integral feature of mobile devices that facilitates communication through exchange of short text messages. (Ahonen, Tomi T., 2011) Informa Telecoms and Media has found that there has been an increase in this exchange from 5 trillion messages being exchanged in 2010 to 10.7 trillion SMS being exchanged in 2015. (Global SMS traffic to reach 8.7 trillion by 2015: study)Also, the revenue generated by SMS has increased from US$132.5 billion in 2010 to US$136.9 billion in 2015.(Global SMS traffic to reach 8.7 trillion by 2015: study)

However, this exchange has been hindered due to submission of unsolicited and unwanted SMS in bulk to recipients without their authorization and consent. Such SMS intends to extract confidential, valuable and personal information out of the recipients. According to Cloudmark Analysis, 92% of such SMS are fraud.(Whitepapers)

The most forms of Spam SMS are (Whitepapers)

* Have won a Gift Card Message,
* Account Phishing Spam Message,
* SMS Service Message,
* Accident Compensation Spam Message, and
* Payment Protection Insurance (PPI)Compensation Spam Message.

Cloudmark Analysis revealed that there has been a growth in reception of Spam SMS of 300% from 2011 to 2012. (Whitepapers) This growth can be accredited to the following two reasons:

1. The availability of affordable and unlimited prepaid SMS packages due to which extraction of confidential, valuable and personal information of the recipients becomes a cost effective opportunity for the spammers. (Khemapatapan, C, 2011)
2. Customers feel more comfortable sharing their confidential, valuable and personal information vis SMS. (Khemapatapan, C, 2011) Therefore, it was revealed in June 2013 statistics that 43% of such messages are everted within 15 minutes of reception. (SMS Marketing Statistics, 2017)

Spam SMS adversely affects consumers and mobile network operators. The reasons that make this problem a significant one are:

1. Mobile network operators bear a heavy loss in order to maintain networks, operations and increased customer care services to the customers. Spamming ruins their reputation making them lose on many valuable customers(Khemapatapan, C, 2010).
2. Customers are also left annoyed and worried as their confidential, personal and valuable information is at stake(Khemapatapan, C, 2010).
3. Many network operators have provided means to their customers to block Spam SMS, which sometimes leads to filtration of legitimate message as a spam due to its characteristics matching to those of a spam message(Khemapatapan, C, 2010).

## **1.2 Related Work and Research Gap**

There have been a number of anti-spam tchniques built in this conetxt to solve this problem like (Khemapatapan, C, 2010)–

* Blacklisting - This technique forbids access to a service if the name is written on the list.
* Simple Filtering - This technique analyses the traffic data and identifies the individual subscriber causing huge volumes of it.
* Spoofing/Faking Detection Techniques

These techniques are brittle, simple and straightforward in nature. They do not consider the core characteristics of Spam SMS. Also, they perform in an ad-hoc an d post-hoc manner. That is, a consumer can only blacklist a number as wand when it sends him a Spam SMS. Blacklisting one number does not guarantee him complete prevention from receiving Spam SMS. Moreover, spamming methods have advanced in a way that they make a Spam SMS appear as a Legitimate one. Therefore, there is an urgent need to build a more sophisticated and appropriate model to eradicate this issue. (Khemapatapan, C, 2010)

Moreover, not much research could be done in this context due to scarce data available on Spam SMS.(Khemapatapan, C, 2010)

### **1.2.1 How this Project Addresses the Problem**

I have worked on bridging the gap by building a predictive model to accurately predict whether an SMS is a Spam SMS or a Legitimate one by working on the publicly available dataset available at Kaggle. It comprises of 5,574 English, real and non-encoded text messages, submitted to Grumble text website. All claims made on this site about the text message being spam are identified and investigated through carefully scrutinizing over a hundreds of webpages. (SMS Spam Collection) All messages have accurately been tagged as legitimate and spam. In total, the dataset consists of 747 Spam messages.

