World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:5, No:8, 2011

Identification of Non-Lexicon Non-Slang Unigrams in Body-enhancement Medicinal UBE

Jatinderkumar R. Saini, Apurva A. Desai

***Abstract*—**Email has become a fast and cheap means of online communication. The main threat to email is Unsolicited Bulk Email (UBE), commonly called spam email. The current work aims at identification of unigrams in more than 2700 UBE that advertise body-enhancement drugs. The identification is based on the requirement that the unigram is neither present in dictionary, nor is a slang term. The motives of the paper are many fold. This is an attempt to analyze spamming behaviour and employment of word-mutation technique. On the side-lines of the paper, we have attempted to better understand the spam, the slang and their inter-play. The problem has been addressed by employing Tokenization technique and Unigram BOW model. We found that the non-lexicon words constitute nearly 66% of total number of lexis of corpus whereas non-slang words constitute nearly 2.4% of non-lexicon words. Further, non-lexicon non-slang unigrams composed of 2 lexicon words, form more than 71% of the total number of such unigrams. To the best of our knowledge, this is the first attempt to analyze usage of non-lexicon non-slang unigrams in any kind of UBE.

***Keywords*—**Body Enhancement, Lexicon, Medicinal, Slang, Unigram, Unsolicited Bulk e-mail (UBE)

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' [4, 10, 20]. 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, 6, 16].UBE incidences range from fake job offers and viruses to pornography. Another area of concern is of spam e-mails that advertise the body enhancement medicinal products. The target areas of these products range from enhancement of male and female organs to loose or gain weight, improve hair growth, increase height and reduce blood-sugar. The dangerous thing about these emails is that they demand a handsome amount of money for delivery of the product, which is never delivered or in worst case a fake product is delivered. But due to the fear of society and feeling of embarrassment, the victim rarely comes out to declare of the way he/she was cheated through non-delivery or delivery of a fake product against a heavy payment of a so-called body enhancement medicine. Further, this kind of UBE mostly targets medicines or drugs like Viagra, Xanax and Phentrimine for the genitals and many times the advertising pharmacies include pictures and textual statements in the emails which are largely pornographic. Even though there are many target areas of such medicinal products as advertised and offered in the UBE, in general this paper refers to this kind of UBE as body enhancement medicinal UBE.In past, researchers have worked in direction of understanding the spam for combating it [9, 12, 26]. We also believe that first step in combating spam is to understand spam and the best way of understanding spam is to analyze it. Most importantly, spam can be differentiated by content [23] and in this paper we target content-based analysis of un-structured UBE documents which advertise fake medicines for body enhancements. This work aims towards identification of specific type 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 non-lexicon non-slang unigrams 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. The lexicon used by us for identification of lexicon words is English language dictionary. Further, we do not use dictionary of technical terms like legal terms, medicinal terms, etc. The present work treats all those

J. R. Saini is with the Sankalchand Patel College of Engineering, Visnagar, Mehsana, Gujarat, India as Associate Professor and Head of Department of Computer Science. 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. He is PhD from Veer Narmad South Gujarat University, Surat, Gujarat, India. (e-mail: desai_apu@hotmail.com).

words, which are not present in lexicon and also can not be called slang words, as non-slang words. The next section on survey of related literature details more about slang words.

