Identification of Most Frequently Occurring Lexis in Winnings-announcing Unsolicited Bulk e-mails

Jatinderkumar R. Saini, Apurva A. Desai

***Abstract*—**e-mail has become an important means of electronic communication but the viability of its usage is marred by Un-solicited Bulk e-mail (UBE) messages. UBE consists of many types like pornographic, virus infected and 'cry-for-help' messages as well as fake and fraudulent offers for jobs, winnings and medicines. UBE poses technical and socio-economic challenges to usage of e-mails. To meet this challenge and combat this menace, we need to understand UBE. Towards this end, the current paper presents a content-based textual analysis of nearly 3000 winnings-announcing UBE. Technically, this is an application of Text Parsing and Tokenization for an un-structured textual document and we approach it using Bag Of Words (BOW) and Vector Space Document Model techniques. We have attempted to identify the most frequently occurring lexis in the winnings-announcing UBE documents. The analysis of such top 100 lexis is also presented. We exhibit the relationship between occurrence of a word from the identified lexis-set in the given UBE and the probability that the given UBE will be the one announcing fake winnings. To the best of our knowledge and survey of related literature, this is the first formal attempt for identification of most frequently occurring lexis in winnings-announcing UBE by its textual analysis. Finally, this is a sincere attempt to bring about alertness against and mitigate the threat of such luring but fake UBE.

***Keywords*—**Lexis, Unsolicited Bulk e-mail (UBE), Vector Space Document Model, Winnings, Lottery

I. INTRODUCTION

WITH the increase in usage and availability of Internet, there has been a tremendous increase in usage of e-mail. It has proved to be an important medium of cheap and fast electronic communication. But the same thing that has increased its popularity as a communication medium has also proved to be a source of non-personal, non-time critical, multiple, similar and un-solicited messages received in bulk. This type of message is called Unsolicited Bulk e-mail (UBE) and is known by various other names like Spam e-mail, Junk e-mail and Unsolicited Commercial e-mail (UCE). The spread of UBE has posed not only technical problems but has also posed major socio-economic threats. Also, the definition of spam e-mail is ‘relative’ [5, 12, 17]. This means to say that all e-mails going to spam folder may not be spam for a person – same as all e-mails going to inbox may not be ham (i.e. non-spam) e-mails. Further, all spam e-mail is not harmful, some is just annoying [2, 8, 14]. UBE incidences range from fake job offers and fake medicines to pornography. Some like virus-infected UBE have been responsible for bringing the entire computer networks and commercial businesses down. Another area of concern is ‘get-rich-quick’ e-mail which is very harmful to innocent persons who may get engulfed in the network of greedy people. A variation of this kind of email is the one that claims to give large amount of wealth by announcing the recipient’s name as winner through schemes like year-end selection, lottery, prize, award and draw. In general this paper refers to this kind of UBE as winnings-announcing UBE.

In past, researchers have worked in direction of understanding the spam for combating it [1, 10, 23]. We also believe that first step in combating spam is to understand spam. A novel idea proposed in this paper is that the best way of understanding spam is to analyze it. Most importantly, spam can be differentiated by content [20] and in this paper we target content-based analysis of un-structured UBE documents which claim to promise huge amount of wealth through announcements of winnings. The present work aims towards identification of lexis occurring in such UBE. The basic structure of spam e-mail message is same as of ham e-mail, consisting of ‘header’ and ‘body’ parts. In this paper, we have treated spam e-mail as un-structured because in addition to consideration of contents of structured ‘header’ part, we propose content analysis of ‘body’ part also. The structure of ‘body’ part is not fixed with respect to number of words, lines, format, etc. and hence we treat UBE as an un-structured document. From a technical perspective, identification of most frequently occurring lexis in UBE documents is a Text Parsing and Tokenization task and we propose to solve it using Bag of Words (BOW) and Vector Space Document Model approach.

II. RELATED WORK

As far as, our study of past and contemporary literature for this field is concerned, this is the first formal attempt for identification of lexis occurring in winnings-announcing UBE.