## **1.3 Aims and Objectives of the Project**

The objective of this project would be to answer the following two questions:

1. What are the characteristics that distinguish Spam messages from Legitimate messages?
2. What is the effectiveness of the classification methods – Support Vector Machine, Decision Trees, Logistic Regression or Bayesian Classifiers in identifying SMS Spam?

Therefore, the purpose of this project is:

1. To analyze the data to understand the differentiating features of SMS Spam.
2. To build a predictive model which can accurately predict whether the SMS is a Spam SMS or a Legitimate SMS.

Particularly, the main aim of this project is

* Carrying out an exploratory analysis on the dataset to explore and learn about the data.
* Statistically predicting and modelling the data
* Result Interpretation

## **1.4 Brief Overview of Methods used in the Project**

In this project, I will develop a data analysis including the investigation of a number of predictive models. This analysis is structured in 4 phases (Guo, P., 2013):

1. **Preparation Phase -**

This phase includes acquiring data from Kaggle, formulating key questions for analysis and cleaning the data.

1. **Exploration Phase -**

This phase includes carrying out exploratory analysis on the dataset to analyze the features that differentiate a Spam SMS from a Legitimate SMS. This analysis starts with analyzing "How Length of Messages and Number of Messages relate to each other for each label", followed by finding words that appear most frequently in Spam SMS.

1. **Data Preparation Phase -**

This phase includes preparing data to be used to build predictive models in Classification phase (Next Phase). In this phase, a corpus is created and cleaned by transforming all the text to lower case, removing punctuations, numbers, stop words and white space. The, the data was split into 70% training set and 30% test set.

1. **Classification Phase -**

This phase includes building 4 different classifiers in 2 settings. The settings are:

* 1. ***Setting 1:*** Considering all features of the data
  2. ***Settings 2:*** Considering manually engineered features

The 4 classifiers used are:

* 1. Naïve Bayes
  2. Decision Tree
  3. Support Vector Machine
  4. Logistic Regression

After having modelled all classifiers for each setting, a comparison would be made to determine the best model and the corresponding setting.

1. **Final Delivery Phase -**

This phase includes submitting outputs from Exploration Phase and Classification Phase in the form of R Markdown and an Analysis Report.

Therefore, the two **target deliverables** of the project would be -

1. R Markdown
2. Analysis Report

## **1.5 Outcome of the Project**

The outcome of this project will facilitate:

* An effective and a deeper knowledge of characteristics and features that make a Spam SMS different from a legitimate SMS.
* A predictive model that can accurately predict whether an SMS is spam or legitimate.

These outcomes would indirectly affect the society in a better way. The learnings and research could be converted into operational products in future that would aid accurate identification and filtration of spam SMS.

# **2.Literature Review of Previous Work**

There have been many researches to build the best technique to prevent growth of Spam SMS. The solutions that have been proposed till date are (Khemapatapan, C, 2010):

1. ***Blacklisting:*** This technique denies permission to a number added to a list, called a blacklist, from sending SMS to other recipients.
2. ***Spoofing/Faking Detection Techniques:*** This technique aims at recognizing a Spam SMS pretending to be Legitimate.
3. ***Implement Email Spam Filters:*** This methodology was adopted with an understanding that the filters used to detect an email Spam could also be used to detect a SMS Spam.

## **2.1 Blacklisting and Spoofing/Faking Detection Techniques**

Researchers suggested that SMS spamming could be avoided and, eventually stopped, by blacklisting the spammers' number. They suggested that mobile network operators should authenticate each subscriber and create a blacklist for SMS Spammers. Therefore, this would facilitate prevention of submission of Spam SMS to other recipients and reception of SMS only from legitimate and trusted users. (Khemapatapan, C, 2010)

Similarly, spoofing detection technique was suggested with a perception that it would be capable of preventing Spam SMS by detecting the manipulated address information.

