Project Report on

EMAIL and SMS Messaging Fraud discovery and Classification

Mallesham Yamulla (40AIML288-21/1)

2022-02-27

Contents

Phase-1: Literature survey & Data Acquisition

1.1 Problem Defination

1.1.1 Introduction, Objective and Goals

We are into Threat Intelligence Unit where we do fight against Spammers who keeps on sending the different kinds of Spams to the mobile subscribers, and these spam attacks can be getting through

- EMAIL TO SMS and EMAIL TO MMS
- SMS
- Grey Routes/SIM Banks
- SS7 Signaling/ Diameters

We have already built up the two products called NPP (Network Protection Platform)and SIGIL(Signalling Intelligence) to control the attacks being happened on the above mentioned messaging systems

In this Project, I'm going to work on EMAIL and SMS messaging traffics only, first I would like to give a small introduction on how the messaging gets routed from Spammers to Subscribers, how our team TIU can block the suspicious, unsolicited , bulk messages out from the networks using the different tools

The mobile subscribers go on receiving the variety of messages (Personal, Business, Etc Etc..), from the known-unknown persons-companies., here It's little difficult for them to figure out a message to be Legit or Spam

In this case, they are supposed to pass these suspicious/doubtful/unwanted messages to our system called 7726 that gets these complaints registered for further investigation, these 7726 complaints would also be sent in to our internal tool called Security Center where we can classify the messages to their respective categories

We have the following 8 different types of Spam Categories in our systems.

- LEGIT-CLEAN Messages
- SPAM- SCAM
- SPAM-LOAN
- SPAM-PHISHING
- SPAM-ADULT
- SPAM-MARKETING
- SPAM-GAMBLING
- SPAM-MALWARE

Our team is given an access to roll though the 7726 logs, carry out an investigation using the listed principles, intelligence and take a necessary action if a message is found to be a spam and these can be of

- SMS
- EMAIL-TO-SMS/EMAIL-MMS

We generally bank upon the following given filtering methods that are part of our NPP product to block out the spams,

- Fingerprinting
- URL Blacklist
- Content Matching
- IP Blacklists
- Domain Blacklists
- Reputation Filters
- Regular Expressions
- And many others

The messaging classification can only be done via Fingerprinting Filter in Security Centre that connects to NPP.

Fingerprinting Filter:

Once messages are sent into Security Center an analyst from Team has to look through them, and apply a fingerprint to a message by moving it to the SPAM or LEGIT Category so that it gets assigned a 13 digit unique hash code, stored in SC and looks for its variants in the traffic that comes in.

If a fingerprinted messages match with any other messages taking an account of its configured similarity(about 80%) its cluster size would go on increasing and here cluster size means the total No.of similar messages.

If a fingerprint is of SPAM category all of these messages would be blocked out and on the other hand legit fingerprinted messages would be getting allowed.

Challenge:

Our Teams works in three different time zones (US-EUROPE-INDIA) 24-7, here weekly, the messaging traffic gets stored in our systems about 3-5B messages and 10-15mil of them are of spam messages. And Per day about 10K messages which have come from 7726 system would be manually classified in Security Center by Analysts.

In this continuous manual messaging marking process, we have to get through the below mentioned issues,

- MisClassifications of Messages

- Higher False Positive and False Negatives
- Time Taking process in classification
- Spam leakages

As the messaging complaints sends in faster from subscriber our team should proactively mark and classify the messages one by one hourly basis, in his/her manual classification there could be a change to get misclassified messages, as explained below,

Example:

- SPAM Phishing category to SPAM Loan
- SPAM Loan Category to SPAM Scam
- LEGIT Message to any one of SPAM category
- Any of Spam Message to Legit Category

And the first step misclassifications leads to improve the rates of False Positive and False Negative rate which can't be acceptable by Businesses.

False Positive Message - It's a case when a LEGIT message is found to classified as SPAM message

False Negative Message - It's a case when a spam message is found to classified as LEGIT message

In the second place, Analyst has to spend a more time to make up his mind to classify and take an action on messages when he/she comes across any new spam campaigns day to day, here there might be spam leakages if they are not actioned on time.

Here our business goal would be that the current email spam blocking rates should be improved about 20% from around 10% on average while meeting the above specified requierement. In addition to it, the following business questions would also be required to look through

1. How are the spam traffic volumes/behaviours over week/days/hours?.
2. are the spam patterns getting changed frequently ?
3. is there any effect of manual messaging classification done by fingerprint filter on controlling spams ?
4. estimating fingerprint filter active blocking time ?
5. any delays in classifying a message once its on Security Center?

1.1.2 Voice of Customers:

A. Classify messages correctly

1. Proper guidelines to be followed in classification of messages
2. Experienced analysts
3. Security Centre (SC) GUI appearance / flexibility / smarter features for manual classification

B. Make sure over blocking of legitimate message/lower blocking of spam messages wouldn't get occurred

1. Carry out daily checks on the classified messages to figure out false positive and false negatives
2. Monitor Cartridge updates

C. Get Spams blocked out as faster as possible and Improve spam blocking rates

1. Look for spams in logs continuously
2. Do regular query searches on Database.
3. Do bulk uploads of better spam candidatures.
4. Get fingerprint fiter algorithm worked more efficient
5. Do experimenting with text analytics

1.1.3. Process Map

A picture is worth a thousand words, considering this statement here a process map is a diagram that provides a visual representation of the process flow, or sequence of activities, or steps that take place in a process from start to finish.,the flow can go from top to bottom or left to right.

A process map enables us to see and understand the process.once a current process has been mapped, the team will also know what's not happening or what's different from what should be happening.

In our process map, It starts from messaging being sent out from spammers towards telecom network through which users receive messages,here it could be personal or business related messages. Users generally want to get a benefit from legitimate messages which give them a useful information regarding a service or a conversational message.

Spammers take an advantage of user's need and go on pumping in unwanted/suspicious/bulk messages to them for doing a fraud, User's wouldn't be aware of these messages or found them to be a harmful for them, So here they would have a **system 7726** where these messages can be reported for furhter investigation.

The reported messages to 7726 system will be pushed to a system called **Security Center** in which they are classified, blocked out and controlled from not being sent on to users if they are really spams.

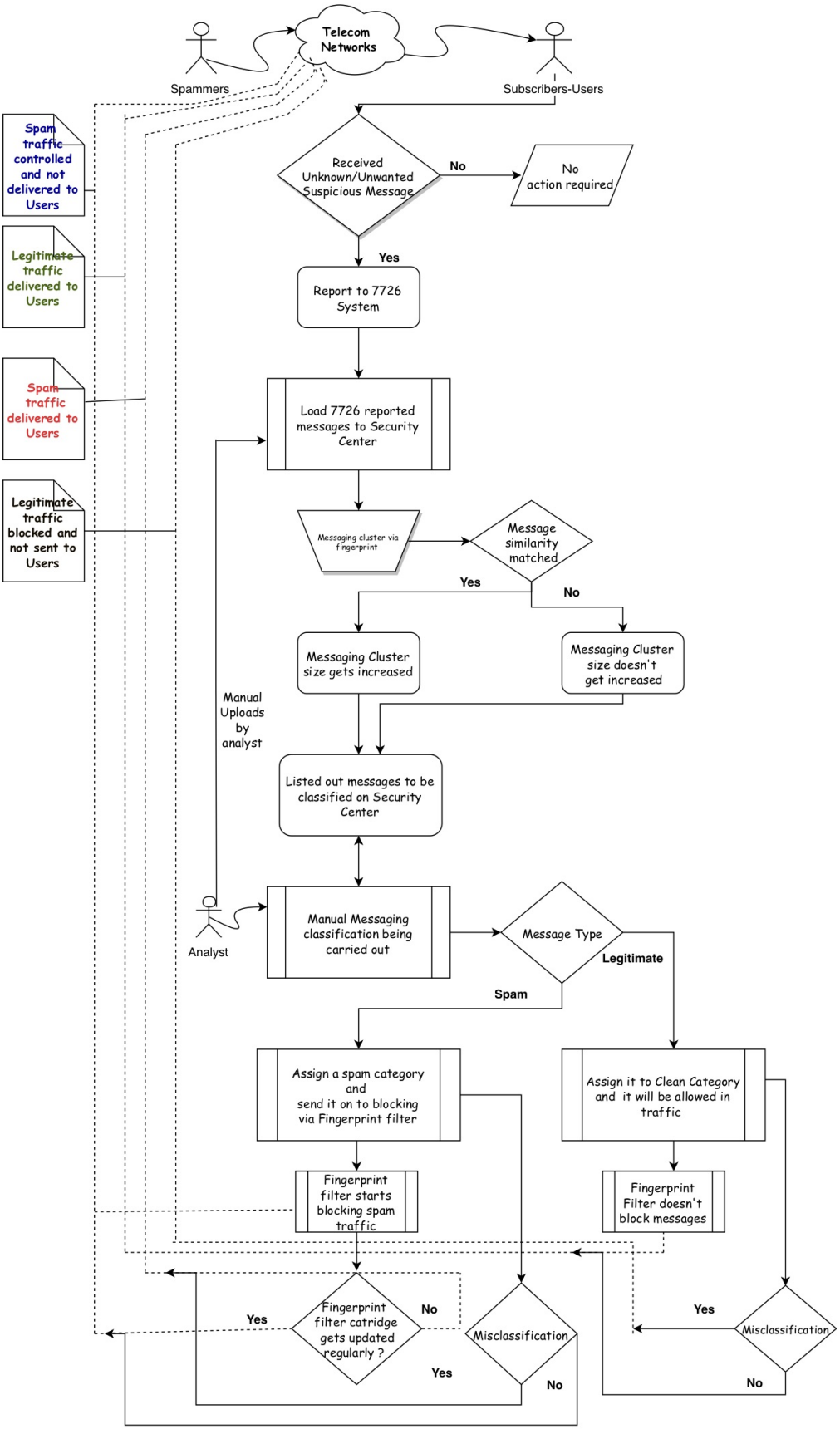
The SC (Security Center) is continuously being looked through by analysts in 24/7 hours, SC mechanism works like in this way: a message gets matched with another message considering the configured similarity percentage in fingerprint filter, and they goes on getting clustered as long as its similarity matches.

Analysts would often go for a method to upload messages that are collected from another system called TSM where the entire messaging traffic gets stored into Security Center to controll spams effectively.

Once these are listed out in SC Analyst would start classifying them manually banking upon his intelligence/techniques,the classified spam messages are being tripped and blocked out by a fingerprint filter, here spam traffic get handled and not sent to users and in another case the classified legitimate messages are being allowed towards users.

There could be a chance to happen a misclassification on messages that reached to SC, Analyst intelligence would go wrong sometimes and it leads to increase rates of false positives and false negatives, spam leakages as well,as explained here, a legit message is wrongly classified to spam, hence these messages wouldn't be delivered to users, on the other hand a spam message is categorized as legit,in this case spams would be flown to users and they get effected.

The process gets stopped with blocking out spams and allowing legit messages towards Users.


