An Analysis of Spam SMS Features

Data Analysis and Research Project

Harshita Jain

n9539361

Supervisor: Dr. Guido Zuccon

Project Background and Context

- 92% of spam SMS are fraud
- Overall rate of receipt grew by 300% from 2011 to 2012.

- Mobile network operators suffer a loss
  - Higher network costs
  - Higher operating costs
  - Increased customer care costs
  - Tarnished reputation

- Annoying for customers
  - Loss of confidential and valuable personal information

Source: http://arrow.dit.ie/cgi/viewcontent.cgi?article=1022&context=scschcomart

Gap in Previous Solutions

- Simple solutions are used - Blacklisting and Spoofing/Faking Detection
  - Brittle by nature
  - Do not take content of messages into account
  - Require on-going management
- Not much data available for research studies
- Available datasets are different

Source: http://arrow.dit.ie/cgi/viewcontent.cgi?article=1022&context=scschcomart

Data Set

v1	v2
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.

Source: Kaggle

Project Purpose and Deliverables

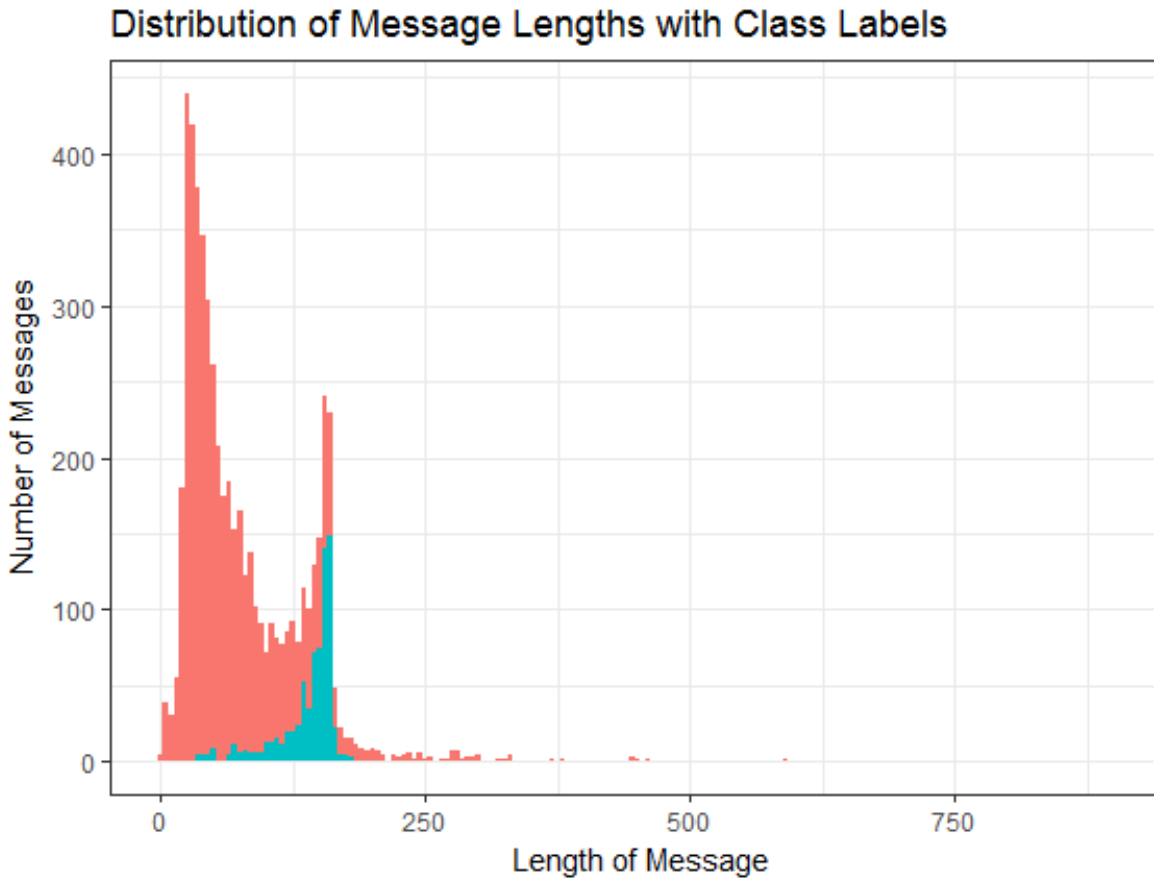
- **Purpose of the Project:**
  - Analyze the data to understand the differentiating features of Spam SMS.
  - Build a predictive model which can accurately predict if an SMS is a Legitimate SMS or a Spam SMS