However, it was experienced that blacklisting and spoof detection techniques were not an efficient way of solving the problem of SMS Spamming. they do not guarantee complete prevention from receiving a Spam SMS as the spammer might resort to change in identity and send new Spam SMS or there might come a new spammer out there in the market. (Khemapatapan, C, 2010)

Thus, there is an urgent need that a system is built which can accurately predict whether an SMS is a Spam SMS or a Legitimate SMS. Therefore, in this project, I aimed at building a predictive model which can accurately identify whether an SMS is a Spam SMS or a Legitimate SMS.

### **2.1.1 Approach towards Building a Predictive Model**

In this project, I aimed at building 4 different classifiers using 2 settings. The 2 settings are:

1. **Setting 1:** Considering all the features of the data
2. **Setting 2:** Considering manually engineered features, which are the most frequent words appearing in Spam SMS.

The 4 classifiers used are:

1. Naïve Bayes
2. Decision Tree
3. Logistic Regression
4. Support Vector Machine

After having modelled each classifier for both the settings, I compared their output of Cross Table to determine the most effective classifier and the corresponding setting. I had specifically focused on ***recall*** for spam class and ***incorrect prediction*** for Legitimate class.

### **2.1.2 Outcome of Predictive Model**

I started with building the model for Naive Bayes but the results were not impressive. (Refer Table 1) The accuracy for Setting 1was just 20.89% and only 3 out 224 Spam SMS were detected in Setting 2.

***Naïve Bayes***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All Features |  | Manually Engineered Features |  |
|  | Legitimate | Spam | Legitimate | Spam |
| Precision | 0.98 | 0.14 | 0.87 | 1 |
| Recall | 0.09 | 0.99 | 1 | 0.013 |
| Incorrect Prediction | 0.86 | 0.016 | 0 | 0.13 |
| Accuracy | 20.89 |  | 86.77 |  |

Table 1: Effectiveness of Naïve Bayes in each Setting

Henceforth, I started building models for Decision Tree, Support Vector Machine and Logistic Regression. The outcome for each is as follows:

***Decision Tree***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All Features |  | Manually Engineered Features |  |
|  | Legitimate | Spam | Legitimate | Spam |
| Precision | 0.97 | 0.94 | 0.94 | 0.83 |
| Recall | 0.99 | 0.78 | 0.98 | 0.61 |
| Incorrect Prediction | 0.06 | 0.03 | 0.17 | 0.06 |
| Accuracy | 96.7 |  | 93.17 |  |

Table 2: Effectiveness of Decision Tree in each Setting

This table reveals that 175 out of 224 Spam SMS were detected accurately and 7 out of 1447 Legitimate SMS were incorrectly detected as Spam SMS for Setting 1. On the other hand, 137 out of 224 Spam SMS were detected accurately and 28 out of 1447 Legitimate SMS were incorrectly detected as Spam SMS.

Therefore, Setting 1 was the best for this model having an accuracy of 96.70%.

***Logistic Regression***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All Features |  | Manually Engineered Features |  |
|  | Legitimate | Spam | Legitimate | Spam |
| Precision | 0.97 | 0.85 | 0.95 | 0.82 |
| Recall | 0.98 | 0.83 | 0.98 | 0.63 |
| Incorrect Prediction | 0.15 | 0.03 | 0.18 | 0.05 |
| Accuracy | 96.17 |  | 92.94 |  |

Table 3: Effectiveness of Logistic Regression in each Setting

This table reveals that 186 out of 224 Spam SMS were detected accurately and 32 out of 1447 Legitimate SMS were incorrectly detected as Spam SMS for Setting 1. On the other hand, 142 out of 224 Spam SMS were detected accurately and 32 out of 1447 Legitimate SMS were incorrectly detected as Spam SMS.

Therefore, Setting 1 was the best for this model having an accuracy of 96.17%.