II. Related Work

As far as, the study of past and contemporary literature for this field is concerned, this is the first formal attempt for identification of non-lexicon non-slang unigrams in body-enhancement advertising medicinal UBE. There are quite a large number of research instances in the scientific literature where the classification problem of emails into spam and non-spam categories has been discussed. The numbers of research instances dealing with classification of spam emails, as such, are quite limited. Evett [5], Lance [17], Ma et al. [18] and Sravan [24] have provided a preliminary classification for UBE. Among the prominent ones, Saini and Desai [21] have worked towards classification of UBE into 14 categories. They have defined 'Medicinal Advertisements (ADV_MED)' as one of the UBE categories and classified messages dealing with genital-enhancement drugs, weight-loss tablets, weight-gain tablets and hair-grow oils, into this category. This paper is an attempt to extend this work by digging deeper in the analysis of unigrams used by spammers in such Medicinal UBE. [11] have treated e-mail classification as a special case of text classification. Gajewski [7] has discussed the use of a naïve Bayesian classifier based on a BOW representation of an e-mail. The usage of tokens, which do not consist of multiple words of documents, has been done specifically for emails by Meyer and Whateley [19] and Gajewski [7]. They have termed these kinds of tokens as 'Unigrams'. Also, based on the review of related literature, it is evident that the usage of slang, by different groups, has been employed by researchers for analysis purpose. According to analysis of Krasny [13] for usage of slang words, English language is constantly changing and slang is increasingly becoming a greater part of our shifting linguistic terrain. He has concluded that most new words come from slang. Astriyani et al. [1] have presented an identification and analysis of slang language related to sex in lyrics of a rap singer Eminem. Thorne [25] in his research article on slang, style-shifting and sociability has remarked that along with other factors, email is responsible for generation of new slangs as also for enormous proliferation of websites designed to celebrate and decode slang.In the more specific fields of computer sciences, [15] in their work on predicting user and message attributes in computer-mediated communication have concluded that the use of slang words and misspellings is frequent during chatting. They have identified various slang words in their work which is mainly focused on chat mining. In another paper,[14] have employed the analysis of slang words for gender prediction in chat data. Based on the slang terms identified this time, they have concluded that males are more dominant and decisive in usage of slang words. [8] have presented a stylometric analysis of bloggers' age and gender. They have employed slang words as a stylometric feature and concluded that teenage bloggers use more slang words than adults. They have advocated that the usage of slang can be a good feature to predict the geographical location or the ethnic group of the user. Moving on these lines, we have attempted to identify those unigrams which are neither present in the usual English dictionary, nor are used as slang words. These words of interest are actually mutated forms of formal words. The next section details the methodology for identification of such mutated words in form of non-lexicon non-slang unigrams.

III. Methodology

In this section, is described the detailed methodology followed towards identification of non-lexicon non-slang unigrams (NNU) in body-enhancement medicinal UBE. For the sake of simplicity and better understanding, the entire section is divided into four major sub-sections as follows. The picture in Figure 1 is the diagrammatic representation of the followed methodology.

A. Data Collection and Clustering
B. Data Pre-processing
C. Feature Extraction & Feature Selection
D. Identification of Non-lexicon Non-slang Unigrams

A) Data Collection and 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 body enhancement medicinal UBE whereas the other cluster did not contain such UBE. The cluster comprising body enhancement UBE was the cluster of interest and the number of instances in it was 2711, which amounted to nearly 178 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 advertise the body enhancement medicines or medicinal products like for genitals, hair, fat and weight. Our task of data collection was eased by the fact that many of this kind of UBE have an explicit 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.