Telecom Networks
Spammers
Subscribers-Users
Spam traffic controlled and not delivered to Users
Legitimate traffic delivered to Users
Spam traffic delivered to Users
Legitimate traffic blocked and not sent to Users
Received Unknown/Unwanted Suspicious Message
No
No action required
Yes
Report to 7726 System
Load 7726 reported messages to Security Center
Messaging cluster via fingerprint
Message similarity matched
Yes
No
Messaging Cluster size gets increased
Messaging Cluster size doesn't get increased
Manual Uploads by analyst
Listed out messages to be classified on Security Center
Analyst
Manual Messaging classification being carried out
Message Type
Legitimate
Spam
Assign a spam category and send it on to blocking via Fingerprint filter
Assign it to Clean Category and it will be allowed in traffic
Fingerprint filter starts blocking spam traffic
Fingerprint Filter doesn't block messages
Fingerprint filter catridge gets updated regularly ?
Yes
No
Misclassification
Yes
No
Misclassification
Yes
No

1.2. Data collections

1.2.1. What type of tools/libraries have been made use of? and its setup

I have decided to get this project done using Python since it's an open source tool whose community has been growing up in Statistical analysis and Machine Learning fields.

- PostgreSQL (data store)
- sqlalchemy (data management using python)
- pandas(EDA)
- pydatatable(EDA faster process)
- spacy and regular expressions (Text Analytics)
- gggplot/seaborn/altair(Data Visualizations)
- scikit-learn(Machine Learning)
- tensorflow(Machine Learning and Deep Learning)
- streamlit (Deployments)
- rmarkdown(Reporting)

Here is a flow diagram that shows a roadmap for creating a messaging classifier that predicts whether a sending messaging is a kind of spams or legit.

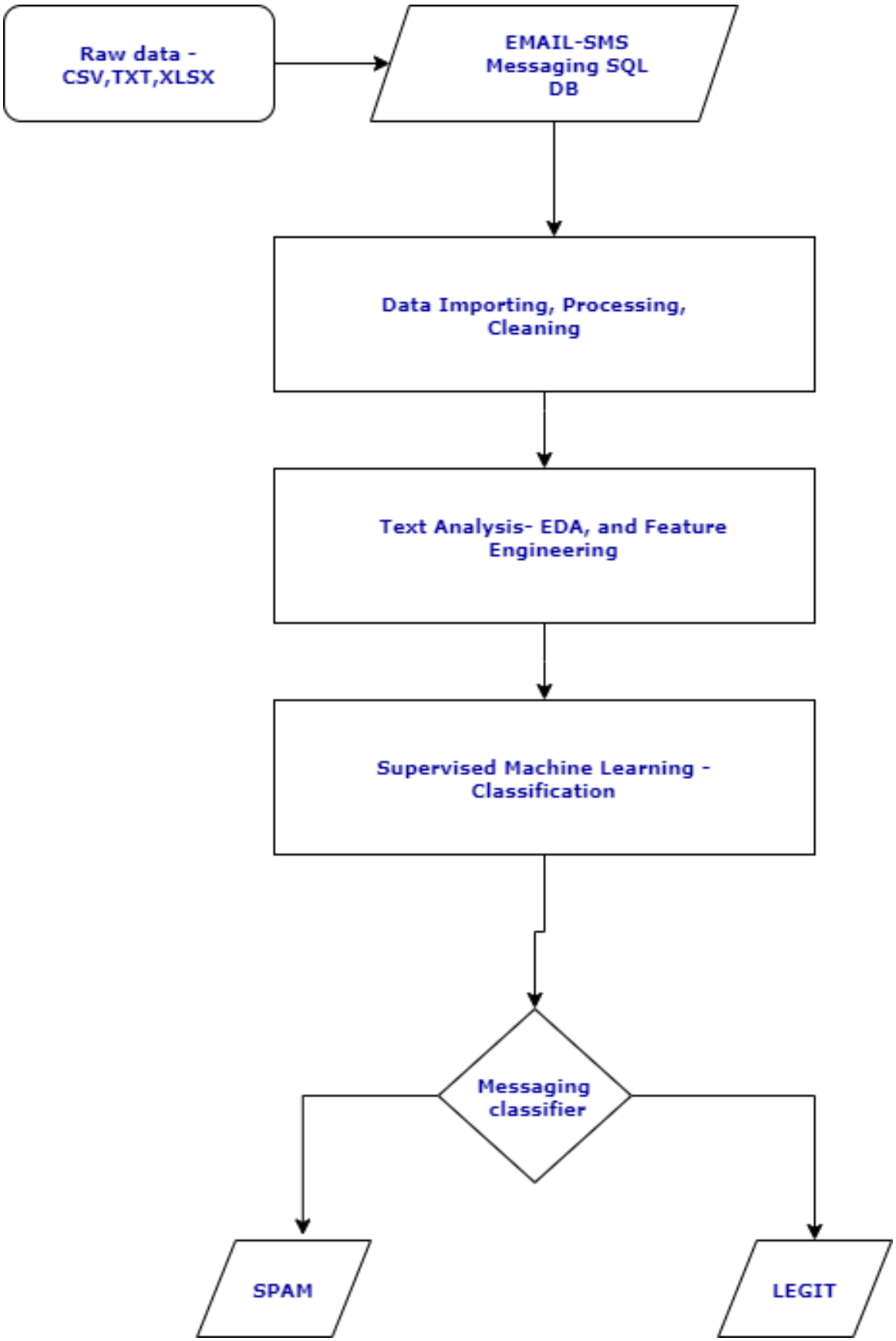


Figure 1: Classification Map

1.2.2. Introduction to Data and ETL

Our team of analyst classify the message the messages reached to the system called Security center, Two expert analysts of our team are dedicated to look through a file that gets generated with the last day messages classified by all the analysts from 3 time zones and reclassify the messages if needed, here they would have to spend at least half of the day (4hrs) for this task itself. On a daily basis the false positive and false negative checks have to be verified so that there wouldn't be an over blocking/lesser blocking of messages.

A single file would consist of about 5-10K classified messages from Security Center, for this project I have gathered these files(in txt, xlsx,csv types) for a period from February 2021 to October 2021.

I have loaded these files into PostgreSQL database, and stored them in a single data source db file and overall there are about 2.5 Million Observations with 3 Variables.

Message_id: A unique hash code for a message

Message_content: message text and it may contain URLs, Call to Actions, notifications etc etc

Message_flag: Indicates whether a message is classified as spam or legit

Here from the field message content the below 4 extra fields are created by making use of regular expressions.

has_URL: It tells whether a message contains any of URL's an

has_CTA: it also tells whether a messages contains any of call to actions such as phone numbers or email id's etc etc

URL_Domain: URL domains are extracted from a messages whose has_URL field is True

URL_tidy: Removed all the unnecessary strings/subdomains from a URL, extracted only a main domain and replaced the symbol . with dot such as (ams.com - amsdotcom)

Table 1: messages table - raw version

message_id	message_preview	message_flag
sc3YawezK7FS	Hi! Anika Schweitzer wants to meet with you on http://sexfriends.mygirlfriends.us	spam
GohJ7Yui00wD	Ok, it's me Natalie http://paper11-almost.center/J5D1/78	spam
GweuBEZyy0CH	Addison Foreman invite you to dating on http://sexfriends.mygirlfriends.us	spam
yglJKUei6Lbn	Eva Latshaw invite you on http://sexfriends.mygirlfriends.us	spam
W28pvGVFvjY2	info K & find my profile (Sheila,36) http://bit.do/ecuCS i am	spam
kfTIHsgCUW5J	Julia from Hartly update her profile pics http://paper11-almost.center/J65T/5a	spam
O4h9BOrJuvDr	(2/3) Intention Here => bit.ly/MarchBlueMoon2018 ? Shine on ~Anne	spam
Ls1H3KO8mI5w	2 hotties from McMinnville http://paper11-almost.center/J5FT/d7	spam

Table 2: messages table - tidy version

message_id	message_preview	message_flag	has_URL	has_EMAIL_CTA	URL_Domain	url_tidy
8ID14pBPqoPX	good afternoon morris number person hated made arrangement pbpicdotsite	spam	1	0	['http://pbpic.site']	['pbpicdotsite']
zyVYzFIZ14YK	alert yvanna kierra gift delayed check address correct 5nsxjmcdotsite	spam	1	0	['http://5nsxjmc.site']	['5nsxjmcdotsite']
pZ8TcMyHxBxD	mentioned photo raul view timeline pyspfdottop	spam	1	0	['http://pyspf.top']	['pyspfdottop']
5FnWCWTUDKIt	sue manager digitextracted hour digitextracted bonus working livonia driveshopdotus reply stop cancel	spam	1	0	['http://driveshop.us/2e54e1ff.']	['driveshopdotus']

1.3. Introduction to KPI

Our business problem can be solved by making an experiments using the supervised machine learning classification techniques.

Here below we are defining the KPI's to be achieved/optimized for the proposed business problem.

- Spam classification accuracy
  - We are seeing the spam classification accuracy about **55-60%** with manual classifications carried out by team of analysts using the human intelligence, and we are looking for improving the spam classification accuracy by making the automated messaging classifiers which are expected to improve an accuracy at least 80-85%
- Improve False Negative Rates(FNR)
  - Make sure not to happen a lot of spam leakages - we are not observing much False Negative Rates while doing the manual classification system, our business are OK to accept the minimal false negative rates.
- Decrease False Positive Rates
  - Avoid legitimate messaging blocks - we are experiencing a lot of false positives are being caused in manual messaging classification process, here business are not accepting them as it causes a lot of economical, reputation losses on organization products and the telecom operators subscribers are also affected much on it as their legit messages are blocked out in networks. so we are trying to decrease the false positive rates by implementing the machine learning classifiers.

4. Failures, contraints,challenges and solutions

1.4.1 FMEA (Failure Mode and Effects Analysis)

It is an engineering technique used to define, identify and eliminate known and /or potential failures, problems, errors and so on from the system, design, process, and or service before they reach the customer.

Any FMEA conducted properly and appropriately will provide the practitioner with useful information that can reduce the risk (work) load in the system, design, process and service.

Used to analyze services before they reach the customer. It focuses on failure modes (tasks, errors, mistakes) caused by system or process deficiencies.

Process step 1: Load messages from 7726 system to Security Center for classification

Potential Failure Modes, effects and causes:

1. There is a chance of getting Late report of suspocious/unwanted messages to 7726 system from Users due to this classification can't be done on time and spam can get leaked towards users, this failure can get occured where there is no proper connectivity in between SC, and 7726 systems
2. API Calls gets down between 7726 and SC when there is a huge flow of traffic in between them, when it happens no message will be appeared on SC, because of it no classification can be done and spam gets leaked towards users. here cause could be connectivity issues in SC and 7726 systems.
3. Security Center Storage Disk gets filled up when there is no memory space in SC, since then no message would get into Security center, it leads to no message gets classified and spam gets leaked towards users, here cause could be insufficient storge system built up in SC.

Process Step 2: Message Clustering

Potential Failure Modes, effects and causes:

Different variants/patterns of messages keep on getting sent out from spammers, in this case analyst has to spend more time to understand a message and get it classified to it's respective category,if he come acrooss 100's of variatns it would definetly be a herculean task to do classification, when there is no on time action taken on a spam message it would get leaked and lower blocking can also happen, here cause could be an Ineffective Fingerprint filter algorithm that we have been using in spam filtering.