***Support Vector Machine***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All Features |  | Manually Engineered Features |  |
|  | Legitimate | Spam | Legitimate | Spam |
| Precision | 0.98 | 0.85 | 0.95 | 0.8 |
| Recall | 0.98 | 0.88 | 0.98 | 0.65 |
| Incorrect Prediction | 0.15 | 0.02 | 0.19 | 0.05 |
| Accuracy | 96.23 |  | 93.24 |  |

Table 4: Effectiveness of Support Vector Machine in each Setting

This table reveals that 196 out of 224 Spam SMS were detected accurately and 35 out of 1447 Legitimate SMS were incorrectly detected as Spam SMS for Setting 1. On the other hand, 145 out of 224 Spam SMS were detected accurately and 34 out of 1447 Legitimate SMS were incorrectly detected as Spam SMS.

Therefore, Setting 1 was the best for this model having an accuracy of 96.23%.

Comparison among models reveals that Support Vector Machine built in Setting 1 is the best classifier that could be used to build a predictive model that could accurately predict whether an SMS is a Spam SMS or a Legitimate SMS.

## **2.2 Implementation of Email Spam Filters to Filer Spam SMS**

Many studies indicate that the proposal of implementing content-based email spam filters, to prevent SMS Spam from being sent to the subscribers, was made with a perception that a Spam Filter would be able to perform its job with all types of Spams. However, the most common length of a Spam SMS is 160 characters and makes it inappropriate to be used by content-based filter systems. (Khemapatapan. C, 2010) Also, the type of language used in the two services is quite different; from abbreviated language, emoticons, bad punctuations, etc. being used in SMS to a formal language being used in emails. (Khemapatapan. C, 2010)

Another important concern with the outcomes of previous work was that many legitimate messages resembling the nature of a Spam SMS were also filtered. (Contributions to the Study of SMS Spam Filtering)

Thus, a system should be built which can accurately predict whether an SMS is a Spam SMS or a Legitimate SMS on the basis of features of the SMS. Therefore, in this project, I aimed at working towards exploring the features that differentiate a Spam SMS from a Legitimate SMS.

### **2.2.1 Approach Towards Analyzing Differentiating Features**

Analysis of features that differentiate a Spam SMS from a Legitimate SMS started with plotting ***'Length of Messages' against 'Number of Messages'*** for each label (Spam and Legitimate). I calculated the number of characters for each SMS and plotted them against Number of Messages for each label.

Henceforth, I started with analysis of words that appear most frequently in a Spam SMS. I explored the data and manually engineered features that appear most frequently in a Spam SMS. Example: Winner, Call, Congratulations, Free, etc. Following this, I produced a word cloud for Spam SMS to verify the manually engineered features. *A word cloud is an image consisting of words belonging to a text, document or a dataset, in which the size of the words depict their importance and frequency in that text. (Thesaurus)*

Now, after having a visualized interpretation of all the words that appear most often in a Spam SMS from a word cloud, I made 6 categories. That is,

1. *winner, win, won, award, selected, prize and claim* were put into one category '*Winner*',
2. *congratulations and congrats* were put into one category 'Congratulation',
3. *xxx, babe, naked, dirty, flirty* were put into one category '*Adult*',
4. *urgent, attention, bonus, immediately, now, stop* were put into one category '*Attention*',
5. *free* was put into one category '*Free*', and
6. *ringtone, call, mobile, text, txt* were put into one category '*Ringtone*'.

After having made these 6 categories, I ran for loop for each to assign a 'y' or an 'n' to messages depending on occurrence of words from these categories in the messages. Now, I plotted these 6 categories on a bar plot to determine the most important category of all. Following this, I verified the output of bar plot by using Importance function of Random Forest which extracts importance of all tokens as assigned by randomForest() and plots it. (https://www.rdocumentation.org/packages/randomForest/versions/4.6-12/topics/importance).

