Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Cosine Similarity for Substituted Text Detection

Sonal N. Deshmukh
Department of MCA
J.N.E.C., Aurangabad
Maharashtra, India

Ratnadeep R. Deshmukh
Department of Computer Science,
Dr. BAM University, Aurangabad
Maharashtra, India

Sachin N. Deshmukh
Department of Computer Science,
Dr. BAM University, Aurangabad
Maharashtra, India

Abstract— Text substitution is generally used for communication of messages amongst a group of people who wants to share the secrete information. Unethical group of people are using text substitution for communicating the details of anti-social acts that they want to carry out. Text substitution helps by hiding the actual meaning of the text written. However, as their group members are well aware of the substitution, they can easily understand actual meaning of the text written. This information is generally shared with the help of emails or also published on the websites. However, as the content of emails or the web contents do not contain any sensitive words rather they look normal, existing algorithms cannot detect the sensitivity of it. This paper deals with detection of substituted text in a sentence using statistical techniques and cosine similarity.

Keywords— cosine similarity, correlation, standard deviation, page count, Enron dataset

I. INTRODUCTION

Communication technology and internet has been proved to be the blessing to human being when used with good intension. However, it may be a curse when the intension of user is destructive. A common example is the use of high end technologies by the terrorist groups for planning as well as execution of their tasks and attacks. They use internet to achieve their goals by various means [1]. In order to hide the information they wants to share with each other, initially they started using encryption and decryption techniques. But as the information encrypted can be detected with the help of available algorithms, a possibility of getting it decrypted by policing agencies is more. Hence they searched for other ways to achieve secrete exchange of information. One such example is use of text substitution for sharing information e.g. plan of attack at certain place or procedure of preparation of bomb, via email or by publishing it on website.

In text substitution, instead of using harmful words like bomb, attack etc., and these words are replaced by innocuous words so that it looks like normal. For e.g. the sentence "there will be a blast at 11a.m. today", after substitution becomes "there will be a show at 11a.m. today". Here 'blast', a sensitive word is replaced by 'show', a normal word, so that it hides meaning of information. There are various measures used by the researchers to detect such type of substitutions.

In this paper we tried to find out the similarity between two words by using cosine similarity measure.

II. RELATED WORK

Illicit groups have adopted multiple ways for substitution. One of the ways is use of synonyms or hypernyms of the sensitive word for substitution. S. Mehta *et al.* [2] presented these hypernyms as an extended measure. But one can recognize such word substitutions since semantic analysis can detect such substitutions easily. Turney *et al.* [3] has presented an algorithm for mining the web for synonyms however this algorithm is not useful for detection of substitution as substitution do not follow any specific rule in general. Word frequency information is readily available on www.wordcount.com, so it is possible that, in ordinary circumstances, 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. Email datasets like Enron email dataset and 20 News group dataset are the best examples of natural human writing and hence are commonly used in text mining application as training and testing data set. However, it is essential to clean the emails before using it in the application [4]. If a list of sensitive words is maintained at Gateway, emails containing harmful words can be traced easily at gateway. But use if substitution in text converts the sensitive sentence into normal sentence e.g. the sentence "there will be bomb blast in a mall today stop it if you can" can be written as "there will be a gift distribution in a mall today stop it if you can". As this sentence does not have any sensitive word, it cannot be categorized as sensitive sentence by detection algorithm. Earlier terrorist group used to adopt some common techniques for substitution, example of which can be substitution by replacing the word with the word having near wordcount rank e.g. the word "missile" have word rank 6316 and "garment" have word rank 6319 [6][8]. Decision tree algorithm was used to categorize such type of suspicious emails containing deceptive text [5][7] where author focused on detecting criminal identity of deception in law enforcement. Aim of this paper to present novel technique for detection of text substitution using cosine similarity.

