A Technique To Differentiate Spam Short Message Service (SMS) From Mobile Data

Session: 2016 – 2017

Submitted by:

Majid Yaseen 2016-MS-CS-11

Supervised by:

Prof.Dr.Muhammad Shoaib

Department of Computer Science and Engineering

University of Engineering and Technology

Lahore Pakistan

A Technique To Differentiate Spam Short Message Service (SMS) From Mobile Data

Submitted to the faculty of the Computer Science and Engineering Department of the University of Engineering and Technology Lahore in partial fulfillment of the requirements for the Degree of

Master of Science

in

Computer Science.

Internal Examiner
Signature:
____
Name:
____
Designation:
____

External Examiner
Signature:
____
Name:
____
Designation:
____

Chairman
Signature:
____
Name:

Dean
Signature:
____
Name:

Department of Computer Science and Engineering

University of Engineering and Technology

Lahore Pakistan

Declaration

I declare that the work contained in this thesis is my own, except where explicitly stated otherwise. In addition this work has not been submitted to obtain another degree or professional qualification.

Signed: ____________________

Date: ____________________

Acknowledgments

Without acknowledgment and mentioning names of those guiding stars who supported through all this research work; it would become a matter of injustice.

I am heartily grateful to my M.Sc. thesis supervisor Prof. Dr. Mohammad Shoaib, who open heartily agreed to my research proposal of this work and extend my sincere vote of thanks for his continuous support, mentorship and guidance at each and every step until the successful completion of this research work.

I would also like to present bundle of thanks to the Department of Computer Science and Engineering for their encouragement, support and timely reply to our queries submitted. My last but not the least ”Thank You” goes to my friends and my family for their direct or indirect help in completion of this research work.

Dedicated to my beloved parents…

Contents

List of Figures

List of Tables

Abbreviations

SMS	**S**hort **M**essage **S**ervice
SPAM	**S**tupid, **P**ointless **A**nnoying **M**essage
WEKA	**W**aikato **E**nvironment for **K**nowledge **A**nalysis
MATLAB	**M**atrix **L**aboratory

Abstract

Today text messages have a significant impact on our lives, but at the same time we face many critical problems due to SMS spamming. Therefore to detect spam messages and distinguish it via accurate filtering is a challenging task for researchers. In current research work content based spam SMS filtering technique is proposed depending on machine learning approach to distinguish spam messages from mobile data while considering the low processing power and limited memory of mobile phones. From each text message; five attributes are extracted and based on these attributes/features, an unknown/unlabelled SMS message can be classified as spam or ham by a trained learning algorithm. These attributes are the length of the text message and presence of repeatedly occurring spam words, count of spam words, combination of spam words and SMS class. It is shown that Decision Tree classifier performance is better than other machine learning algorithms investigated. The other learning algorithms explored in this work are Nave Bayes and neural network architecture-Multilayer Perceptron Algorithm.

Chapter 1

INTRODUCTION

1.1 Overview

Spam is generally defined as unrequested and unwanted email or SMS, usually sent to a huge number of recipients. From economic point of view SMS spam has a very important impact on mobile phone users and also service providers. It not only degrades the quality of service of cellular operators but also harm the privacy of users. This problem of Spam short message service is becoming critical day by day around the globe and has encouraged the researchers to develop different filtering techniques to fight this evil [1]. Like email SPAM, SMS spam is also a serious threat, resulting in resource consumption and annoyance at the recipient. There is a burning need to detect and remove such messages with minimum false positives. Researchers have proposed different filtering techniques for SMS spam that are based on two different approaches i.e. content based SMS spam detection and spam SMS detection using non-content features. Classification techniques for SMS-Short Message Service that are based on non-content attributes observes and scan metadata of the message; for instance length of the message, white spaces it contains, total number of characters, time-of-day and network features etc. while on other hand filtering techniques based on contents of the message mainly focus on the detail description of the contents of the message. In this approach content of the message is matched to the user defined characteristics.

1.2 Motivation and Problem Statement

The approaches adopted while developing any filtering technique for SMS spam from mobile data are a bit different from that of email. It is not only because of exclusive requirements of SMS-Short Message Service but also mobile phones. Following are the major reasons:

- *The unavailability of real and public datasets*: which may results in misleading conclusions while evaluating different approaches, thus real time technique for spam SMS discrimination is getting harder as compared to email.
- *Lower processing power*: low processing power of mobile phones as compared to computer systems is also one of the main challenge researchers are facing while developing efficient SMS spam filters. As spam SMS filtering technique requiring a lot of processing may not produce desired results on time.
- *Limited memory*: mobile phones limited memory must be taken into account while working on any filtering technique for spam SMS.

Other trouble in designing SMS spam discriminator is that the performance of already established and developed email spam filtering techniques drastically degrades when employed to mobile SMS spam, the reason is standard Short Message Service size is bound to 140 bytes that is equal to 160 characters of English alphabet when translated. In addition SMS text is excessively rich in abbreviations, acronyms and idioms.

From the literature review presented in chapter 2; it has been noticed that a lot of work has been done in the field of text classification based on non-content features, but content based spam SMS filtering is not very much discussed yet. An efficient spam SMS filter will be introduced through this proposed solution while considering the low processing power and limited memory of mobile phones and also changing nature of short message service. It will help to filter the spam text messages from mobile point of view. The proposed work will be considered as a good step towards the field of SMS spam detection.

1.3 Contribution

To propose a spam SMS differentiation technique using a content-based approach with low false positive rate has been the aim of this research work. The utilization of Machine Learning algorithms for the goal of Spam Short Message Service- SMS filtering from mobile data with the training parameters like SMS size, presence and count of Spam words in the text message as the key parameters is a novel aspect in this work.

1.4 Organization

In the next chapter 2 a detail literature review of spam SMS filtration techniques based on content-based approaches and non-content based approaches is presented.

In chapter 3 of this thesis proposed methodology with detail step by step description and flow diagrams are presented. In chapter 4 Results and Discussions including experimental setup and empirical results are shown. Also results comparison of different Machine learning algorithms while using same set of extracted features are shown in the same chapter. Finally in last chapter 5; Conclusions are summarized. Future work is also discussed in the same chapter.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Today among the different services of data communication available on mobile phones, one of the most wanted and widely adopted service is Short Message Service (SMS). What makes Short Message Service (SMS) so popular are its ease of use for mobile phone users, lower rates and big revenues service providers are generating from it. With the rapid increase in the use of Short Message service, world is also facing a critical problem of SMS spamming. SMS spam is a kind of spamming carried out at the mobile phones short messaging service. It not only reduce the quality of service of cellular operators but also harm the privacy of users. Different techniques based on both content and non-content approaches have been proposed by researchers for differentiating Spam SMS.

The purpose of this review is to examine the different techniques presented by the researchers for the detection and differentiating Spam SMS from valid or non-spam SMS. The compatibility of these techniques with the mobile phones limited memory resources, lower processing power of mobile phones and power consumption is also noted in this review.

2.2 Techniques Based on Non-Content Features of Text Message

Classification techniques developed for differentiation of spam SMS-Short Message Service that are based on non-content attributes examines and investigate meta data of the message; for instance message size and network features like number of recipients, clustering coefficient and route of the message [2].

In the following section a number of studies on techniques developed for Spam SMS differentiation from mobile data and based on non-content characteristics of the message are discussed.

2.2.1 Analysis of different temporal attributes of network used for SMS communication

In [3] researchers of Hong Kong University of Science and Technology presents SMS spam countering algorithm while utilizing different attributes of the network used by the short message service for communication. Also the characteristics of various groups of SMS metadata including temporal attributes along with static features are inspected in this proposed methodology. Feature Description is summarized in table 2.1.

Network feature	Cluster coefficient Number of recipients
Static attributes	Reply ratio Total size of messages for specific time duration Total count of messages for a particular time period
Temporal features	Average SMS delivery time for seven days Average sending time gap of messages for seven days Average count of recipients for seven days

Table 2.1: Features Set

SVM-Support vector machine classification algorithm is then populated with the above features to assess its efficiency on real dataset of short message service. While evaluation on a dataset of real SMS it is shown that incorporation of network attributes and temporal features to construct an SVM classifier, 8% improvements on area under the curve (AUC) can be achieved as compared to only using static metadata features. In this work KNN classifier also trained on the same attributes, and on the basis of results comparison shown in this research study, SVM classifier surpass KNN classifier when SMS spammers are predicted. However the major loophole in this work is, the effect of change of network connected with a node in a certain period of time is not discussed.

2.2.2 Use of Message Discrimination Module (MDM) at routing node

In .[4], a filtering mechanism for spam SMS is proposed by Rick L. Allison and Peter J. Marsico, in which they suggest to include a discrimination module in the routing node for SMS messages. The SMS differentiating module is basically a database build-up of information that is utilized in decision making whether a sent message to the recipient is wanted or positive SMS for called party or recipient. The MDM-message discriminating module also contains processing instructions for differentiating spam SMS. This differentiating module is also capable of examining the information about sender and recipient and the information about routing path contained in the SMS header in order to determine and detect the repeated spam messages originator source. In this way it would be possible for the proposed module for differentiating spam SMS to determine the specific network, communication peripheral(s) and elements of the network from which unwanted or spam messages are generated and flooded into the network. Once the originator of the spam messages is recognised then network operator(s), through a new short message service can be notified about the spamming event so that the access of the spam SMS originator to the network and use of resources of the network can be blocked through suitable action.

In the figure 2.1, a network architecture diagram including a node for routing messages having proposed SMS-DM (short message service discrimination module) is shown.

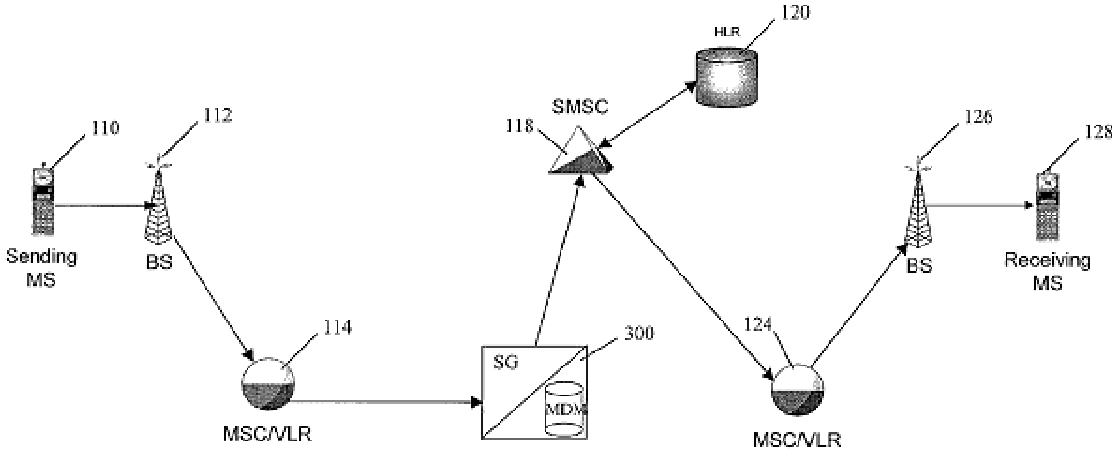


FIGURE 2.1: Network diagram containing SMS differentiating module at routing node.

