Email Hoax Detection System Using Levenshtein Distance Method

Yoke Yie Chen , Suet-Peng Yong and Adzlan Ishak
Computer & Information Sciences Department
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
chenyokeyie@petronas.com.my, yongsuetpeng@petronas.com.my, adzlan89@gmail.com

***Abstract*—Hoaxes are non-malicious viruses. They live on deceiving human's perception by conveying false claims as truth. Throughout history, hoaxes have actually able to influence a lot of people to the extent of tarnishing the victim's image and credibility. Moreover, wrong and misleading information has always been a distortion to a human's growth. Some hoaxes were created in a way that they can even obtain personal data by convincing the victims that those data were required for official purposes. Hoaxes are different from spams in a way that they masquerade themselves through the address of those related either directly or indirectly to us. Most of the time, they appear as a forwarded message and sometimes from legit companies. This paper addresses this issue by developing a hoax detection system by incorporating text matching method using Levenshtein Distance measure. The proposed model is used to identify text-based hoax emails. Sensitivity and specificity are used to evaluate the accuracy of the system in identifying hoax emails.**

***Index Terms*—hoax detection system, Levenshtein distance, text analysis**

I. INTRODUCTION

E-mail hoaxes are known not to be malicious towards any created systems. Hoax is considered to be a junk which can give misleading information to the users or readers of the email. According to Blanzieri and Bryl [1], spam is a junk mail which is mostly unwanted or unsolicited which sent either directly or indirectly by personnel who has no current relationship to the email user. Hoaxes are unsolicited and unwanted emails which directly or indirectly sent by personnel who has current relationship with user of the email [2]. In short, hoaxes are the 'smarter' version of spams which masquerade themselves well via the personnel that are present as ones' contacts. Usually hoaxes come in a form of forwarding messages from various sources. Misleading information from hoaxes may cause financial damage and irritating individual users. Worse of it, hoax has the ability to gather information and could possibly convince e-mail receivers act on non-exist events.

Some email hoaxes had caused a lot of false alarms which lead to negative impact on certain parties. For example, PayPal had to undermine and took responsibility after a Nigerian scam which was considered as a type of hoax was forwarded to most of its recipients by an unknown source that masquerade itself by using the email address similar to PayPal in year 2003. Another false alarm on the existence of acid rain due to the explosion of nuclear reactor in Japan had also caused chaos to the South East Asians in which this hoax was forwarded to many after a week from the earthquake incident in Japan. Although these hoaxes had yet to be proven as frauds, their contents were misleading [2-4].

Even though some e-mail hoaxes are not malicious, it is often a disturbance to the society. Furthermore, based on Meme's theory [5], when a person's mind is being infected by the wrong information, it will be hard for that infected person to be convinced otherwise of the real truth. The problem can only be solved once the users were exposed to the real fact of these hoaxes. Nevertheless, it will not be easy to convince a person on the real truth unless the source of the truth presented is valid to that person's view [5]. Therefore, it is crucial that hoaxes should be identified primarily to avoid further misunderstandings which can bring significant losses as well as negative impact to a party's growth. In view of this, the development of automatic hoax detection system will serve the purpose to protect e-mail receivers from mislead by false information. In the literature, there are not many hoax detection system created thus far. The work related to email hoax detection systems are stated in [2-4]. Most of its detection patterns are the same but different by definition [1].

Previous works in this area suffer from the ability to update their hoax database in order to automatically identify the latest hoaxes existence. In this paper, we present our proposed hoax detection system by using text matching approach to determine the hoax suspicious level of an incoming e-mail. The main objective of this project is to develop a Hoax Detection System which is able to

Manuscript received January 2014
Yoke Yie Chen is with Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Perak, Malaysia (email: chenyokeyie@petronas.com.my)
Suet-Peng Yong is with Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Perak, Malaysia (email: yongsuetpeng@petronas.com.my)
Adzlan Ishak was with Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Perak, Malaysia (email: adzlan89@petronas.com.my)

detect new and old email hoaxes presented in today's world. The main concern is to create higher awareness on the existence of email hoax and enlighten the users on the validity of every email's content that has been forwarded.