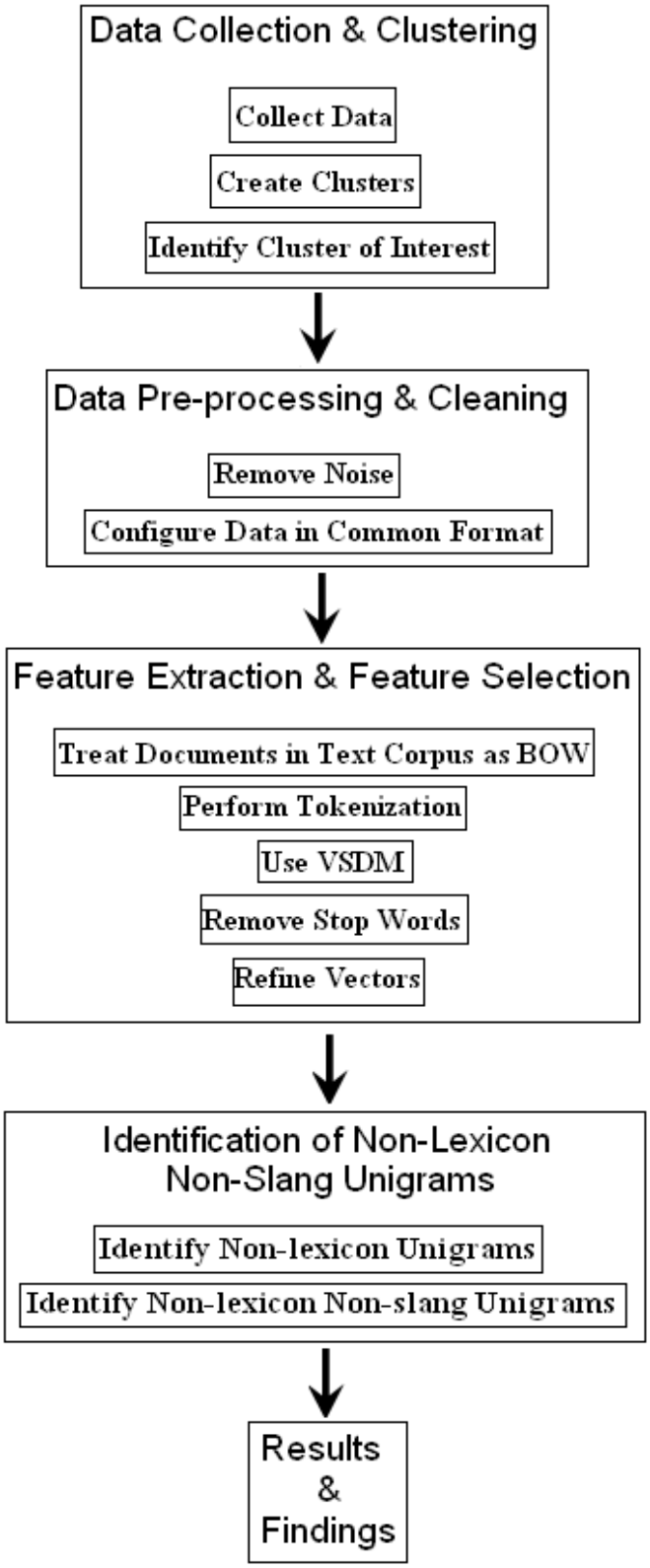

Fig. 1 Block Diagrammatic Representation of Methodology

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 body enhancement medicinal 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) [22]. We then performed Syntactic Text Analysis by Parsing the UBE document, for extraction of Tokens.

In English language the tokens are words [27] 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 2711 vectors, one each for the 2711 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 is 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 body enhancement medicinal UBE and the cluster formed of non body enhancement medicinal UBE. Hence, the entire stop-list considered by us, consists of following three types of stop-words:

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

As a final step towards simplification of data processing, we created a single vector from the 2711 vectors of UBE documents. This 2-dimensional vector consisted of 16879 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 2711 vectors. The frequency count for a given token in this vector is the aggregate sum of the frequency count of the token in the 2711 vectors. This means to say that those vectors which do not contain the given token, contribute a value of zero towards the aggregate sum.

TABLE I
VARIOUS TYPES OF UNIGRAMS

Sr. No.	Unigram	Type of Unigram
1	AHEZMCJI	N
2	AHEZWXADUTXOZS	N
3	AIRPLANE	L
4	ALWAYSDREAMT	NNU
5	ATUTHOUY	N
6	AVZYNMG	N
7	CANADIANPHARMACY	NNU
8	CHEAAP	S
9	CIAALIS	S
10	CIALIS	T
11	COCK	S
12	DIC	S
13	DRUGS	L
14	ENLARGEMENT	L
15	FUCKINGG	S
16	HERBAL	L
17	INCHES	L
18	LEVITRA	T
19	LOOOOOOOSERS	S
20	LOWESTPRICE	NNU
21	MALE	L
22	PENIS	L
23	PENISGROWTH	NNU
24	PENNISS	S
25	PHARMACY	L
26	VIAGRA	T

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 2711 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 8 and 8366 respectively. The removal of such words resulted in the highly refined selected lexis set of 8505 unigrams.