Process step 3: Manual Messaging classification by analysts

Potential Failure Modes, effects and causes:

- Misclassification

  - Over blocking/lower blocking caused by Inadequate training/Knowledge transfer
- Delayed in classification
  - Spam leakages caused by Inadequate training/Knowledge transfer
- Delayed in processing classified messages
  - Spam leakages caused by Ineffective Fingerprint filter algorithm
- Security Center crashes down
  - Spam leakages caused by Improper System design/Architecture
- Inflexible Security Center UI/UX visibility/options
  - Time consuming for manual classification caused by Improper System design/Architecture
- No proper monitoring of logs during busy shift hours
  - Spam leakages caused by Shortage of analysts in team
- Deprioritize the classification task
  - Spam Leakages caused by Inadequate training/Knowledge transfer
- Left messages unclassified for longer time in Security Center
  - spam Leakages caused by shortage of analysts in team

Process step 4: Manual uploads of messages from TSM system to SC

Potential Failure Modes, effects and causes:

Analyst would have to look for spams on a system TSM where all the messaging traffic gets stored, sometimes users might not be complaining about unknown/unwanted messages to 7726, that's why the non reported messages are not appeared on to SC. in this case analyst would have an option to get the messages from TSM and upload it into SC via .csv files. there is a chance of misclassification of messages when a manual upload is carried out that leads to messaging traffic overblocking/lower blocking, here a cause could be Ineffective Fingerprint filter algorithm.

Process Step 5: Regular Interval Fingerprint filter catridge updates

Potential Failure Modes, effects and causes:

Once messages are classified as SPAM, they go on getting blocked out via fingerprinting filter, here fingerprinting filter gets updated in an time intervals i.e for every 3-5mins.In case of any delay occured in this update time there would be spam leakages as fingerprinting filter stop blocking spams. here cause could be Ineffective Fingerprint filter algorithm

Process Step 6: Review of Classified messages

Potential Failure Modes, effects and causes:

On everyday, a .csv file gets exported automatically with the messages classified on earlier day to figure out if there are any false positives or false negatives caused. this file can be review by an expert analyst without fail, here there is a duplication of work for the same task, and even an expert analyst might misclassify messages manually if he has messed up with other taks or doen't have an idea about it. here cause could be Inadequate training/Knowledge transfer of analyst

Process/ Function	Potential Failure Mode	Potential Effects of Failure	SEV	Potential Cause(s) Mechanism(s) of Failure	OCC	Current Process Controls	DET	RPN
Load messages from 7726 system to Security Center for classification	Late report of suspocious/unwanted messages to 7726 system from Users	No classification, Spam leakage	9	Connectivity issues	1		2	18
	API Calls gets down between 7726 and SC	No classification, Spam leakage	9	Connectivity issues	1		2	18
	Security Center Storage Disk full	No classification, Spam leakage	9	Insufficeint Storage	1		1	9
Message Clustering	different variants/patterns of messages	time consuming for manual classification, lower blocking	9	Ineffective Fingerprint filter algorithm	9		9	729
							2	0
Manual Messaging classification by analysts	misclassification	over blocking / lower blocking	9	Inadequate training/Knowledge transfer	9		3	243
	Delayed in classification	Spam leakages	8	Inadequate training/Knowledge transfer	8		9	576
	Delayed in processing classified messages	Spam Leakages	9	Ineffective Fingerprint filter algorithm	8		9	648
	Security Center crashes down	Spam leakages	6	Improper System design/Architecture	3		3	54
	Inflexible Security Center UI/UX visibility/options	time consuming for manual classification	2	Improper System design/Architecture	2		2	8
	No proper monitoring of logs during busy shift hours	Spam leakages	6	Shortage of analysts in team	5		5	150
	Deprioritize the classification task	Spam Leakages	7	Inadequate training/Knowledge transfer	5		5	175
	left messages unclassified for longer time in Security Center	spam Leakages	8	shortage of analysts in team	5		5	200
Manual uploads of messages from TSM system to SC	misclassification	over blocking	9	Ineffective Fingerprint filter algorithm	6		3	162
	Security Center crashes down	Spam leakages	6	Ineffective Fingerprint filter algorithm	5		3	90
Regular Interval Fingerprint filter catridge updates	Delayed in catridges updates	Spam leakages	7	Ineffective Fingerprint filter algorithm	6		5	210
	Shorter time-to-live of catridge in blocking	lower blocking/spam leakages	9	Ineffective Fingerprint filter algorithm	7		8	504
Review of Classified messages	misclassification	Over blocking, time consuming	7	Inadequate training/Knowledge transfer	6		5	210
	Duplication of work	time consuming	5	Inadequate training/Knowledge transfer	5		5	125

Figure 2: FMEA

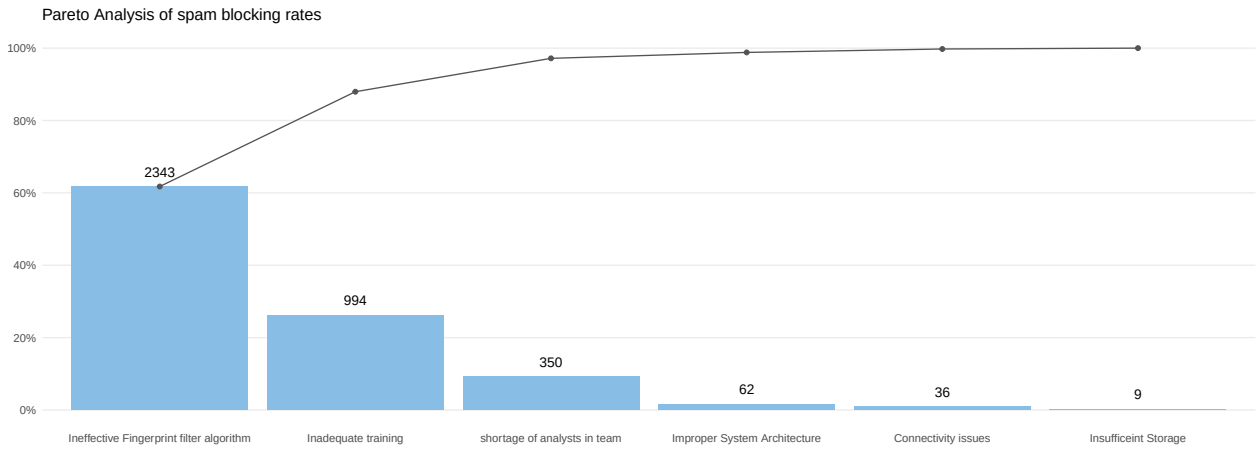
1.4.2 Pareto Analysis

Pareto analysis is a statistical technique that is used in decision making for the selection of the limited number of tasks that produce the most significant overall effect. It uses the concept based on identifying the top 20% of causes that need to be addressed in order to resolve 80% of the problems.

The potential causes for having recorded lower/improper spam blocking rates are identified as

1. *Insufficient storage*
2. *Connectivity issues*
3. *Improper system architectures*
4. *Shortage of analysts*
5. *Inadequate training*
6. *Inefficient fingerprint filter algorithm*

These causes can be drawn on Pareto chart to investigate which one of them have occured more to make spam blocking rates down or let spams allowed.



From the above Pareto diagram it is insight that about **95% of spam leakages/improper blocking has happens because of having ineffective fingerprint filter algorithm, inadequate training to analyst and shortage of analysts in team**.

The machine learning based classifiers are required to be replaced the rule based fingerprinting algorithm so that the accurate spam classifications are achieved which leads to improve the spam blocking rates as well.

1.5. Approaches and Literatures

We have developed our anti-spam products with rule based algorithms which get fed from text similarities, frequencies, volumes of messages, we have a larger messages corpus which would be helpful to build a machine learning algorithms and in built in our anti-spam products. Our business objective is to make a decision on a message whether it is a SPAM or LEGIT, so its a kind of classification problem to which there are many solutions are avail in supervised machine learning techniques. we would make use of the below classification algorithms,

- Logistic regression
- Naive Bayes
- Decision tress and Random forest classifiers
- Word2Vectors and Gradient Boosting Decision Trees

Phase-2: Exploratory Data Analysis and Feature Extraction

Exploratory Data Analysis (EDA) is used on the one hand to answer questions, test business assumptions, generate hypotheses for further analysis. On the other hand, we can also use it to prepare the data for modelling. The thing that these two probably have in common is a good knowledge of our data to either get the answers that we need or to develop an intuition for interpreting the results of future modelling.

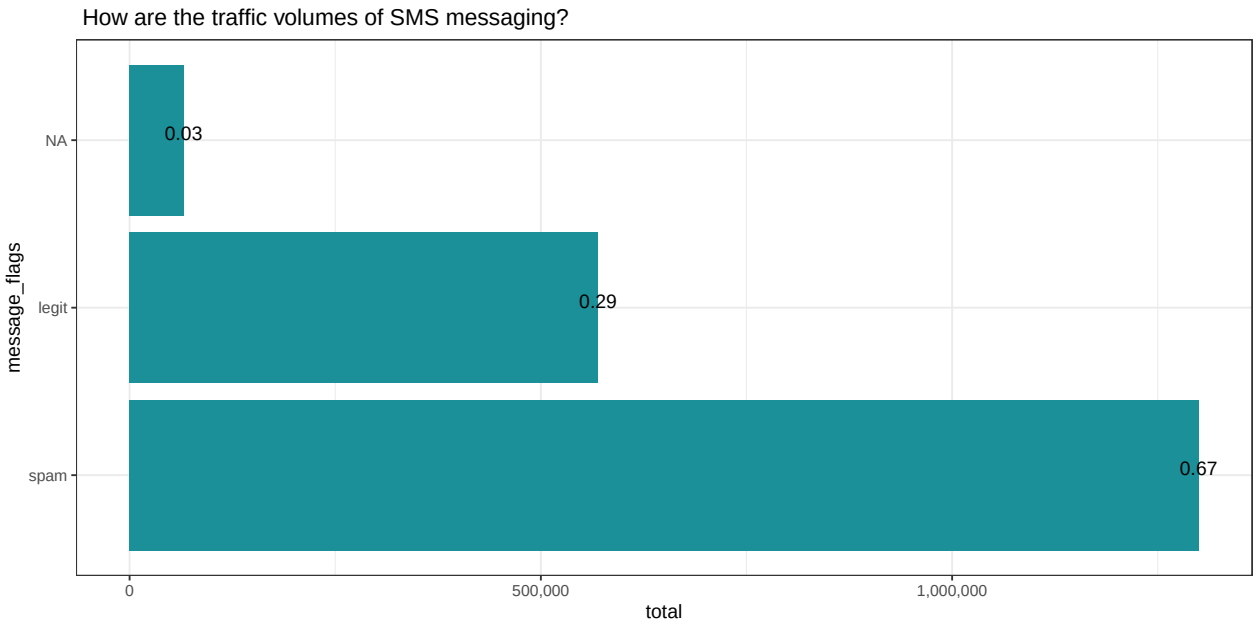
There are a lot of ways to reach these goals: we can get a basic description of the data, visualize it, identify patterns in it, identify challenges of using the data, etc.

One of the things that we will often see when we're reading about EDA is Data profiling. Data profiling is concerned with summarizing our data set through descriptive statistics. We want to use a variety of measurements to better understand our data set.

The goal of data profiling is to have a solid understanding of our data so we can afterwards start querying and visualizing our data in various ways. However, this doesn't mean that we don't have to iterate: exactly because data profiling is concerned with summarizing our data set, it is frequently used to assess the data quality. Depending on the result of the data profiling, we might decide to correct, discard or handle our data differently.

Here I have framed up about 10-11 EDA questions which help us to understand how have been the messaging traffics and what kind of patterns are used in texts?