In figure 2.1, a sending end terminal of mobile communication network usually pointed by a reference numeral 110 is shown that construct and frame an SMS to send. The base station indicated by a reference numeral 112, manages the interface between the physical layer of the network and a wireless medium, to make the wireless transmission of SMS messages reliable. The reference numeral 114 represents main switching centre-MSC. Since it has the ability to forward an SMS received from a wireless communication network to the suitable SMSC, it is also sometimes known as SMS-IWMSC (short message service interworking main switching centre). According to the proposed methodology the reference numeral 300 indicates a SG-signalling gateway with an MDM i.e. discrimination module for short messages. The unit that manages forwarding, relaying and storing of SMS between receiving and sending peripherals is SMSC, indicated by reference numeral 118. In order to organize profiles of mobile subscribers and to manage location information of mobile subscribers, a database is used as a tool for permanent storage as represented by reference numeral 120 HLR-Home Location Register. VLR-Visitor location register that is a part of database is used for the temporary storage of information about those subscribers who are roaming in the VLR serviced area. In the shown network diagram a base station is indicated by a reference numeral 126 and receiving mobile end by 128.

2.3 Techniques Based on Message Content Features

Another approach used by the researchers for differentiating spam SMS is content-based filtering. Filtering techniques based on contents of the message mainly focus on the detail description of the contents of the message. In this approach content of the message is matched to the user defined characteristics. Some of the research studies presenting different SMS discrimination techniques based on content of the message are discussed below.

2.3.1 Using Artificial Immune System

In research work [1] Tarek and Ahmad suggested anti-spam filtering methodology that is based on the principles of artificial immune system. The idea is motivated by Biological immune system which is a system of different and connected tissues, organs, cells and process that struggles and fights against infections and prevents attacks in future by growing immunity in advance. The ability to differentiate between self and non-self-bodies is the distinctive attribute of the immune system [5]. The proposed anti-spam SMS filtering technique is combination of three filtering layers. In first layer all those phone numbers are listed that the mobile phone user desires to block. In this way, the SMS messages sent from these numbers will be blocked in black and white by the proposed filtering technique. In the second layer a list of all those possible spam words is implemented that might contain a spam SMS message. In this case the proposed SMS differentiating technique will filter out all the incoming messages that contains such words. In the third filtration layer, a black list containing all detectors is used. The detectors are generated from the feedback of mobile phone users and training process. At this layer, the incoming SMS is analyzed by the proposed technique to make decision whether it is spam or not based on the resemblance ratio between the black list detectors and the incoming message. Through this mechanism the incoming SMSes that match these black list detectors will be blocked. A tokenizer is used to break the incoming message into informal pieces. These tokenized pieces can be separate single words or a combination of words. This tokenized data is then passed to analysis engine to analyze the incoming message and to reach a reasonable conclusion about the spamminess of the incoming SMS message. Based on the results shown in this research study, the authors believe that the proposed anti-spam filtering technique performs better as compared to Naive Bayesian algorithm.

2.3.2 Finding correlation between the SMS sender, receiver and message contents

In a research study [6] a team of researchers presents a technique for differentiating spam SMS from non-spam messages. The whole idea is built on the correlation between the message sender and receiver and message text. If the proposed system finds no mutual relationship between the SMS sender party and message recipient and the message contents contains the spamming material then SMS message is labelled with spam by the system and message is put into the spam box. On the other hand, if the system finds and correlation between the sender and receiving entities and message contents have no spam material then the incoming SMS message is allowed to pass to the receiver inbox.

The proposed spam SMS detection system is based on four phases. In first phase, the goal is to design and build a communication system for text messages that facilitates the mobile phone users to register and begin to use short message service-SMS. In second phase, the objective is to develop an application server at mobile service provider side that is capable of allowing the registered user to connect and communicate. It will be responsible for managing the transfer of SMS message properly from sender to recipient.

Third phase is the designing of database to store the all SMS logs and predefined spam messages contents in the indexed format properly so that reliable access and faster retrieval of information can be made possible any time. The most important phase 4 is the development of analysis service for database in order to discover the correlation or direct relation between the message sender and receiving party and to examine the message contents. Based on the results calculated by this analysis service the incoming SMS is then allowed to forward to the receiver with a tag of spam or normal.

In the suggested system presented in this research work, the text message will be first unicasted or multicoated by the sender party which will be arrived at the server of mobile service provider. The analysis service module is then used to get the conclusion result in spam/negative or positive format, once the message lands at the server. In order to find any mutual relationship between the message sender and receiver, the SMS log between the SMS sender and receiver is looked into detail by the relation analyser. The system will also examine and scan for duplicate messages and the contents of the message will also be checked. Once the successful results are calculated in proper format by the relation analyser, tag of normal or spam is applied to the message and then the message is forwarded to the receiver or discarded by the system on the basis of calculated results from the relation analysis system. The authors of this work are of the view that by deploying this system the problems like balance deduction in some countries and wastage of SMS memory can be solved.

2.3.3 Text categorization for SMS Spam Filtering based on online Indexing

In this work Liu and Wang proposed a mechanism for text classification for spam SMS filtering that is based on online indexing [7]. Two index model structures i.e. index model for document-level and index model for word-level are presented for index based online text classification for spam SMS filtering. A pair of index i.e. [SpamIndex, HamIndex] is constructed in document-level index model. In this set of index, an inverted list data structure is denoted by each index, which stores [Term/Token, set of Message ID] where the set of message ID will store the ID of the Term contained in the body of the text message. Two index pairs i.e. [SpamPositionIndex, HamPositionIndex] and [SpamFrequencyIndex, HamFrequencyIndex] are included in the word-level model for indexing. Conditional probabilities for both spam and ham messages are calculated and then based on these conditional probabilities, the message category is allocated in accordance with bigger probability of the index category. For practical deployment SS-Spamminess Score is used to demonstrate the spam probability of the message that is being processed. The proposed framework is made up of two sections working parallel. In incremental indexing part, a labelled message is passed through the feature extractor to analyze it and extract different attributes of the text message. The Index updater then adds the tokens into ham index or spam index accordingly. In the second part i.e. the online filtering, the unlabelled incoming message is passed to feature extractor for analysis to output various features. The authors of this research work believe that the proposed methodology efficiently reduces time cost and shows high precision in differentiating Chinese spam text messages.

2.3.4 Use of Graph Based Learning Model at Application Layer

In this work, a learning model is presented for spam SMS discrimination that is graph-based and the basic approach used is; modelling the message contents and SMS syntax pattern into a weighted graph [8]. Syntactical attributes of the message are used to differentiate spam SMS in the suggested approach.

In the proposed methodology, the messages are transformed into free of space tokens without out applying any technique of deep parsing. This space delimited set of tokens builds the graph node-set. What constitutes the directed edges between the graph nodes is based on the order of tokens occurrence. The graph is constructed from the training data containing labelled text messages. Once graph is built then the probability of appearance and correlation among words present in the spam and ham messages is calculated. After training phase i.e. construction of graph and probabilities computation; these calculated probabilities are used as input to the differentiation module, which act as a binary discriminator to differentiate the message as either benign or spam. The decision to classify the incoming message as being spam or ham is based on divergence between the nodes probabilities and probabilities of links or edges in the graph. In this research work; relative entropy or KL-Divergence; a mostly adopted non experimental measure is employed for the detection of spam messages in real-time. KL-Divergence makes the core of statistical theory and is used for distance computation between distributions of two probabilities. Authors of this study believe that the proposed methodology is capable of differentiating spam messages efficiently including advertisement messages, promotions and offers from service providers and telecom companies and also push messages. Also the nature of the proposed architecture of the spam detection mechanism is modular and thus independent of the architectural design of the smart phones. In addition to modular nature, the proposed methodology is syntax independent and resource efficient.

2.3.5 Use of Hidden Markov Models (HMMS) Utilizing Byte-level Distribution

In [9] Zubair and Mudassir propose a methodology for SMS spam detection while utilising the underneath byte-level pattern of data coding scheme. This proposed scheme is designed to operate on the mobile phones access layer. In this methodology first byte-level distributions of spam and ham messages is deeply analyzed and then using HMMS-Hidden Markov Models; models for benign and spam messages is constructed. Hidden Markov Model is build-up of five quantities
M = $(\pi, T, E, \theta, \Phi) : i.e.$

$\pi = initialprobabilities$
$T = statetransitionprobabilities$
$E = outputemissionsymbolprobabilities$
$\theta = asetofoutputsymbols$
$\Phi = asetofstates$

Hidden Markov Model works under the following two conditions:

1. State transition probabilities satisfy the Markov property i.e. the current state or the active state depends only on the immediate previous state and not on all previous states.
2. Output emission probabilities of a symbol depends only on current state.

After computing transition probabilities, Spam(score) for each SMS in training data is calculated so that to have a threshold for SMS spam detection. According to the authors; the framework proposed in this research study is lightweight as its memory requirements are less than 512 KBs and it operates at access layer of mobile phones; capable of detecting spam SMS in as short time as less than 1 millisecond with a detection rate of more than 95% and false rate of zero.

2.3.6 Use of Evolutionary Learning Classifiers

In a research work [10], a methodology for SMS spam detection is proposed in which two novel features of spam SMS are analyzed in order to differentiate it from non-spam i.e. wanted SMS. And all this analysis phase is performed at the access layer of mobile phone and the following two attributes are extracted; pattern distribution of octets and bigrams of octet in hexadecimal structure.

The whole process of Spam SMS detection is designed to complete in two steps. In first step, In order to train the algorithm the system is loaded with a set of messages containing both spam and benign SMS messages. Then different machine learning algorithms including Fuzzy Ada Boost, C-SVM, Naive Bayes, UCS etc. are provided with a set of three different attributes; only frequency distribution of octets, only octet bigrams, and a combination of these features i.e. a set that is composed of both octet bigrams and frequency distributions of octets. Each of the learning classifier, based on the pre-defined features, extracts the knowledge and construct a model of ham and spam during its learning phase. Once the process of training the algorithm is completed, the system is then utilized to discriminate the incoming SMS in real time. At the end classifier showing best performance is taken into account.