III. COSINE SIMILARITY

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [9]. In Information Retrieval, each term is notionally assigned a different dimension and a document

is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. We can use cosine function to find the similarity between target word and replaced words.
The cosine of two vectors can be derived by using the Euclidean dot product formula:

$$a.b = \|a\|\|b\|\cos\theta \qquad [1]$$

Given two vectors of attributes, *A* and *B*, the cosine similarity, *cos(θ)*, is represented using a dot product and magnitude as

$$simiarity = \cos\theta = \frac{A.B}{\|A\|\|B\|}$$

$$= \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^{\,2}} * \sqrt{\sum_{i=1}^{n} B_i^{\,2}}} \qquad [2]$$

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.
For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.
In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

TABLE I.
COSINE OF A SENTENCE FOR BAG OF WORDS

Original Sentence: "you will get bomb at Delhi"			**Sentence after substitution: "you will get chocolate at Delhi"**			**Cosine**
Original Sentence without Stopwards	Bag of Words(original)	Freq of Original Sentence(A)	Substituted Sentence without Stopwards	Bag of Words (B)	Freq of Substituted Sentence(B)	
get bomb Delhi	get	8860000000	get chocolate Delhi	get	8860000000	.9991978
	bomb	400000000		chocolate	756000000	
	Delhi	446000000		Delhi	446000000	
	get bomb	538000000		get chocolate	574000000	
	get Delhi	311000000		get Delhi	311000000	
	bomb Delhi	19300000		chocolate Delhi	15500000	
	get bomb Delhi	19300000		get chocolate Delhi	9290000	

IV. EXPERIMENT

We performed experimentations on the training dataset for testing the use of Cosine Similarity for detecting text substitution. We selected emails from Enron Dataset having content size less than 18 words. The count of 18 is considered to get handful amount of emails for testing. Then stop words were removed from the content of the emails. In order to test the substitution, we substituted some specific words in the email with another words. In the first experimentation, we used the emails where only one word is substituted. For example consider the sentence "you will get bomb at Delhi". After removing the stop words, we get the bag of words as "get bomb Delhi". Various combinations of the words in the bag of words were generated Table 1. For every combination, a Google hit count is retrieved using Google API. Similarly, Google hit count for the substituted sentence was also retrieved. Based on the values for combinations of both the sentence, Cosine similarity was calculated.

In this experiment, we also used the dataset where we searched for the n grams containing target words only. N grams without target words are excluded from the calculation. Following table 2 shows the result of such combinations.

TABLE II
COSINE OF A SENTENCE FOR BAG OF WORDS WITH TARGET WORD ONLY

Original Sentence without Stopwards	Bag of Words with only Target Words(original)	Freq of Original Sentence(A)	Substituted Sentence without Stopwards	Bag of Words (B)	Freq of Substituted Sentence(B)	Cosine
get bomb Delhi	bomb	400000000	get chocolate Delhi	Chocolate	756000000	.96026
	get bomb	538000000		get chocolate	574000000	
	bomb Delhi	19300000		chocolate Delhi	15500000	
	get bomb Delhi	19300000		get chocolate Delhi	9290000	

Cosine values thus calculated are all around one. In practical, we need to match a normal word with a list of sensitive words. We tried that approach in second experiment. For example, for the statement "we expect that marriage will happen tonight", the normal word "marriage" is checked with all the sensitive words and accordingly cosine value is calculated. Another variation "we expect that flower will happen tonight". The word marriage can be a suitable word for substitution than flower. This can be easily verified from the cosine values obtained and shown below. Following Fig 1.shows the result for four different statements and the cosine values.