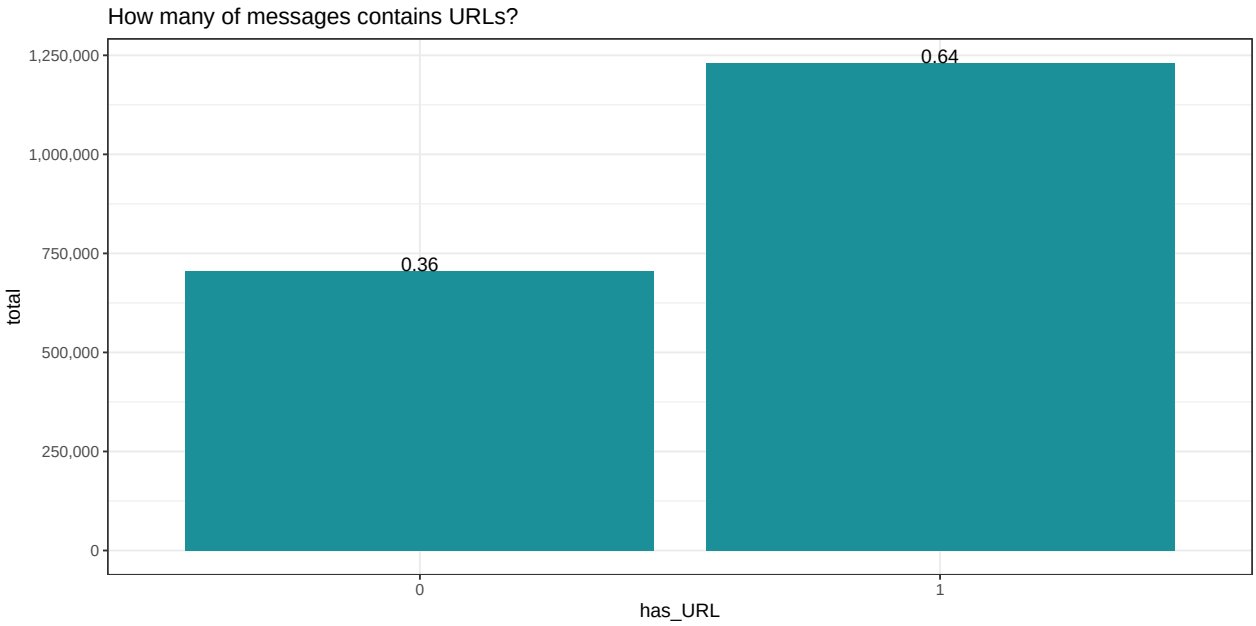
EDA-1 How are the traffic volumes of SMS Messaging?



Key Points

1. There are about 2.5Million messages collected.
2. Our team of analysts have classified the messages banking upon the rules and human intelligence as Spam or Legit
3. Manually classified Spam messages are about 67% and Legitimate messages are about 29%
4. Very few amount of messages are unclassified whose message flag is filled as NA

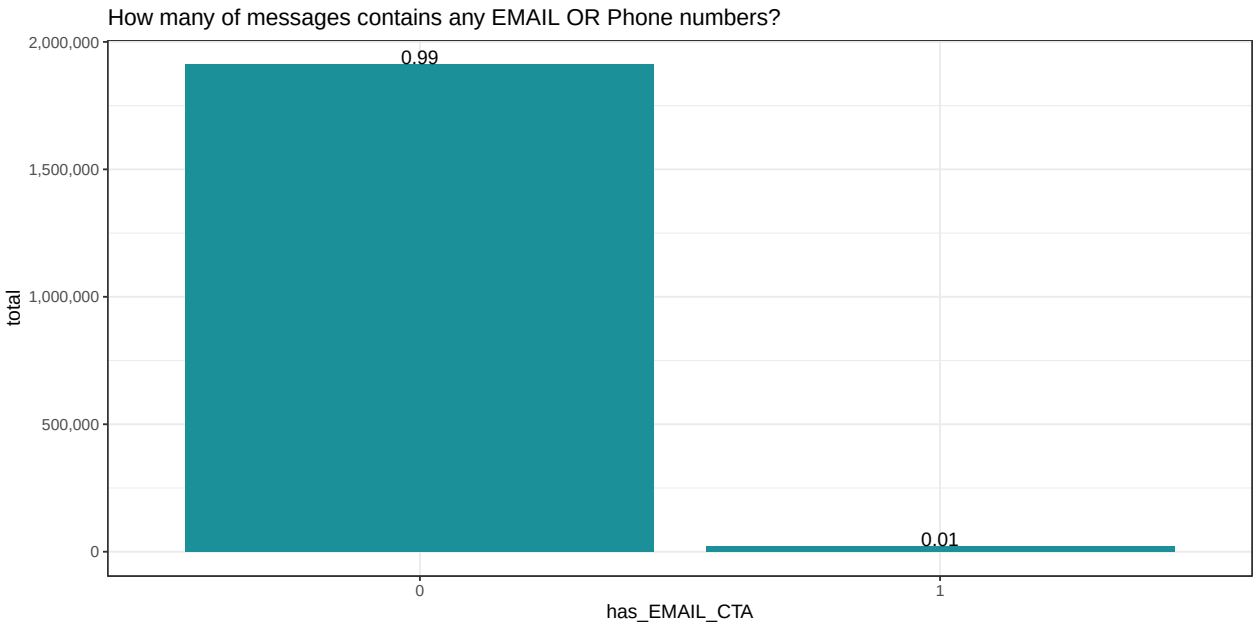
EDA-2 How many number of messages contains URLs ?



Key Points

1. The SMS messages are sent out with URLs as part of the service messages to the users in telecoms, here about 64% of messages have URLs contained in the text.

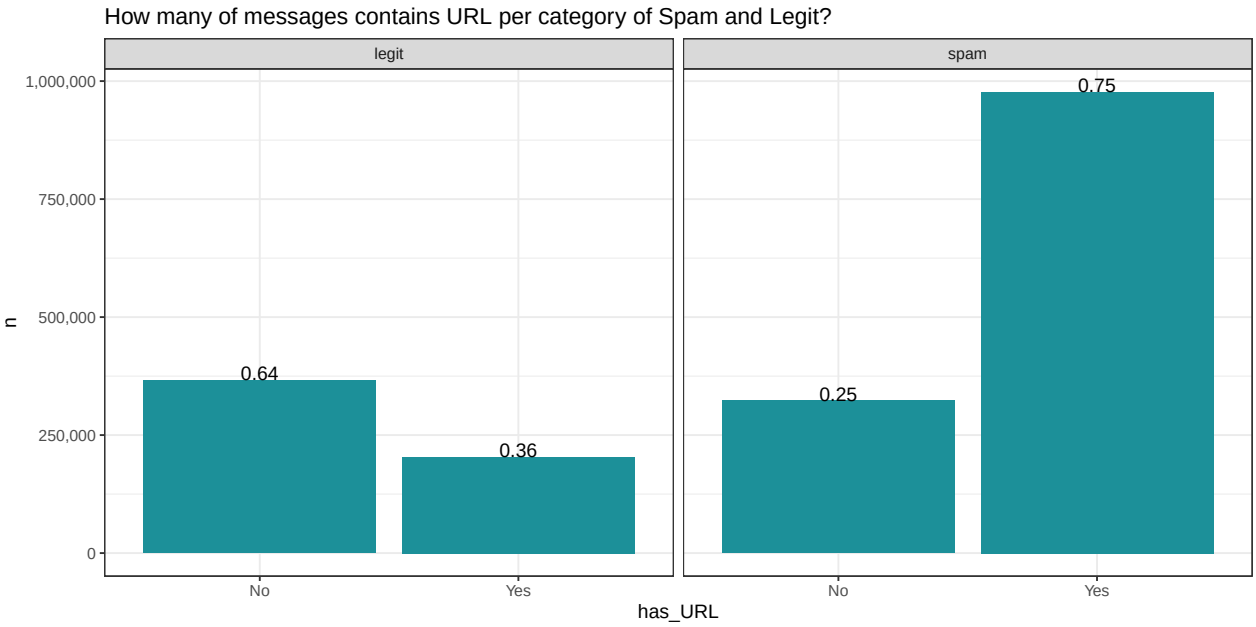
EDA-2-1 How many number of messages contains EMAIL or Phone Numbers?



Key Points

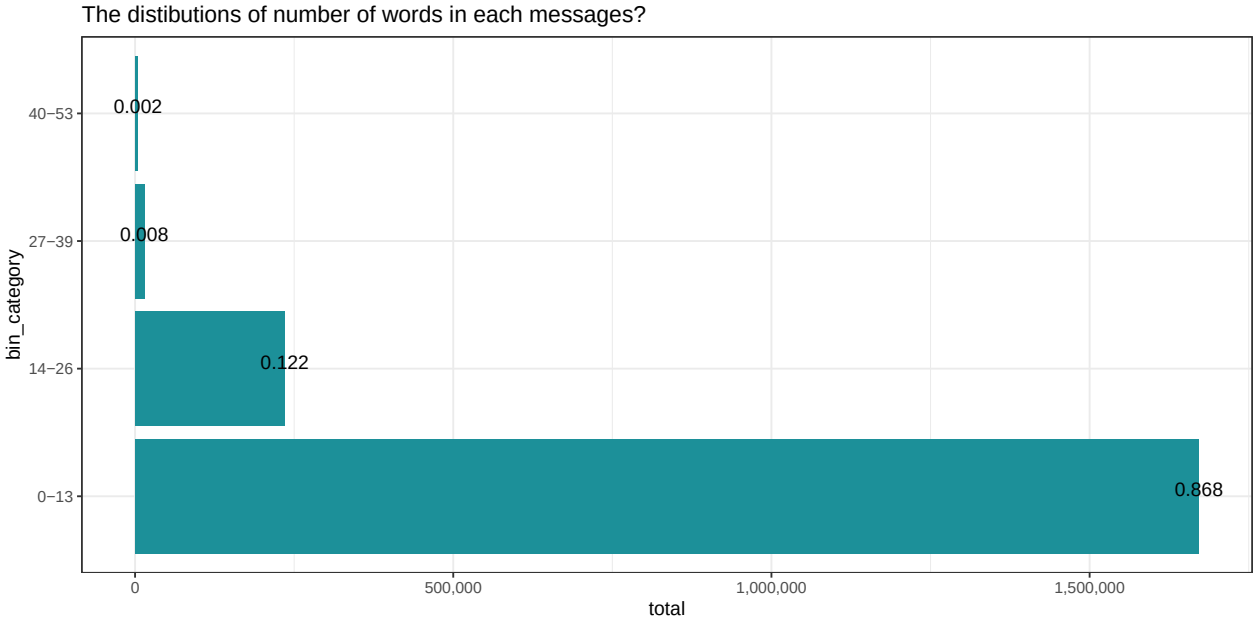
1. The SMS messages are sent out with some of phone numbers or email id as part of the service messages to the users in telecoms, here about just 1% of messages have CALL TO ACTIONS in the text.

EDA-3 How many number of messages contains URL in Spam and Legitimate messages?



