ISSN:0975-9646

Automated Identification of Substituted Words in Text

Amit Kulkarni, Mayur Patil, Nitish Mali, Rajesh Kharche Prof. Mrs. G.S.Navale.

Abstract— **Text substitution is a common practice used by the criminals and terrorists for communication between them. Text substitution is helpful as it hides the actual meaning of the sentence. The substitutions are known only by their group members! In this paper we have considered detecting the substituted words with similar word frequencies to that of the original word. As these messages do not contain any sensitive words so they look normal. The algorithms such as kgram and oddity are individually not effective. So by using these algorithms together and that too in logical order (in sequence) which provides an effective detector.**

Index Terms—k-gram, oddity, word frequencies.

I. INTRODUCTION

There is a vast development in communication media, like internet especially in India, in last few years. This includes use of telephones, mobiles, E-mails. Groups (terrorists or criminals) that are involved in unethical acts communicate with one another via Internet. Contents of their communication provide evidence for thinking and actions for their targets. If the interception is being done by automated scanning of large numbers of messages, for example, by government Intelligence programs or organizational analysis of e-mail, obfuscation of content may be a better technique. The Terrorists are aware that their communication is likely to be scanned by system such as Echelon. One way to conceal content is encrypt the message but encryption draws attention to messages, making techniques such as traffic analysis easier to apply. Emails containing sensitive text can be separated by scanning every email for occurrence of sensitive words and then processing it using another level of data mining algorithms. However, illicit groups started substituting the sensitive word in the email by a normal word in order to hide the meaning of the sentence so that it can be interpreted as a normal mail. A terrorist or criminal group might adopt a standard set of substitutions, in which the words they do not wish to use are replaced by other words with similar frequencies. This system addresses the problem of detecting such substitutions of words with words of similar frequency.

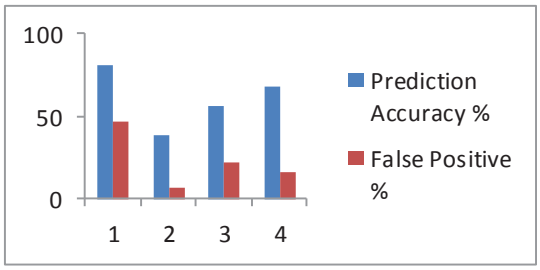
When a terrorists and criminals send messages they replace certain words by other words or locutions. In this paper we consider ways to detect replacements that have similar frequencies to that of original words. In this paper we have implemented cosine similarity measures for the effective results.

II. RELATED WORK

Substituted words in sentences are generally difficult to find. Generally terrorist use non-toxic words in place of sensitive words to conceal the meaning of actual communication there are three different ways to detect substituted words in text. But when these three methods are executed individually the accuracy is not satisfactory. Therefore when these three methods are implemented together, the output obtained is of high accuracy. SW.Fong and D.B. Skillicorn have proposed K-grams,Oddity, Hypernym oddity for detection of substituted words along with their accuracy rate and false rate [1].

In the graph 1 the combined predictor has prediction accuracy of 68% for sentences with substitution with a false positive rate of 16%.

Graph 1



In the graph 2 the combined predictor has prediction accuracy of 82% for sentences with substitution with a false positive rate of 30%.

Manuscript received 23 September 2014.

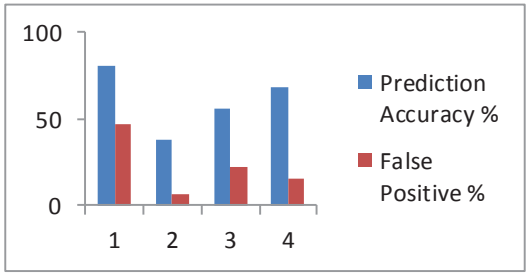
Amit Kulkarni is with Sinhgad Institute Of technology and science, Pune Maharashtra India (e-mail: amkul2011@gmail.com).

Mayur Patil is with Sinhgad Institute Of technology and science, Pune Maharashtra India (e-mail: patilmayur161@gmail.com).

Nitish Mali is with Sinhgad Institute Of technology and science, Pune Maharashtra India (e-mail: nitishmali@yahoo.com).

