An Analysis of Spam SMS Features

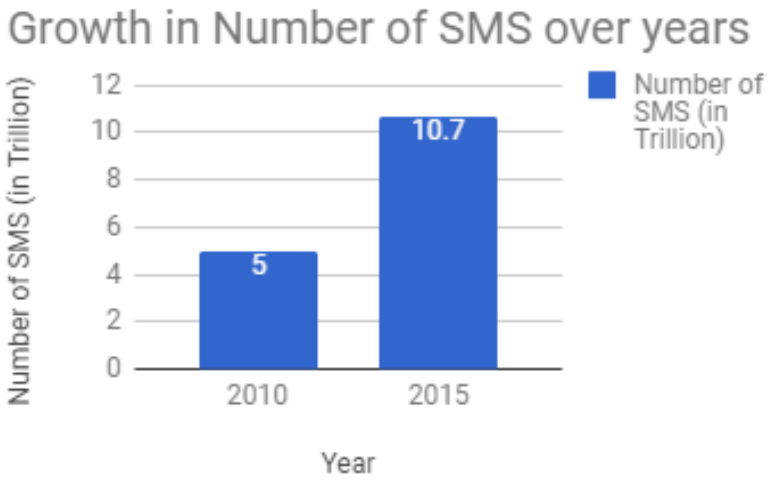
Data Analysis and Research Project

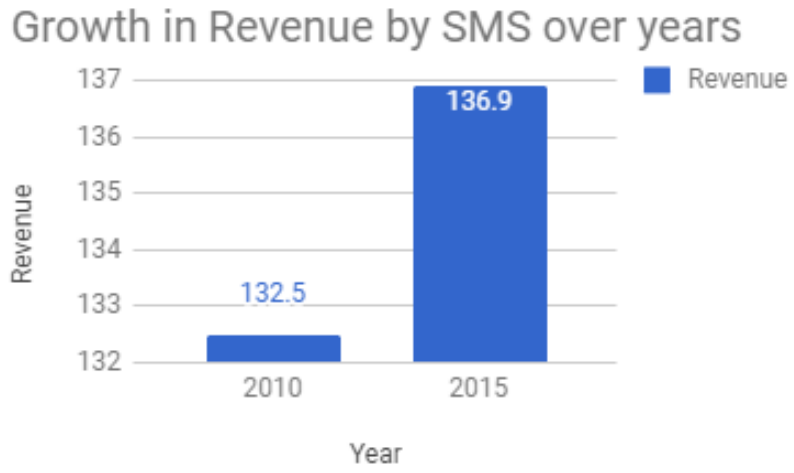
Harshita Jain

n9539361

Supervisor: Dr. Guido Zuccon

Context





- Good proportion of these SMS are Spam
  - Found that 13% are spam SMS through representative data
- 92% of spam SMS are fraud
- Overall rate of receipt of spam SMS is increasing.

Reasons and Effects

Reasons for Proliferation of Spam SMS

- The availability of affordable unlimited prepaid SMS packages
- Customers being more comfortable with sharing their confidential information via SMS
  - 43% of messages were responded in the first 15 minutes of receiving them

Effects

- 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

Gap in Previous Solutions

- Simple solutions are used - Blacklisting and Spoofing/Faking Detection
  - Brittle by nature
  - Do not take content of messages into account
  - Perform in Ad-hoc and Post-hoc manner
- Not much data available for research studies

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 features that make Spam SMS different from a Legitimate 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

APPROACH

Data Analysis and Research

Preparation Phase

Input	Output
Acquired data from Kaggle	Key question for Analysis. Clean data.

Exploration Phase

- Analyzed Length of Messages v/s Number of Texts for each Label
- Manually Selected Differentiating Features of Spam SMS
  - Verified by Producing Word Cloud for Spam SMS
  - Visualized Uni-Grams using Bar-Plots
  - Visualized Bi-Grams and Tri-Grams using Venn Diagram

Input	Output
Clean data	Analysis of features that make a Spam SMS different form a Legitimate SMS

Data Preparation Phase

- Prepared Data to be used to Build Predictive Models
  - Created a Clean Corpus by Transforming Text to Lower-Case, Removing Numbers, Stop Words, Punctuation and White Space.
  - Split the data into 70% Training Set and 30% Test Set

Input	Output
Clean data	Data ready to be used to build predictive models.