In the proposed system, a new incoming e-mail will go through a statistical based filtering algorithm. Based on the statistical outcomes, an email which identified as a hoax that was not found in the hoax database will be stored as new hoax in the database. In other words, our proposed system is able to perform automatic update while detecting hoaxes.

II. Related Works

A. Hoax Detection System vs. Spam Filters

Hoax detection system is not commercialized as good as spam detectors [1, 2]. Spam detectors and spam filters are similar to one another [6]. However, the concept of spam is pretty much simple and easy to understand. If the spam is not traceable, the spam is an effective spam while if the spam is recognizable, the spam filter is definitely an effective spam filter [6]. Spam filter has a lot of resemblance to a hoax detection system. In example, both hoax and spam email content contains clues such as capital letter words and exclamations marks. Perhaps, the only differences in these two subjects are the signatures of the targeted filter subjects, the definition of their own and also the method of extraction. Therefore, this suggests that the approach used in spam filtering system can also be applied to hoax detection system.

The process of approving whether an email is a spam or not are as follows [1, 2]:

- Compare email to spam database.
- If spam database contains the spam which was detected in the email, flag it as spam and remove them to the junk folder automatically (used by all email service provider today).
- Un-flagged spams can either be flag as neutral (safe) or suspicious depending on the probabilities of the spam and presented to the user to decide on which email should they remove as spam manually later on.

The most common approach in spam filtering systems is Naive Bayes algorithm, cluster-based approach, cross-regulation method, and text mining techniques. Naive Bayes classifier is the most common methods used in today's spam filtering system. According to Paliouras et al. [7], spam filtering is suitable with a machine learning classification such as Naive Bayes (NB) as it has been proven to cope well with this task despite of its simplicity.

On the other hand, cluster-based classification approach is another interesting approach for spam filtering. Often spams are being distributed more than the ordinary legitimate mails in which it tackles the skewed class distribution concept while concept drift is based on the user's preferences which may change over time or spam topic may vary according to fashionable trends [8]. A cluster-based classification approach is able to handle concept drift and skewed class distribution fairly well.

Cross-regulation method is a newly found method for spam filtering [9]. It was introduced with the aim to further detect spams based on an immune system concept. Thus far, the research has proven positive result but have very little similarities to hoax detection system based on its architecture. Text mining techniques has the ability to tackle key phrases as part of its mined texts [8]. Researches on text mining [10-12] show promising effort in comparing between texts. Text mining concept is a data mining tool that is able to extract text based contents based on the desired outcome. However, there are several limitations to this concept and the processing method might be a little slower.

B. Comparison of Previous Works on Hoax Detection Systems

Research has shown that it is possible to automate a hoax detection system. Hoaxes have certain characteristics in which most of its content can be found bolded, underlined, a lot of exclamation marks and often use words which require instant decisions to be made. Therefore, it is possible for hoaxes to be identified automatically [13].

Hoax detection research began actively in the year 2002. This model began its research after noticing that hoaxes are indeed a disturbing matter in the society. According to Hernandez et al. [3], knowledge is essential to human's growth and that misleading information can be costly. Having said so, they began by using the rule-based method as a start for this particular research. In 2002, rule-based method is one of the leading text extraction method used for anti-spams. Rule-based methods are methods in which extraction are made based on certain rules and conditions. This traditional filter method uses simple and straightforward methods which classify spams by matching email fields with key words. The failure to this method is due to the fact that the complexity of spams today has overridden the benefits that this method has yet to offer.

In 2004, Croatian CERT took a step further by developing another version of a hoax detection system [4]. This system has proven to be successful due to low Internet traffic over in the nation. However, this system has shown some drawbacks in certain area mostly on the idea of using fuzzy logic and Levenshtein method. These two methods have shown promising results in detecting hoaxes but the speed of processing the hoaxes is deemed to be slow added with heavy traffic of Internet and large volumes of data. Furthermore, the hoax database is not able to update on its own.