Key Points

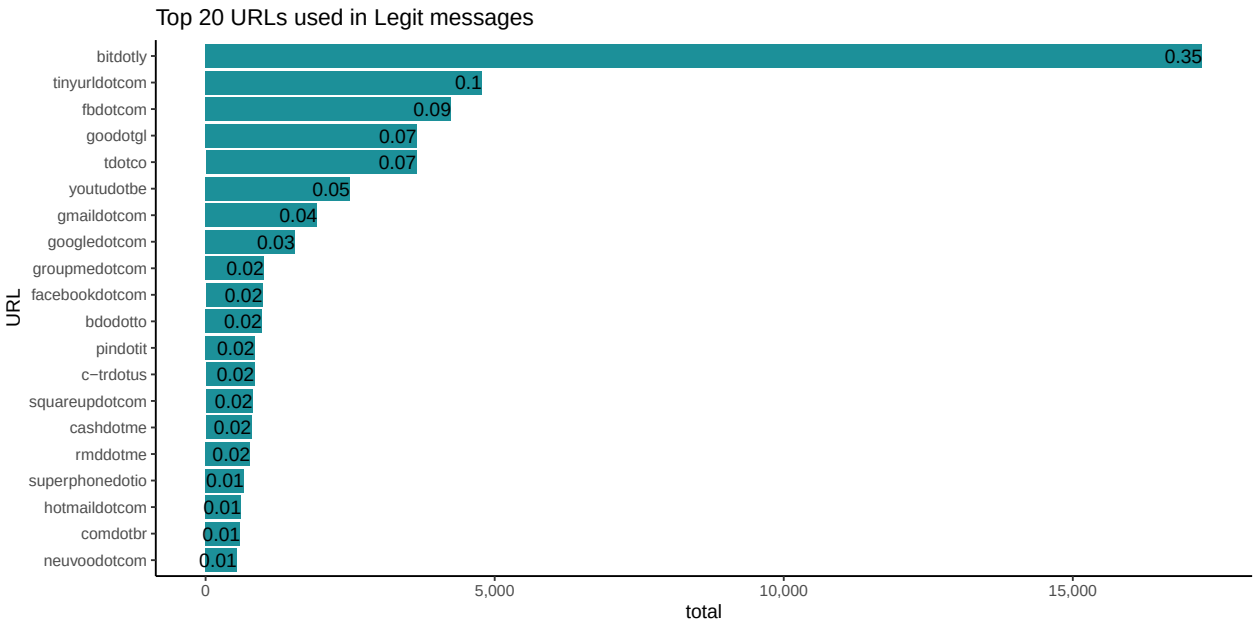
1. Spammers would always try to take away some useful information from users by asking get registered or enter in your credentials etc etc, here they make use of URL services and we can see that about 75% of spam message contains URLs where as legitimate messages is about 36%.



Key Points

1. Two of new features created from a text message such as no of words and length of a message, generally the solicited messages would have the required words and length, if the messages are being spammed they would go with some random words or length patters.
2. A list of bin category created on the field no of words, here in the above visualizations we can see about 4 bins which lumps of no of words from 1 to 53.
3. Interestingly 86% of messages are only with in the range of 1-13 words, 12% of messages are with words 14-26.

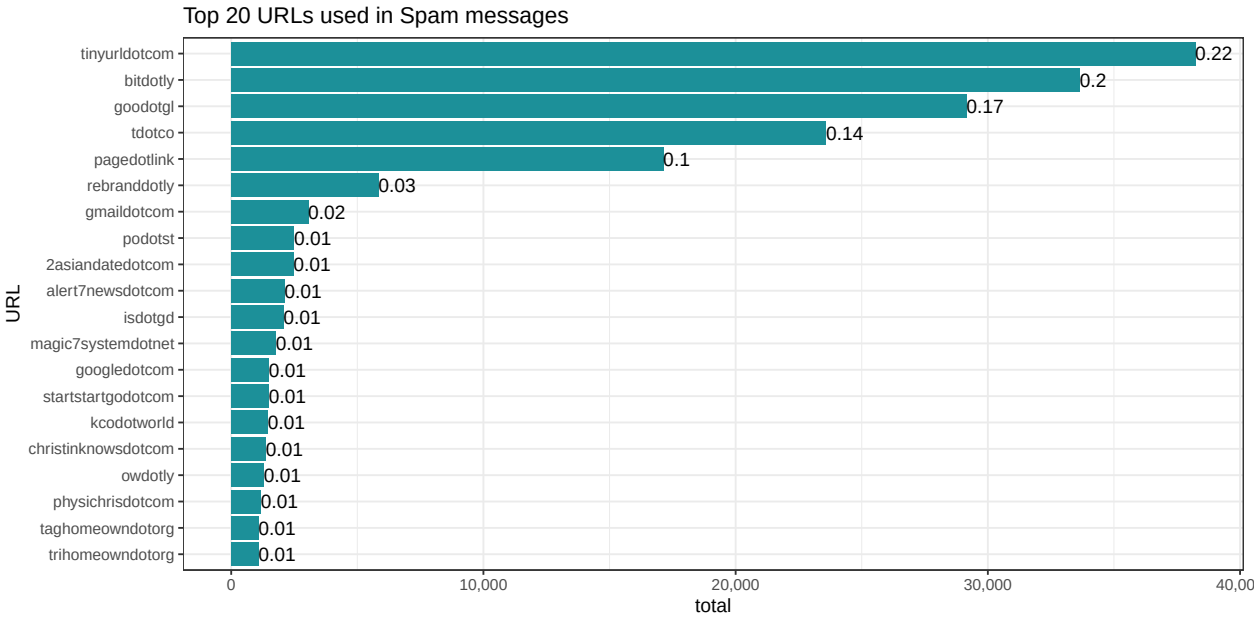
EDA-5 What are most used URLs in legit messages?



Key Points

1. Usually service organizations buy out a URL shortner services to serve their customers. we have extracted a URL from messages wherever included.
2. The URL services bitly, tinyurl,facebook, google, twitter have come in the top of Legitimate URL

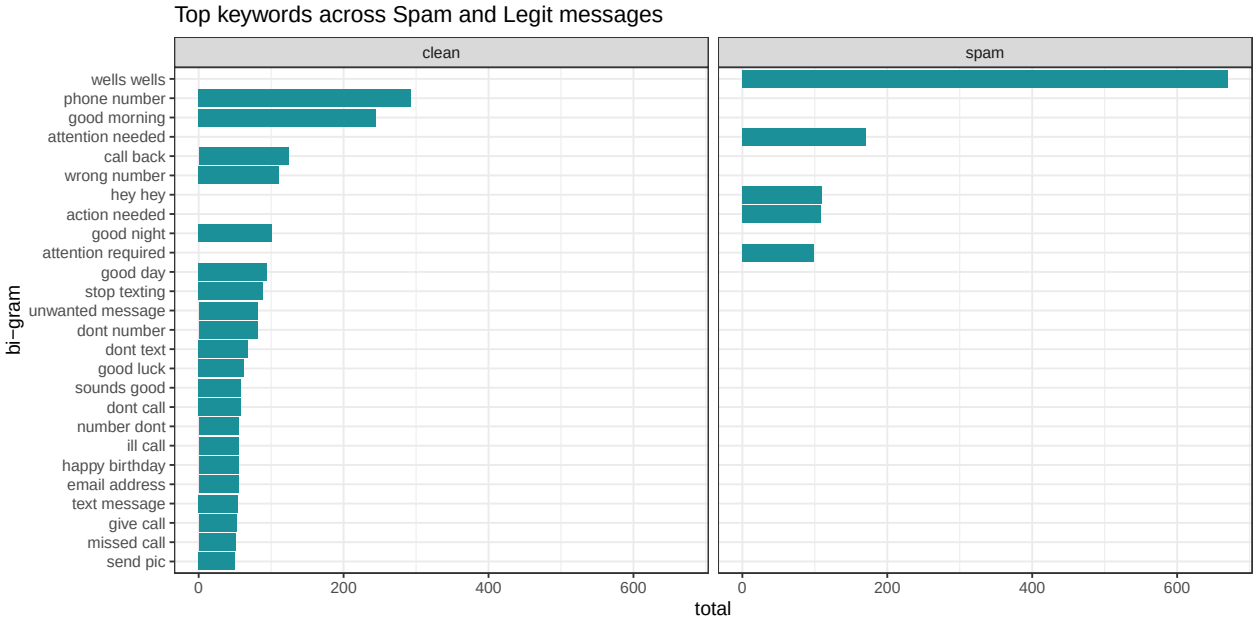
EDA-5-1 What are most used URLs in Spam messages?



Key Points

1. Spammer would always try to do spamming with a url services which are most used in legitimate messages to confuse the users.
2. We can see that the mix of tinuyrl, bitly, google, twitter and some of fake URLs are also listed out in the top URLs used in spam messages.

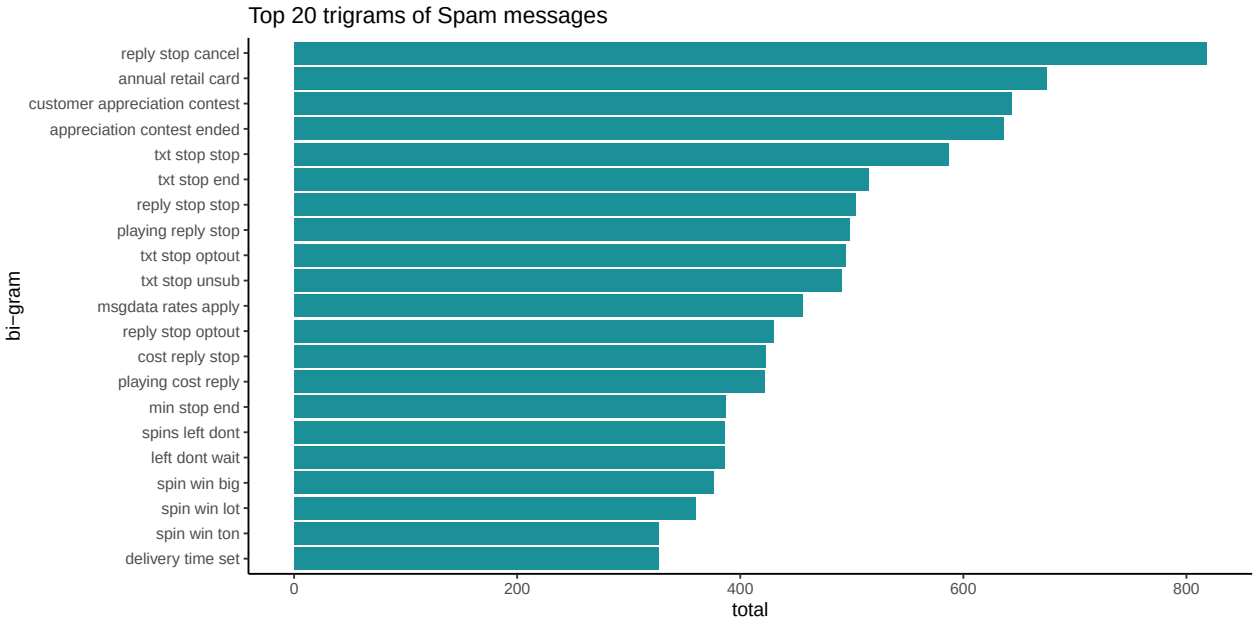
EDA-6 What are the short keywords(bigrams) are used in messages?