J. R. Saini is with Sankalchand Patel College of Engineering, Visnagar, Mehsana, Gujarat, India as Associate Professor and Head of Department of Computer Science (MCA). He is PhD from Veer Narmad South Gujarat University, Surat, Gujarat, India. (phone: +91-9426861815; e-mail: saini_expert@yahoo.com).

A. A. Desai is with the Veer Narmad South Gujarat University, Surat, Gujarat, India as Professor and Head of Department of Computer Science (MCA). He is PhD from Veer Narmad South Gujarat University, Surat, Gujarat, India. (e-mail: desai_apu@hotmail.com).

The survey of related work shows that the researchers have made many attempts to classify e-mails into ham and spam groups. The numbers of attempts targeted towards classification of spam e-mails are also there, but very scarce, per se. Kiritchenko et al. have treated e-mail classification as a special case of text classification [13]. Martin et al. have laid emphasis on individual user's behavior for identifying spam messages [16]. Fette et al. have presented and evaluated a classification method for spam and non-spam e-mails based on Training-data Set approach [7]. Gajewski has discussed the use of a naïve Bayesian classifier based on a BOW representation of an e-mail [9].

Many researchers have listed winnings-announcing UBE as a kind of scam or financial scam [6, 11, 22]. Lance James in his book on Phishing has shown that such scams-containing spam emails account to 9% of the total spam emails [15]. Threat Research and Content Engineering (TRACE) group of Marshal Ltd. has broadly classified spam into nine categories. The first category called 'scams' includes Lotteries (“you have won $10,000!”) and get rich quick schemes such as the 419 Nigerian fraud spam [21]. The ScamBuster Editors at the scamdex website provide an instance where winnings-announcing spam emails are explicitly categorized under a separate heading of 'Lottery' [18].

Under the heading of ‘E-mail fraud’, Wikipedia also has given the description of various email frauds. They believe that the winnings-announcing lottery scam is a contemporary twist on the variously known scam called “Nigerian scam”. The common thing between these scams is that both are variants of advance fee fraud. Advance fee fraud is the one in which the spam email sender first lures the recipient by an attractive amount of money and later, once the victim falls prey to the bait, the sender asks to first deposit some fee in advance for getting the money. According to Commtouch report [4] of June 2004, advance fee fraud spam may be sent by a single individual or even an organized group called “spam gang”.

Lambert [14], in his report on 'Analysis of Spam', has discussed about spam scams and offered suggestions to assist e-mail users in avoiding these spam traps. He adds that the advance fee e-mail scam is probably the most pervasive e-mail scam in existence. We share our concern with Lambert and express the gravity of problem of winnings-announcing UBE by presenting the textual analysis of such e-mails.

III. Methodology

In this section, we describe the detailed methodology followed by us for the identification of most frequently occurring lexis in winnings-announcing spam e-mails. For the sake of simplicity and better understanding, the entire section is divided into three major sub-sections for Data Collection & Clustering, Data Pre-processing and Feature Extraction & Feature Selection.

A)Data Collection & Clustering

We first collected various UBE documents of all types together. We used 40 e-mail addresses for collecting the required data. Another 18 websites providing online archives of UBE were also used for data collection. This formed a text corpus amounting to approximately 1.5 GB of data-size and consisted of 30074 UBE documents. To prevent the data from 'contributor bias' [3], it was sourced from different locations and at different times from e-mail addresses owned by different persons.

As a next step, we identified the data clusters. For this, we used hierarchical divisive clustering approach in which initially all the UBE documents formed one text corpus of a single cluster. The process of clustering was based on the analysis of the contents of UBE documents in the text corpus. This text corpus was processed to yield 2 clusters in such a way that one cluster contained the winnings-announcing UBE whereas the other cluster contained the UBE which did not announce the winnings. The cluster comprising winnings-announcing UBE was the cluster of interest and the number of instances in it was 2979, which amounted to nearly 34 MB of data size. Given the inherent in-secure nature of UBE documents, a noteworthy thing here is that the collection of such UBE is a difficult process. Our intention was to create a corpus of UBE which claimed to provide large amount of money through announcing of winnings by schemes like year-end selections, various kinds of online and offline lotteries, prizes, awards and various forms of draws. Our task of data collection was eased by the fact that most of these kind of UBE have an explicitly subject line which makes it easy to identify the category of UBE under question. Besides our naïve approach for categorization of UBE, the spam filters provided by the e-mail providers also helped us confirm the categorization by actually classifying the UBE under the spam folder.

