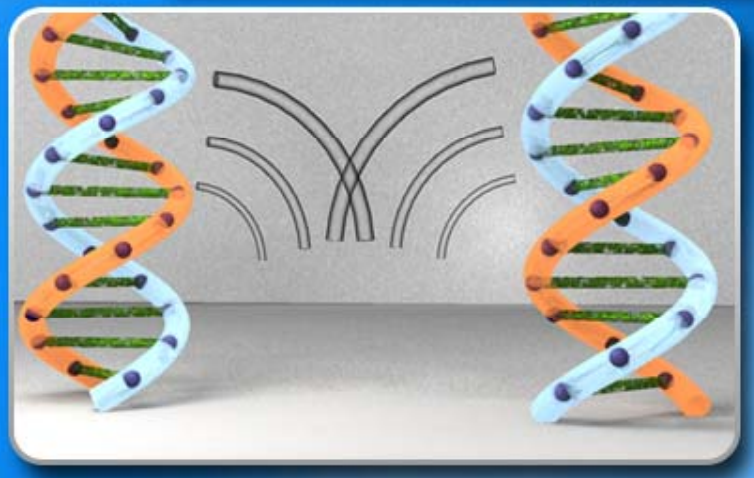
SMS SPAM DETECTION BY OPERATING ON BYTE-LEVEL DISTRIBUTIONS USING HIDDEN MARKOV MODELS (HMMs)

(***Be Liberal in What you Receive on Your Mobile Phone***)



M. Zubair Rafique and Mudassar Farooq

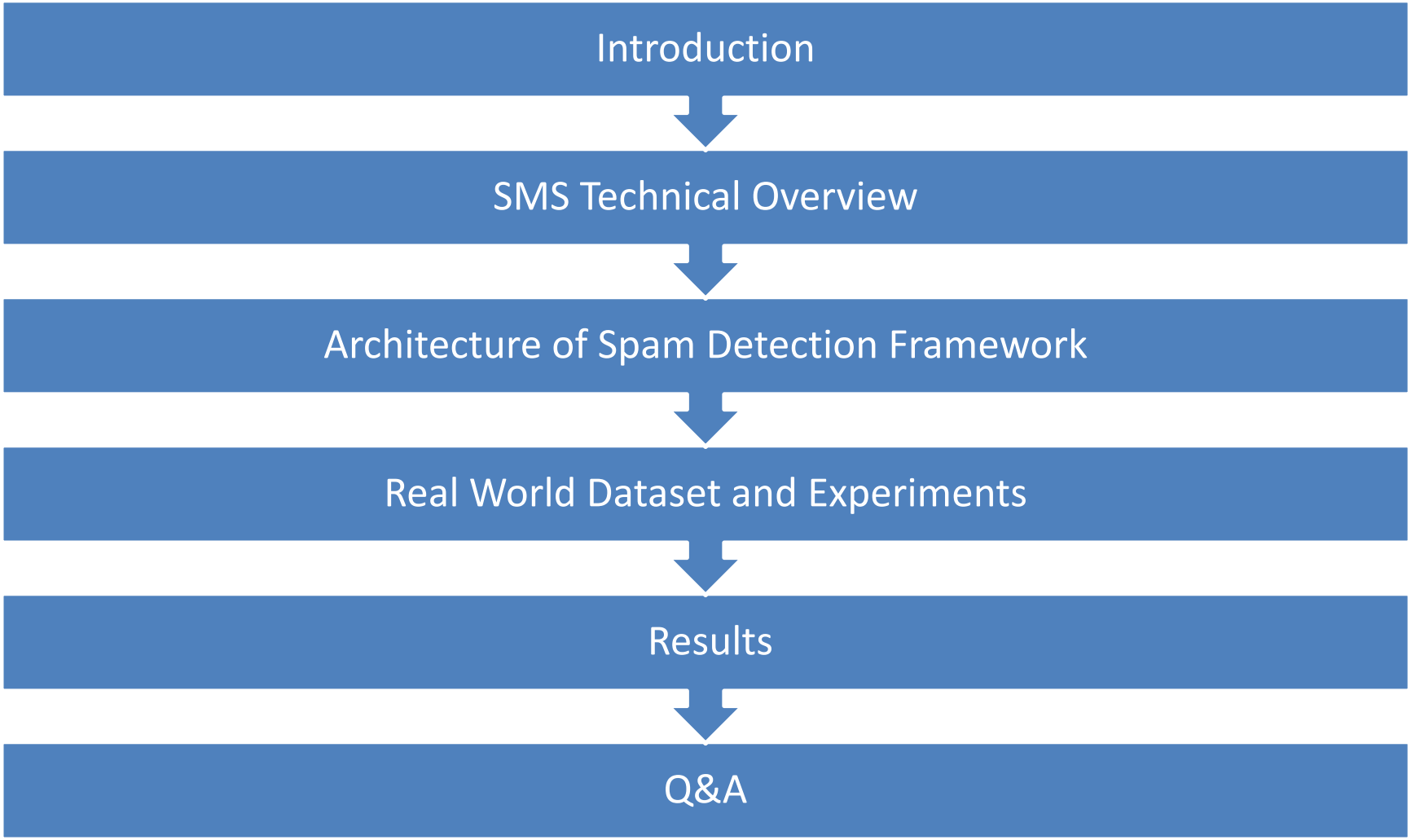
Next Generation Intelligent Networks Research Centre (nexGIN RC)