Key Points

1. Bi-grams are extracted from text messages, in spam category we can see that a kind of phishing attack keywords are in top place such as wells fargo, attention needed or attention required.
2. In Legitimate message categories lots of salutations, greetings, appointments , text back services message keywords are appeared

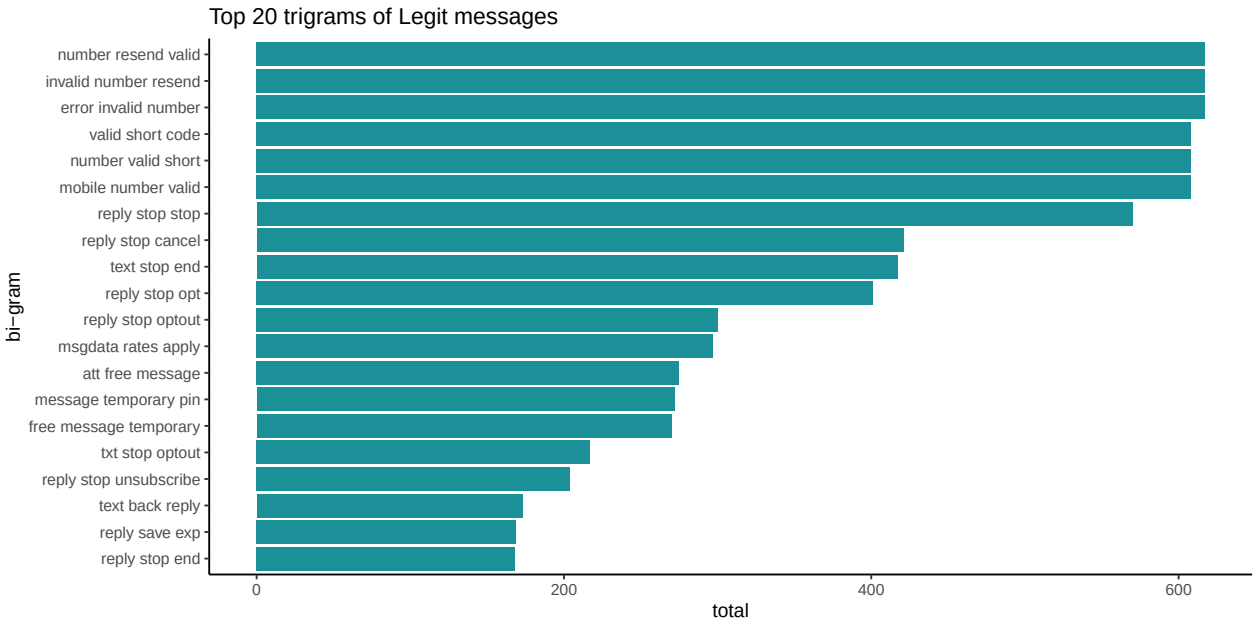
EDA-7 What are the keywords(tri-grams) are used in Spam messages?



Key Points

1. Tri-grams are extracted from text messages,in spam category we can see that a kind of phishing attack, giveaways, reply back keywords are in top place s

EDA-7-1 What are the keywords(tri-grams) are used in Legitimate messages?

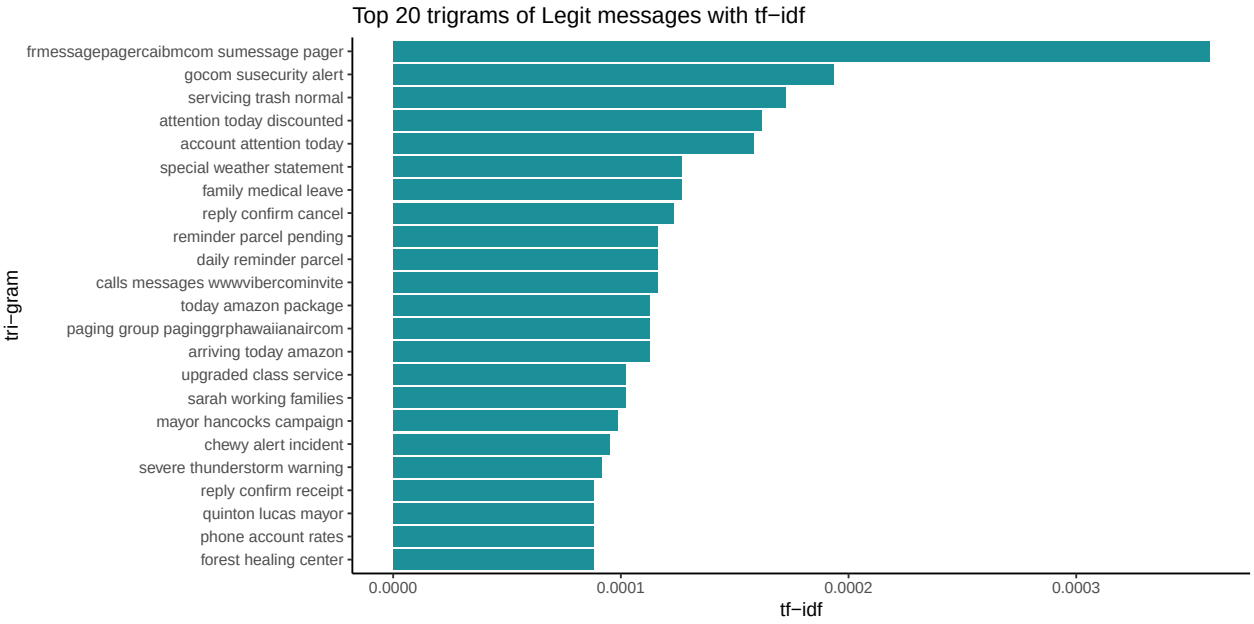


Key Points

1. Tri-grams are extracted from text messages,in spam category we can see that a kind of

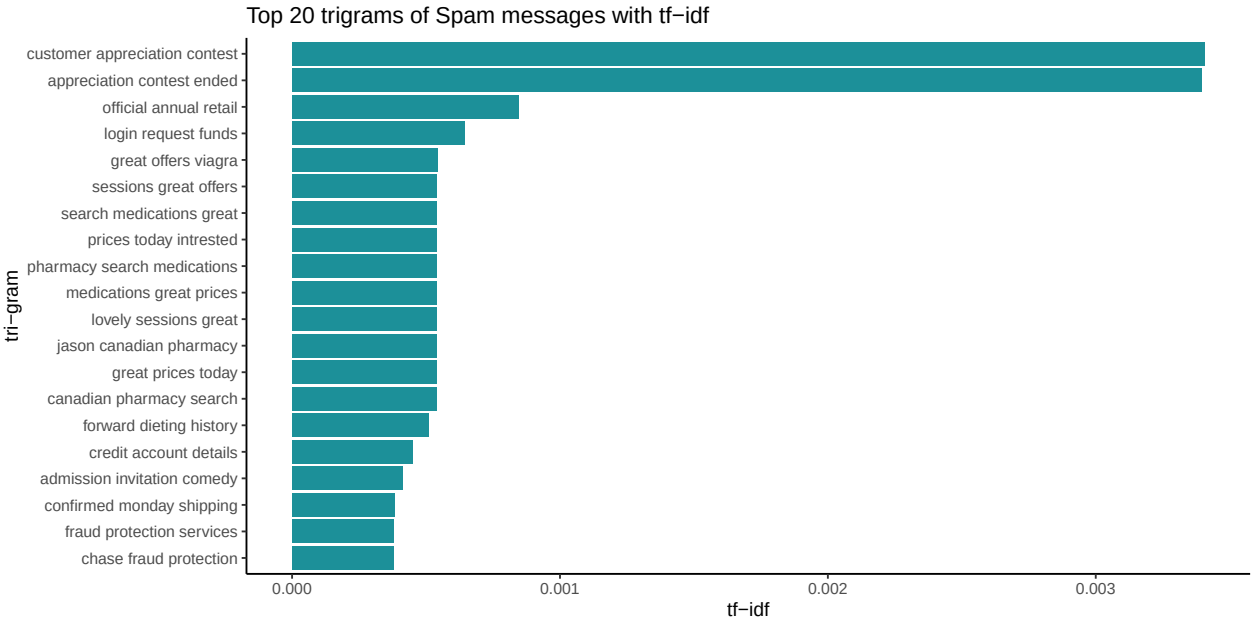
phishing attack, giveaways, reply back keywords are in top place

EDA-8 Trigrams with the tf-idf values.



Key Points

1. Tri-grams are extracted from text messages and tf-idf values are caluclated - the keywords related to notifications or alerts from banking. pharma, ecommorce,logistics, social media platforms



Key Points

1. Tri-grams are extracted from text messages and tf-idf values are caluclated - the

keywords related to spamy behaviour ones such as bank accounts, login failed requests, fake medicals, gift winnings and claimings, credit card frauds and some of abusive contents

EDA-9 What are the common trigrams in Spam messages?

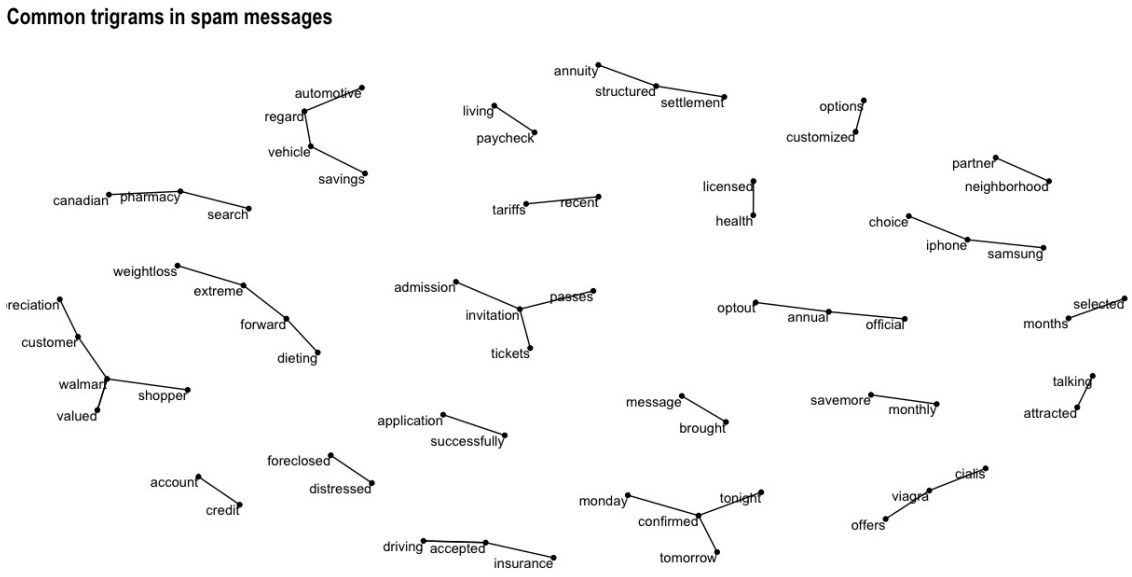


Figure 3: Common tri-grams network

EDA 10 : Is there any linear relationship between no of words and length of a message?

Null Hypothesis(H_o): The variables no of words and length of messages are not linearly associated in positive direction.

Alternate Hypothesis (H_a): The variables no of words and length of messages are linearly associated in positive direction.

