BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
Vol. 1, Issue 2, Nov 2013, 7-12
© BEST Journals

FRAMEWORK FOR SURVEILLANCE OF MULTILINGUAL CONTENTS FROM EMAILS

MOHD MAHMOOD ALI[1], MOHAMMED KHAJA MOIZUDDIN[2] & LAKSHMI RAJAMANI[3]

[1] Department of CSE, Muffakham Jah College of Engineering & Technology, Banjara Hills, Hyderabad, India

[2] Department of CS & SE Hail, College of Computer Science, University of Hail, Hail, Saudia Arabia

[3]University College of Engineering, Osmania University, Hyderabad, India

ABSTRACT

The new trend of spammers and criminals in emails, for communicating spam and suspicious messages is via images embedded with multilingual text. We present an anti-spam and suspicious filtering framework for text-based and image-text messages with multilingual support. These suspicious messages are further classified that aid in predicting suspected cyber crime with details of culprits. Experimental result shows the precision efficacy of proposed system achieves best results with minimum false positive, when compared with existing state of spam filters.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering – Retrieval model, Search process, Selection process

General Terms

Design, Algorithms, Experimentation

KEYWORDS: Image-Text, Spam, Suspicious, Multilingual, Bayesian Classifier

INTRODUCTION

The anti-spam filter in emails are divided into two categories, one is server-side filters[1,2] which make use of manually derived rules from historical spam messages that were marked as spam by experts based on blacklist addresses, spam words, signature filters, URL filters, and Content analysis and the other is client-side filters[3] trained by classifiers at run-time based on user preferences with collaborative filtering techniques. Pattern recognition and computer vision had contributed a lot for detecting image-spam. However, earlier solutions proposed are prone to exhibit several weaknesses and their effectiveness is not, yet, investigated thoroughly. Abdolrahman Attar and et al. [1], classified Header-based, Content-based, and Text-based. Among these image-spam types are of advertisements from Adult, Financial, Products, Internet, Leisure, Health, Political, Education, and spiritual. These tricks are named as Template construction and randomization techniques in literature [7], but Multilingual Trick and Ontology Trick are not yet, focused by research community, except, for few specific languages [9]. We are arguing; in the context of a single image-text, consisting of multiple languages.

PROBLEM STATEMENT & RELATED WORK

The two major aims, of our research are to block the Spammers tricky techniques and blacklist them from sending

[1]www.spamassassin.apache.org/
[2]http://spambayes.sourceforge.net/
[3]http://www.spamhaus.org/

multilingual image-spam messages in emails, and the second aim is to trace the culprits, who perform cyber crimes[4], which were undetected earlier in emails. The first problem in email is image-spam messages with multilingual text are sent purposely which are undetected by current anti-spam filters as shown in Figure 1. We studied the different features of text which is embedded in images, and techniques to retrieve multilingual text, and multilingual text from images. To detect multilingual Text-and-Image spam, we have gone through the functioning of existing anti-spam filters, like SpamAssasin[5] and others. Few Anti-spam filters extracts text from images using image filtering techniques like Haar DWT wavelet, Histogram filter, Morphological filters, MRF technique, DCT filter for compressed images, and orientation filters. Cormack proposed Anti-spam filter, for segmenting batches of text regions from pictures and background for online and offline messages[6]. Object detection from RGB images, proposed using three different techniques when used sequentially, a) Edge detection, b) connected components detection, and c) Morphological filtering. In our approach we use globally matched wavelets (GMWs), and Markov random field (MRF) to extract text regions from pictures and background in Image-text from attached messages of email [2].

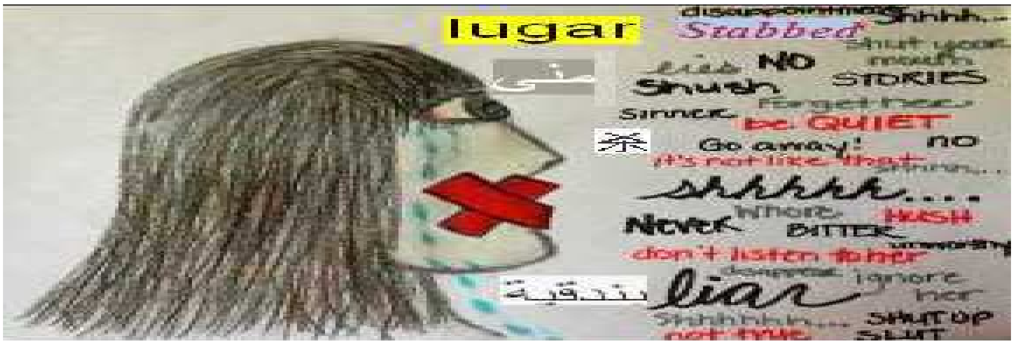

Figure 1: Shows an Original Multilingual Image and Text with Words of Different Languages (English, Chinese, Spanish, Urdu and Arabic)