National University of Computer and Emerging Sciences (NUCES)

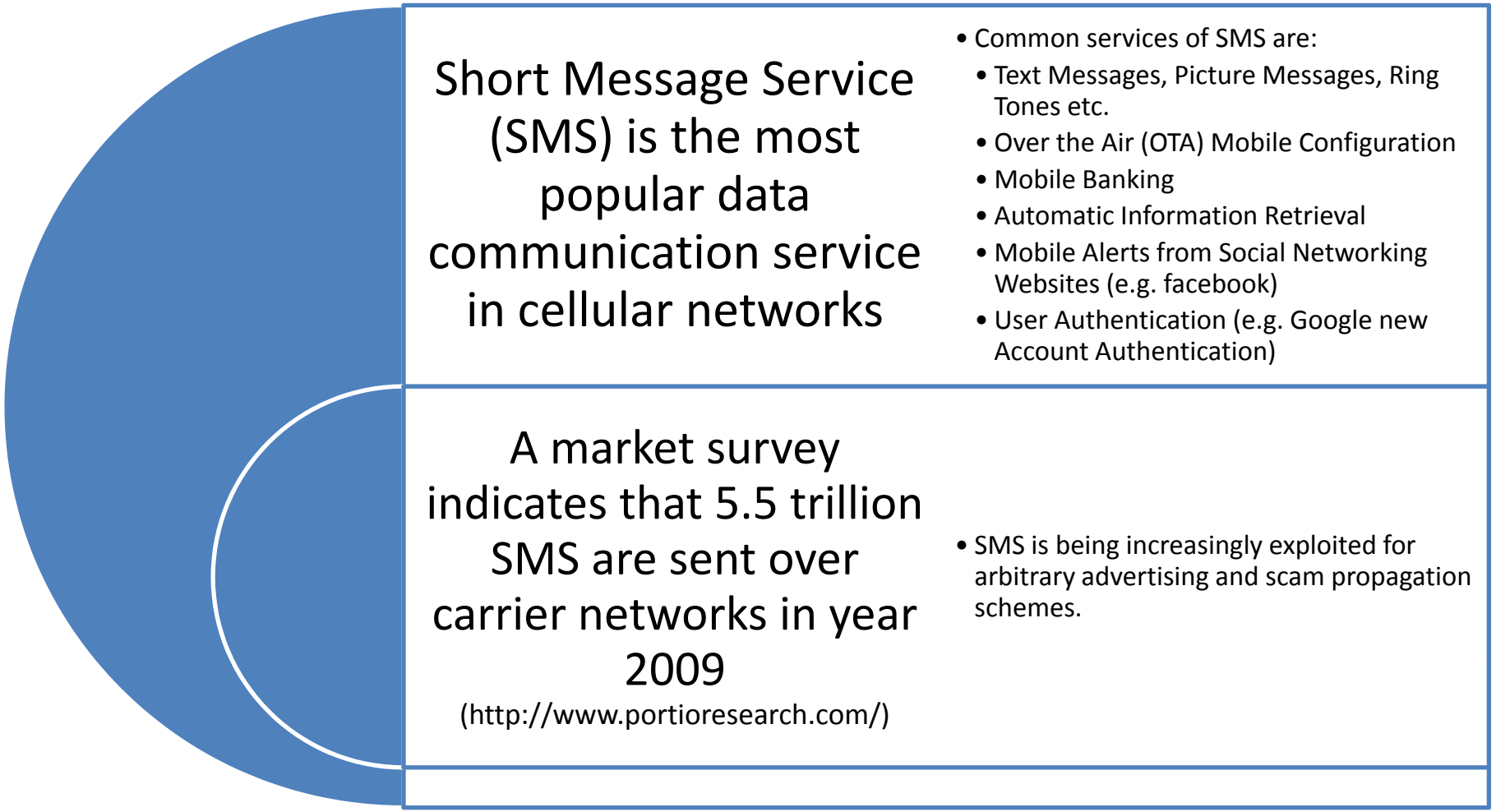
Islamabad, Pakistan

{zubair.rafique,muddassar.farooq}@nexginrc.org

Agenda


Introduction
SMS Technical Overview
Architecture of Spam Detection Framework
Real World Dataset and Experiments
Results
Q&A

SMS Usage


Short Message Service (SMS) is the most popular data communication service in cellular networks
• Common services of SMS are:
• Text Messages, Picture Messages, Ring Tones etc.
• Over the Air (OTA) Mobile Configuration
• Mobile Banking
• Automatic Information Retrieval
• Mobile Alerts from Social Networking Websites (e.g. facebook)
• User Authentication (e.g. Google new Account Authentication)
A market survey indicates that 5.5 trillion SMS are sent over carrier networks in year 2009
(http://www.portioresearch.com/)
• SMS is being increasingly exploited for arbitrary advertising and scam propagation schemes.

The Increasing Trend in SMS Spam

The number of SMS spam messages accounts for more than **50%** of the total SMS messages received by users.