B)Data Pre-processing & Cleaning

The main motive of this phase was to clean the data. At this stage, we pre-processed the collected text-files in the UBE corpora by removing 'obvious noise' from them and converting them in a common format. By 'obvious noise', we mean the location and site specific data slipped into the UBE documents when sourced from different locations, e.g. website name. This data-cleaning is also required for making the data ready for further processing – specifically, easing the subsequent phase of feature extraction.

C)Feature Extraction & Feature Selection

This is the most important and bulkiest phase of data-processing. The types of operations done during this phase are often referred to as 'Feature Extraction' and 'Feature Selection' by the research literature of text analysis and text mining. Here, we picked the corpus of winnings-announcing UBE. The corpus under consideration is actually formed of UBE which are eventually text documents. For each text document, we performed sentence-splitting in order to treat it as a Bag Of Words (BOW). In BOW representation of a text document, lexis or terms or tokens in the document are identified with words in the document. Hence this representation is also called Set of Words (SOW) [19]. We then performed Syntactic Text Analysis by Parsing the UBE document, for extraction of Tokens.

TABLE I
FREQUENCY AND PERCENTAGE OF PRESENCE OF MOST FREQUENTLY OCCURRING TOP 100 LEXIS

Sr. No.	Lexis	Frequency	Percentage of Presence
1	LOTTERY	22765	764.18
2	BATCH	8629	289.66
3	LOTTO	7184	241.15
4	BALLOT	3875	130.08
5	EUROS	3623	121.62
6	NETHERLANDS	3389	113.76
7	COORDINATOR	3379	113.43
8	WINNINGS	3232	108.49
9	PROCESSED	3022	101.44
10	DRAWS	3011	101.07
11	UNCLAIMED	2631	88.32
12	UNWARRANTED	2617	87.85
13	COMPLICATIONS	2605	87.45
14	UNNECESSARY	2540	85.26
15	LUMP	2481	83.28
16	CREDITED	2001	67.17
17	RANDOMLY	1679	56.36
18	AMSTERDAM	1598	53.64
19	IDAHO	1336	44.85
20	PAYOUT	1272	42.70
21	FIDUCIARY	1239	41.59
22	LOTERIA	1088	36.52
23	SWEEPSTAKES	1049	35.21
24	JACKPOT	1042	34.98
25	UNSCRUPULOUS	1008	33.84
26	SWEEPSTAKE	969	32.53
27	OXFORD	958	32.16
28	PROMOTED	868	29.14
29	RATINGS	828	27.79
30	CHRONOLOGICALLY	828	27.79
31	CATALOG	828	27.79
32	SCHEMAS	795	26.69
33	XMAS	728	24.44
34	PRECAUTIONARY	720	24.17
35	BOOKLET	710	23.83
36	CORRESPONDENCES	704	23.63
37	PRIMITIVA	691	23.20
38	CONGRATULATION	682	22.89
39	STAATSLOTERIJ	670	22.49
40	BALLOTING	627	21.05
41	PARTICULARS	619	20.78
42	NATIONALITY	618	20.75
43	FREELOTTO	601	20.17
44	LOTERIJ	596	20.01
45	CONGRATULATE	493	16.55
46	OCEANIA	477	16.01
47	MARITAL	452	15.17
48	CHEQUE	432	14.50
49	BREACH	418	14.03
50	STAKES	416	13.96
51	LOTTERIES	415	13.93
52	EASTER	397	13.33
53	NETHERLAND	388	13.02
54	NOTARIZED	385	12.92
55	FIDUCIAL	385	12.92
56	VIRGILIO	378	12.69
57	DAYZERS	374	12.55
58	CONSOLATION	352	11.82
59	INFORMATIONS	349	11.72
60	UNIONS	340	11.41
61	VERIFICATIONS	339	11.38
62	ASSOCIATIONS	338	11.35
63	ANNOUNCEMENT	337	11.31
64	EUROMILLION	335	11.25
65	ISSUANCE	335	11.25
66	POWERBALL	334	11.21
67	ZONAL	329	11.04
68	TRANSFERRING	323	10.84
69	PROGRAMMED	296	9.94
70	EUROMILLIONS	292	9.80
71	RESIDENCE	290	9.73
72	SIXTY	283	9.50
73	PRIVILEGED	280	9.40
74	IDAHOLOTTO	268	9.00
75	LUCKYDAY	264	8.86
76	SWISS	263	8.83
77	ANNOUNCED	262	8.79
78	COLA	257	8.63
79	STATUTORY	256	8.59
80	ZOMERLOTERIJ	255	8.56
81	TREASURY	252	8.46
82	MINISTERIO	244	8.19
83	AFRO	243	8.16
84	LEGALISATION	242	8.12
85	COCA	241	8.09
86	ECONOMIA	241	8.09
87	CYBER	240	8.06
88	KIN	240	8.06
89	WISHING	238	7.99
90	STERLINGS	235	7.89
91	HACIENDA	235	7.89
92	WIDOW	228	7.65
93	ORPHAN	228	7.65
94	ACKNOWLEGEMENT	227	7.62
95	MILLONES	226	7.59
96	APRAISAL	225	7.55
97	PROCEEDINGS	225	7.55
98	NOTARIZATION	218	7.32
99	HOORNWIJCK	213	7.15
100	TWOHUNDRED	207	6.95