Classification Phase

- Built 4 Different Classifiers for 2 Different Settings

Setting 1: Considering all features

Setting 2: Manually engineered features

- Compared each Classifier for each Scenario

Classifiers used:

  - Decision Tree with Random Forest
  - Support Vector Machine
  - Logistic Regression
  - Naïve Bayes

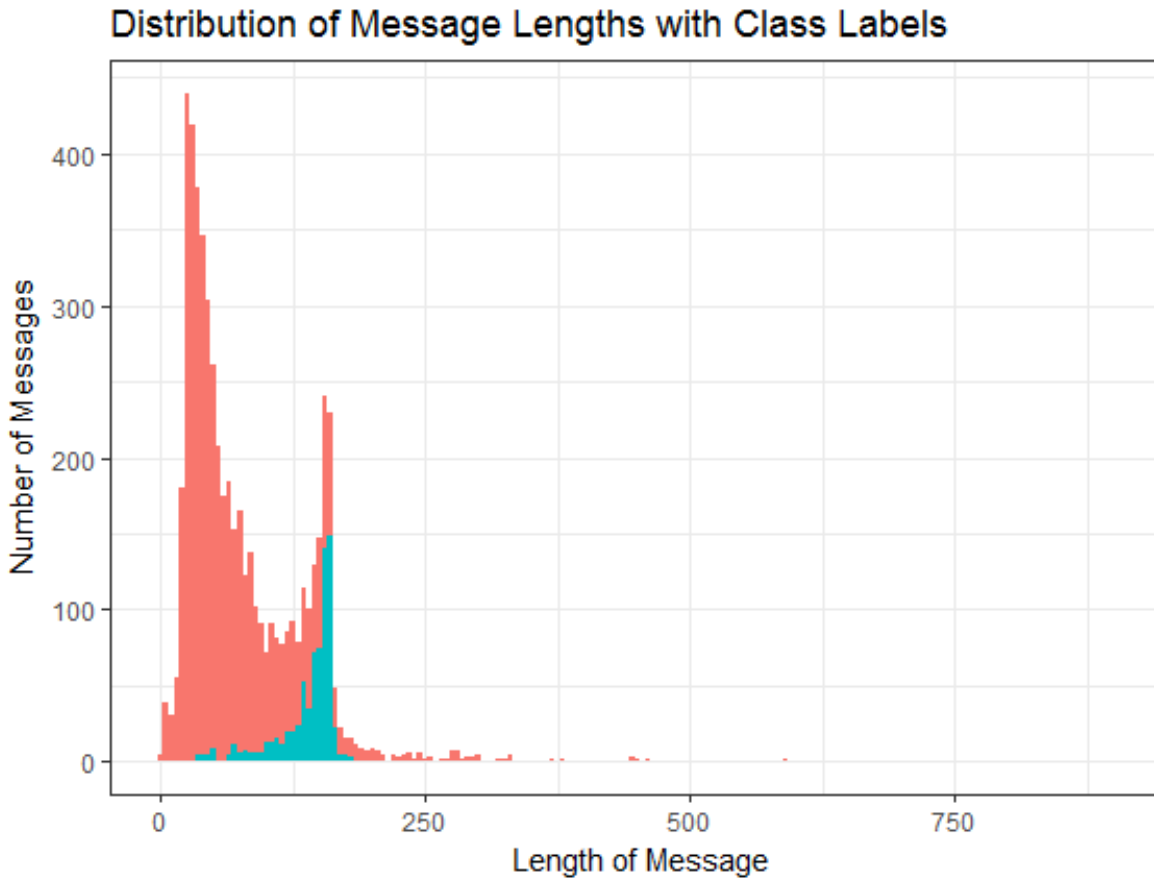
Input	Output
Data from Data Preparation Phase	Precision, Recall, F1 and Accuracy Measures for each model in each scenario Best Scenario for each Model