In 2009, the study for an automatic hoax detection system became much more feasible than before [2]. Instead of using rule-based as well as fuzzy logics, this research uses n-grams as text extraction method. N-grams is one of the data mining tool which predict words within the email. N-grams allow the system to predict the subsequent words in the analyzed email in which it can process the words faster (assumed and deemed to be by the inventors) However, given the study for n-grams in

spams [14, 15], the complexity of the probability in executing this concept is pretty complicated for hoax detection system. According to Vukovic et al. [2], n-grams approach is able to speed up the process of hoax detection. However, the drawback of using n-grams approach is that it is unable to train the hoax identifier on the latest hoaxes found based on the characteristics of a hoax. Table I shows the similarities and differences in existing works that we have composed from the literature.

TABLE I.
COMPARISON OF WORKS

Previous Works	2002 Hoax Detector [3]	Croatian CERT (2004) [4]	2009 Automatic Hoax Detection System [2]
Hoax Detection	Yes. Automated but not really intelligent. Expand to anti-virus software	Yes. Automated and rather intelligent (embed with a database)	Yes. Both automated and intelligent system integrated together
Target	More to Organizational level	Personal Based	Personal Based
Detection	Server based	Server Based	Server based
Form	Text	Text	Text
Approaches / Data mining Concepts highlighted	Not available	Levensthein + Fuzzy Logic + modified nearest neighbour algorithm	n-grams
Result	Able to detect but at minimal effect	Able to detect with probabilities analyzed	Able to detect with probabilities analyzed
Time Require to process	Not Available	Slower with tons of data to process	Light and faster (based on few samplings)
Claims	Time Require to Process made claim and have no specific results to prove on its similarity of comparison thus far		
Gap	It requires an intelligent mechanism to read and compare hoaxes at personal level	Gap	It requires an intelligent mechanism to read and compare hoaxes at personal level

Based on the analysis as shown in Table I, there are still rooms for improvement on the current approaches in developing hoax detection system. Other sort of close-related subjects as part of the investigation related to hoax detection are spams and text mining tools. The architecture of spams filters are similar to the hoaxes detection systems created over the years with different definitions. Spams can be traced based on the sender's address of a particular email [1, 6] while hoaxes need to be addressed by its content [2-4]. The methods used in hoaxes detection [2-4] have also been used by spam filters before in the literature [7, 8, 14, 16].

To the best of our knowledge, there are no hoax detection systems that are able to self update its database on the existence of new hoaxes. This means that the system is unable to recognize newly created hoaxes. This is the biggest challenge faced in this research. Therefore, our proposed system is to close the gap in hoax detection by identifying the characteristics of hoaxes based on an online hoax data.

The ultimate aim of this project is to increase the awareness of the existence of hoaxes and to avoid unnecessary misleading information that may cause misunderstanding among the society. Having said that, in order to close the gap with the previous researches, our proposed system includes features of identifying the characteristics of a hoax in helping to classify an incoming e-mail into different hoax suspicious levels.

III. METHODOLOGY

A. System Architecture

The proposed hoax detection system comprises of three main components: text pre-processing, hoax detection and new hoax detection. Text pre-processing gathers the email to be tested for its validity of being a genuine email or a hoax. Email header and signature will be removed from the original email. The system uses Levenshtein Distance to identify potential hoax from the email content by comparing it with the hoax database. Hoax database consists of email hoaxes that collected from the Hoax-Slayers.com, which was recognized by the National Library of Australia (PANDORA) as a good source of hoaxes. The proposed system design flow is shown in Fig. 1.