D) Identification of Non-lexicon Non-slang Unigrams

The lexis set consisting of 8505 unigrams had two types of words, listed as follows:

a. Lexicon Words (L)
b. Non-lexicon Words

We removed all the lexicon words from this list and created a vector of 5618 non-lexicon unigrams. This vector of non-lexicon unigrams, in turn, was further found to have four types of words, listed below:

a. Slang Words (S)
b. Non-slang Words (NNU)
c. Noise Words (N)
d. Technical Words (T)

Table I presents a snap-shot of randomly selected unigrams of various types from the vector consisting of 8505 unigrams. The motive of present work was identification of non-slang words. Hence, we ignored the slang words, noise words and technical words too. It is noteworthy to state that words like Viagra, Cialis, Vicodine, Valium and Levitra are non-lexicon but technical words of medical domain. We have not considered these words as NNU.

IV. RESULTS AND FINDINGS

Based on the processing of more than 2700 body enhancement medicinal UBE, we obtained a vector containing 8505 lexis. This vector is a set of words contained in the body enhancement medicinal UBE. This vector was further analyzed and processed to yield a vector containing 5618 non-lexicon unigrams of which 132 unigrams were also non-slang. An exhaustive listing of all such 132 unigrams is presented in Table II. It was found that non-lexicon words constitute around 66% (5618 / 8505) of the total number of lexis in body enhancement medicine-advertising UBE corpus. The non-slang words constitute nearly 2.4% (132 / 5618) of non-lexicon words in this corpus.

A typical characteristic of unigrams presented in Table II is that these unigrams are formed of more than one literal word. This is also the reason of their being getting eligible for identification as non-lexicon non-slang unigrams. The fact that they are not present in dictionary makes them non-lexicon. The usage of these words is neither vernacular, nor do they form vocabulary of jargon English. Hence these words are also non-slang. Further, each word is used as a single word in the UBE corpus. Hence they are termed as unigrams. We tried to analyze the composition of unigrams presented in Table II and found that they are formed of any of 2, 3, 4 and 5 lexicon words. This means to say that their composition is done from lexicon words but absence of punctuation marks (like hyphen, comma, full-stop, semi-colon and colon) as well as special characters (like space, hash and dollar sign) makes them classified as non-lexicon. Table III presents the frequency distribution of words used for composition of non-lexicon non-slang unigrams.