Following three matrices are used to define the classifiers performance i.e. negative alarm rate, detection rate and testing time. The experimental results shown in this research study suggests that UCS-Uniform-cost search algorithm is feasible to implement it on real mobile phone devices as it provides average 93% spam detection performance with 0% false negatives. Also time it takes to classify single SMS is one second approximately with low memory requirements.

Authors of this research work believe that this novel mechanism for spam SMS detection leads over other methodologies that search for contents or specific keywords in message. I.e. acronyms and abbreviations are ignored; independent of geo-location, culture and community.

It is also independent of local languages and also ignores punctuation or exclamation marks as hexadecimal format of octet values is used in all this process of spam SMS detection.

2.3.7 Use of SVM- Support Vector Machine and Nave Bayesian

In [11] a content base Spam SMS classification technique is presented while keeping in mind the dynamic nature of message contents and unstructured format of

Short Message Service as compared to email that consist of specific structure information; for instance mail header, senders address, subject etc.

In the proposed methodology a supervised learning algorithm Nave Bayesian is used to model the changing behaviour of text messages while utilizing probability theory and SVM- Support Vector Machine represents those using different attributes. In first phase the SMS data is pre-processed to transform the text messages to uniform SMS format by analyzing the collection of documents semantically or syntactically. The main objective of applying algorithm at this stage is the removal of stop words, number removal stemming and to make the text free of white spaces. Algorithms are trained on pre-defined attributes set of real data. Incoming SMS is first tokenized and then passed through the trained system. On the basis of calculated spamming probability the incoming SMS is declared as ham or spam. Experimental results shown in this research study, presents the following findings; the filtering accuracy is 92.74% while Nave Bayes classifier is used and 87.15% classification accuracy was achieved for Support Vector Machine classifier. It must be noted here that all this practice was carried out on Nepali SMS dataset.

2.3.8 Probabilistic Topic Modelling and Stacked DE noising Auto encoder approach

A recently developed well known text mining methodology is adopted for Spam SMS classification in this research work [12]. Probabilistic topic modelling approach is used to extract hidden features or topics that are related to SMS statically and then SDA -Stacked DE noising Auto encoder is used to develop a complex data model. What makes topic modelling the best choice is its capability to handle text of any size robustly and seamlessly. After creating topics for each SMS, they are used as input to unsupervised learning approach, SDA -stacked denoising auto-encoders; in order to construct a comprehensive data model. The choice of using probabilistic topic modelling solves the problem of features selection that may be very sparse due to small size and limited text in the SMS messages. Also the pre-defined selected attributes are very hard to adapt to new emerging SMS spam patterns as they are usually hard coded and not dynamic nature. On the other hand topic modelling approach based on probability is a text mining method automatically trace and point out topics within a group of messages and models pattern or signature in the text. The only requirement of using this approach is to define the maximum number topics. Also only very simple steps of pre-processing are required i.e. removal of stop words and tokenisation. This approach also solves the problem of unavailability of labelled data as deep neural network; an unsupervised learning technique is used to utilise unlabelled data that was considered useless previously.

Authors of this work believe that by proposing topic modelling approach as an attribute extraction technique provides solution to handle several drawbacks of the previous proposed methodologies for SMS spam detection.

2.3.9 Two Layer Spam SMS Filtering Technique for Mobile Communication

In this research study [13], a hybrid framework for Spam SMS classification is presented i.e. filtering based on content of the message and challenge-response. In this proposed methodology when the system stuck in taking decision to declare the incoming SMS spam or ham by analyzing contents of the message, then the incoming message is further investigated by sending a challenge message to the message originator. Since an automated bulk spam SMS generator is unable to respond correctly to the challenge message, in that case, the SMS is marked as spam message.

The core functionality of the proposed hybrid framework is as follows; the incoming SMS messages are divided into three separate categories by using content based filtering technique: spam, uncertain and ham. Since filtering mechanism fails in case of uncertain messages, then another approach of challenge-response is used to further differentiate uncertain SMS messages into spam and ham categories.

The interaction of hybrid framework with other stockholders i.e. message recipient, SMSC- Short Message Service Center and message sender is shown in the high level block diagram Fig. (2.2). As shown in the overview diagram the mobile operator SMSC sends a challenge response query in case of uncertain message to test whether the message originator is spammer machine or human. The sender responds the SMSC generated query by answering it and the SMSC then make a comparison of the senders response with the known actual value. If the value is matched, the SMS is categorized as not-spam/ham otherwise it is marked as spam.

Authors of this research work presents the following arguments in favour of running this hybrid framework at short message service centre:

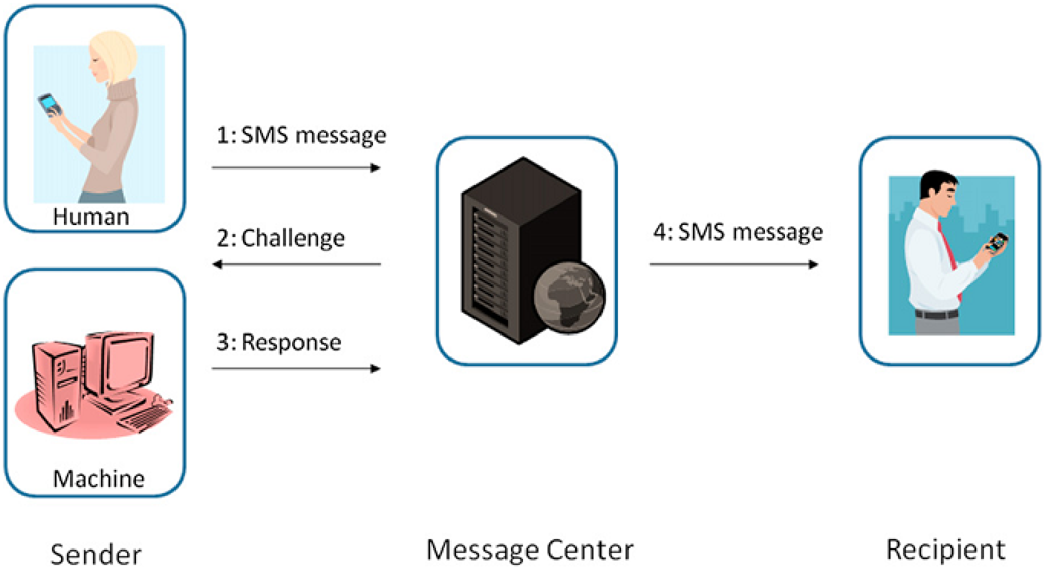


FIGURE 2.2: Hybrid Spam Filtering Overview

- By deploying the suggested hybrid framework at short message service centre will help in reduction of the SMS traffic usage as spam SMS messages will be filtered out before delivered to the bulk target recipients.

- Through the use of challenge-response mechanism [in which a challenge response message, for instance; an image or text CAPTCHA is sent to the sender in case of uncertain SMS messages: a spamming programme will not be able to reply with a correct answer while a legitimate user will most probably answer correctly], a real time sample data in huge amount will be collected at short message service centre; these can then be utilized in developing highly efficient spam SMS classifiers and will help in improving the filtering algorithms performance..

- Another argument by the researchers of this study in favour of deploying this hybrid framework at short message service centre is; it will provide one solution to all users connected to the SMSC instead of performing a large scale practice of installing and maintain an anti-spam filtering software on every single mobile device.

Chapter 3

PROPOSED SOLUTION

3.1 Proposed Spam Short Messaging Service-SMS Filtering Technique

In the proposed solution 3.1 for differentiating spam messages from mobile data ; the very first step after collection of data set is the labeling of messages as spam or ham which is a basic requirement for supervised machine learning techniques. It has been noticed that there are some attributes that are specifically related to spam messages. In the proposed solution; once we have a data set of labelled messages, the next important step is the extraction of some pre-defined features i.e. **s**ize of message, **s**pam words frequency, **s**pam words pattern etc. These extracted features are then used to train the learning algorithm. In 3.1.3; the training and learning process of algorithm i.e. construction of Decision-Tree learning algorithm is explained in detail. Any incoming message whether it is spam or ham in nature is first tokenized and same features are extracted which were used in the training data and then these features are passed through the trained learning algorithm. Once all calculations are performed on the incoming message; it is declared as spam or ham message by the trained learning algorithm.