Final Delivery Phase

Input	Output
Output from Exploration Phase Output from Classification Phase	Code in R Markdown Analysis Report

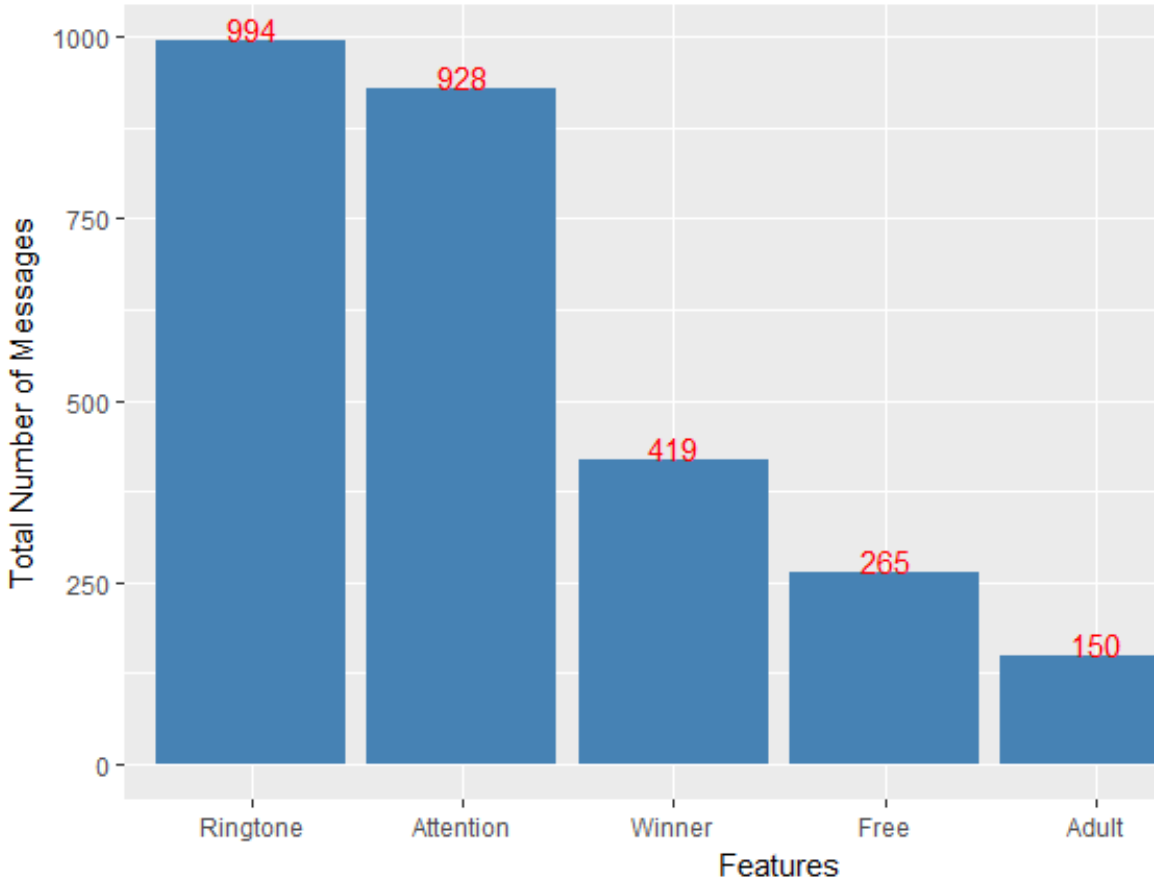
Work Done (1)