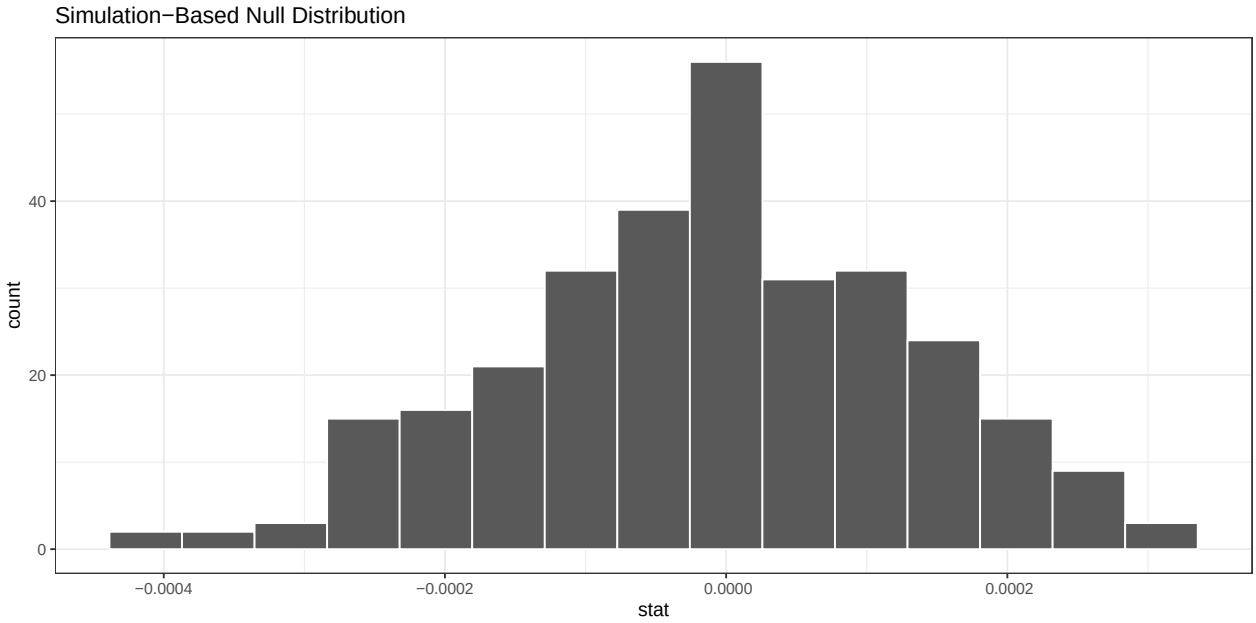
linear model summary:

term	estimate	std.error	statistic	p.value
(Intercept)	3.368415	0.1827401	18.43282	0
no_of_words	8.241210	0.0137767	598.20040	0

key points

1. The intercept(B_o)=2.7 is the average length of messages which have zero no of words

2. The slope(B_1)=8.29 is summarizing the relationship between the no of words and message length i.e For every increase of one unit in no of words there is an associated increase of an overage 8.29 units of no of words.



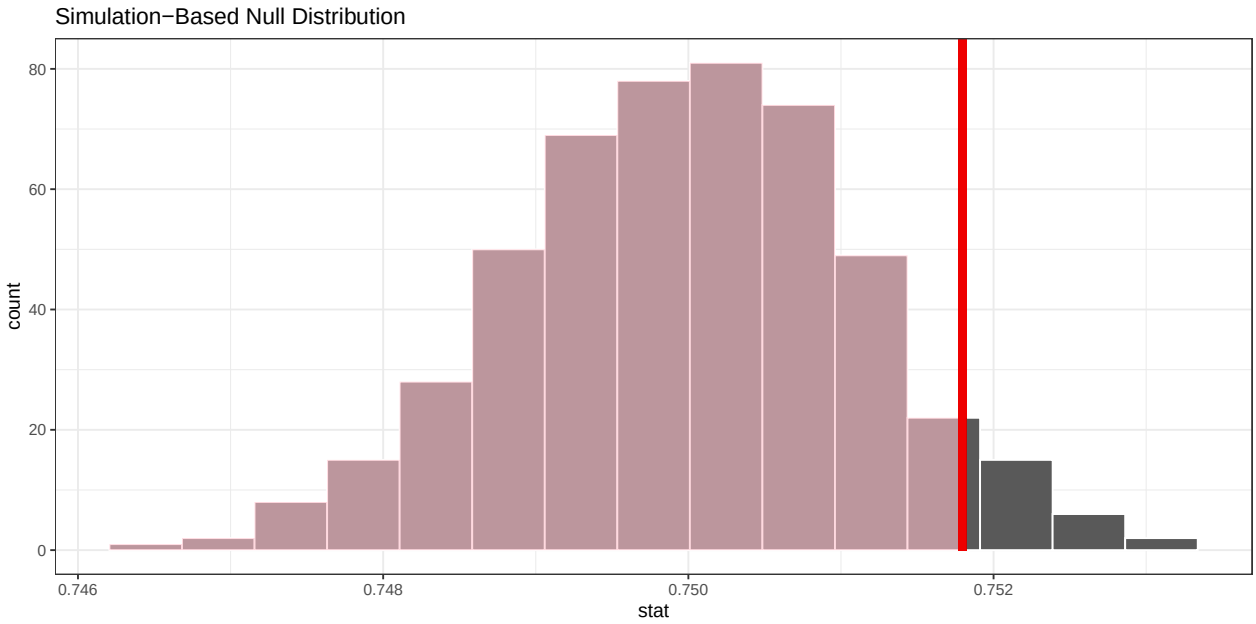
p_value
0

Here P-value calucated and it is found to be zero, Hence we reject the null hypothesis - and there is a significant relationship between no of words and the length of message

EDA 11 : One proportion Hypothesis test on Spam messages URLs.

Null Hypothesis(H_o): The true proportion of Spam messages which contains is about 0.75

Alternate Hypothesis (H_a): The true proportion of Spam messages which contains is less than about 0.75



p_value
0.5

Here P-value calucated and it is found to be 0.5, We therefore fail to reject the null hypothesis and here the true proportion of spam messages URLs are 0.75

Phase 3 and 4: Supervised Machine Learning - Classification Modelling and Error Analysis

Our business problem is a kind of classification i.e figure out a given message is spam or legitimate, as we have two categories in a target variable this classification is said to be binary classification.

Supervised Machine Learning classification algorithms that we are going to make use and build a classifier to automate the process of messaging classifications.

1. Logistic Regression and SGD
2. Naive Bayes
3. Random Forest
4. Gradient Boosting-XGB

Let us look at data points and how are the spam and clean messages distributions, and after dividing how are the training and testing datapoints.

Messages table observations are as follows:

message_flag	total
Spam	806,664
Legit	288,854

Training and Testing dataset observations are as follows:

type	total
Train	824,633
Test	274,878

3.1. Base Approaches

The classifier are applied on the data points and caluclated the ROC-ACUs per each model. these details are found in the below table.

ROC - AUC scores per model:

Technique	Logistic	Naive	RF	XGB	SGD
Train	0.78160	0.74709	0.86160	0.67808	0.81383
Test	0.78066	0.74656	0.78066	0.67653	0.81282

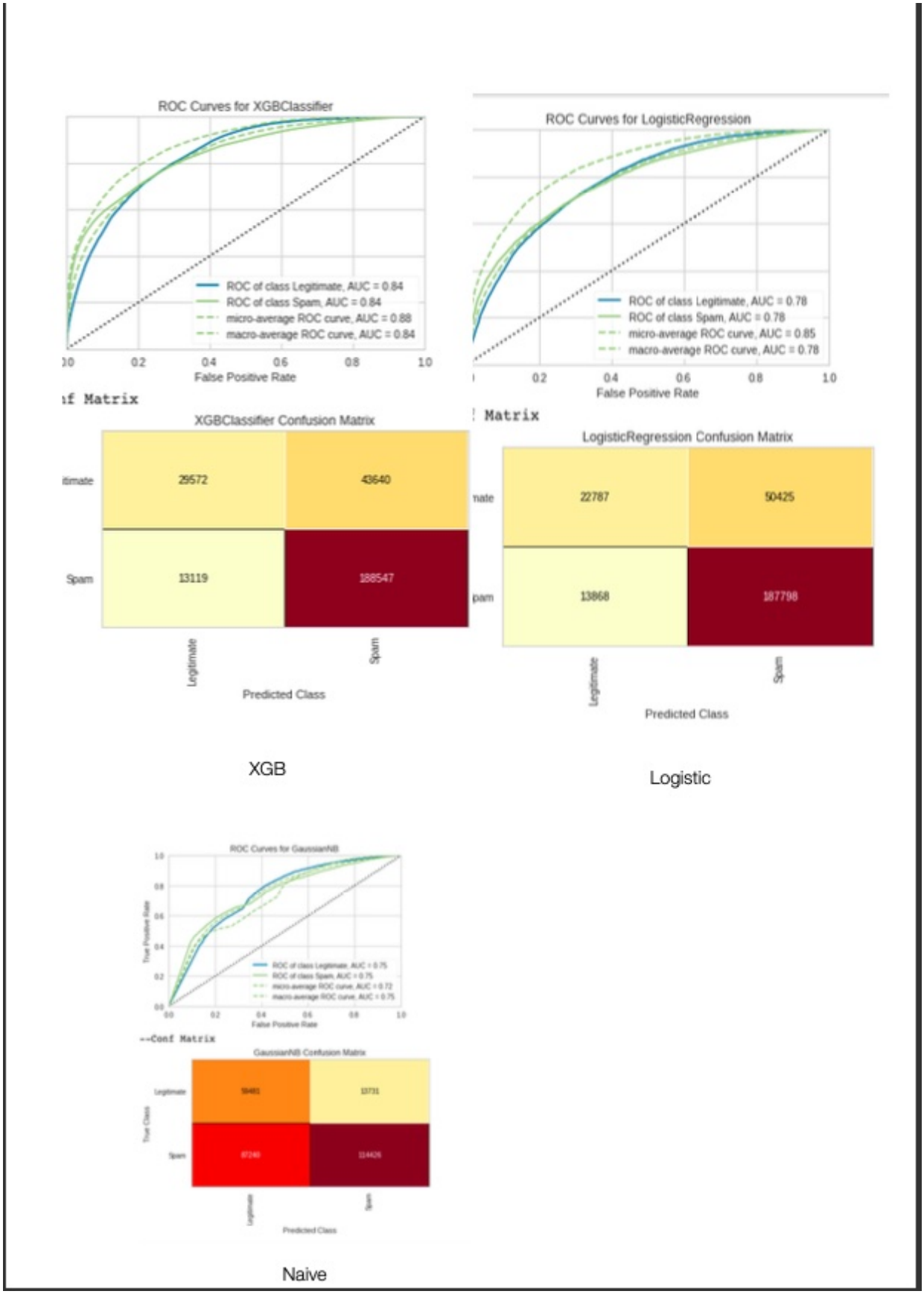


Figure 4: Confustion matrix and ROC Curces

It seems some of models are over-fitting and some models performs poor. as we have seen that there is a class imbalance in the dataset.if we sort out the class imbalance these may be resolved further.

I have also presented the other metrics(Precision, Recall and F1 Scores) per each model.

Metric type	Naive	Logistic	XGB	RF
Precision	0.893	0.788	0.812	0.79
Recall	0.567	0.931	0.935	0.83
F1-Score	0.694	0.854	0.864	0.79

3.2. Class Imbalances - Samplings

I will list out what kind of techniques are available to deal with class imbalance problems.

Under-Sampling

- Random
- Condensed Nearest Neighbour
- Tomek Links
- Near Miss

Over-Sampling + Random + SMOTE + ADASYN

Here are the distributions of data points across the random over and under sampling techniques.

type	total	Spam	Legit
Over	1,209,996	604,998	604,998
Under	439,270	219,635	219,635

The classifiers Logistic, Naive, Random forest, XGB and SGD are played around on each of the above mentioned techniques using 3 fold cross validations with a metric ROC-AUC score.

ROC - AUC scores per model:

Sampling Technique	Logistic	Naive	RF	XGB	SGD
Random Under	0.780993	0.74700	0.746732	0.83937	0.76342
Random Over	0.781068	0.74707	0.743213	0.84118	0.76234

Precision,Recall and F1 Score metrics per random over and under samplings :

Metric type	XGB Over	XGB Under
Precision	0.904	0.881
Recall	0.714	0.724
F-1 Score	0.798	0.768

Random over tech with XGB found to be classifying the messages better with the specified params and we try now we can improve performance tuning the hyperparams of XGB model.