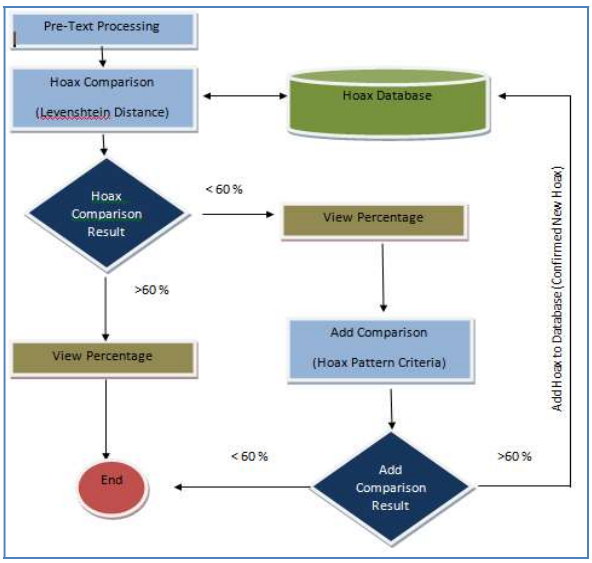

Figure 1. Proposed System Design Flow

B. Levenshtein Distance

Levenshtein Distance (LD) is a metric for measuring the number of differences between two strings. The following is the notation for Levenshtein Distance:

$$LD(s,t)$$

Where s = source string
t = target string

The distance between strings is measured by the number of insertion, deletion or substitution required to transform string s to string t. For example:

- If the source string (s) is "bold" and target string (t) is also "bold", then LD(s,t) = 0, because no transformations are needed. The strings are identical to each other.

- If original source string (s) is "bold" and target string (t) is "bonk", then LD(s,t) = 2, because two substitution (changes of "l" to "n" and "d to "k") is necessary to transform the original string to the compared string.

Hence, the higher the Levenshtein distance, the more different the strings are and the less likely that the email is a hoax. The email content input by the user will treat as the source string and target string are the hoax email stored in the hoax database. The final distance count will be converted into percentage using the following formula:

*Distance = (100 – total distance count) * 100*

As mentioned in the previous section, most of the hoax detection systems mentioned in the literature is not able to detect new hoaxes. To address this issue, the proposed system added another component to identify potential new hoaxes. Therefore, the following rules are applied in the system.

1. **Distance is greater than 60%, the e-mail is a hoax**

2. **Distance is less than 60%, the e-mail is suspected being a hoax. The email content will be sent for final filtering.**

C. Final Filtering

Based on our observations, hoaxes have certain similar patterns that can be easily detected such as:

- They often use uppercase letters
- Title comprises of words such as URGENT! Please FORWARD this message and etc.
- Includes extreme words such as WORST, VERY SERIOUS and etc

We maintain a bag of words that consists of common words that are frequently found in hoax emails. Table II is the sample collected bag of words found from hoaxes.

TABLE II.
COLLECTION OF HOAX KEYWORDS

Num	Words	Num	Words
1	URGENT	11	VIRUS
2	ATTENTION	12	FOLLOW
3	PLEASE	13	INSTRUCTIONS
4	FORWARD	14	IMMEDIATELY
5	SEND	15	MESSAGE
6	INFECTED	16	NOT
7	WORST	17	ADD
8	CLOSED	18	TELL
9	EVERYONE	19	EVERYBODY
10	LOVED	20	SPREAD

Hence, we calculate the percentage of words that are written in uppercase over the total number of characters found in the e-mail.

*Uppercase Percentage = (Total number of Uppercase Character / Total number of Characters in e-mail) * 100*

*Keywords Percentage = (Total number of Keywords found / Total number of words in e-mail) * 100*

We apply the 60% rule mentioned above on the formulas to determine whether or not the email is a new email hoax. New email hoaxes identified will be added into the hoax database.

IV. RESULTS AND DISCUSSIONS

A. System Prototype

As shown in Fig. 2, users who want to test whether an email content is a hoax will need to copy the email content to the system. Once user has click on 'Calculate Hoax Percentage' button, the hoax suspicious level will be calculated. If the suspicious level is less than 60%, the system will suggest user to determine if there is new hoax detected.