TABLE II
NON-LEXICON NON-SLANG UNIGRAMS (NNU) IDENTIFIED IN BODY-ENHANCEMENT MEDICINAL UBE

Sr. No.	NNU
1	ABOUTWHO
2	ACADEMICTUTORINGSERVICES
3	AERIALPHOTOS
4	AFFORDHERE
5	AFTEROPERATE
6	AGENCYPHONE
7	AGILISADVISORY
8	AGROBOX
9	AHOMEOFYOUROWN
10	AIDEYE
11	AIMANCOLLEGE
12	AIRBOX
13	ALASKABLUEGRASS
14	ALFASTUDIO
15	ALLBANKTURTLE
16	ALLENSDIAMONDS
17	ALLFRESHSEAFOOD
18	ALLIEDMEDICALTRANSPORT
19	ALLINTOGETHER
20	ALLPRODUCTS
21	ALONGSYRUP
22	ALWAYSDESIRED
23	ALWAYSDREAMT
24	ALWAYSLONGED
25	ALWAYSWANTED
26	AMAZINGMAKEOVERS
27	AMBIEMVALIUM
28	AMERICANCOLLEGE
29	AMERICANIDIOT
30	BESTLIFESYSTEMS
31	BESTPRICECANADA
32	BESTSELLERSBESTQUALITYVIAGRA
33	BESTWESTERNSEAPORTINN
34	BETTEREVENTS
35	BITSSIZE
36	BODYCOUNT
37	BOYTOYS
38	CANADIANPHARMACY
39	CANADIANPHARMCYONLINE
40	CHERISHPRODUCTS
41	CIALISVIAGRA
42	COMPLETELOVE
43	COMPLETETHEIR
44	CONSTANTCONTACT
45	COUNTRYBOY
46	COURSEADVISOR
47	ENLARGEPENIS
48	FEMALEVIAGRA
49	FREEVIAGRAPILLS
50	GAININGINCHES
51	GENERALSTICK
52	GENERICDRUGS
53	GETBIGTODAY
54	GIVINGDRUGS
55	GLASSLEAST
56	GLASSPRETTY
57	GLASSTHESE
58	GREATNOW
59	HAPPINESSWHO
60	HARDERECTIONS
61	HAVETHEM
62	HEALTHLINK
63	HEARTSTHROUGHHISTORY
64	HEATGREAT
65	HIGHDESERTBEVERAGE
66	HIGHPERFORMANCECONSULTANTS
67	HOSPITALTHE
68	HOTELMILANO
69	HOTMAILNEWS
70	HOTSECOND
71	INCREASEPENISSIZE
72	INSPIREYOU
73	LICENSEDPHARMACY
74	LIFEGOESON
75	LIFETEENWORLD
76	LIFEWOOD
77	LOVEFOREVER
78	LOVINGBRIDE
79	LOWESTPRICE
80	LOWMOLECULE
81	MASTERCARD
82	MEDICALHAIRRESTORATION
83	MEDICALHAIRRESTORATIONOFFER
84	MEDICALNEWCHARACTER
85	MYCANADIANPHARMACY
86	MYCANADIANSTORE
87	NAMESERVER
88	NANOPARTICLES
89	NEVERPERSONNEL
90	NEWLEVELMINISTRIES
91	NEWMASTER
92	NOFILL
93	NOINFECTION
94	ONLINEBLOGSPOT
95	ONLINEPHARMACY
96	OVERJOY
97	PENISGROWTH
98	PENISWIDTH
99	PRESCRIPTIONDRUGS
100	REALBEHIND
101	SEXDRIVE
102	SMALLBREAST
103	SMALLSHIPCRUISES
104	SOMETHINGHE
105	SPOKESMAN
106	SPOKESPERSON
107	SUPERSIZE
108	SUPERSIZEZME
109	THEACCOMPLISHMENTS
110	THEALSO
111	THECONCLUSION
112	THEDAY
113	THEEMPLOYMENTGUIDE
114	THEFIRSTPET
115	THEGOODLOVER
116	THERECORDGUY
117	UNLUCKYGIRL
118	UNLUCKYLOVER
119	UNLUCKYPINKS
120	UNLUCKYREDNECK
121	VIAGRACAILIS
122	VIAGRAPILL
123	VIAGRAPILLS
124	VIAGRAPROFESSIONAL
125	VIRTUALINTERFACE
126	WANDMOVE
127	WATCHCAUSE
128	YESCAME
129	YESTERDAYBEING
130	YOURPENIS
131	YOURSTRULY
132	YOUTUBE

The first record of Table III could be interpreted by saying that there are 94 words (out of 132) which are formed of 2 lexicon words and such a bulk consists of more than 71% of the total number of words distributed according to the number of lexicon words used for composition of NNU. The remaining records of Table III could be interpreted likewise. An example of a NNU composed of 2 lexicon words is 'ALWAYSDREAMT'.