Rajesh Kharche is with Sinhgad Institute Of technology and science, Pune Maharashtra India (e-mail: kharcherajn@gmail.com).

Graph 2



III. COSINE SIMILARITY

Cosine similarity is used for getting or testing the similarity of two vectors of inner product space which measures cosine angle between them. Cosine similarity identifies how similar two documents are, in term of the matter. We are going to measure to find similarity between the original and replaced words [2].

Cosine of the two vectors given by:-

$$a.b = ||a|| ||b|| COS\theta \qquad [3]$$

$$\text{Similarity} = cos\theta = \frac{V1.V2}{||V1|| ||V2||}$$

Consider two vectors V1 and V2, the cosine similarity is given by:-

$$= \frac{\sum_{i=1}^{n} (V1_i * V2_i)}{\sqrt{\sum_{i=1}^{n} V1_i^2} * \sqrt{\sum_{i=1}^{n} V2_i^2}} \qquad [4]$$

The possibilities of the results are as mentioned in the table 1

Table 1

Similarity Value	Remark
-1	Exactly Opposite
1	Exactly Same
0	Independent/Between similarity and dissimilarity

The attribute vector V1 and V2 are the terms frequencies vector of documents. Since frequencies cannot be negative, the cosine similarity ranges from 0 to 1.

IV. EXPERIMENTS

Consider the example "The Attack will be tomorrow" as original sentence and "The complex will be tomorrow", The figures in the table indicates the google Hits received when words are provided in " " [5].

Table 2 shows some retrieved Frequencies Of bag of words,

Table 2

Original Sentence without stop words	Bag of words with only target words (original)	Frequency of original sentence(A)	Substituted sentence without stop word	Bag of words (B)	Frequency of original sentence(B)
Attack will tomorrow	Attack	68500000	Complex will tomorrow	Complex	76100000
	Attack will	510000		Complex will	4650000
	Attack will tomorrow	4		Complex will tomorrow	7

By using above results we get,

Cosine similarity = 0.99993 [7]

While finding the frequencies of the bag of the words we have removed the stop words. For example we have considered the sentence "the attack will be tomorrow", for it after removing the stop words it becomes "attack will tomorrow". In table 2 we have derived various word combinations. For every combination, a google hit count is received from googling them and the cosine similarity is calculated.

In above sentence "the attack will be tomorrow", "attack" word is replaced by "complex" and we have the cosine value as 0.99993. From these we can detect that the original word can be "attack".

V. CONCLUSION

The task of detecting substitutions is becoming important since terrorists and criminals or illicit groups are using substituted words in their document or mail by a normal word in order to hide the meaning of the sentence the use of certain words (e.g. 'bomb', 'gun', 'attack', etc.). Our technique proposes work to overcome from such a situation. This paper uses cosine similarity measure for the probable detection of substitution of text. Each of the measures performs poorly on its own, although we are exploring improvements to each, but together they begin to become practically effective.

REFERENCES

[1] Szewang Fong, Dmitri Roussinov, and David B. Skillicorn, "Detecting Word Substitutions In Text", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 20, No. 8, August 2008

[2] Sonal N. Deshmukh, Ratnadeep R. Deshmukh, Sachin N. Deshmukh, "Cosine Similarity for Substituted Text Detection" *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 4, Issue 1, January 2014.

[3] "The use of the internet for Terrorist purposes, Report of United Nations office on drugs and crime", Vienna

in collaboration with the United Nation‟s Counter Terrorism implementation task force 2004 published by united Nation Newyork Sep 2012.

[4] Mrs. Shilpa Mehta, Dr. U Eranna, Dr. K. Soundararajan, "Surveillance Issues for Security over Computer Communications and Legal Implications", *Proceedings of the World Congress on Engineering 2010 Vol I WCE 2010*, June 30 - July 2, 2010, London, U.K.

[5] S. Fong, D. Skillicorn, and D. Roussinov, "Measures to Detect Word Substitution in Intercepted Communication,"Proc. IEEE Int'l Conf. Intelligence and Security Informatics (ISI '06), May 2006.

[6] http://en.wikipedia.org/wiki/Cosine_similarity.

[7]www.appliedsoftwaredesign.com/archives/cosine-similarity-calculator/.