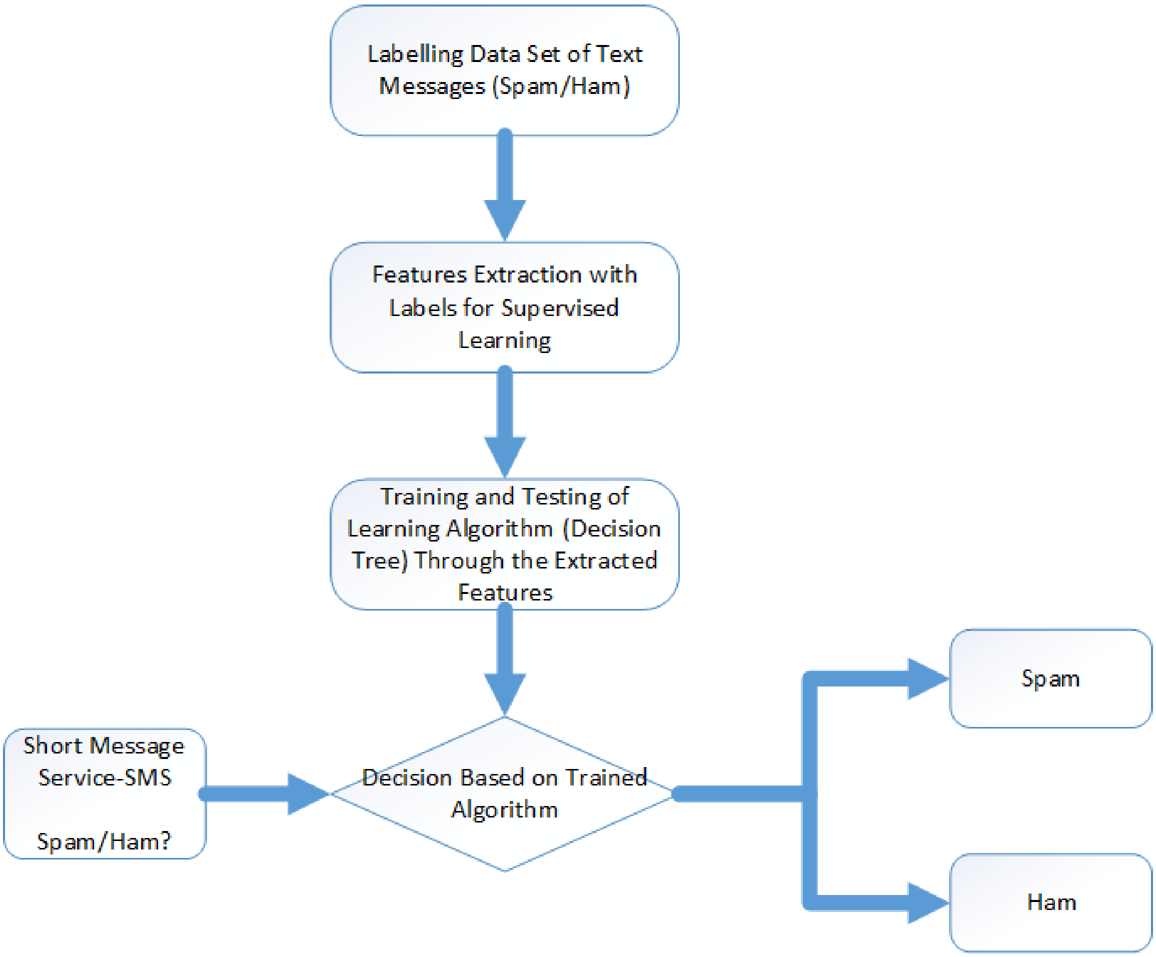

Labelling Data Set of Text Messages (Spam/Ham)
Features Extraction with Labels for Supervised Learning
Training and Testing of Learning Algorithm (Decision Tree) Through the Extracted Features
Short Message Service-SMS Spam/Ham?
Decision Based on Trained Algorithm
Spam
Ham

FIGURE 3.1: Proposed SMS Spam Filtering Technique

3.1.1 Labelling Data Set of Text Messages for Training of Learning Algorithm

As supervised learning is based on labelled data so the entire data set containing both spam and ham messages is labelled with the out attributes i.e. SPAM or HAM according to the nature and contents of the messages. Flow diagram is shown in fig. 3.2 .

Input:
Collection of Short Message Service-SMS containing both SPAM and HAM (not spam) messages.

Processing:
If (Short Message Service-SMS is Spam)
Label the SMS as SPAM

Else-if (Short Message Service-SMS is Not Spam)
Label the SMS as HAM

Output:
Labelled Data Set of Short Message Service-SMS

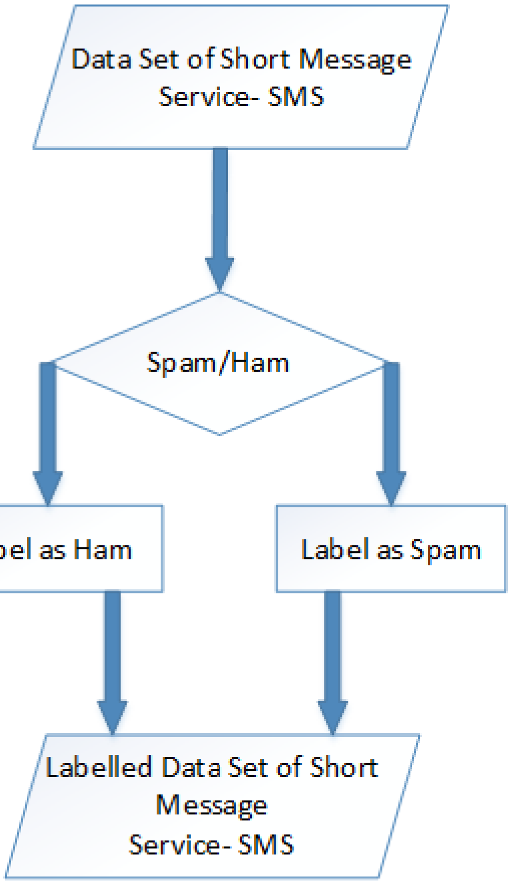

Data Set of Short Message Service- SMS
Spam/Ham
Label as Ham
Label as Spam
Labelled Data Set of Short Message Service- SMS

FIGURE 3.2: Labelling of Data Set

3.1.2 Features Extraction from the Labelled Data

Another aspect of learning algorithm is that; it does not accept raw data meaning that learning algorithm can be trained on explicitly defined features that are extracted from the collection of raw data. Since the proposed solution for differentiating spam SMS from mobile data is solely based on the contents of the message; the following contents attributes are taken into account while extracting features: (1) size of the message; usually spam messages are lengthy as compared to non-spam messages, this attribute can play a vital role in differentiating spam messages from ham (non-spam) messages (2) occurrence of spam words; spam messages mostly contains spam words and characters like FREE, OFFER, @, +, !, tax, Rs. Discount etc. (3) frequency of combination of spam words; only depending on single spam words while deciding whether an incoming message is spam or ham is a bit risky and it may result in increased false negatives so along with stand-alone spam words, pattern and combination of spam words is also considered i.e. @Rs., +tax , FREE! etc.

Input:
Labelled Data Set of Short Message Service-SMS

Processing:

- Calculate size of each message by counting total no. of characters containing the message.
- Tokenize each labelled Short Message Service-SMS using space as delimiter.
- Count total no. of words in the message that matches with the pre-defined spam words.
- Count total no. of combination of words in the message that matches the pre-defined combination of spam words.

Output: Data Set of Labelled Extracted Features/Attributes

3.1.3 Training Learning Algorithm

The data set of extracted features prepared in the previous step along with the output labels i.e. SPAM or HAM, is used to train the learning algorithm. Flow diagram is shown below in fig. 3.3.

In the proposed solution Decision-Tree learning algorithm is constructed and trained through the learning process. The construction of decision tree follows top-down fashion, but matter of concern is how to select which feature to split each node? The answers lies in finding the feature that doesnt contain mix of both SPAM and HAM i.e. feature that is capable of differentiating the target class explicitly into the best possible pure child nodes. This degree of purity (measure of certainty) is called ***information***. It basically shows the expected amount of required information needed to identify whether a new incoming short messaging service-SMS should be declared as SPAM or HAM.

On the other hand Entropy is the degree of impurity (the uncertainty; opposite of purity). For a class with binary values SPAM/HAM; it can be defined as:

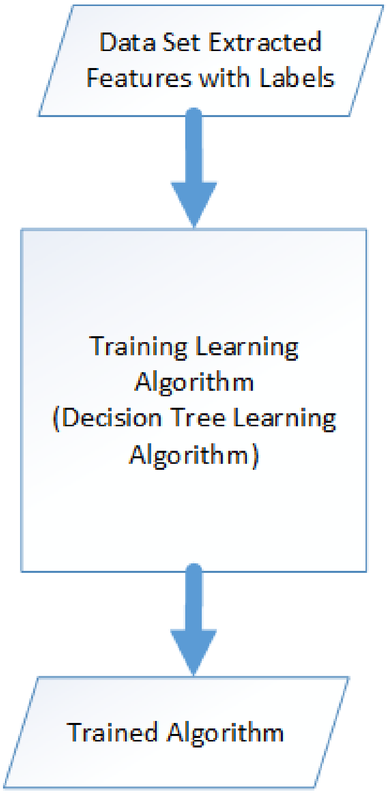

Data Set Extracted Features with Labels
Training Learning Algorithm (Decision Tree Learning Algorithm)
Trained Algorithm

FIGURE 3.3: Training learning algorithm

Entropy = - p(a)*log(p(a)) - p(b)*log(p(b))

In figure 2.8 the binary entropy function graph is shown (random variable can be one of two values i.e. SPAM or HAM). It touch peak value when probability is p=1/2 i.e. p(X=SPAM) = 0.5 or similarly p(X=HAM) = 0.5 having a chance of 50%/50% of being declared as either SPAM or HAM (the degree of uncertainty is at peak). The value of entropy function is at minimum zero when probability p=0 or p=1 with zero uncertainty [p(X=SPAM)=0 or p(X=HAM)=1 respectively] as shown in fig. 3.4

For more than just two outcomes i.e. N outcomes; the generalized form of entropy definition for a discrete random variable X can be written as:

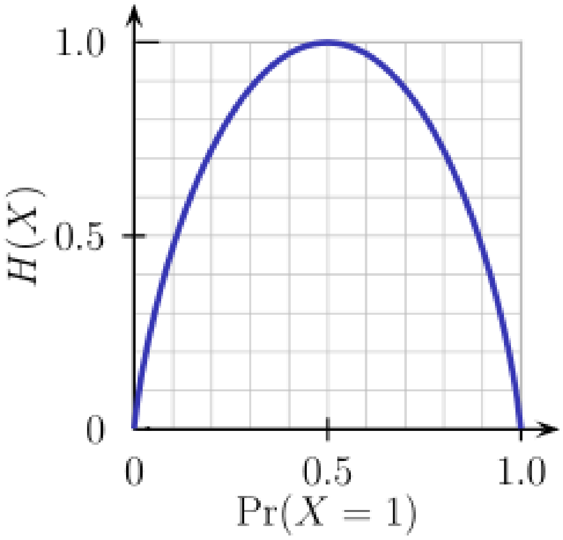


FIGURE 3.4: Binary entropy function graph

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

FIGURE 3.5: Generlized form of Entropy formula

(Commonly logarithm to the base 2 is used as log in the above formula)
Now back to our goal of classifying messages into SPAM and HAM; first we calculate Entropy before split at some point during the tree construction process. Next we make a comparison of this computed Entropy before split with Entropy of two children branches calculated after the split. After splitting the node; the final entropy is computed by combining the left and right entropies, taking the number of samples down each branch as a weight factor.

Now we can obtain degree of ***information gain*** by making a comparison of entropies before and after split or in other words it can be said how much information we obtained by performing the split using that specific attribute/feature:

Information Gain = Entropy(BEFORE) Entropy(AFTER)

This whole process of calculating entropy before and after split and then information Gain is performed for every attribute at each node of tree and then in greedy fashion; the feature with the least uncertainty or largest information gain is selected for the split. This process is performed in a recursive manner starting from rood node down, and terminates when leaf or external nodes contains samples all having the same group/class. Flow diagram of the Decision tree construction process is shown in Figure 3.6.