(http://www.ironport.com/pdf/ironport_case_study_wireless.pdf.)

It has been witnessed that more than **200 million** cell phone users were hit by SMS spam in a single day in China on March 2008.

(http://www.sophos.com/pressoffice/news/articles/2008/03/china_sms.html.)

SMS Spam Provocation

More annoying than E-mail spam

- Notification through a ring tone or vibration alert
- Can not delete a spam SMS without opening it

Majority of SMS spam are sent directly by operators or on behalf of third-party providers

- SMS Spam detection not effective on operator side
- Demand of intelligent spam detection on mobile devices

Limitations of Current Techniques

Resource Constrictions

- Requires large memory resources (Features like words and character bi-grams or tri-grams)
- Requires large processing power (Content Based Analysis)

Non-Conformance with SMS Writing Styles

- E-mail based approaches easily evadable (Spam SMS are mostly written in local languages or in *romanized* English [7])

Real-World Deployment

- Not in accordance with underline reception mechanism of SMS on mobile devices

Our Contribution

Access Layer Detection

- Analysis and quantification of byte-level distributions of SMS.
- Hidden Markov Models (HMMs) for benign and spam messages,
- Robust to word adulteration techniques and language transformations.
- New learning algorithm for the classification of spam based on the probabilistic variation from the trained models.

Effective and Efficient Detection

- More than 97% detection rate with zero false alarm rate.
- Lightweight: requires only 256KB of memory.
- Less than 1 millisecond to detect spam message.

Real World Dataset

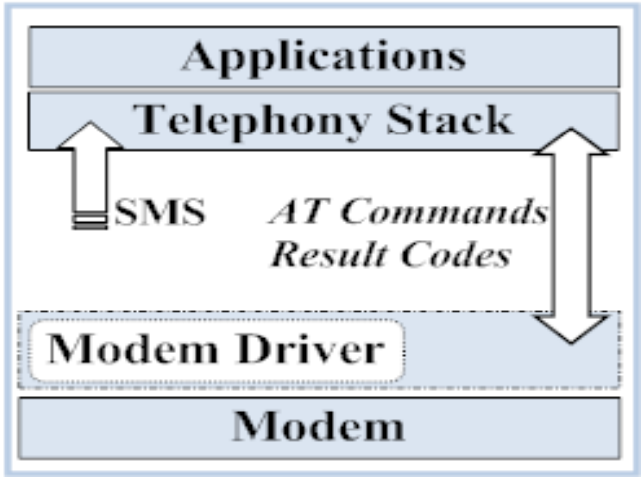
- More than 5000 benign messages collected from volunteers.
- More than 800 spam messages collected from Grumbletext.
- More than 300 spam messages collected from volunteers.

SMS Technical Overview

SMS Reception on Mobile Phones

Mobile Phone Architecture

- SMS is received on Base band (GSM modem) of mobile phone from Short Message Service Center (SMSC)
- AT Commands are used to read SMS from modem and deliver to Application processor through Telephony stack
- SMS is delivered in form of SMS-DELIVER PDU format from modem to OS of mobile device



SMS-DELIVER PDU Format

Bit no		7	6	5	4	3	2	1	0	
Oct. no										
Address of SMSC max. (12 bytes)	1	Length of SMSC Address Information								Address Length
	1	1	Type of Number			Numbering Plan Identification				Type-of-Address
	1	SMSC Number in Semi Octet Representation								Address Value
	2									
	-									
	X									
	1	TP-RP	TP-UDHI	TP-SRI	X	X	TP-MMS	TP-MTI		First-Octet(M)
Address of Sender max. (12 bytes)	1	Length of Sender Address Information								Address Length
	1	1	Type of Number			Numbering Plan Identification				Type-of-Address
	1	Sender Number in Semi Octet Representation								Address Value
	2									
	-									
	X									
	1	Bits 7-6 TP-PID		Bit 5 TP-PID	Bits 4...0 TP-PID					TP-PID(M)
	1	Bits 7-4 TP-DCS				Bits 3-0 TP-DCS				TP-DCS(M)
Time Stamp 7 bytes	1	Year								TP-SCTS(M) in Semi-Octet Format
	2	Month								
	-	Day								
	-	Hour								
	-	Minute								
	-	Second								
	7	Time Zone								
	1	User Data Length								TP-UDL(M)
User Data max(140 bytes)	1	User Data								TP-UD(O)
	-									
	-									
	-									
	1									

SMS-DELIVER PDU

SMS TP-UD (User Data)

- Maximum user data transferred in single SMS can be of 140 bytes in hexadecimal octets
- TP-DATA-CODING-SCHEME is used to indicate the underline encoding of user data

SMS Encoding Schemes

- 7-bit
  - Default encoding scheme for text messages
  - Maximum of 160 characters
- 8-bit
  - Usually data is not viewable (if not used for text messages)
  - Used in Smart messaging like picture SMS, ring tones and OTA configuration
  - Maximum of 140 characters
- 16-bit
  - Unicode (UCS2) encoding of text messages
  - Maximum of 70 characters

Architecture of Spam Detection Framework

Requirements

SMS Spam Detection at Access Layer

- In order to silently move spam SMS messages into a spam folder without disturbing the user through ring tone or vibration alerts