TABLE III
FREQUENCY DISTRIBUTION OF WORDS USED FOR COMPOSITION OF NNU

Sr. No.	No. of Lexicon Words Used for Composition of NNU	Frequency	Percentage
1	2	94	71.21
2	3	33	25.00
3	4	2	1.52
4	5	3	2.27
Total	-	**132**	**100.00**

There are many nnu, presented in table ii, whose meaning and presence in body-enhancement ube corpus of present work is obviously in accord. 'online pharmacy', 'sexdrive','penisgrowth','prescriptiondrugs and 'smallbreast' are a few instances of this. Further, there are many nnu, presented in table ii, which do not manifest an apparent association between their meaning and their presence in body-enhancement ube corpus of present work. These nnu provide a better comprehension of this association only when they are treated contextually. For instance, presence of four similar nnu, viz. 'alwaysdesired','alwaysdreamt','alwayslonged', and 'alwayswanted' seems surprising. However, when the meaning of these nnu is deduced based on the context of their presence in following statements, their presence is well justified.

a. "Just click here and have the penis you alwayswanted!"
b. "Just click here and have the penis you alwayslonged for!"

V. CONCLUSION

Based on the textual content analysis of more than 2700 UBE containing advertisement of body-enhancement medicinal products, we conclude that it is possible to identify the lexis which are occurring in these UBE. We identified such 132 lexis based on the criteria of 'Non-lexicon and Non-slang Unigrams' in the data set under consideration. These words were abbreviated as NNU. The non-lexicon words constitute nearly 66% of total number of lexis of corpus whereas non-slang words constitute nearly 2.4% of non-lexicon words. Based on the analysis of NNU, we conclude that they are formed of combination of lexicon words. We also conclude that the number of such lexicon words contributing towards composition of NNU ranges from 2 to 5. It is also concluded that NNU composed of 2 lexicon words form more than 71% of the total NNU. This is followed by NNU composed of 3 lexicon words with a distribution share of 25%. The NNU composed of 4 and 5 lexicon words had a share of nearly 1.5% and 2.25%, respectively. Finally, it is concluded that there are many NNU whose meaning is evidence of their obvious presence in the UBE corpus under consideration. Also, there are many other NNU whose meaning in present context is justified only when their presence is analyzed in detail in context of body-enhancement medicinal UBE.

The current work can be extended to implement a sophisticated textual content based anti-UBE fighter specifically for filtering such fake medicinal product 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 body enhancement medicinal product announcing UBE. We just present the identification of lexis which are non-dictionary and non-slang and occur in such UBE. 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 very significantly provided an insight into behaviour of spammers' preference for selection of lexicon words and their combinations for designing fake body enhancement medicinal-product-announcing UBE. The present work also serves to provide an understanding of the 'word mutation' technique used by the spammers. Finally, we sincerely believe that only awareness and alertness can help protect the general masses against the fake and sometimes lethal-consequences bearing net of greedy persons. Such persons are always looking for victimizing the innocent persons through their luring offers targeting the psychologically vulnerable points like enhancement of genitals.

REFERENCES

[1] Astriyani, Sutjiati R. and Purwaningsih D. E. "An Analysis of Slang Language Related to Sex in Eminem's Rap Songs' Lyrics", *Repository of Gunadarma University, Jakarta*, 2007. ISSN: 1987-4783
[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., and Vigna S. "A. Reference Collection for Web Spam", *ACM SIGIR Forum*, v. 40, n. 2, p. 11-24, December 2006, ISSN: 0163-5840
[4] Crucial Web Hosting Ltd. "How Consumers Define Spam". Available: http://www.crucialwebost.com/blog/how-consumers-define-spam/, March 06, 2007
[5] Evett D. "Spam Statistics 2006", TopTenREVIEWS Inc. Available: http://spam-filterreview.toptenreviews.com/spam-statistics.html
[6] 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
[7] Gajewski W. P. "Adaptive Naïve Bayesian Anti-spam Engine", in *Proceedings of World Academy of Science, Engineering and Technology (PWASET 2005)*, Pages 45-50 Volume 7 August 2005 ISSN 1307-6884
[8] Goswami S., Sarkar S. and Rustagi M. "Stylometric Analysis of Bloggers' Age and Gender" in *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM – 2009)*, San Jose, California, May 2009
[9] Gyongyi Z., Garcia-Molina H. "Web Spam Taxonomy", *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb, 2005)*, Chiba, Japan, April 2005