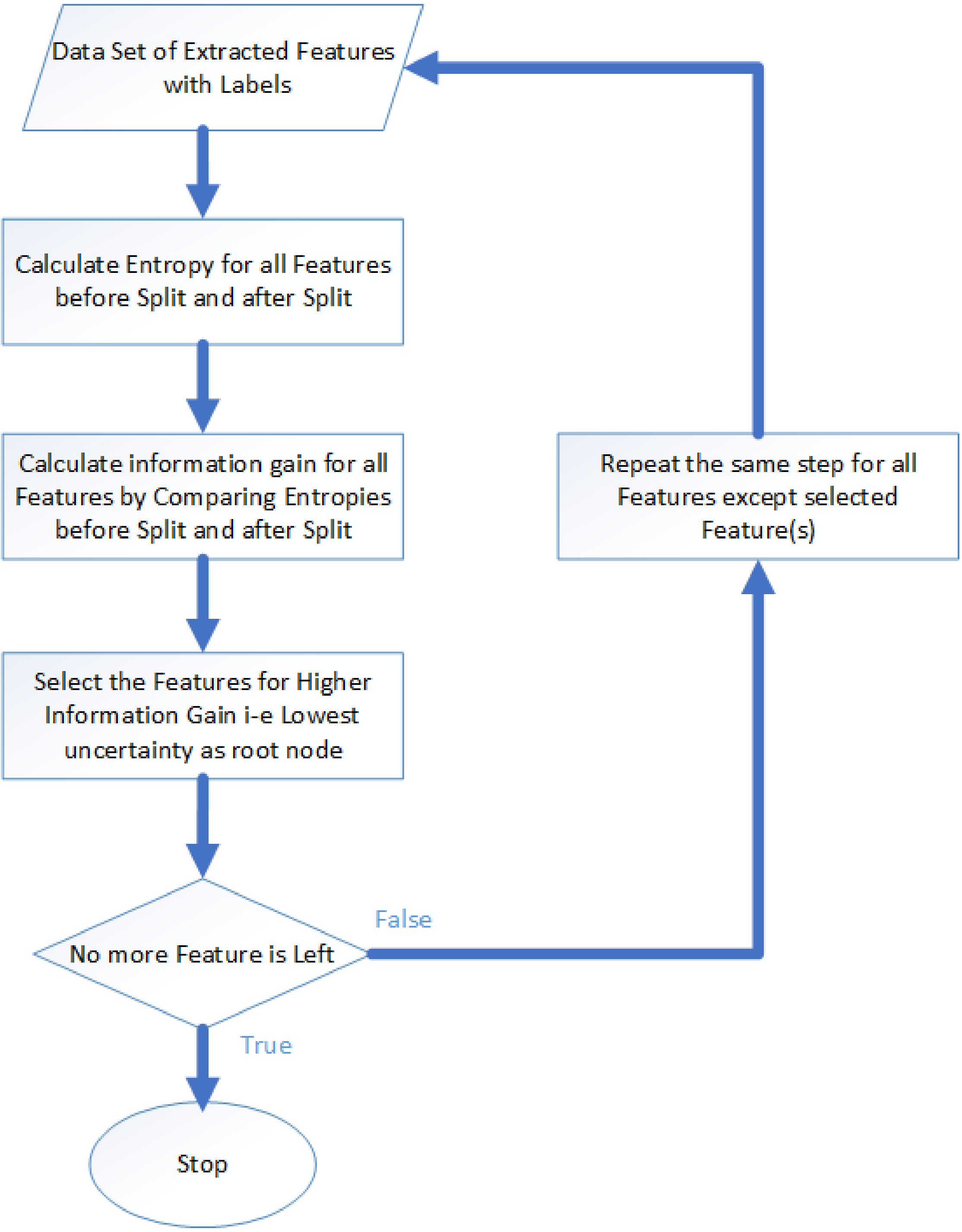

Data Set of Extracted Features with Labels
Calculate Entropy for all Features before Split and after Split
Calculate information gain for all Features by Comparing Entropies before Split and after Split
Select the Features for Higher Information Gain i-e Lowest uncertainty as root node
No more Feature is Left
False
True
Stop
Repeat the same step for all Features except selected Feature(s)

FIGURE 3.6: Construction of Decision Tree Flow Diagram

3.1.4 Testing of Trained Learning Algorithm

The trained learning algorithm build on pre-defined features in then tested on un-labelled incoming text messages to differentiate them as SPAM or HAM message. Flow diagram of the process is shown in fig. 3.7. As soon as the incoming message is received, it is first tokenized and then predefined features are extracted and these features are passed through the trained learning algorithm in order to declare it as SPAM or HAM.

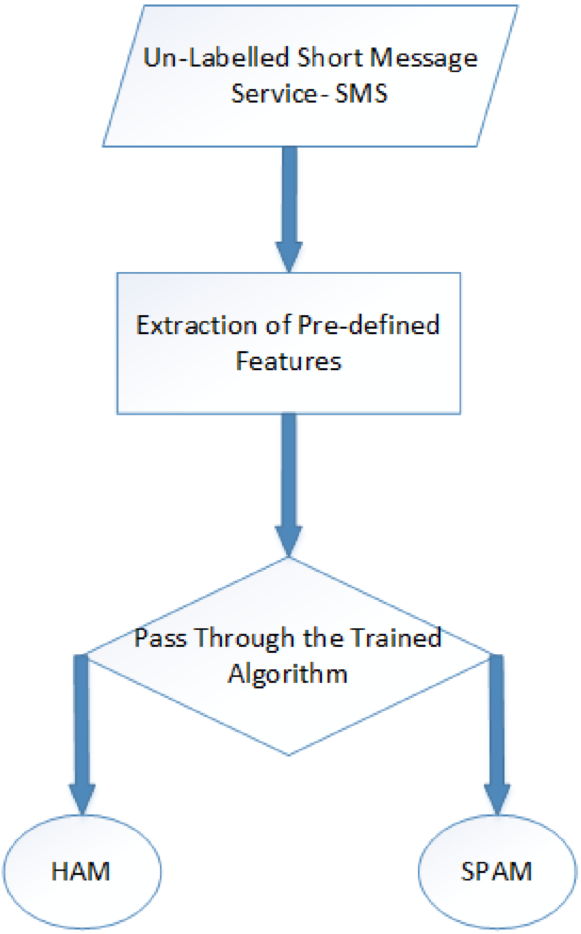

Un-Labelled Short Message Service- SMS
Extraction of Pre-defined Features
Pass Through the Trained Algorithm
HAM
SPAM

FIGURE 3.7: Testing of Trained Learning Algorithm

Input: Un-Labelled Short Message Service-SMS
Input data in records of the form;
(a,B) = a1, a2, a3 . ak, B

Where:

a: Short Message Service-SMS samples

B: Target Output i.e. Spam/Ham

Processing:

- Extraction of pre-defined features
- Pass through the Trained Learning Algorithm

Output:
Classified Short Message Service-SMS as SPAM or HAM

Chapter 4

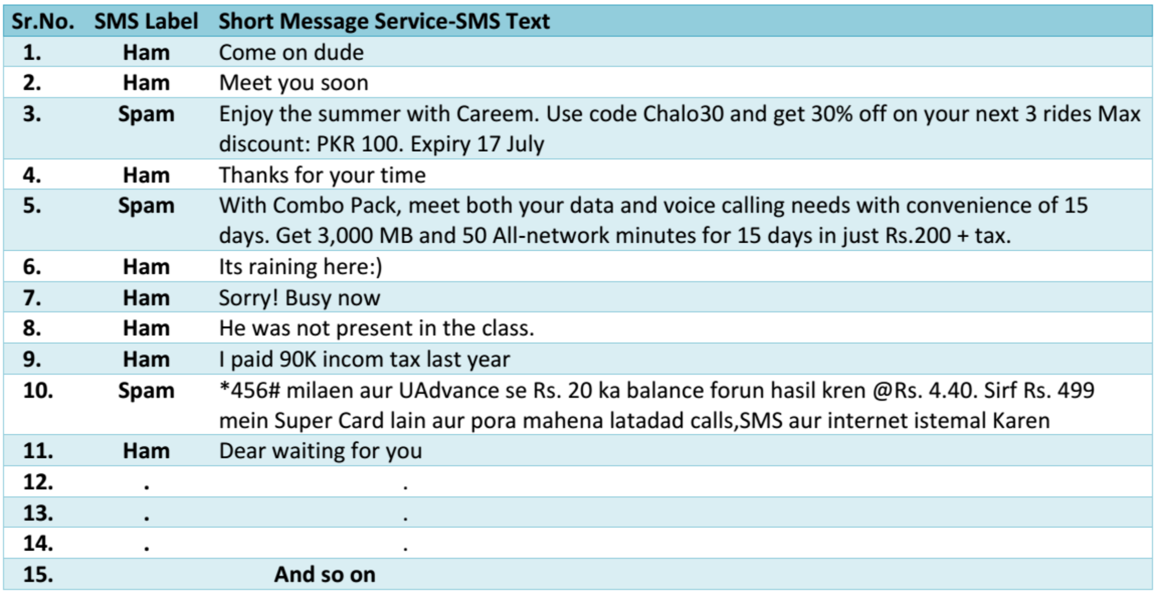
RESULTS AND DISCUSSIONS

4.1 Experimental Set Up

As discussed in previous chapter; the approach adopted in this work for differentiating Short Message Service-SMS from mobile data is based on machine learning. Following the proposed methodology; each step is evaluated experimentally in this chapter.

4.1.1 Collection of SMS data set

The collection of relevant data leads to effective classifier based on machine learning algorithm. In the first step a real Short Message Service-SMS dataset was built by collecting text messages using different mobile phone networks with several volunteers who gave consent to take part in this work and share their SMSes. In this way a SMSes collection of about 7000 text messages containing both Spam and Ham SMSes was organized. In order to train the algorithm; the basic requirement for supervised learning is to make sure the availability of labelled data as to which category/class it belongs to. So each message is labelled with the output attribute as SPAM or HAM. The below figure 4.1 shows a snapshot of the messages dataset.



Sr.No.	SMS Label	Short Message Service-SMS Text
1.	Ham	Come on dude
2.	Ham	Meet you soon
3.	Spam	Enjoy the summer with Careem. Use code Chalo30 and get 30% off on your next 3 rides Max discount: PKR 100. Expiry 17 July
4.	Ham	Thanks for your time
5.	Spam	With Combo Pack, meet both your data and voice calling needs with convenience of 15 days. Get 3,000 MB and 50 All-network minutes for 15 days in just Rs.200 + tax.
6.	Ham	Its raining here:)
7.	Ham	Sorry! Busy now
8.	Ham	He was not present in the class.
9.	Ham	I paid 90K incom tax last year
10.	Spam	*456# milaen aur UAdvance se Rs. 20 ka balance forun hasil kren @Rs. 4.40. Sirf Rs. 499 mein Super Card lain aur pora mahena latadad calls,SMS aur internet istemal Karen
11.	Ham	Dear waiting for you
12.	.	.
13.	.	.
14.	.	.
15.		And so on

FIGURE 4.1: A snapshot of the SMS dataset

4.1.2 Extraction of Features/Attributes

Any machine learning algorithm is not directly applied on dataset rather it uses the characteristics of each item in the dataset. So after having a labelled dataset of SMSes; the next phase is to extract unique features/attributes from both spam and ham text messages. While keeping in mind the low processing power and memory of the mobile phones and to ensure the system is fast and simple, the following selection of features was made for the classification purpose. The results clearly depicts that these attribute are specific and unique enough to distinguish the spam and ham text messages aside. A set of the following four features represents each text message.