Semantec Independence

- It must not use specific words, character bi-grams and tri-grams of a specific language

Lightweight Framework

- in the sense that it requires less than 512KB of memory

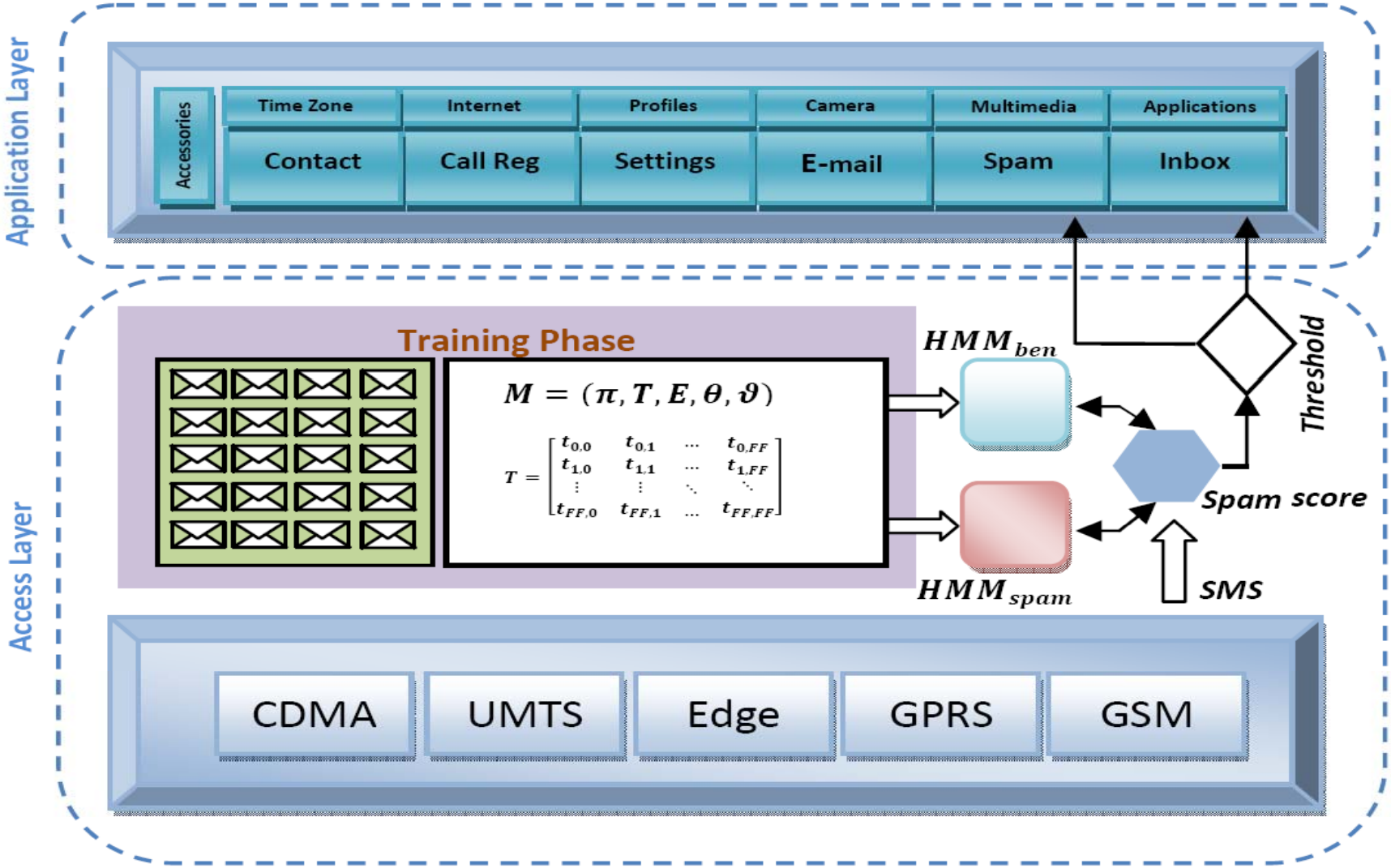
Efficient Performance

- It must classify an SMS in less than 1 millisecond

Effective Detection

- It must provide a greater than 95% detection rate with a zero false alarm rate

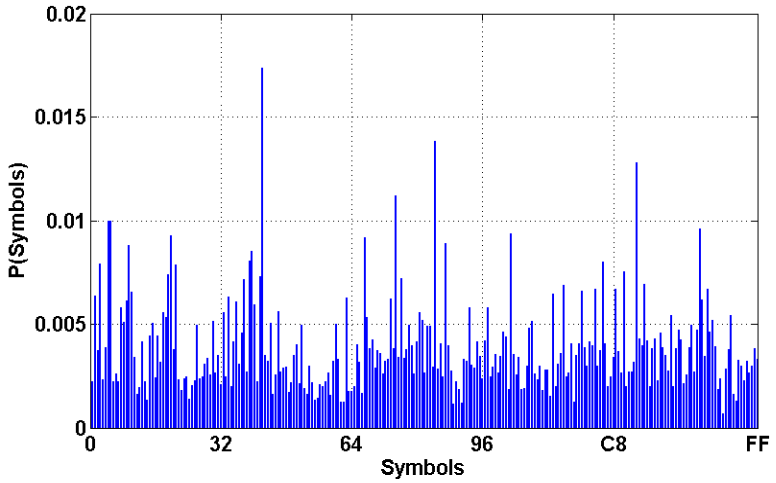
Architecture of Spam Detection Framework