3-4.3. Model Selections and Hyperparameter tuning

XGB model parameters are going to be given in with below parameters and passed in to GridSearchCV hyperparameter tuning methods so that it will return the best parameters to be specified in a xgb model that gives better performance in classifying the spam messages.

Parameters and values
n_estimators:[900,1000,1500,2000]
max_depth:[5,7,10]
learning_rate:[0.1,0.3,0.5,0.8]
booster:[dart,gbtree]
gamma:[0.1,0.3,0.5]
subsample:[0.5,0.9]
colsample_bytree:[0.5,0.9]
colsample_bylevel:[0.5,0.9]
colsample_bynode:[0.5,0.9]
reg_lambda:[1,10,20]

The better params are drawn and given in a xgb model as below.

```
model_xgb = xgb.XGBClassifier(objective='binary:logistic',
                              n_estimators=1000,
                              eval_metric='auc',
                              use_label_encoder=False,
                              max_depth=7,
                              learning_rate=0.898568,
                              booster='dart',
                              gamma=0.010000,
                              subsample=0.614947,
                              colsample_bytree=0.5,
                              colsample_bylevel=0.5,
                              colsample_bynode=0.5,
                              reg_lambda=10)
--INSERT--
```

Phase 5: Deployment of E-Mail and SMS Messaging classification

I have developed an algorithm XGB which found to be classifying the E-mail and SMS messages more accurately than other classifier, this algorithm is going to be deployed in a web app where end users can check what type of a message is.

To deploy this algorithm an open source web framework called **Streamlit** will be made use so that it can be shared on line anyone can have an access to it as long as they have an internet connections.

XGB model file and text vectorizer- TFIDF are packed in a pickle file and they can be loaded into an streamlit app program is developed with all the required input sources, when any text is given it will figure out a message is Spam or Clean.

This classification web app is accessible on this link and please do click on it

Spam classifier Web App

Here are couple of screenshot of this application.

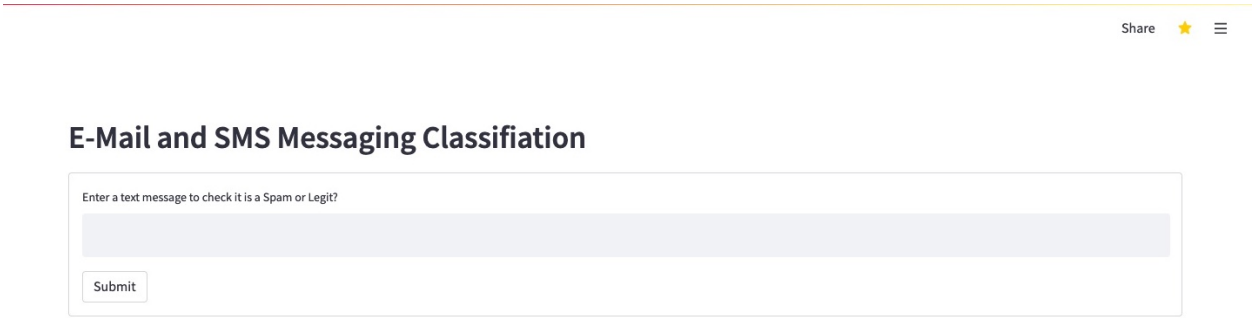
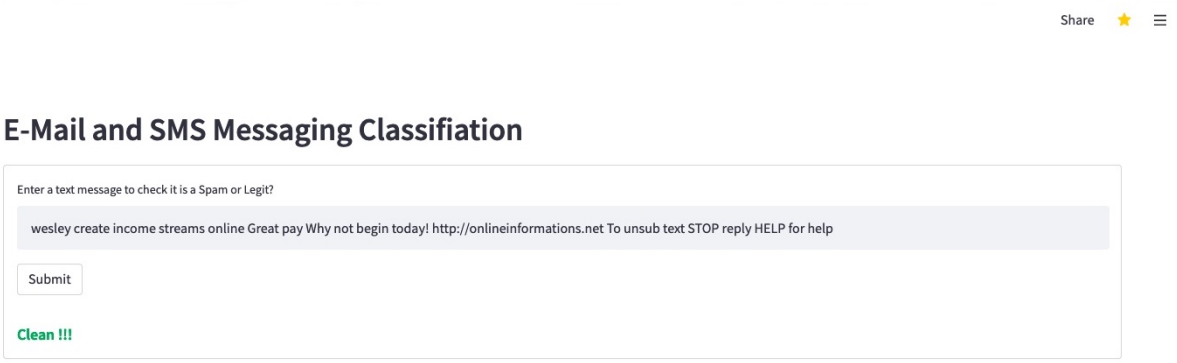
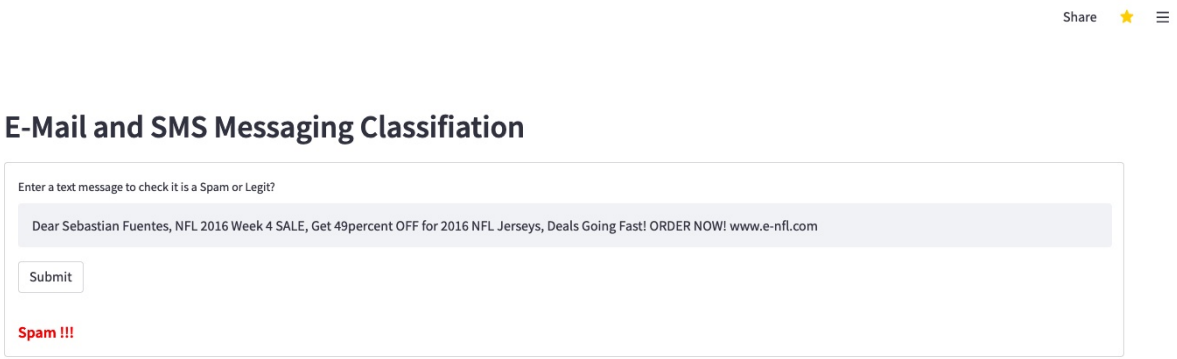


Figure 5: Main page of an app

Here we are taking an input message from end user in a text box, this message will be given to text vectorizer after having processed from a function i.e unnecessary things are removed from a text, made it as a standard form, once the text message is transformed, it will be passed on to XGB model object so that it will predict whether it is a Spam or Legit message. Here text submission is done on clicking a button **Submit**. The respective labels(Spam or Clean) will be displayed on the text label as showed in the below images.



Here the given message is a legitimate message and the model predicts as a Clean correctly.



Here the given message is a Spam message and the model predicts as a Spam correctly.

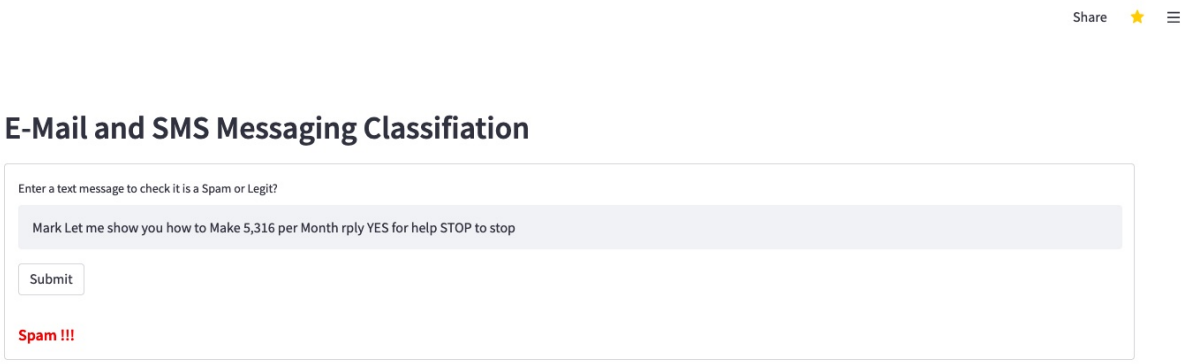


Figure 6: Spam message

Here the given message is a spam message and the model predicts as a Spam correctly.

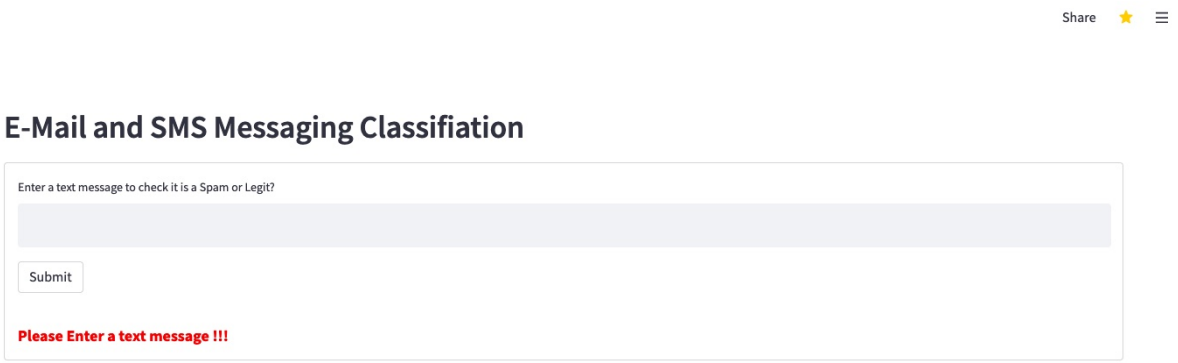


Figure 7: Empty message

If a user does not give in a message and click on submit button it will throw an error: saying that text message is not sent in.

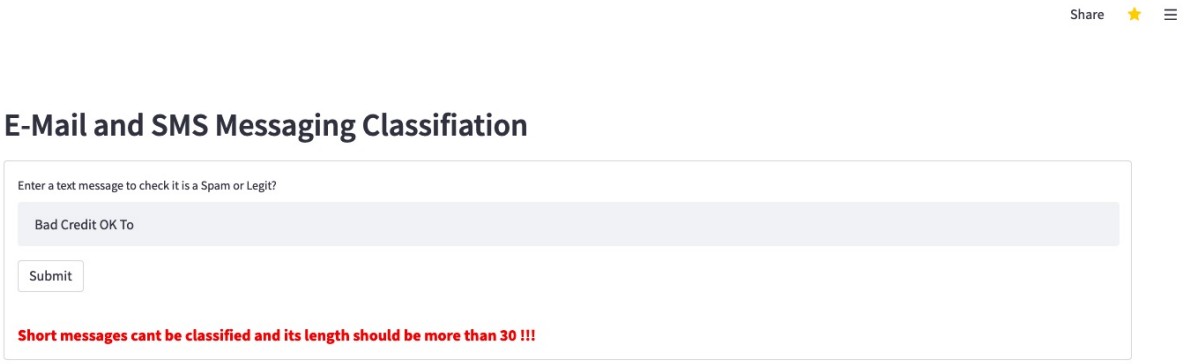


Figure 8: Short message

As a business logic we should consider a message to be classified when its length is more than 30 characters. if the messsage does not meet this requirement it should say a message to end user to give in recommended lenth message.

Project Github Repo.

All work related to this project has been uploaded in the below given github link and please do click on it to view.

Github Repo Link

References

reference-1: Improving Phishing URL Detection Using Transformers

reference-2: A systematic framework to discover pattern for web spam classification

reference-3: Machine Learning for E-mail Spam Filtering: Review,Techniques and Trends

reference-4: Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification

reference-5:Streamlit Web apps

reference-6:R-Markdown

reference-7:Scikit-MachineLearning

reference-8: tidy text mining