[10] Infinite Monkeys & Co. "Spam Defined". Available: http://www.monkeys.com/spam-defined/definition.shtml, 2011
[11] Kiritchenko S. and Matwin S. "Email Classification with Co-Training", in *Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative Research*, Toronto, Canada, pp. 8, 2001
[12] Knujon.com "Categorizing junk eMail". Available: http://www.knujon.com/categories.html, 2011
[13] Krasny M. "Analysis: Usage of Slang Words", *article from Talk of the Nation (NPR)*, August 7, 2000. Available: http://www.highbeam.com/doc/1P1-30383388.html
[14] Kucukyilmaz T., Cambazoglu B. B., Aykanat C. and Can F. "Chat Mining for Gender Prediction", in *Lecture Notes in Computer Science*, Springer Berlin, Heidelberg vol. 4243/2006, pp. 274-283,. ISSN: 0302-9743
[15] Kucukyilmaz T., Cambazoglu B. B., Aykanat C. and Can F. "Chat mining: Predicting user and message attributes in computer-mediated communication" in *Information Processing and Management: An International Journal*, vol. 44, issue no. 4, pp. 1448-1466, July 2008. ISSN: 0306-4573
[16] 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
[17] Lance J. "Phishing Exposed", *Syngress Inc.*, ISBN:159749030X
[18] Ma W., Tran D. and Sharma D. "Filtering Spam Email with Flexible Preprocessors", *Advances in Communication Systems and Electrical Engineering*, Lecture Notes in Electrical Engineering Volume 4 Pages 211-227, ISBN 978-0-387-74937-2
[19] Meyer T. and Whateley B. "Spambayes: Effective Open-Source, Bayesian Based, Email Classification System", in *Proceedings of the First Conference on Email and Anti-Spam (CEAS, 2004)*, Mountain View,California, USA 2004
[20] Roth W. "Spam? Its All Relative". Available: http://www.imediaconnection.com/content/7581.asp, Published online on December 19, 2005
[21] Saini J. R. "Self Learning Taxonomical Classification System using Vector Space Document Analysis Model for Web Text Mining in UBE", Ph.D. Thesis guided by Desai A. A., accepted by Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India, September 2009
[22] Sebastiani F. "Machine Learning in Automated Text Categorization", in *ACM Computing Surveys*, Vol. 32, No. 1, pp. 1-47, March 2002. ISSN: 0360-0300
[23] Sen P. "Types of Spam". Available: http://ciadvertising.org/sa/fall_04/adv391k/paroma/spam/types_of_spam.htm, Interactive Advertising, Fall 2004
[24] Sravan "Types of Spam Mail". Available: http://www.thatdamnpc.com/types-ofspam-mail, November 18, 2008
[25] Thorne T. "Slang, Style-shifting and Sociability", *Multicultural Perspectives on English Language and Literature*, Tallinn/London, 2004. Available: http://www.kcl.ac.uk/content/1/c6/03/08/16/Slang/%20Style-shifting%20and%20Sociability.doc
[26] Youn S. and McLeod D. "Spam Email Classification using an Adaptive Ontology", *Institute of Electrical and Electronics Engineers (IEEE) Journal of Software*, April 2007
[27] Zhang T. "Predictive Methods for Text Mining", *Machine Learning Summer School - 2006*, Taipei. Available: videolectures.net/mlss06tw_zhang_pmtm