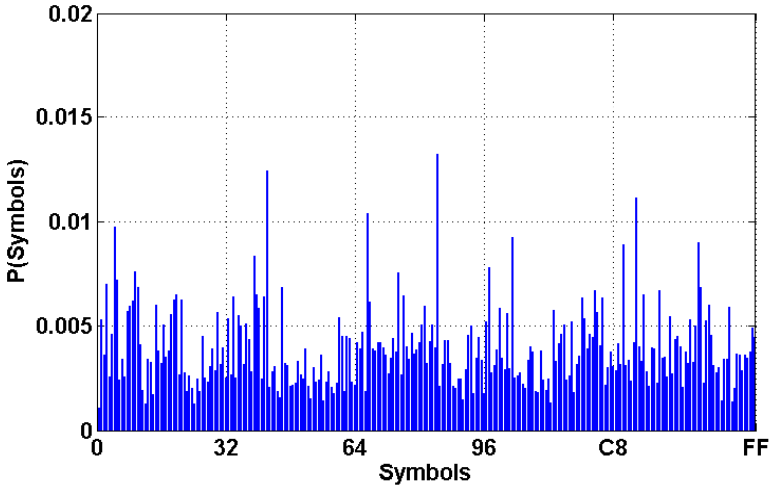
Byte-level Analysis (1/2)

Comparison of Benign and Spam SMS

- Spam messages are ‘intelligently crafted’ to make them appear as benign messages.
- No discernable difference exists in byte level distribution of spam and benign SMS at access layer.
- Not possible to classify an SMS message as benign or spam on the basis of byte-level distributions in any encoding format (7-, 8-, or 16-bit) at the access layer of a mobile phone.

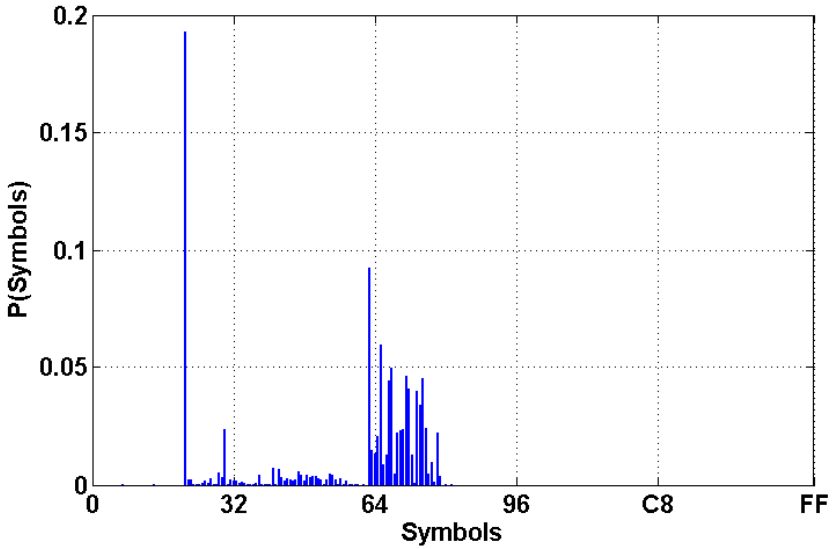


7-bit benign

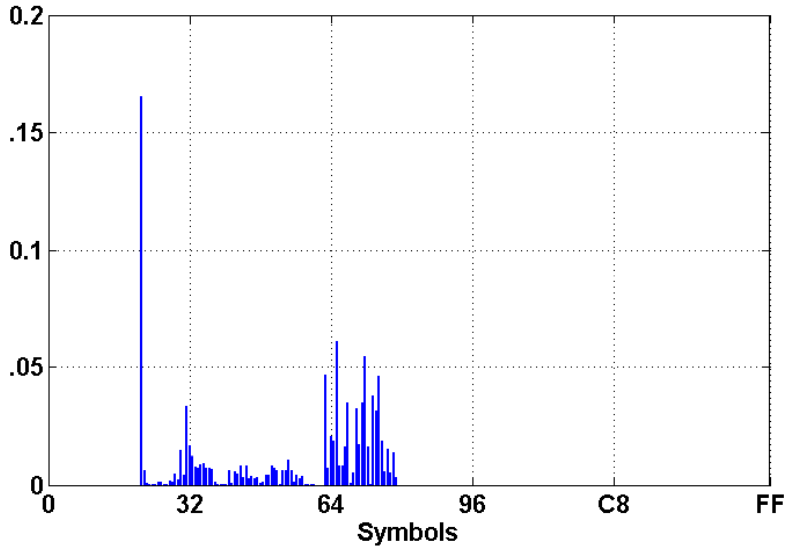


7-bit spam

Byte-level Analysis (2/2)