1. Size of each SMS (In terms of no. of characters in each message): Based on common observation that as compared to ham messages, spam messages contains more characters; this feature represents a count of the total characters a message is composed of.

2. Matching spam words list: A list containing spam words, for instance 60 words having peak frequency in the spam text message is compiled. Then the words in the text messages used for training purpose are matched with the words in the spam list. A match of single word would return this feature as Y; in case of matching no word would render N on this feature. The spam list i.e. the list of frequently appeared words in spam messages looks like:

GUARANTEED http:// WINNER! FREE! @Rs. Discount PRIVATE! Off-net URGENT! Chat Subscription Cash offer retailer +tax order /minute download /day www prize shop

3. Count of Spam words matched: This feature returns the matching frequency of words in the training set messages with that of spam list i.e. the number representing the no of words in the message that matched with the spam words list.

4. Matching spam phrases list: In this feature, a combination or a group of words in the text message are matched with a list containing frequently occurring phrases compiled from the text messages training dataset. The spam phrase list looks like:

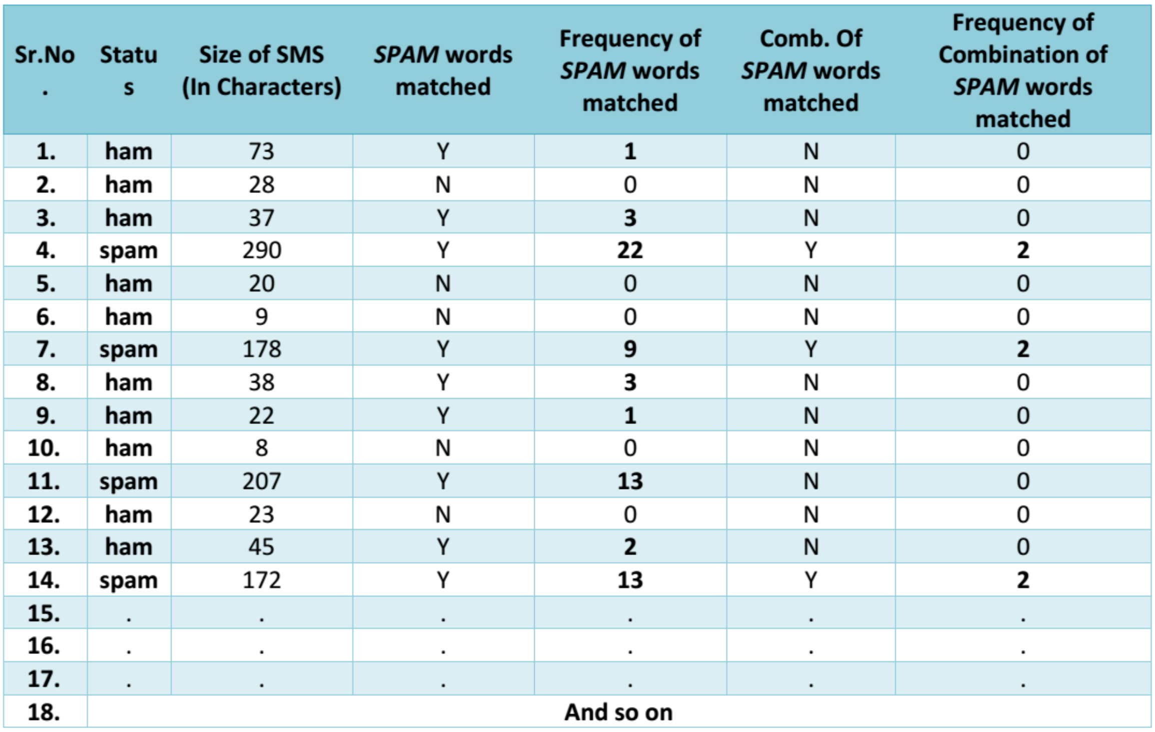
Unlimited SMS — Reply with — GPRS users— just send — SMS alert — Dial now — Unlimited internet — for subscription — Unlimited calls — MMS Package — to subscribe — Activate now — SMS portal — Buy now — Free minutes — Unlimited SMS — Contact us .

5. Count of Spam phrases matched: This feature shows how many combinations of words in the given text messages matched with the spam phrases (composed of two or more spam words) list frequently appeared in the text messages bank used for training.

6. SMS Category/class (Output feature):

- SPAM represents SMS with Spam nature.
- HAM represents Non-spam/valid SMS

MATLAB (matrix laboratory) software is used to extract these features as it provides a user friendly numerical computing environment and high-level language for engineering and scientific computations. The following figure 4.2 shows a snapshot of the features extracted from the SMS dataset.



Sr.No.	Status	Size of SMS (In Characters)	*SPAM* words matched	Frequency of *SPAM* words matched	Comb. Of *SPAM* words matched	Frequency of Combination of *SPAM* words matched
1.	**ham**	73	Y	**1**	N	0
2.	**ham**	28	N	0	N	0
3.	**ham**	37	Y	**3**	N	0
4.	**spam**	290	Y	**22**	Y	**2**
5.	**ham**	20	N	0	N	0
6.	**ham**	9	N	0	N	0
7.	**spam**	178	Y	**9**	Y	**2**
8.	**ham**	38	Y	**3**	N	0
9.	**ham**	22	Y	**1**	N	0
10.	**ham**	8	N	0	N	0
11.	**spam**	207	Y	**13**	N	0
12.	**ham**	23	N	0	N	0
13.	**ham**	45	Y	**2**	N	0
14.	**spam**	172	Y	**13**	Y	**2**
15.	.	.	.	.	.	.
16.	.	.	.	.	.	.
17.	.	.	.	.	.	.
18.	**And so on**					

FIGURE 4.2: A snapshot of the extracted features

4.1.3 Applying Machine Learning

Based on labelled data, the machine learning algorithm learns the correlation between the set of attributes and output class, and therefore it becomes capable of classifying the unlabelled data using the established pattern and leaned relationship. Process diagram of supervised learning algorithm is shown in Figure 4.3.

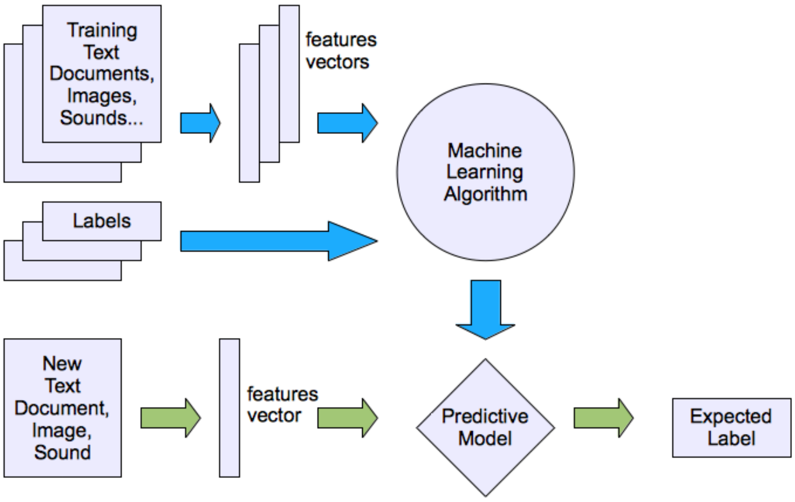

Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
Labels
New Text Document, Image, Sound
features vector
Predictive Model
Expected Label

FIGURE 4.3: Supervised Learning Schematic

4.1.4 Training of Machine Learning Algorithms

In this section, it is demonstrated that differentiation of SPAM messages through machine learning classifiers or classification algorithms is possible while utilizing this data. Several classification algorithms were trained on this data and their effectiveness and classification performance was analyzed. Keeping in mind the people who are not well aware of machine learning concepts and have no expertise in data mining; a machine learning software: WEKA - Waikato Environment for Knowledge Analysis is used. Besides being freely available (GNU public licence) and open source, WEKA software facilitates us with several different machine learning data mining algorithms at one place. The following classification algorithms are used in the experimentation process: nave Bayes, C4.5 decision Tree and Neural Network (Multilayer Perceptron). Since the performance of these classifiers proved satisfactory enough to establish this work.

4.2 Results and Performance of Machine Learning Algorithms

Percentage split Test Method is used to provide data to machine learning algorithm for training and testing purpose. In this method a portion of dataset is selected for training of the machine learning algorithm and the remaining set is used for the testing of the algorithm. In common practice, 66% of the dataset elements are used for the training of the algorithm and 34% is used for the testing of the machine learning algorithm. The results and efficiency of the above mentioned three machine learning algorithms applied on this data is presented empirically here in this section.

4.2.1 Classification of Spam SMS using Nave Bayesian Algorithm

Nave Bayes classifier is based on Bayes theorem which gives us the probability of happening an event based on the two or more independent events probabilities. A Bayes classifier basically use likeliness in the training data and prior knowledge to allocate probability to a class.

The results of applying Bayesian Algorithm can be summarized as:

Features/Attributes: 05

- Status (Spam/ham)-Output feature/attribute
- Size of text message (In terms of No. of Chars)
- Spam words matched
- No. of Spam words matched
- Comb. of spam words matched

Total No. of Instances: 6699

Training Instances: 4,421 (66%)

Testing Instances: 2,278 (34%) Instances Correctly Classified: 2057 (90.2985%)

Instances Incorrectly Classified: 221 (9.7015%)

Time taken to build model: 0.09 seconds

4.2.2 Classification using neural network architecture-Multilayer Perceptron Algorithm

The neural network architecture- Multilayer Perceptron uses back propagation to train itself and classify examples. "The basic building block of Multilayer perceptron algorithm is feed forward artificial neural network model in sets/groups of input data is mapped to appropriate set of output data. Multilayer perceptron is a modified version of the standard neural network architecture-linear perceptron which is a combination of three or more layers of nodes (neurons) along with non-linear activation functions and is more efficient is a sense that it can differentiate data that is linearly not separable .[2]". The flow of data through hidden layers from input to output layer in one direction is termed as feed forward.