PROPOSED FRAMEWORK

Generic multilingual framework for surveillance of email messages in cyberspace is shown in Figure 2, numbered in sequence from 1 to 9 This Framework detects and filter multilingual image spam, multilingual text and multilingual text embedded in images that have suspicious messages, along with criminal details. In our proposed approach we have used techniques which include: a) multilingual text extraction from images using GMW & MRF filters[2], b) Optical Character Recognition system (OCRs), c) Data mining (improved C4.5)[4] d) Ontology guided with pre-defined axioms[3] and e) Tree Alignment and GSHL algorithms[4]. Earlier, spam filters used OCRs, Image processing technique, Machine learning and Data Mining technique (Bayesian classifier) [5], few of them partly make use of Ontology techniques and are restricted to English language discussed [7]. Among these spam filters, one or the other technique is missing therefore are susceptible to filter spam messages efficiently.

[4] http://www.cybercrimelaw.net/
[5] www.spamassassin.apache.org/

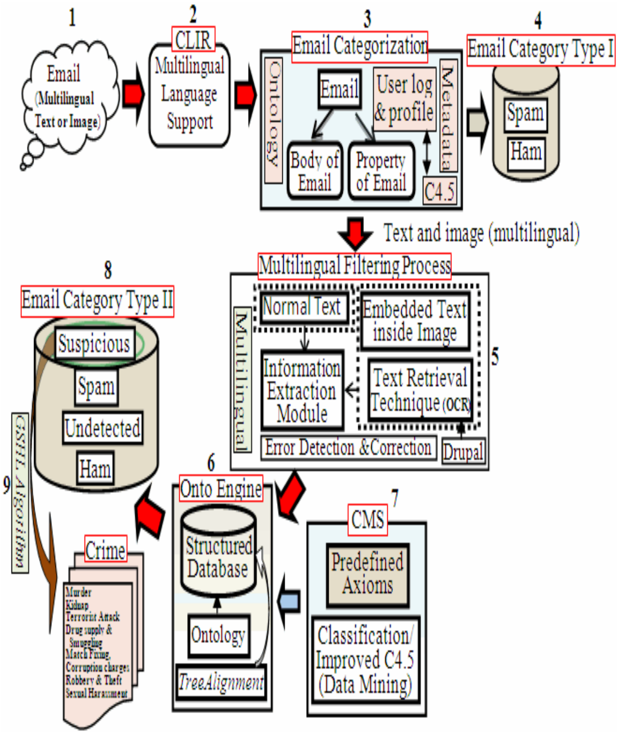

Figure 2: **Framework for Spam and Suspicious Email Detection from Multilingual Text-and-Image Based Contents**

EXPERIMENTAL RESULTS AND DISCUSSIONS

In this Section, we explore results obtained using our framework that categorize email messages from cyberspace. This Framework detects and filter multilingual spam messages from simple English text, multilingual text, and multilingual text embedded in images, and also, for detecting suspicious messages from emails. The output obtained of our framework using predefined axioms of table 1, is depicted in Figure 3 [9]. The Precision and Recall for email messages as spam using improved C4.5 classifier for various datasets is shown in Table 2 [9].

Table 1: Pre-Defined Set of Axioms & Learned Classification Rules

Rule 1 (a): SPAM Domain Has	
Type of Spam (Sub-Domain)	**Collected from**
Black list (IP addresses, URLS,&email Addresses)	http://www.barracudacentral.org/report/
Spam Datasets (FromEmail addresses, IP addresses, URLs & few Spam Keywords are extracted)	mail.diee.unica.it during 2004-2007 (updated 2010) by **Pattern Recognition & Applications Group; CSMINING; & Trec2007**
Rule 1 (b): SPAM Domain Has	
Considers the user-defined preferences of Rule 3	
Rule 2: SUSPICIOUS Domain Has	
Type of Threat Activity(Sub-Domain)	**Stem words**
Murder →	assault, kill, assassinate, eliminate, gun,stab, dagger, weapon location, money, location
RULE 3: User preferences learned using improved C4.5 Classifier	