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



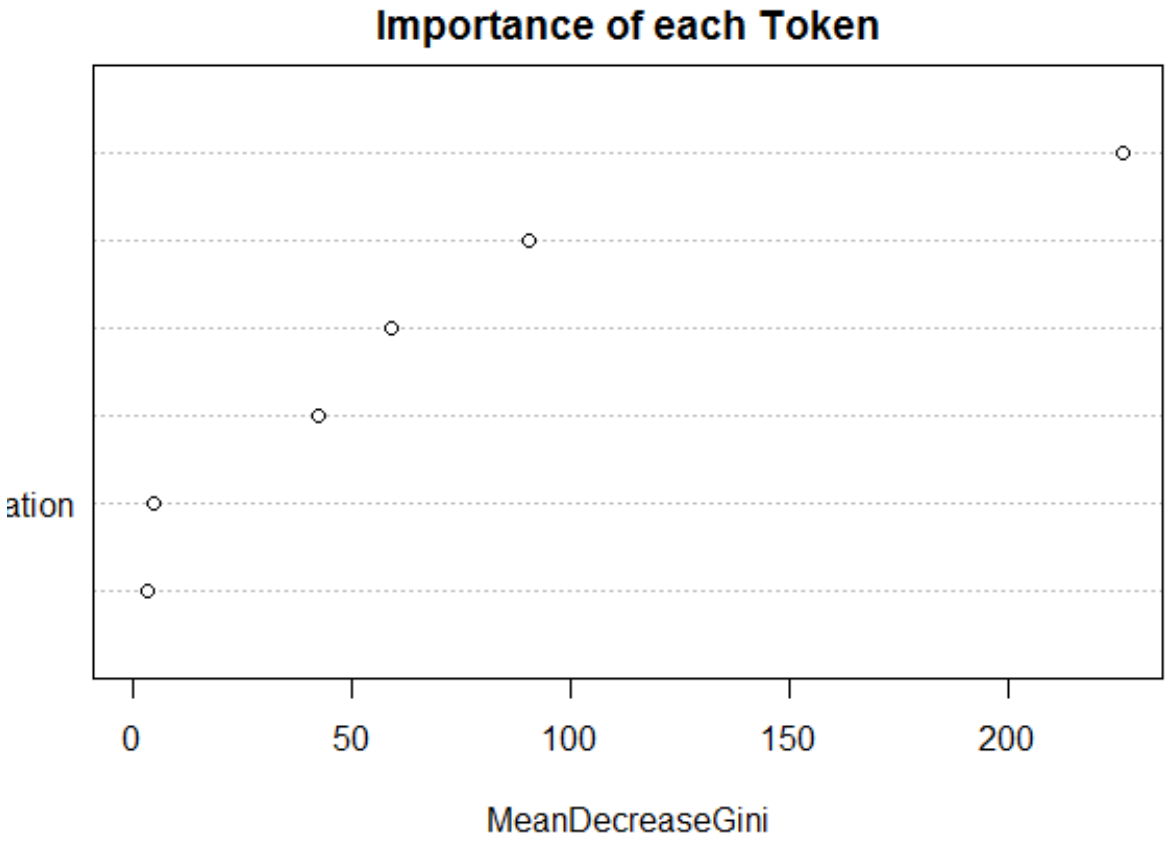
Distribution of Message Lengths v/s Number of Texts

Word Cloud for Messages tagged as Spam



Features v/s Number of Messages

Importance of each Feature



Work Done (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 settings:*
    - Considering all features
    - Manually engineered features

Naive Bayes

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

Support Vector Machine

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

Generalized Linear Model

	Manually Selected Features		All Features	
	Legitimate	Spam	Legitimate	Spam
Legitimate	1415	32	1415	32
Spam	79	142	38	186
Accuracy	92.94		96.17	

Decision Tree - Random Forest

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

Implications of Work Done

Improvements in Filter System

- Replace old solutions like Blacklisting, Spoofing and Faking detection Techniques
- More dynamic in nature: Will only allow ham SMS to reach the recipient, and not the spam SMS.

Benefits to Stakeholders

- Mobile Network Operators
  - Implement SMS Spam Filters
  - Improve SMS Quality and Services to Customers
  - No Overhead Costs to Maintain the Quality
- Consumers
  - Protected Confidential Personal and Valuable Information

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