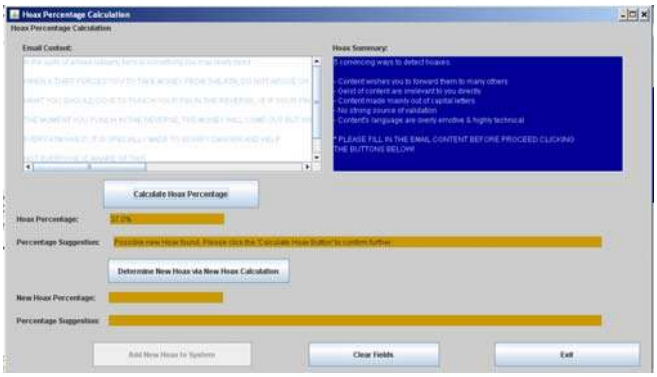

Figure 2. System Prototype

In the case when the system suggests to determine whether a new hoax detected, users are required to click

on 'Determine New Hoax via New Hoax Calculation' button. Once clicked, a new hoax suspicious level will be calculated. As shown in Fig. 3, the hoax suspicious level is 74%. Hence, a new hoax has been detected and the email content will be added into hoax database.

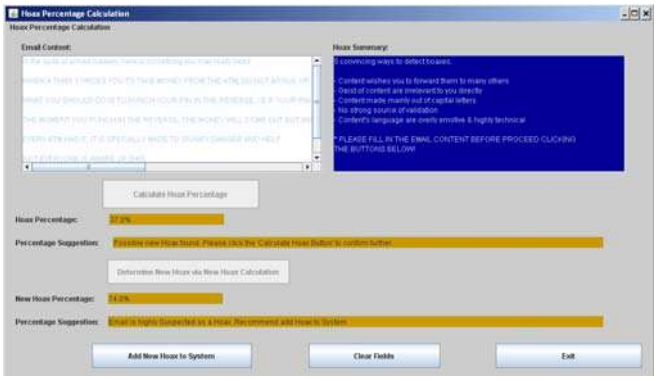

Figure 3. Detect New Hoax

B. System Evaluation

The accuracy of the system in identifying hoax emails need to be assessed in order to ensure the proposed system is effective in identifying hoax emails. The performance measure used to evaluate the accuracy of the system is positive predictive value, negative predictive value, sensitivity and specificity. A total number of 40 sample emails were tested using the system. Table III summarizes the evaluation results.

TABLE III.
SUMMARY OF EVALUATION RESULTS

		Email	
		Hoax	**Genuine**
System Results	**Hoax**	TP = 25	FP = 1
	Genuine	FN = 10	TN = 4

true positive (TP) = a true hoax email is identified by the system
false positive (FP) = a real email is identified as a hoax email
true negative (TN) = a real email is identified
false negative (FN) = a hoax email is identified as a real email

Based on Table 3, the system performance evaluation is as follows:

$$\textit{Positive Predictive Value (PPV)} = TP / (TP + FP) = 0.96$$

$$\textit{Negative Predictive Value(NPV)} = TN / (FN + TN) = 0.29$$

$$\textit{Sensitivity} = TP / (TP + FN) = 0.71$$

$$\textit{Specificity} = TN / (FP + TN) = 0.80$$

The high positive predictive value (PPV = 0.96) indicates that many true hoax emails has been identified by the system. The low negative predictive value (NPV = 0.29) indicates that the system is weak in identifying genuine emails. Hence, there is a high likelihood that the system will treat hoax email as genuine emails. This conforms to the high value of sensitivity (0.71) which indicates that the probability of the system identifying hoax email as genuine email is high. The system has a high specificity value (0.80); this suggests that the probability of the system identify genuine email as hoax is high. Although the system is able to produce good results in identifying hoax emails, it lacks the ability to identify a hoax email written with the format of a genuine email. During system testing, we can observe that most of the hoax emails that have been written professionally have not been identified as hoax by the system. This is because emails that have been written professionally help to increase reader's confidence and make them believe that the story that they tell is true. Hence, by using the format of the email (uppercase letters) as one of the factors to identify hoax email may not be an absolute effective method.