In English language the tokens are words [24] and the act of breaking the text into tokens is called Tokenization. A noteworthy thing here is that our tokenization is not case-sensitive. This means that a word appearing in any combination of lower-case or upper-case letters is treated as the same word. As a next step we counted the number of unique tokens in each UBE. This resulted in each document being represented as sub-set of Vector Space Document Model (VSDM). A vector corresponding to each UBE in this model is 2-dimensional, consisting of unique tokens and their frequency and is sorted on frequency column in descending order. This resulted in a total number of 2979 vectors, one each for the 2979 UBE in the cluster of interest.

Further, the UBE vectors are designed not to include stop-words. A special kind of stop-words considered by us are Domain stop-words. These are the words which are statistically irrelevant in the context of current research work because of their presence in both clusters, i.e. cluster formed of winnings-announcing UBE and the cluster formed of non winnings-announcing UBE. Hence, the entire stop-list considered by us, consists of following four types of stop-words:

a. HTML stop-words e.g. html, body, img
b. Generic stop-words e.g. his, thus, hence
c. Noise stop-words e.g. isdfalj, asdfwg
d. Domain stop-words e.g. salary, academy, phone

As a final step towards simplification of data processing, we created a single vector from the 2979 vectors of UBE documents. This 2-dimensional vector consisted of 7453 unique tokens and was sorted on the frequency count of tokens in descending manner. The number of tokens in this single vector was naturally less than the sum of number of tokens in each of 2979 vectors. The frequency count for a given token in this vector is the aggregate sum of the frequency count of the token in the 2979 vectors. This means to say that those vectors which do not contain the given token, contribute a value of zero towards the aggregate sum. Next, our motive was to keep only the desired lexis in this vector of extracted lexis. As the stop-words were already removed, this was a second level of refinement of the vector. For this we removed all lexis of length greater than 30, as we did not deemed them to be of statistical relevance. The frequency of 1 in the aggregated vector is an indication that the token has appeared only 1 time in 2979 documents. As a result we also removed all those tokens with a frequency of 1. The number of lexis with length greater than 30 and with frequency of 1 was 6 and 3270 respectively. The removal of such words resulted in the highly refined selected lexis set of 4177 lexis.

IV. Results And Findings

Based on the processing of nearly 3000 winnings-announcing UBE, we obtained a vector containing 4177 lexis. This vector is a set of words contained in the winnings-announcing UBE. On the analysis of this vector, we were able to identify the most frequently occurring lexis in such UBE. The identification of lexis with highest frequency is possible from this vector as it is sorted in descending manner on the frequency count of the lexis. A snap-shot of listing of such top 100 lexis is given in Table I. The third column of Table I depicts the frequency of the word in the set of 2979 UBE whereas the fourth column is the ratio of frequency of the word to the number of UBE. In Table I, this is expressed in terms of percentage of the value and is called 'Percentage of Presence'.