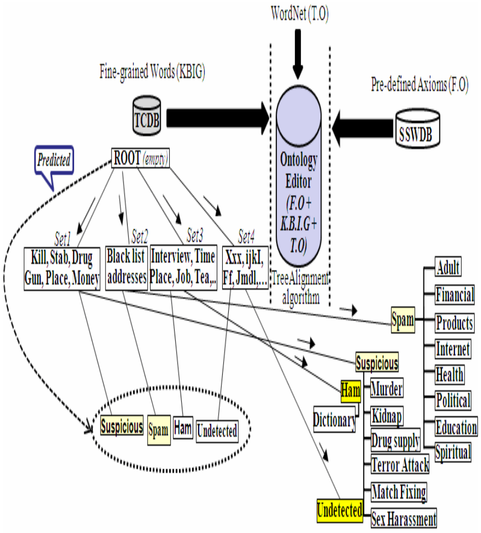

Figure 3: Shows Taxonomic Structure for Spam, Suspicious, Ham, and Undetected Words Mapped Dynamically Using Pre-Defined Axioms

Table 2: Shows Precision & Recall for Spam Using Improved C4.5 Classifier for Different Datasets

Dataset	PRA Group	CEAS2008	CLEF2012	TREC2007	CSDMC2010
Precision(%)	88.23	91.27	90.77	92.0	94.11
Recall(%)	96.89	98.99	98,21	99.80	99.95

CONCLUSIONS

The results of our work offer many perspectives of further research for anti-spam filters are:

- Spam Filters to be enhanced to detect spam which has Text-and-image contents embedded using multimedia.
- SMS in mobile phones is vulnerable of receiving unnecessary messages and phone calls from unknown numbers.
- Surveillance of Cyberspace messages to predict terrorist attacks and cyber crimes from Social Networking sites (Facebook, Twitter, & others).

REFERENCES

1. Applications and Future Directions (Jul.2013) in pressAttar, A., Rad, R. M., and Atani, R. E. 2011. A survey of image spamming and filtering techniques. In Proceedings of Artifificial. Intelligence Review. (2011, Vol. 40. pp. 71-105, Springer). DOI=http://doi.springer.org/10.1007/s10462-011-9280-4.

2. Kumar, S., Gupta, S., Khanna, N., Chaudhur, S., and Joshi S. D. 2007. Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model, IEEE Trans. on Image Processing, Vol. 16, no. 8, (Aug. 2007), 2117 - 2128. DOI=http://doi.IEEE.org/ 10.1109/TIP.2007.900098.

3. Ali, M. M., Rajamani, L. 2013. Framework for Surveillance of Instant Messages. International Journal of Internet Technology and Secured Transactions, Inderscience publisher (2013) in press.

4. Lang, H. J. 2011. Data Extraction for Deep Web Using WordNet. IEEE Transactions on Systems, Man, and Cybernetics, Part C 41(6), 854-868 in 2011. DOI=http://doi.IEEE.org/ 10.1109/TSMCC.2010.2089678.

5. Byungi, B., Chin-Hui, L., Steve W., Irani, D., and Calton, P. 2009. An Anti-spam filter combination frmework for Text-and-Image Emails through Incremental Learning. In Proceedings of 6th Conference on Email and Anti-Spam (CEAS), California, USA, (Jul. 2009).

6. Cormack,G.,and Lynam, T.2007.Spam Track Guide line 2005-2007," link: http://plg.uwaterloo.ca/~gvcormac/spam/.

7. Wong W., Liu, W., and Bennamoun, M. 2011. Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, published by Information Science Reference, IGI Global in 2011. DOI= http://doi.igi-global.com/ DOI: 10.4018/978-1-60960-625-1.

8. Kim, J., Dou, D., Liu, H., and Kwak, D. 2007. Constructing a User Preference Ontology for Anti-spam Mail Systems. In Proceedings of the 20th conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Springer-Verlag Berlin, Heidelberg, pp 272 – 283, 2007. DOI=http://doi.springer.org/10.1007/978-3-540-72665-424.

9. Ali, M. M., Rajamani, L. 2013. Framework for surveillance of emails to detect multilingual image spam and suspicious messages. In Proceedings of IEEE Workshop On Computational Intelligence: Theories