V. CONCLUSION AND RECOMMENDATIONS

Being aware of the existence of hoaxes create greater awareness for finding the right knowledge which can be used to further develop a human's perception as they grow. Misleading information can only cost losses to certain groups of people including tarnishing their image in the eye of the public. In this paper, we have developed a hoax detection system using Levenshtein Distance. The extracted e-mail is compared with the hoax database to support the decision on hoax identification. On the other hand, a value added feature is suggested in the proposed system by presenting the likeliness of hoaxes to the recipients. This makes a difference from the previous system. The system is able to provide high positive predictive value of 0.96 but it lacks the ability to identify genuine emails. Hoax emails does not only sent in a text-based format but also with images. The collection of hoax keywords is limited. Hence, these are the limitation of the system and there is still room for improvements.

For future work, the system can be improved by integrating artificial intelligence techniques such as fuzzy logic. With this technique, the system should be able to better classify hoax emails and genuine emails. Previous works in hoax detection is not adopting text processing theory except n-grams. However, it is believed that other text mining tool using machine learning algorithm [10] can be used as text pre-processing to extract texts from e-mail.

REFERENCES

[1] Blanzieri, E. and A. Bryl, "A Survey of Learning-based Techniques of Email Spam Filtering," *Artif. Intell. Rev.*, vol. 29(1), p. 63-92, 2008.

[2] Vukovi, M., K. Pripu, and H. Belani, "An Intelligent Automatic Hoax Detection System", in Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part I. 2009, Springer-Verlag: Santiago, Chile. p. 318-325.

[3] Hernandez, C.J., Sierra, J.M. and Ribagorda, A., "A First Step towards Automatic Hoax Detection", in Proceedings. 36th Annual 2002 International Carnahan Conference on Security Technology, 2002.

[4]Petković, T., Kostanjčar, Z., and Pale, P., "E-Mail System for Automatic Hoax-Recognition", in XXVII. International Convention MIPRO 2005. 2005: Croatia. p. 117-121.

[5] Brodie, R., "Virus of the Mind: The New Science of the Meme", Integral Press 1995, USA.

[6] Cormack, G.V., "Email Spam Filtering: A Systematic Review", *Foundations and Trends in Information Retrieval*, vol.1 (4): p. 335-455, 2008.

[7] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. and Stamatopoulos, P., "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach", in Proceedings of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France, pp. 1-13, 2000.

[8] Hsiao, W.-F. and T.-M. Chang, "An incremental cluster-based approach to spam filtering", *Expert Syst. Appl.*, vol. 34(3), p. 1599-1608, 2008.

[9] Abi-Haidar, A. and Rocha, L.M., "Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics: A Study of Concept Drift", in Proceedings of the 7th International Conference on Artificial Immune Systems, Springer-Verlag Berlin, Heidelberg 2008.

[10] Isa, D., Lee, L.H., Kallimani, V.P. and Rajkumar, R., "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine", *IEEE Transactions on Knowledge and Data Engineering*, vol. 20(9), p. 1264-1272, 2008.

[11] Viola, P. and Narasimhand, M., "Learning to Extract Information from Semi-Structured Text Using a Discriminative Context Free Grammar", in Proceedings of the 29^{8h} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.330-337, Univerrsity of Washington, Seattle, WA, USA

[12] Rozilawati, D. and Aono, M., "Ontology Based Approach for Classifying Biomedical Text Abstracts", *Internationa Journal of Data Engineering*, vol.2(1), 2011.

[13] M.Christensen, B., "Latest Email Hoaxes - Current Internet Scams - Hoax-Slayer", Available from http://www.hoax-slayer.com

[14] Kanaris, I., Kanaris, K., Houvardas, I and Stamatatos, E., "Word versus Character N-Grams for Anti-Spam Filtering". *International Journal on Artificial Intelligence Tools*, vol. 16(6), p. 1047-1067, 2007.

[15] Ifrim, G., Bakir, G. and Weikum, G., "Fast Logistics Regression for Text Categorization with Variable-Length N-Grams. Saarbrucken", in Proceedings of the 14th ACM SIGKDD International Conference on Knowldege Discovery and Data Miing, pp.354-362, 2008.

[16] Buckingham, J.T., Geoffrey, J.H., Goodman, J.T. and Rounthwaite, R.L., U.S. Patent No. 7899866. Washington D.C: U.S. Patent and Trademark Office.