8-/16 bit benign



8-/16 bit spam

Quantification of byte-level Information

Autocorrelation of byte-level distributions

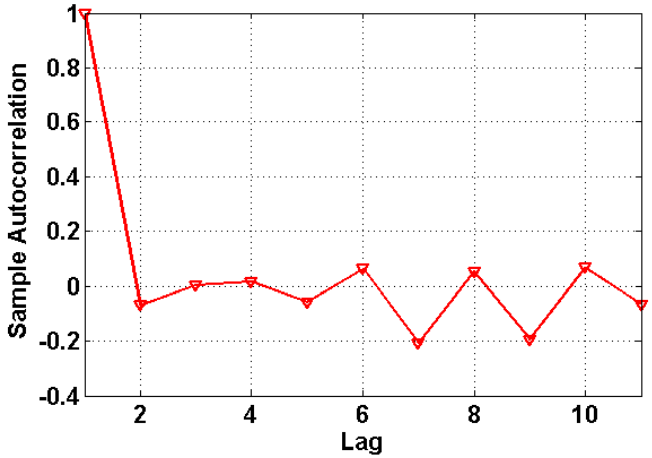
- Autocorrelation is used to study the correlation between the random variables in a stochastic process at different points in time or space.

$$\rho[e] = \frac{E\{X_0 X_z\} - E\{X_0\}E\{X_z\}}{\rho_{X_0}\rho_{X_z}}$$

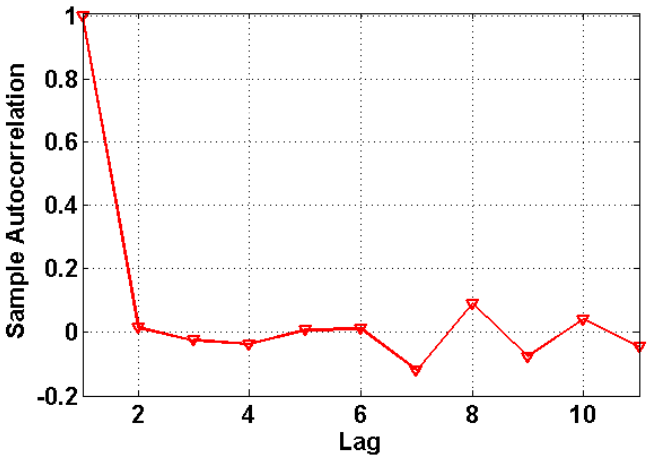
Autocorrelation value lie between -1 and 1. E{.} presents expected value of random process at given lag `e'. X_z presents a stochastic process where z is the space/time lag.

- Autocorrelation results on SMS datasets show that the byte sequences in SMS have first-order dependence.
- It shows that if an octet *k* appears in an SMS, it is more likely that it will immediately be followed by octet *l*.

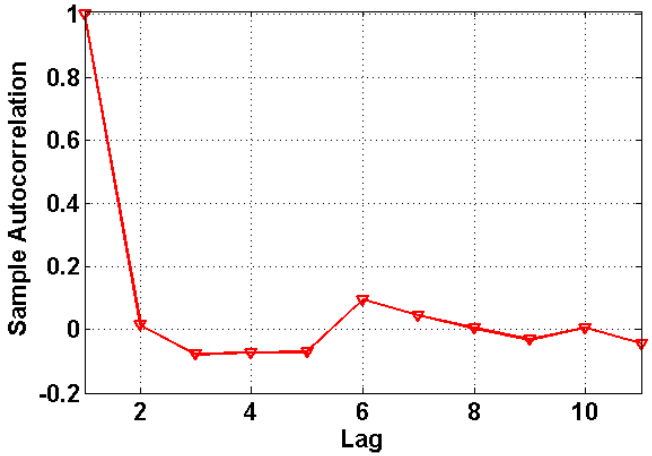
Autocorrelation Results



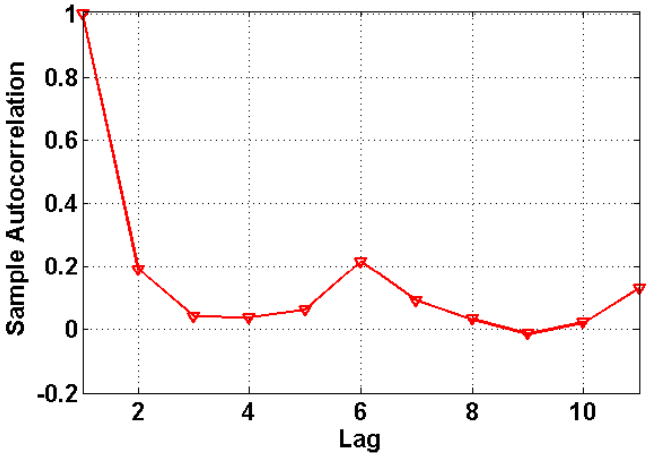
7-bit benign



7-bit spam



8-/16 bit benign



8-/16 bit spam

Hidden Markov Model for Benign and Spam Messages

SMS Byte Level Modeling of HMM

- Autocorrelation results show 1st order reliance in byte level distribution to model SMS using a 1st order discrete time Markov process.
- A byte-level distribution of Markov representation simply implies 2^8 = 256 conditional probability distributions.
- Transition probabilities are computed by counting the number of times hexadecimal octet *k* is followed by hexadecimal octet *l* in an SMS.

$$T = \begin{bmatrix} t_{0,0} & t_{0,1} & \dots & t_{0,FF} \\ t_{1,0} & t_{1,1} & \dots & t_{1,FF} \\ \vdots & \vdots & \ddots & \ddots \\ t_{FF,0} & t_{FF,1} & \dots & t_{FF,FF} \end{bmatrix}$$

t(k,l) presents the probability of moving from octet *k to l* .