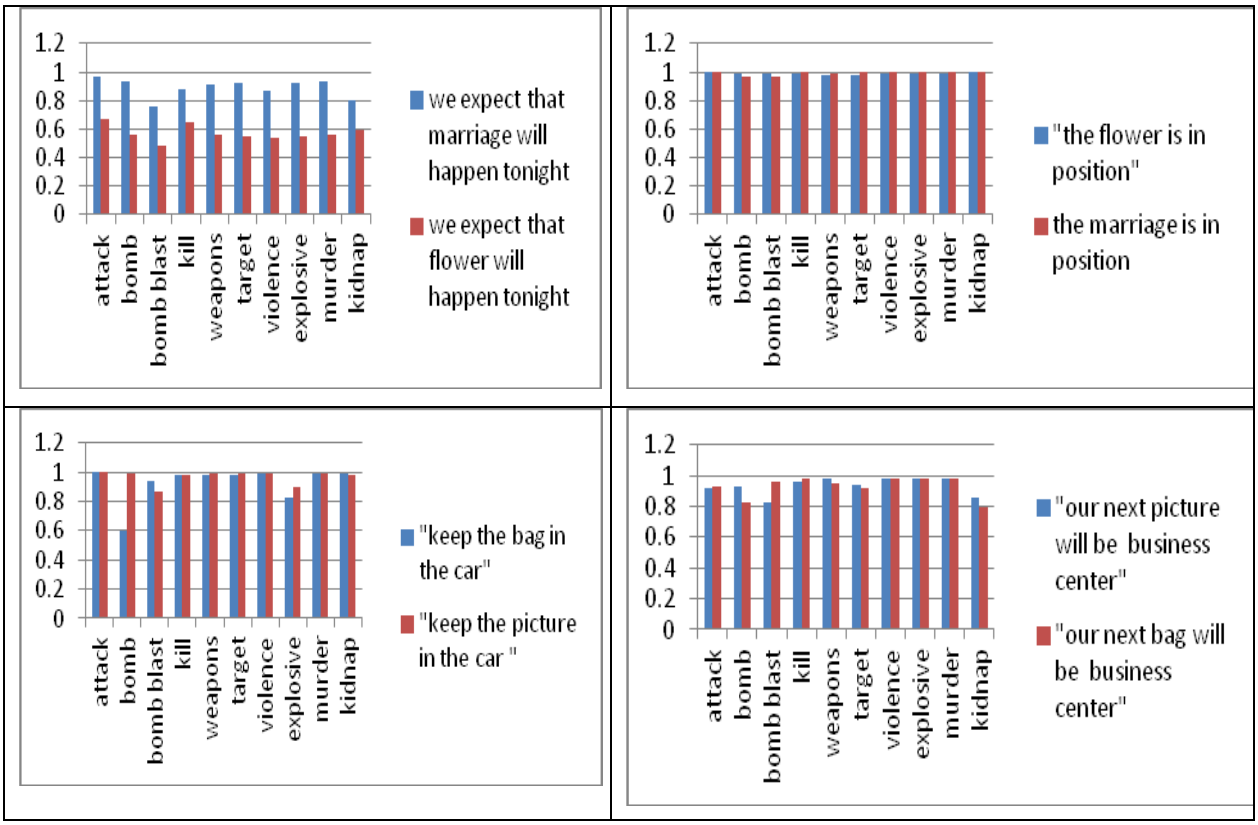

Fig. 1 Cosine values for four different statements

In a sentence "We expect that marriage will happen tonight" considered by [8], original word is 'attack' which is replaced by 'marriage'. When compared with all sensitive words, a sentence having word 'attack' replaced by 'marriage' had the highest cosine value 0.962935. From this we can infer that the original word can be attack. When we used another substituted word like 'flower' instead of 'marriage', cosine values for set of sensitive words which were not near to 1 but had highest cosine value for attack i.e. 0.666638, which is again highest as compared with other target words.

We consider another sentence like "Our next picture will be business centre", here word 'picture' is used instead of word 'target'. Cosine value for this sentence is 0.94479312 which is also near to one. If we replace word 'picture' by a word 'bag' in the substituted sentence, cosine value is more for picture than bag. Very few sentences didn't match the result in this experiment. Following table shows data used for the experiment and fig.2 shows graph of cosine values for a set of sensitive words. When tested for a set of 250 different sentences taken from Enron dataset, we got an accuracy of around 0.78.

