-1. Executive summary

Mobile or SMS spam is a real and growing problem primarily due to the availability of very cheap SMS packages, SMS also generates higher response rates as it is a trusted and personal service. SMS Spam filtering however poses its own specific challenges. Objective of this report is to investigate different validation and machine learning algorithms on a dataset of SMS data which are previously classified as spam and ham (non spam). We have used the algorithms such as Support Vector Machine, Random Forest, and Naïve Bayes. The performances of the algorithms are checked based on its accuracy for predicting the spam messages from the testing data set (30% of the original dataset).

2. Introduction

Network security agencies provide smooth communication channels and it gives firewalls safety mechanism to its customers. Recently the most critical topic in telecommunication industry is the uncontrolled growth of non-legitimate text messages. With readily available databases of contact number and mail ids, companies target its potential customers through messages or emails. Customer may face security problems due to such non-legitimate or spam text messages. Hence network security services aim to provide spam SMS detection mechanism to its customers. For this, companies may develop a prediction model for identifying such spam SMS.

Since SMS or text messages contain raw data in string format, NLP techniques can be applied for spam message detection. Text mining classification can be achieved with the help of Naïve Bayes or Random forest or Support Vector Machine. This project aims to analyze the dataset using all the three classifiers and the best classifier will be selected based on the highest Cohen's kappa score. Stepwise process is mentioned below-

1. Read the data and split it into training and testing sets
2. Create document-term metrics
3. Use Naïve Bayes classifier with 10-fold cross-validation method and monitor its Kappa score
4. Use Random Forest classifier with 10-fold cross-validation method and monitor its Kappa score
5. Use Support Vector Machine classifier with 10-fold cross-validation method and monitor its Kappa score
6. Select the best classifier based on the highest Kappa score and use it for testing dataset. Track the classification report of this model.

Globally, short messaging service (SMS) is one of the most popular and most affordable telecommunication service packages. However, mobile users have become increasingly concerned regarding the security of their client confidentiality. This is mainly due to the fact that mobile marketing remains intrusive to the personal freedom of the subscribers. SMS spamming has become a major nuisance to the mobile subscribers given its pervasive nature. It incurs substantial cost in terms of lost productivity, network bandwidth usage, management, and raid of personal privacy. Thus, in short spamming threatens the profits of the service providers. Mobile SMS spams frustrate the mobile phone users, and just like e-mail spams, they cause new societal frictions to mobile handset devices. Spam can be described as unwanted or unsolicited electronic messages sent in bulk to a group of recipients. The messages are characterized as electronic, unsolicited, commercial, mass constitutes a growing threat mainly due to the following factors: 1) the availability of low-cost bulk SMS plans; 2) reliability (since the message reaches the mobile phone user); 3) low chance of receiving responses from some unsuspecting receivers; and 4) the message can be personalized. Mobile SMS spam detection and prevention is not a trivial matter. It has taken on a lot of issues and solutions inherited from relatively older scenarios of email spam detection and filtering. Unsolicited SMS text messages are a common occurrence in our daily life and consume communication time, bandwidth and

resources. Although the existing spam filters provide some level of performance, the spams misinform receivers by manoeuvring data samples.

2. Data Set:

Kaggle source dataset will be used for this analysis. This dataset contain the raw text data with unique ids and its classification. Each record is classified under either spam or ham SMS category. The dataset contains total 5572 lebelled records. Below are the properties of each field in the dataset-

Field name	**Variable type**	**Example**
V1	Categorical	ham or spam
V2	String	1. Ok good night… 2. 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, å£1.50 to rcv

The dataset will be divided into 70:30 ratios. Random 70% records will be used to train the model and then the model will be tested against the remaining 30% records. With this, the reliability and robustness of the model can be ensured.

3. Data Pre-processing

The data set has 5572 labelled records with two attributes V1, V2 Column. Where V1 column contained the class of the SMS and V2 contained the actual message. Out of 5572 records 747 are spam messages, thus making up 13% of the dataset as SPAM message.

Word Count Frequency

Ham Message

Following table shows the term frequency of the extracted word from the message text which are classified as Ham.

Words	Count of Occurrence
'u'	730
i'm	298
<#>	234
'get'	225
'2'	224
'ur'	199
'go'	196
.	188
'got'	188
'come'	173
'call'	172
'like'	169
'know'	164
?	153
'good'	150
'...'	141

'going'	135
'ok'	128
'4'	126
"i'll"	126
'love'	121
'still'	120
'want'	119
'time'	119
'one'	116

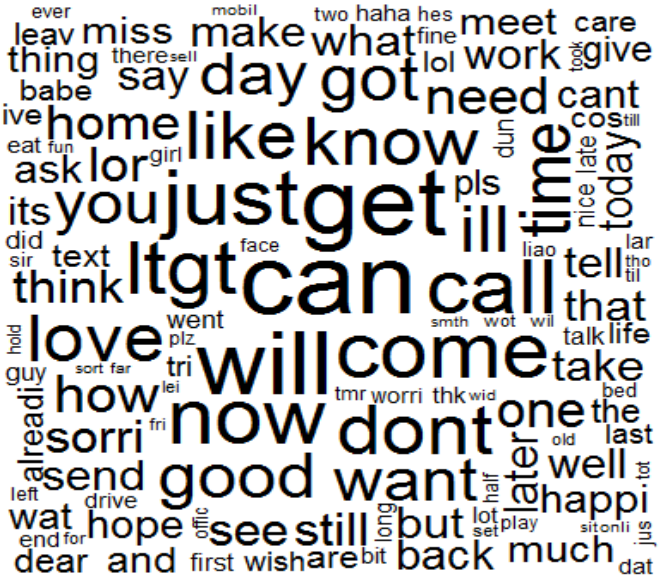

Figure:3.1

The wordcloud for the Ham message text is shown in the figure 3.1, the size of the words reflect the frequency in the dataset.