- **Project Deliverables:**
  - R Markdown
  - Analysis Report

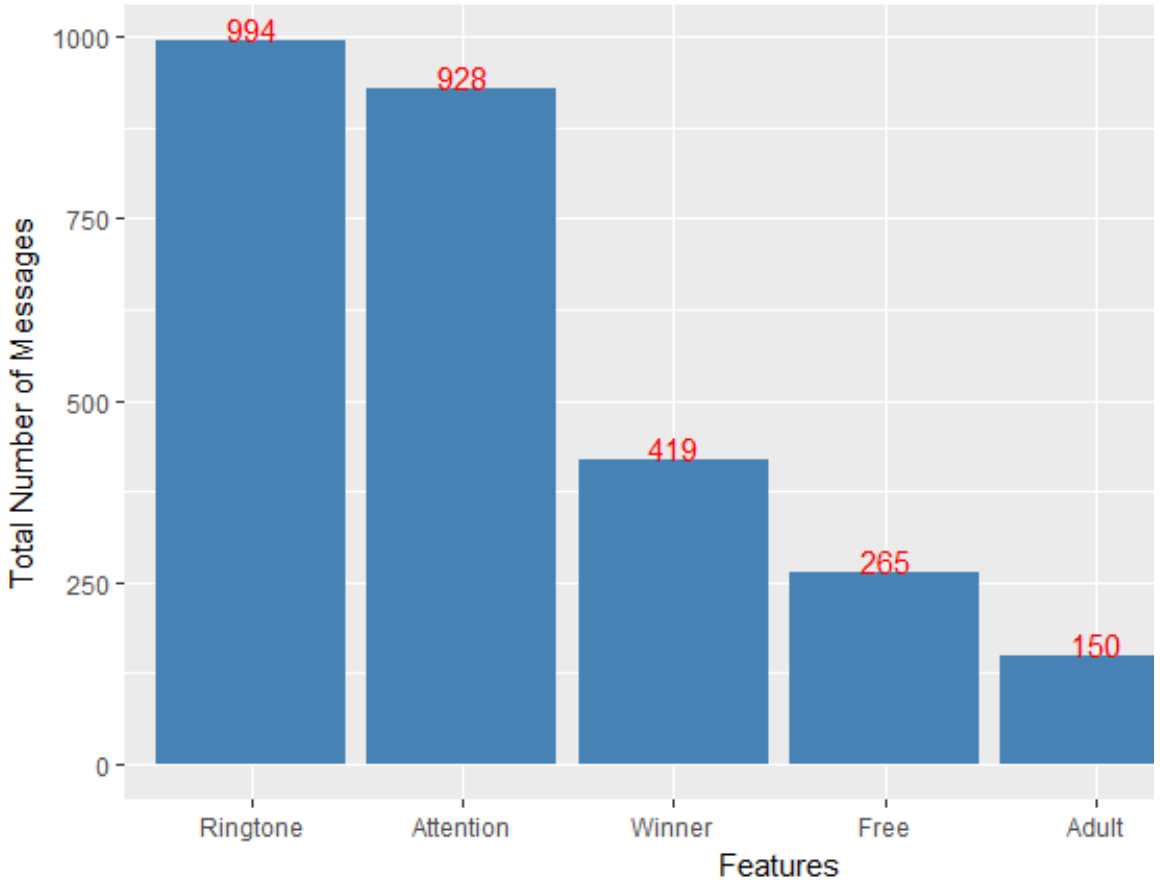
Source: https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