The results of applying MLP algorithm while using same features/attributes can be summarized as:

Total No. of Instances: 6699

Training Instances: 4,421 (66%)

Testing Instances: 2,278 (34%)

Instances Correctly Classified: 2075 91.111 %

Instances Incorrectly Classified: 203 08.889 %

Time taken to build model: 3.06 seconds

4.2.3 Classification of Spam SMS using Weka Decision Tree

It was developed and presented by Quinlan [14]. Weka Decision Tree algorithm like other tree based algorithms is built on divide and conquer rule. The data is divided recursively into parts until we have each tree leaf containing only one class instance or further division/partitioning is not possible. In figure 4.4; a comparison of the Machine Learning Algorithms used in this work is shown. The following results obtained by applying J48 decision tree on this data set:

Total No. of Instances: 6699

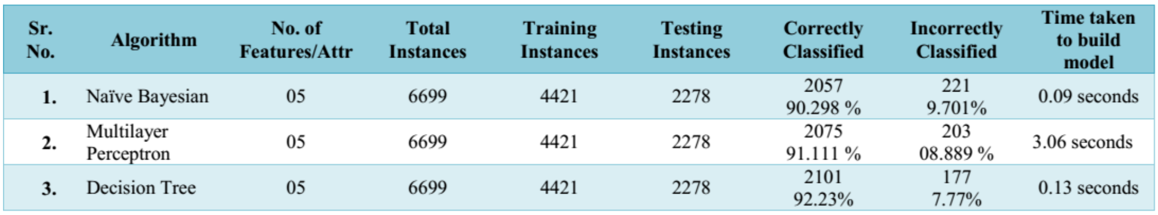
Training Instances: 4,421 (66%)

Testing Instances: 2,278 (34%)

Instances Correctly Classified: 2101 92.23%

Instances Incorrectly Classified: 177 7.77%

Time taken to build model: 0.13 seconds



Sr. No.	Algorithm	No. of Features/Attr	Total Instances	Training Instances	Testing Instances	Correctly Classified	Incorrectly Classified	Time taken to build model
1.	Naïve Bayesian	05	6699	4421	2278	2057 90.298 %	221 9.701%	0.09 seconds
2.	Multilayer Perceptron	05	6699	4421	2278	2075 91.111 %	203 08.889 %	3.06 seconds
3.	Decision Tree	05	6699	4421	2278	2101 92.23%	177 7.77%	0.13 seconds

Figure 4.4: Comparison Table of the Machine Learning Algorithms used in this work

4.2.4 Selection of Decision Tree Algorithm for differentiating spam Short Message Service (SMS) from mobile data

Decision Tree learning algorithm is selected on the following grounds to perform this work:

- After detail analyses of classification performance of all the mentioned machine learning algorithms (presented above in section 4.2); it can be deduced from the empirical results that Decision Tree Learning algorithms performance is better than the remaining algorithms used in this work. So it will better perform in differentiation of spam text messages from mobile data as compare to the other two machine learning algorithms used in this work.

- Keeping in mind the implementation of this work on real mobile devices at users end in future; Decision Tree algorithm requires less computations and memory and is best choice to be implemented on real mobile devices with low processing power and memory having any plate form installed; for instance Android or IOs.

Trained Decision Tree is graphically shown below in figure 4.5:

As discussed above; while using percentage split method, when 34% of the same dataset is used to test this trained decision tree; we have maximum accuracy of 92.23% in classifying spam SMS messages and only 7.77% messages were classified incorrectly that is the highest performance shown by any of the mentioned machine learning algorithms.

Number of Leaves: 23
Size of the tree: 45

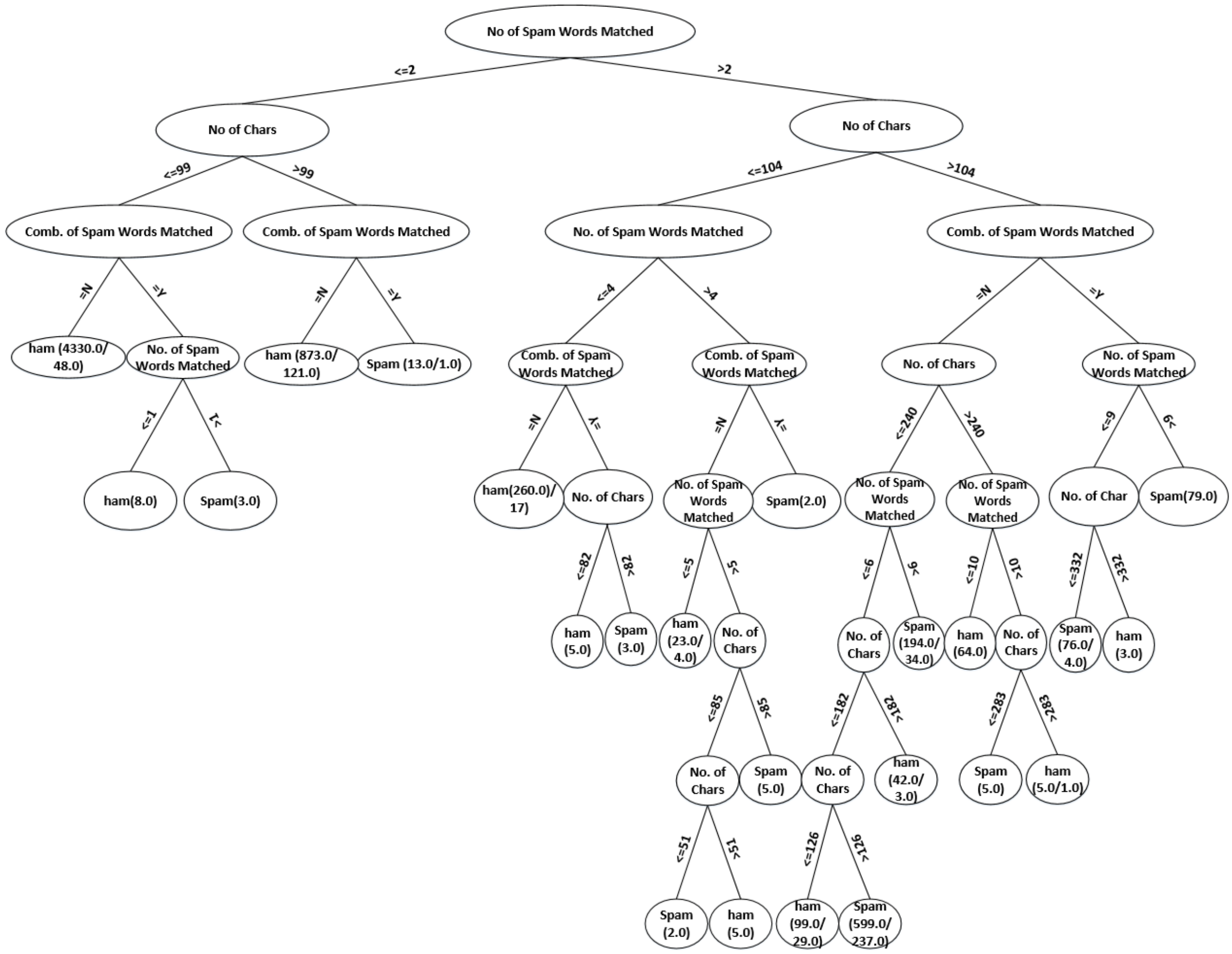


FIGURE 4.5: Trained Decision Tree used in this work

Description of the Trained Decision Tree:

```
No. of Spam words matched <= 2
|   No. of Chars <= 99
|   |   Comb. Of spam words matched = N: ham (4330.0/48.0)
|   |   Comb. Of spam words matched = Y
|   |   |   No. of Spam words matched <= 1: ham (8.0)
|   |   |   No. of Spam words matched > 1: spam (3.0)
|   No. of Chars > 99
|   |   Comb. Of spam words matched = N: ham (873.0/121.0)
|   |   Comb. Of spam words matched = Y: spam (13.0/1.0)
No. of Spam words matched > 2
|   No. of Chars <= 104
|   |   No. of Spam words matched <= 4
|   |   |   Comb. Of spam words matched = N: ham (260.0/17.0)
|   |   |   Comb. Of spam words matched = Y
|   |   |   |   No. of Chars <= 82: ham (5.0)
|   |   |   |   No. of Chars > 82: spam (3.0)
|   |   No. of Spam words matched > 4
|   |   |   Comb. Of spam words matched = N
|   |   |   |   No. of Spam words matched <= 5: ham (23.0/4.0)
|   |   |   |   No. of Spam words matched > 5
|   |   |   |   |   No. of Chars <= 85
|   |   |   |   |   |   No. of Chars <= 51: spam (2.0)
|   |   |   |   |   |   No. of Chars > 51: ham (5.0)
|   |   |   |   |   No. of Chars > 85: spam (5.0)
|   |   |   Comb. Of spam words matched = Y: spam (2.0)
|   No. of Chars > 104
|   |   Comb. Of spam words matched = N
|   |   |   No. of Chars <= 240
|   |   |   |   No. of Spam words matched <= 6
|   |   |   |   |   No. of Chars <= 182
|   |   |   |   |   |   No. of Chars <= 126: ham (99.0/29.0)
|   |   |   |   |   |   No. of Chars > 126: spam (599.0/237.0)
|   |   |   |   |   No. of Chars > 182: ham (42.0/3.0)
|   |   |   |   No. of Spam words matched > 6: spam (194.0/34.0)
|   |   |   No. of Chars > 240
|   |   |   |   No. of Spam words matched <= 10: ham (64.0)
|   |   |   |   No. of Spam words matched > 10
|   |   |   |   |   No. of Chars <= 283: spam (5.0)
|   |   |   |   |   No. of Chars > 283: ham (6.0/1.0)
|   |   Comb. Of spam words matched = Y
|   |   |   No. of Spam words matched <= 9
|   |   |   |   No. of Chars <= 332: spam (76.0/4.0)
|   |   |   |   No. of Chars > 332: ham (3.0)
|   |   |   No. of Spam words matched > 9: spam (79.0)
```

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusion

Spam Short Message Service- SMS has become a serious threat to users privacy and smooth operation of mobile communication networks. It has also a very bad impact on mobile phone network service providers from economic point of view because spam SMSes greatly damage the service quality of cellular operators. Like email spam, SMS spam is also a critical problem, resulting in resource consumption and annoyance at the recipient. The identification and differentiation of spam SMS is important for mobile phone users satisfaction and better quality of mobile communication networks.

From deep literature review presented in chapter 02; it is clear that most of the research work proposes spam Short Message Service-SMS filtering techniques that are based on static non-content features and very few uses content features of the text messages. Secondly machine learning techniques are very rarely utilized in spam SMS discrimination techniques as compared to Email spam filtering techniques.