Hidden Markov Model for Benign and Spam Messages

Introduction

- HMMs are commonly used as a probabilistic modeling technique for linear problems like sequences or time series and can be automatically estimated, or trained from unaligned sequences.
- HMMs provide a straightforward solution to estimate the probability of occurrence of a sequence, given that a trained model of sequences is already computed.
- HMMs have been widely used in speech recognition applications, computational sequence analysis and protein structural modeling.

Classification of Spam Messages

HMMs Learning from Training Data

- HMM_ben: the sequence probabilities in a benign SMS.
- HMM_spam: the sequence probabilities in a benign SMS.
- Probabilities that a given SMS (S) is generated by a benign HMM (HMM_ben) and by a spam HMM (HMM_spam) are calculated using Viterbi algorithm as:

$$P_{r1}(S/HMM_{spam}) = \sum_{\theta \in valid(\theta)} \prod_{i=1}^{|S|} t_{\theta_i - 1, \theta_i} e_{\theta_i}(s_i)$$

$$P_{r2}(S/HMM_{ben}) = \sum_{\theta \in valid(\theta)} \prod_{i=1}^{|S|} t_{\theta_i - 1, \theta_i} e_{\theta_i}(s_i)$$

The *Pr_1 and Pr_2 represent the* probabilities that a given SMS (S) is generated by a benign HMM (*HMM_ben) and by a spam HMM (HMMs_pam) respectively. |S| is the number of octets in an SMS and valid(θ) are the* valid state sequences.

Spam Threshold Score Calculation

Calculation of Spam Score

- The spam score for each SMS in training data is computed as a function of Pr_1 and Pr_2 using the following formula:

$$spam_{score} = \frac{(P_{r1})^{1/|S|}}{(P_{r1})^{1/|S|} + (P_{r2})^{1/|S|}}$$

- Squashing the probability values by the length (number of octets) of an SMS amplifies higher probability values compared with low values.

Threshold Calculation

$$threshold = max(spam_{score_v}), 1 \leq v \leq Z$$

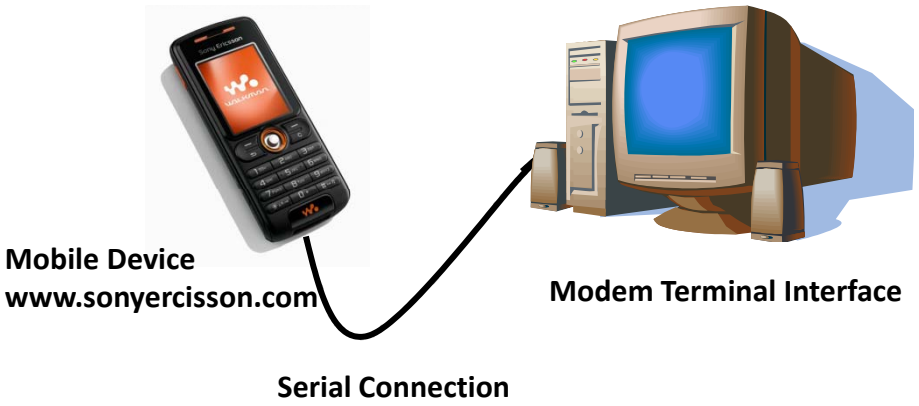
Z corresponds to the total number of spam messages used to calculate the threshold value.

Real World Dataset and Experiments

Real World Dataset

Modem Terminal Interface

- Accesses SMS from the memory of the base band processor of a mobile phone in an SMS-DELIVER format.
- Interacts serially with the modem of a mobile device through AT commands.
- Configures the modem to operate in the PDU mode by giving the AT+CMGF=0 command.
- Using AT+CMGL=ALL, all messages in the memory of the base band processor of a mobile phone are redirected to the terminal.




```
AT+CMGF=0                          Setting Mobile in PDU mode
OK
AT+CMGL=ALL                        Reading SMS in SMS-DELIVER format
+CMGL: 0,,26
07912943550000010 40C91294352429505
0000012052715365020 8C834885C279743
OK
_ _ _ _ _
_ _ _ _ _
```

Real World Dataset

Benign Dataset

- 30 mobile phone users volunteer for this study.
- 5000 benign messages were collected in SMS-DELIVER format.
- Subject of study belongs to different socio economic background:
  - Teenagers
  - Corporate executives
  - Researchers
  - Students
  - Housewives
  - Software developers
  - Senior citizens

Spam Dataset

- 800 spam messages from Grumbletext: UK consumer complaints – post online and via SMS text. http://http://www.grumbletext.co.uk/.
- 300 spam messages collected from volunteers.