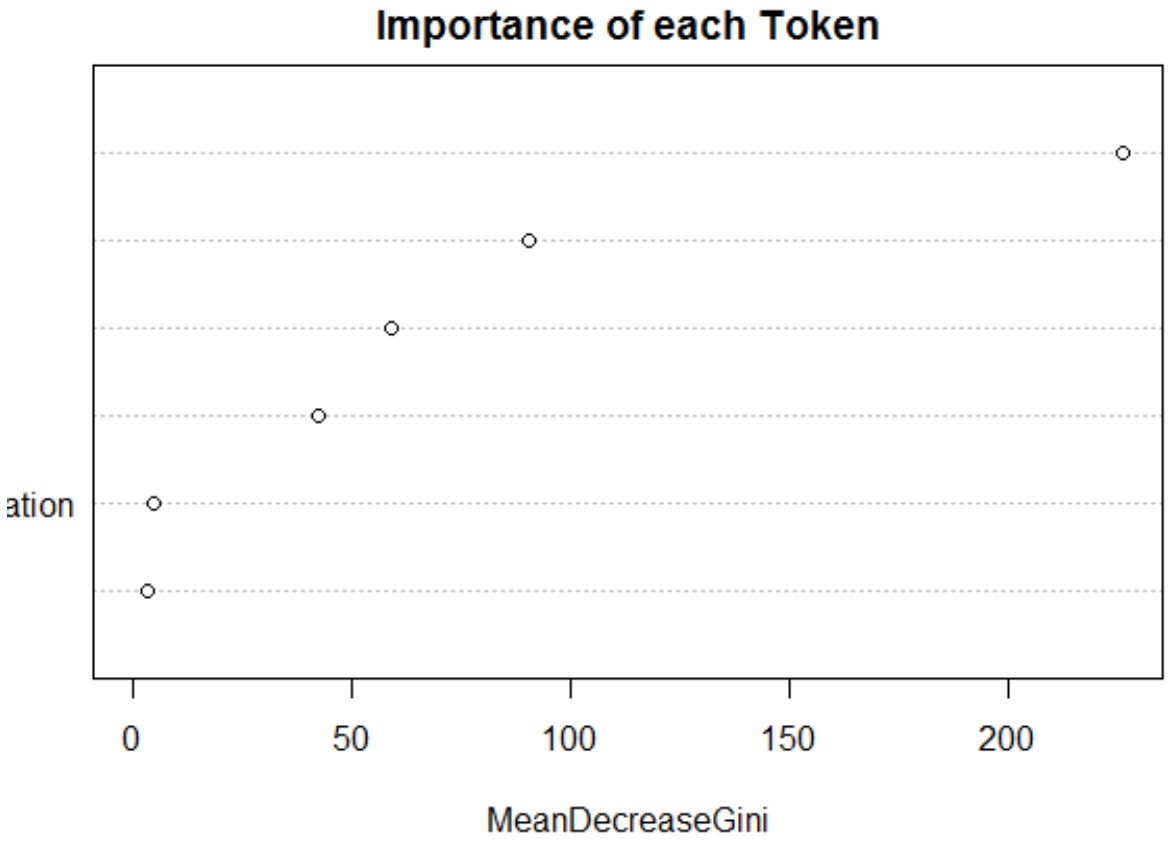
Justification to Claims (1)

- *Analyze the data to understand the differentiating features of Spam SMS.*
  - Explored words that occur most frequently in Spam SMS.
  - Explored Length of Messages