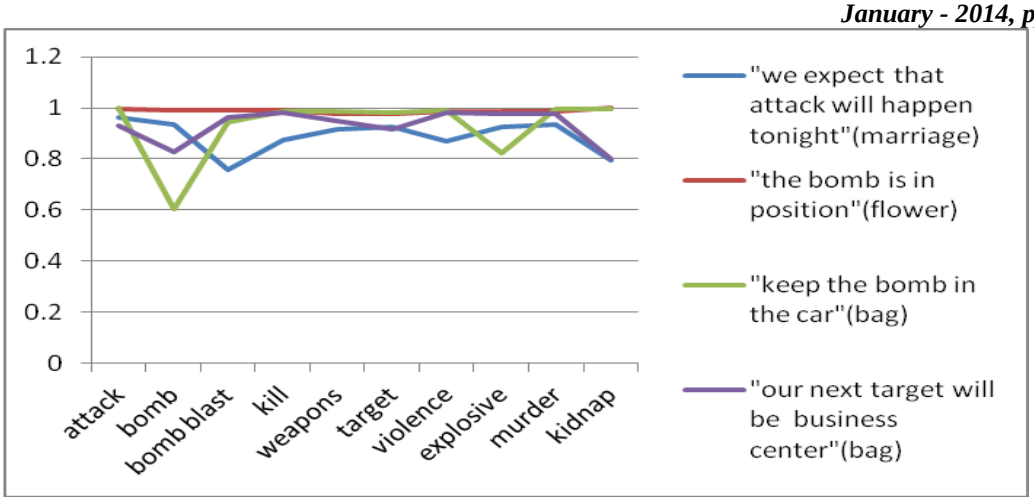

Fig. 2 Cosine for list of sensitive words

TABLE III
COMPARISON OF MEAN, STANDARD DEVIATION AND CORRELATION

Sentences	**Sentences without stopwards**	**Mean (original)**	**Mean (substitution)**	**Std Dev (original)**	**Std Dev (substitution)**	**Correlation**
You will get bomb at Delhi (Chocolate)	get bomb Delhi	244150000	338697500	265676552	384046250.7	0.91774554
We have to do murder in Mumbai(Felicitation)	do murder Mumbai	245350000	2445500	260765712	2675027.913	0.49597115
Ramesh will come to collect explosive material(cotton)	come collect explosive material	23832222.2	81203333	35474393.5	154266045.4	0.96874393
We expect that attack will happen tonight(Marriage)	expect attack happen tonight	198357143	113900000	112941340	77457558.27	0.86267527
Give training to attack on the city(Rain)	give training attack city	389600000	190580000	245229077	253373078	0.87944198
The bomb is in the position(Flower)	bomb position	263500000	575500000	191625938	501338707.9	1
Burn the train tomorrow(Colour)	Burn train tomorrow	88233333.3	99566667	22744303.3	58336638	0.78899684
Spread violence as soon as possible(Happiness)	Spread violence soon possible	143814286	71528571	69559864.6	75386995.22	0.90348047
Our next target will be business centre(Picture)	next target business centre	670000000	2.651E+09	681643846	1533225669	0.93611812
Keep the bomb in the car(Bag)	Keep bomb car	681666667	655000000	683587839	232213264	-0.7945758

In third experiment, we tried to find out relationship between original and substituted word in the sentences using correlation and standard deviation. Standard deviation is used to get dispersion of values. Values of standard deviation and correlation are shown in the Table 3. In this table, the sentence "The bomb is in the position" has correlation 1, it means that the relation between original word 'bomb' and replaced word 'flower' is very strong. However, in order to detect the substitution of the text, this measure is not helping much. This is because the randomness in the values of variations that we have tested.

V. CONCLUSION

Problem of text substitution has become a point of concern for Governments as well as researchers. Multi fold research is started recently and various ways are being used to detect it. In this paper, we discussed use of Cosine similarity as a measure for detection of text substitution. Accuracy that we achieved here is 0.78. However, there is a need to enhance the accuracy of detection. Use of Ontology may enhance the accuracy, provided the domain of communication is well known. As a future scope, we are considering domain specific hit counts with pre-processing of inputs to the similarity measure.

REFERENCES

[1] *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.

[2] 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.

[3] Peter D. Turney, *Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL*, Proceeding of the 12th Europium Conference on Machine Learning ,pages 491-502,Springer-Verlag, UK,2001

[4] J.Tang, H.Li, Y. Cao, Z.Tang, *Email Data Cleaning*, Proceedings of KDD, Chicago, USA, (2005).

[5] Appavu alias Balamurugan and Ramasamy , *Suspicious E-mail Detection via Decision Tree: A Data Mining Approach*, Rajaram Thiagarajar College of Engineering, Madurai, India

[6] www.wordcount.org/main.php

[7] Gang Wang, Hsinchun Chen and Homa Atabakhsh, *Criminal Identity Deception and Deception Detection in Law Enforcement*, Group Decision and Negotiation, Mar 2004, Vol. 13 Issue 2, p111-127. 17p. Department of Management Information Systems, University of Arizona, 430 McClelland Hall, Tucson, AZ

[8] 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

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