Experiments

Validation Procedure

- Stratified 10-fold cross validation procedure is used in all of the experiments.
- Standard representations of detection accuracy and false alarm rate:
  - Detection of spam message, True Positive (TP).
  - Detection of benign message, False Positive (FP).
  - Does not detect a spam message, False Negative (FN).
  - Does not detect a benign message, True Negative (TN).
- Detection Rate (DR) is defined as:

$$DR = \frac{TP}{TP + FN}$$

- False Alarm Rate (FAR) as:

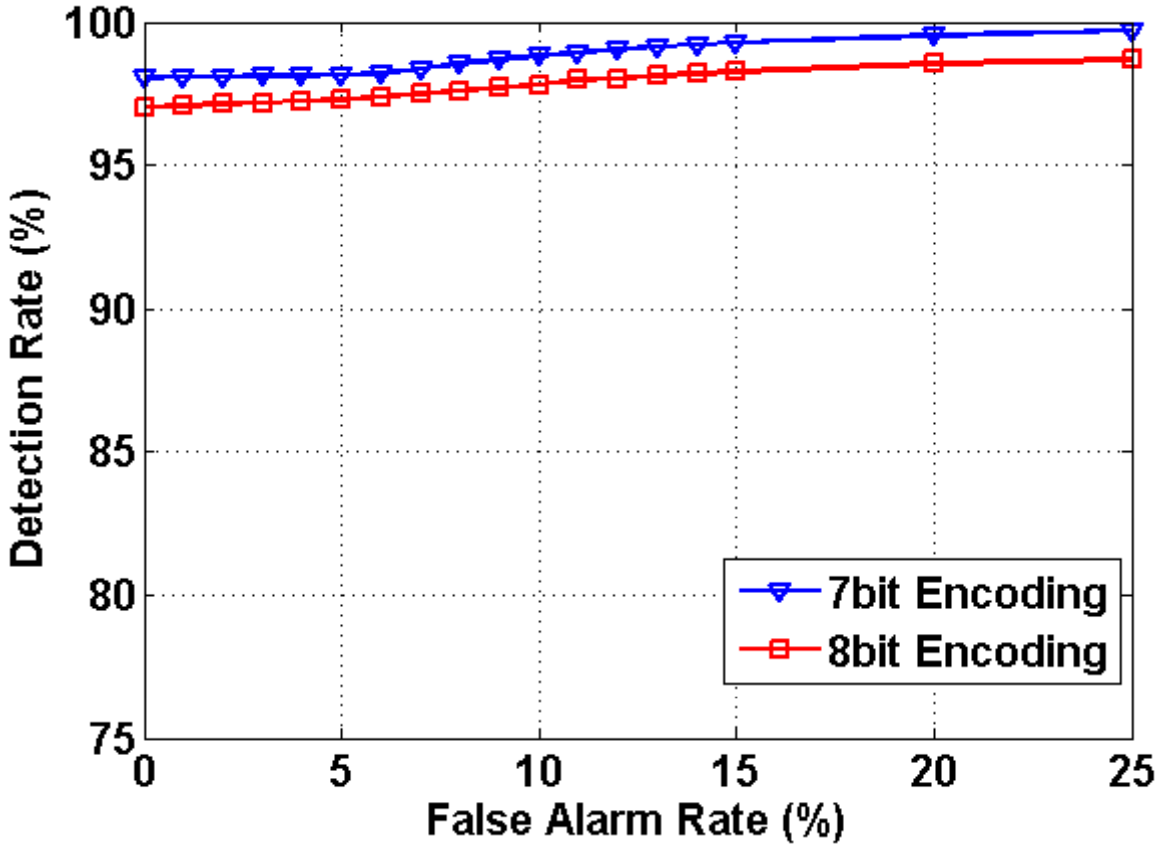
$$FAR = \frac{FP}{FP + TN}$$

Results

Results Summary

- Receiver Operating Characteristic (ROC) curves to show trade off between detection rate and false alarm rate.
- More than 98% detection rate with a 0% false alarm rate for SMS messages encoded in 7-bit.
- 97% detection rate with a 0% false alarm rate for SMS messages encoded in 8-/16 bit.
- Transition and Emission matrix for benign and SMS models needs only (4 * 65536) = 256KB of memory.
- Framework tested on an old 200MHz computer (the approximate speed of the processors of most mobile phones) proves testing time for a single SMS is less than 1 millisecond.

ROC Curve for Spam Detection Framework



Conclusion

- Novel spam detection framework that uses autocorrelation of underlying byte-level distributions of an SMS to detect spam messages.
- Robustness to word adulteration techniques and language transformations as scheme works on the access layer of a mobile phone.
- Byte-level distributions of benign and spam messages to train Hidden Markov Models (HMMs).
- New learning algorithm for classification of SMS spam based on the probabilistic variation from the trained models.
- Collection of real world dataset from volunteers and Grumbletext..
- More than 97% detection rate with a 0% false alarm rate with 256KB memory and testing time less than 1 millisecond.

What next ?

- SMS SPAM datasets in other languages
  - Russian
  - Arabic
  - Chinese
- Implement it on Symbian smart phones
- Model is generic for Emails, IM
- See for details about the front end company: **http://www.hikmahtech.com**

Q&A