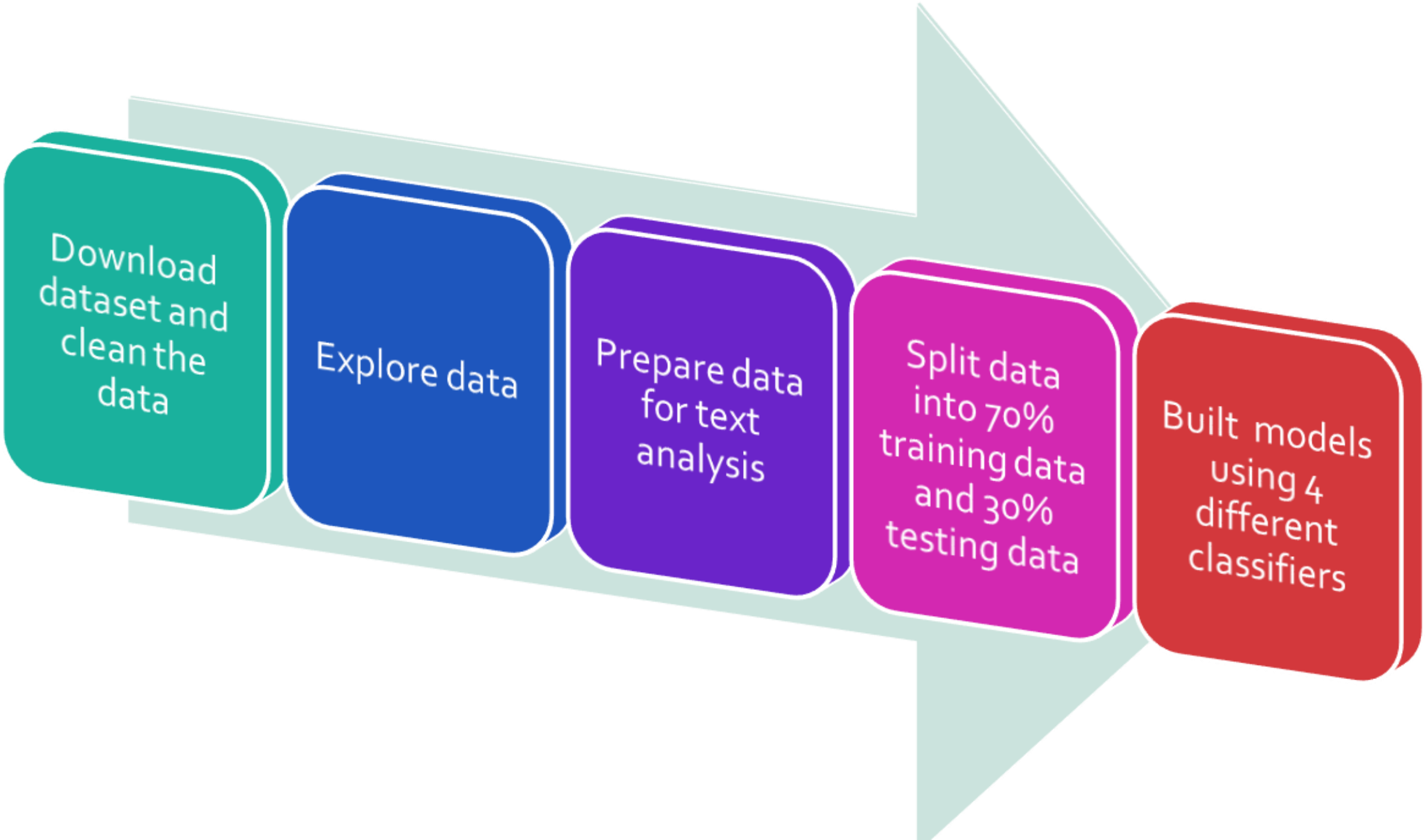
Click to add text


994
928
419
265
150
1000
750
500
250
0
Total Number of Messages
Ringtone
Attention
Winner
Free
Adult
Features


Importance of each Token
ation
0
50
100
150
200
MeanDecreaseGini

Justification to Claims (2)

- *Build a predictive model which can accurately predict if an SMS is a Legitimate SMS or a Spam SMS*
  - *Built 4 models using different classifiers:*
    - *Decision Tree with Random Forest*
    - *Support Vector Machine*
    - *Logistic Regression*
    - *Naïve Bayes*
  - *Built on two types of data:*
    - *Manually explored and selected 6 most important features of the Spam SMS*
    - *All features of the data*


Download dataset and clean the data
Explore data
Prepare data for text analysis
Split data into 70% training data and 30% testing data
Built models using 4 different classifiers

Thank You!!

Any questions?

References

- Team, A. V., Shaikh, F., Jain, K., Gupta, A., & Gupta, D. (2016, October 11). A Complete Tutorial to learn Data Science in R from Scratch. Retrieved August 09, 2017, from https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/
- Guo, P. (2013, October 30). Data Science Workflow: Overview and Challenges. Retrieved August 09, 2017, from https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext
- Delany, S. J., Buckley, M., & Greene, D. (2012). SMS Spam Filtering: Methods and Data. Retrieved from http://arrow.dit.ie/cgi/viewcontent.cgi?article=1022&context=scschcomart
- Whitepapers. (n.d.). Retrieved August 09, 2017, from https://www.cloudmark.com/en/s/resources/whitepapers/sms-spam-overview
- (n.d.). Retrieved August 09, 2017, from https://archive.ics.uci.edu/ml/datasets/SMS Spam Collection
- The DSDM Agile Project Framework (2014 Onwards). (2017, April 18). Retrieved August 09, 2017, from https://www.agilebusiness.org/resources/dsdm-handbooks/the-dsdm-agile-project-framework-2014-onwards