The first record of Table I can be interpreted to say that the word 'LOTTERY' appears for 22765 times in 2979 UBE with a presence percentage of {(22765 / 2979) x 100 =} 764.18%. The other records of the Table I can be interpreted similarly. If a word appears for more than 2979 times in a set of 2979 UBE, evidently the word registers a presence of more than 100% in the UBE set under consideration. We were able to find that the top 10 words namely, 'LOTTERY', 'BATCH', 'LOTTO', 'BALLOT', 'EUROS', 'NETHERLANDS', 'COORDINATOR', 'WINNINGS', 'PROCESSED' and 'DRAWS' register a presence of more than 100% in the UBE set. An important interpretation of this result is that the presence of these words is a clear indication of a high probability of the UBE under consideration to be one announcing a winning. Similarly, the next 8 words have registered a presence of more than 50% in the UBE set.

TABLE II
Range of Percentage of Presence of Lexis and Frequency of Lexis in that Range

Sr. No.	Range of Percentage of Presence of Lexis	Frequency in the Range
1	0-24	66
2	25-49	16
3	50-74	3
4	75-99	5
5	>=100	10
Total	5	100

Moving on these lines, we divided the entire presence percentage data of Table I into 5 ranges. The pertinent data is presented in Table II. The third column of Table II depicts the frequency of the values of percentage of presence of the fourth column of Table I. The most important interpretation of Table II is that as we move from first record to the fifth record in this table the probability of a given UBE being a winning announcer increases. This increase is proportional to the occurrence of word(s) from the number of words listed corresponding to the range. For instance, the probability of a given UBE being a winning-announcer is more if a word 'LOTTERY' occurs in it, compared to occurrence of word 'COCA' in the UBE. This is so because the word 'LOTTERY' forms an element of the set with range '>=100' whereas the word 'COCA' forms an element of the range set '0-24'.

V. Conclusion

Based on the textual analysis of nearly 3000 winnings-announcing UBE, we conclude that it is possible to identify the lexis which are occurring in these UBE. We identified such lexis based on the criteria of their frequency of occurrence in the data set under consideration. We also attempted to analyze the identified lexis of UBE of interest. Based on the analysis of such top 100 most frequently occurring lexis in the winnings-announcing UBE, we conclude that the presence of words, ‘LOTTERY’, ‘BATCH’, ‘LOTTO’, ‘BALLOT’, ‘EUROS’, ‘NETHERLANDS’, ‘COORDINATOR’, ‘WINNINGS’, ‘PROCESSED’ and ‘DRAWS’ is an indication of a high probability that the given UBE will be a winning-announcer. We advocate that our results could be put to use for text-based identification of winnings-announcing UBE. We further conclude that the identification of presence of combination of lexis presented in this paper can be put to use for further research work concerning the winnings-announcing UBE.

We believe that the best way to fight spam is to understand it. The current paper is an attempt to understand the UBE which claim to provide large amount of wealth through announcements of fake schemes like lotteries, prizes, selections and draws. The current work can be extended to implement a naïve anti-UBE fighter for such winnings-announcing UBE.

Our results are best reported on the dataset used. We do not promote or discourage either the use of specific word or of lexis in the designing of winnings-announcing UBE. We just present the identification of lexis which occur most frequently in the UBE announcing fake winnings. The current work is having a wide range of general applicability to other text domains including the other categories of UBE. On the sidelines of the current study, we advocate that it has also provided an insight into behavior of spammer’ preference for selection of lexis for designing winnings-announcing UBE. Finally, we sincerely believe that only awareness and alertness can help protect the general masses against the fake and lethal-consequences bearing net of greedy persons who are always looking for victimizing the innocent persons through their luring offers of winnings-announcing UBE.