The aim of this work has been to present a spam SMS differentiating mechanism that uses content attributes of the text message with low false positive rate. The utilization of Machine Learning algorithms for the goal of Spam Short Message Service- SMS filtering from mobile data with the training parameters like SMS size, presence and count of Spam words in the text message as the key parameters is a novel aspect in this work. As a basic requirement for using supervised machine learning algorithm; the relevant training dataset of SMS (Containing both SPAM and HAM messages) has been labelled with the output attributes i.e. spam or ham. Spam messages bear some specific characteristics that have been used to differentiate them from Ham messages in this work. Once a data set of labelled messages is prepared; in the next step predefined features of text messages are extracted. The defining of content based features of text messages has been one of the most important step in this work because based on these features the spammicity of the SMS is decided. The more relevant and features with less uncertainty are selected, the greater accuracy level of the spam SMS filter is achieved. Following features have been extracted from the training dataset of about 7000 text messages; size of text message, Spam words matched (Y/N), count of Spam words matched, combination of spam words matched (Y/N).The selected learning algorithms used in this work were then trained by these set of extracted features. After training of the learning algorithms, they were tested on a set of SMS messages from mobile data. Also the sub part of this work was to compare the results of the proposed mechanism with other machine learning algorithms using the same methodology and same content features of the text message and choose the best from mobile phone point of view.

From the results presented in chapter 04; it has been shown that the accuracy of SMS spam differentiating technique proposed in this research work is above 90% while using any of the three machine learning algorithms selected for this research work on same set of content features of text messages. But Decision J48 tree learning algorithm leads in spam SMS classification performance that is 92.23% which is higher than Multilayer Perceptron learning algorithm that is 91.111% and Nave Bayesian Algorithm that is 90.2985%. Hence a simple content based and a lightweight spam SMS filtering technique is introduced in this research work which has 92% around spam SMS differentiating accuracy. The proposed spam SMS filtering technique also meets the requirements for implementation on real mobile phone devices i.e. the proposed mechanism is feasible for residing on mobile phones with low processing power and limited memory in the form of mobile phone application.

5.2 Future Work

The same set of extracted features can be used to explore and investigate other machine learning algorithms as a further augmentation of this research work and compared with those used in this work. Also the same methodology can be programmed to develop application for real mobile phone devices having different OS i.e. Android, Windows or IOs etc.

The scope of this work may be enlarged by applying the same methodology to text messages written in other languages like Urdu, Chinese etc. except English. Also to further reduce the false positive rate and improve the Spam SMS differentiating efficiency; non-content static features may be combined with the content features extracted in this research work.

Appendix A

MATLAB CODE FOR FEATURES EXTRACTION FROM SMS DATA SET

```
Excel = actxserver ('Excel.Application');
File='C:\Users\majid1\Desktop\2-ExtractedFeaturesofSMS.xlsx';
if ~exist(File,'file')
ExcelWorkbook = Excel.workbooks.Add;
ExcelWorkbook.SaveAs(File,1);
ExcelWorkbook.Close(false);
end
invoke(Excel.Workbooks,'Open',File);
sheet=1;
[ndata, text, alldata] = xlsread('C:\Users\majid1\Desktop\SMS_DataSet.xlsx');
%#Reading Excel file
text=lower(text); %#Converting all the characters
from Upper case to Lower case
A={'FREE! discount jazz indigo < > Mobilink MNP @PKR
Rs. Re. @Re. @Rs. GUARANTEED WINNER! tax/ /SMS /minute /day URGENT!
http:// PRIVATE! representative shop Now!
 Minutes @ %  PKR OFF MB Reply Bundle SMS subscribe
 prepaid RS tax + # offer! Subscriber postpaid * charges
 Dial Customer retailer per collection online order Buy
 subscription call Free won paisa chat claim
 Cash daily Type on-net talk off-net Simply
 prize Ringtone
 unsubscribe MMS SUB Unlimited download
 Talkshawk telenor +tax www.
 Warid zong Zem '};
C={'Activate now|Advertisement portal|just dial|SMS portal|SMS package|
all networks|SMS alert|Contact us|just send|Buy now|
To activate|internet package|to subscribe|Moaziz sarif|
Unlimited internet|Visit now!|postpaid cutomerS|
Unlimited SMS|Dial now|Subscribe now|prepaid customers|
free sms|Free Calls|Call now|Reply now|Unlimited Calls|
Customer service|SMS package|All networks|free minutes|
friends & Family|Only Re.|GPRS users|Non-GPRS users|
Only Rs.|on-net|off-net|reply with|Just visit|per call|
per mint|/minute|MMS Package|plus tax|/mint|to subscribe|
Subscribe no|for subscribtion|subscription to sms|
subscription to ringtone|subscribe to|charged Rs.|
calling customer|call customer|call our customer|free camera|
free mobile|free message|Mobilink customers|SMS portal|
SMS alerts|SMS Package|Telenor customers|Unlimited calls|
Unlimited SMS|Ufone prepaid|Ufone postpaid|Warid prepaid|
web portal|Warid postpaid|Warid telecom|Warid customers|
Zong subscribers|Zong customers|'};
A=lower(A);
C=lower(C);
fprintf('\t')
disp('Sr.#  ham/spam  No. of Char   SPAM Words  Comb.of SPAM words')
for i=2:6700 %# Outer loop equal to the size of Data base

disp(i)
%# Writing SMS Sr.# into Excel file
xlRange = sprintf('A%i', i);
xlswrite1(File,text(i),sheet,sprintf('A%i', i));

fprintf('\b\b')
TF=strcmp(text(i), 'ham');
if(TF==1)
fprintf(' ')
end
disp(text(i))
```

```
%# Writing SMS status into Excel file
if(i>1)
xlRange = sprintf('E%i', i);
xlswrite1(File,text(i),sheet,xlRange);
end
fprintf('\b\b')
fprintf('     ')
z=cellfun('prodofsize',text(i,2));
disp(cellfun('prodofsize',text(i,2)))
%#Writing Total No. of Characters into Excel file
if(i>1)
xlRange = sprintf('B%i', i);
xlswrite1(File,z,sheet,xlRange);
end
B=text(i,2);

count = zeros(numel(A),numel(B));

%# for each string
for x=1:numel(A)
%# split into words
str = textscan(A{x}, '%s', 'Delimiter',' .'); str = str{1};
%# for each word
for y=1:numel(str)
%# count occurences
count(x,:) = count(x,:) + cellfun(@numel, strfind(B,str{y}));
end

fprintf('\b\b')
fprintf('     ')

disp(count)
if(i>1)
xlRange = sprintf('D%i', i);

%#Writing No. of SPAM words Matched

xlswrite1(File,count ,sheet,xlRange);
```

```
end

if(i>1)
xlRange = sprintf(’C%i’, i);

%#Writing "Y" for SPAM words Matched or "N" for SPAM words NOT Matched
if(count~=0)
xlswrite1(File,’Y’,sheet,xlRange);
else
xlswrite1(File,’N’,sheet,xlRange);
end
end

end
D=text(i,2);
count1 = zeros(numel(C),numel(D));
%# for each string
for a=1:numel(C)
%# split into words
str = textscan(C{a}, ’%s’, ’Delimiter’,’|’); str = str{1};
%# for each word
for b=1:numel(str)
%# count occurences
count1(a,:) = count1(a,:) + cellfun(@numel, strfind(D,str{b}));
end
fprintf(’\b\b’)
fprintf(’          ’)

disp(count1)

%#Writing No. of Combination of SPAM words Matched
if(i>1)
xlRange = sprintf(’F%i’, i);

xlswrite1(File,count1,sheet,xlRange);

end
```

```
%#Writing "Y" for Combination of SPAM words
Matched or "N" for Combination of SPAM words NOT Matched
if(i>1)
xlRange = sprintf('E%i', i);
if(count1~=0)
xlswrite1(File,'Y',sheet,xlRange);
else
xlswrite1(File,'N',sheet,xlRange);
end
end
end
end

invoke(Excel.ActiveWorkbook,'Save');
Excel.Quit
Excel.delete
clear Excel
```

References

[1] Tarek M Mahmoud and Ahmed M Mahfouz. Sms spam filtering technique based on artificial immune system. *IJCSI International Journal of Computer Science Issues*, 9(1):589–597, 2015.

[2] Michael W Berry, Azlinah Hj Mohamed, and Bee Wah Yap. *Soft Computing in Data Science: Second International Conference, SCDS 2016, Kuala Lumpur, Malaysia, September 21-22, 2016, Proceedings*, volume 652. Springer, 2016.

[3] Qian Xu, Evan Wei Xiang, Qiang Yang, Jiachun Du, and Jieping Zhong. Sms spam detection using noncontent features. *IEEE Intelligent Systems*, 27(6): 44–51, 2012.

[4] Rick L Allison and Peter J Marsico. Methods and systems for preventing delivery of unwanted short message service (sms) messages, November 16 2004. US Patent 6,819,932.

[5] Anil Somayaji, Steven Hofmeyr, and Stephanie Forrest. Principles of a computer immune system. In *Proceedings of the 1997 workshop on New security paradigms*, pages 75–82. ACM, 1998.

[6] Suraj J Warade, Pritish A Tijare, and Swapnil N Sawalkar. An approach for sms spam detection. *Int. J. Res. Advent Technol*, 2(12):8–11, 2014.

[7] Wuying Liu and Ting Wang. Index-based online text classification for sms spam filtering. *Journal of Computers*, 5(6):844–851, 2015.

[8] Muhammad Zubair Rafique and Muhammad Abulaish. Graph-based learning model for detection of sms spam on smart phones. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International*, pages 1046–1051. IEEE, 2013.

[9] M Zubair Rafique and Muddassar Farooq. Sms spam detection by operating on byte-level distributions using hidden markov models (hmms). In *Proceedings of the 20th virus bulletin international conference*, 2010.

[10] Muhammad Bilal Junaid and Muddassar Farooq. Using evolutionary learning classifiers to do mobilespam (sms) filtering. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 1795–1802. ACM, 2011.

[11] Tej Bahadur Shahi and Abhimanu Yadav. Mobile sms spam filtering for nepali text using naïve bayesian and support vector machine. *International Journal of Intelligence Science*, 4(01):24, 2013.

[12] Noura Al Moubayed, Toby Breckon, Peter Matthews, and A Stephen McGough. Sms spam filtering using probabilistic topic modelling and stacked denoising autoencoder. In *International Conference on Artificial Neural Networks*, pages 423–430. Springer, 2016.

[13] Ji Won Yoon, Hyoungshick Kim, and Jun Ho Huh. Hybrid spam filtering for mobile communication. *computers & security*, 29(4):446–459, 2010.