Spam Message

Following table shows the term frequency of the extracted word from the message text which are classified as Spam.

Words	Count of Occurrence
'call',	269
'free',	146
'2'	138
'ur'	115
'txt'	111
'u'	92
'claim'	88
'mobile'	88
'text'	82
'&'	82
'reply'	80
'4'	74
'get'	71
'stop'	71

'now!	56
'new'	54
'nokia'	51
'send'	49
'cash'	48
'prize'	47
'please'	45

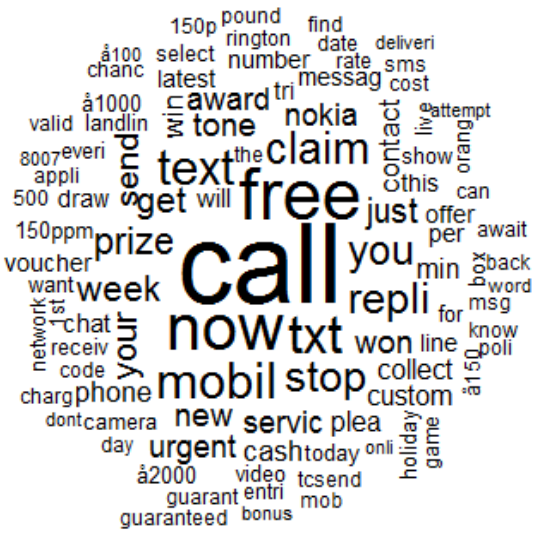

Figure 3.2

The wordcloud for the spam message text is shown in the figure 3.1, the size of the words reflect the frequency in the dataset.

The dataset was divided into training and testing dataset, with 70% used for training the model while 30% was reserved for testing. 10 fold cross validation technique was used to build the model.

The text messages string was transformed into vectors by using three different methods of transformation.

- Document Term Matrix
- Term Frequency-Inverse Document Frequency Matrix
- Term Frequency Matrix

Each transformed vector matrix was then used as an input parameter in building the classification model.

3. Methods used for Classification:

The goal is to apply different machine learning algorithms to SMS spam classification problem, compare their performance.
We have used three different classification techniques on the dataset to identify which technique is

performing the best

3.1 Naïve Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. The speed and simplicity along with high accuracy of this algorithm makes it a desirable classifier for spam detection problems.

3.2 Decision tree:

A random forest is an averaging ensemble method for classification. The ensemble is a combination of decision trees built from a bootstrap sample from training set. Additionally, in building the decision tree, the split which is chosen when splitting a node is the best split only among a random set of features. This will increase the bias of a single model, but the averaging reduces the variance and can compensate for increase in bias too. Consequently, a better model is built.

3.3 SVM (Linear):

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier

4. Model Selection:

Due to the imbalance in classes of spam and ham message, Cohen's kappa was used as the performance metric, to account for chance agreement between actual and predicted values. Cohen's Kappa is a statistics measure which measures inter-rater agreement for qualitative items and is a more robust measure than simple percentage agreement calculation, since Kappa takes into account the possibility of agreement occurring by chance

	n-gram	**Naïve-Bayes**	**Random Forest**	**Support Vector Machine**
Document Term Matrix	1	Kappa Stat	Kappa Stat	Kappa Stat
	2	0.889	0.871	0.894
	3	0.893	0.862	0.898
Tf-Idf Matrix	1	0.892	0.869	0.896

	2	0.872	0.879	0.898
	3	0.873	0.878	0.898
Tf Matrix	1	0.873	0.882	0.896
	2	0.879	0.878	0.896
	3	0.88	0.889	0.90

Based on the Cohen's Kappa The best model is using TF matrix with 3-gram and SVM as the modeling algorithm.

5.0 Results and Observations

The results obtained on the application of SVM Linear 3-gram with TF matrix on the vectorized testing data are given in the *Table 5.1.1*. As observed from the table, it is clear that the best classification model could be built by using SVM algorithm with 3 gram vectorized data with an accuracy of 99.79% (True Ham) 82.14% (True Spam) and 97.24% overall accuracy in prediction. The Cohen's Kappa on the testing dataset came out to be 0.87, which is much less than 97.24% due to the imbalance in the classes.

Table 5.1.1 Accuracy of classifiers in Testing Datasets

Confusion Matrix	Ham	Spam	Accuracy
Ham	1442	6	99.79%
Spam	40	184	82.14%
	97.3%	96.84%	97.24%

The classification report for the testing dataset is shown below.

Classification Report				
	Precision	**Recall**	**F1 Score**	**Support**
Ham	**0.97**	**1.00**	**0.98**	**1448**
Spam	**0.97**	**0.82**	**0.89**	**224**
Avg/Total	**0.97**	**0.97**	**0.97**	**1672**

5. Conclusions

- The Cohen's Kappa result improves with n-gram's, increase of n upto n=3 for SVM (linear).
- SVM, Naïve Bayes and Random Forest algorithms gives different result, out of which SVM (Linear) has the highest Cohen's Kappa of 0.90.
- Term Frequency Vectorized Matrix improves Cohen's Kappa with SVM (Linear) algorithm of which 3-gram has the best Cohen's Kappa.