REFERENCES

[1] Anonymous, “Categorizing junk eMail”, Available: http://www.knujon.com/categories.html, 2008

[2] Berry R. “The 100 Most Annoying Things of 2003”, Available: http://www.retrocrush.buzznet.com/archive2004/annoying2003/, January 18, 2004

[3] Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., Vigna S. “A Reference Collection for Web Spam”, *ACM SIGIR Forum*, vol. 40 (2), pp. 11-24, December 2006, ISSN: 0163-5840

[4] Commtouch Software Ltd. “Spam Trends For First Half of 2004”, Commtouch Report, Available: http://www.commtouch.com/Site/News_Events/pr_content.asp?news_id =45&cat_id=1, Press Release, 30 June, 2004

[5] Crucial Web Hosting Ltd., “How Consumers Define Spam”, Available: http://www.crucialwebost.com/blog/how-consumers-define-spam/, March 06, 2007

[6] CUED, “Junk e-mail”, Cambridge University Engineering Department, Available: http://www.cam.ac.uk/cs/email/junk, 2008

[7] Fette I., Sadeh N. and Tomasic A. “Learning to Detect Phishing Emails”, Institute for Software Research International School of Computer Science (ISRI), Carnegie Mellon University (CMU), CMU-ISRI-06-112, June 2006

[8] Frederic E. “Text Mining Applied to Spam Detection”, *Presentation given at University of Geneva* on January 24, 2007, Available: http://cui.unige.ch/~ehrler/presentation/ Spam%20Filtering.pdf

[9] Gajewski W. P. “Adaptive Naïve Bayesian Anti-spam Engine”, *Proceedings of World Academy of Science, Engineering and Technology (PWASET 2005)*, Pages 45-50 vol. 7 August 2005 ISSN 1307-6884

[10] Gyongyi Z. and Garcia-Molina H. “Web Spam Taxonomy”, *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb, 2005)*, Chiba, Japan, April 2005

[11] Indiana University. “What is spam?”, University Information Technology Services, Knowledge Base, Indiana University, Pennsylvania, November 11, 2008. Available: http://kb.iu.edu/data/afne.html

[12] Infinite Monkeys & Co., “Spam Defined”, Available: http://www.monkeys.com/spam-defined/definition.shtml, 2008

[13] Kiritchenko S. and Matwin S. “Email Classification with Co-Training”, *Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research,* Toronto, Canada, Page 8, 2001

[14] Lambert A. “Analysis of Spam”, *Dissertation for Degree of Master of Science in Computer Science*, Department of Computer Science, University of Dublin, Trinity College September 2003

[15] Lance J. “Phishing Exposed”, Syngress Inc., Chapter 1 Page 2 ISBN: 159749030X; 2005

[16] Martin S., Sewani A., Nelson B., Chen K. and Joseph A. D. “Analyzing Behaviorial Features for Email Classification”, *Proceedings of the Second Conference on Email and Anti-Spam (CEAS, 2005)*, Stanford University, California, U.S.A. July 21-22, 2005

[17] Roth W. “Spam? Its All Relative”, Available: http://www.imediaconnection.com/content/7581.asp, published online on December 19, 2005

[18] ScamBusters Editors “Email Scam Analysis”. Available: http://www.scamdex.com/MHON/E/msg08805.php, Scamdex, Scambusters Online - Issue No. 292

[19] Sebastiani F. “Machine Learning in Automated Text Categorization”, *in ACM Computing Surveys*, vol. 32 (1), pp. 1-47, March 2002. ISSN 0360-0300

[20] Sen P. “Types of Spam”, Interactive Advertising, Fall 2004, Available: http://ciadvertising.org/sa/fall_04/adv391k/paroma/spam/types_of_spam .htm

[21] Threat Research and Content Engineering (TRACE) “Spam Type Descriptions”. Available: http://www.marshal.com/TRACE/Spam_Types.asp, TRACE Blog, 2008

[22] Wikimedia Foundation Inc. “E-mail”, Available: http://en.wikipedia.org/wiki/Email, 2010

[23] Youn, S. and McLeod D. “Spam Email Classification Using an Adaptive Ontology”, *Institute of Electrical and Electronics Engineers (IEEE) Journal of Software*, April 2007

[24] Zhang T. “Predictive Methods for Text Mining”, *Machine Learning Summer School - 2006*, Taipei. Available: videolectures.net/mlss06tw_zhang_pmtm