### **2.2.2 Outcome of Analysis of Differentiating Features**

Analysis of 'Length of Messages' against 'Number of Messages' for each label *(Refer Fig. 1)* revealed that the most common length of a Spam SMS is 160 characters while the most common length of a Legitimate SMS is 20 characters. But, the length of Legitimate SMS can vary from being as short as just an "OK" in an SMS, making it just 2 characters in an SMS, to being as long as containing around 900-1000 characters in an SMS. Therefore, drastic variance in the length of Legitimate SMS and overlapping of Legitimate SMS with Spam for the entire range of Spam SMS made analysis of differentiating features more important.



Figure 1: Length of Message VS Number of Messages for each Label

Therefore, I produced a word cloud for Spam SMS, which revealed that the most important and frequent words occurring in a Spam SMS are: Call, Free, Now, Text, Txt, etc. *(Refer Fig. 2)*



Figure 2: Word Cloud for Spam SMS

Plotting the six categories (Winner, Congratulation, Adult, Free, Attention and Ringtone) on a bar plot revealed that the most important category is ***Ringtone,*** with words belonging to this category being present in ***602 out of 747*** Spam SMS, and the least important categories being ***Congratulation*** and ***Adult***, with words belonging to these categories being present in ***24 out of 747*** Spam SMS and ***36 out of 747*** Spam SMS respectively. *(Refer Fig. 3)*

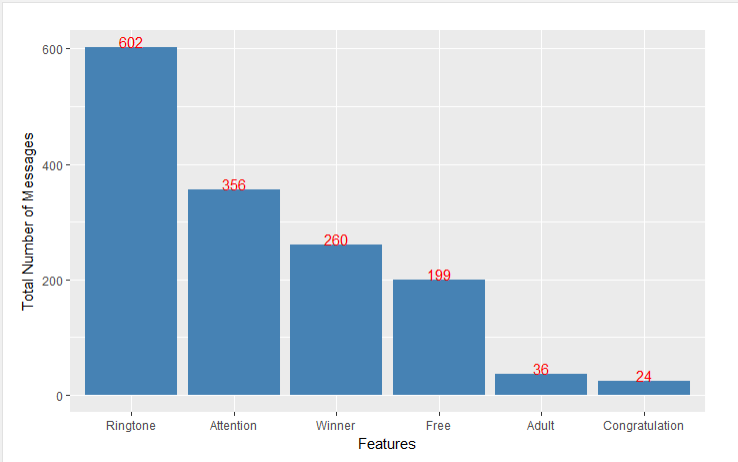


Figure 3: Bar-Plot Depicting Importance of each Category in Spam SMS

The output of bar plot was then verified using Importance function of Random Forest. It substantiated the output of bar plot and confirmed that the most important category for Spam SMS is ***Ringtone***, while the least important are ***Congratulation*** and ***Adult.*** *(Refer Fig. 4 and Fig. 5)*

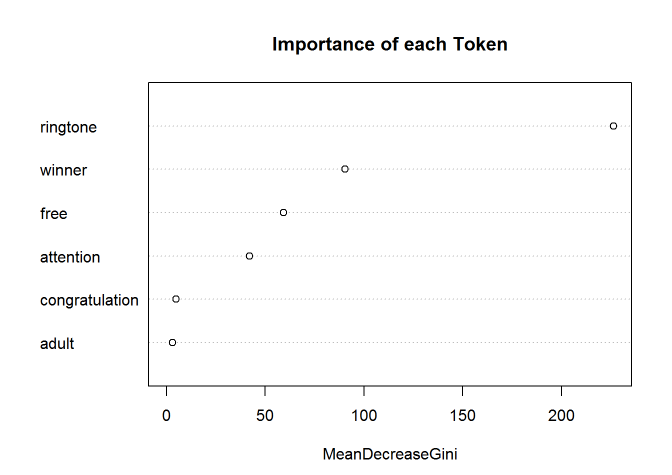


Figure 4: Importance Plot of each category of Spam SMS

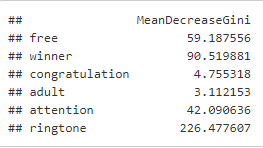


Figure 5: Importance of each Category in Tabular Form