International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 3, Issue 3(March 2014), PP.20-25

Performance Analysis of Different Sentence Oddity Measures Applied on Google and Google News Repository for Detection of Substitution

Sonal N. Deshmukh[1], Ratnadeep R. Deshmukh[2], Sachin N. Deshmukh[3]
[1](*Department of MCA, JNEC, Aurangabad, Maharashtra, India*)
[2](*Department of CS and IT, Dr. BAM University, Aurangabad, Maharashtra, India*)
[3](*Department of CS and IT, Dr. BAM University, Aurangabad, Maharashtra, India*)

Abstract:- Substituted text detection is now great challenge for the antiterrorism agencies. It plays important role in terrorist activities. Now-a-days criminals are using internet through various devices for communications. As they wanted to hide their information from others, they started to use some code so that other person cannot understand the meaning of their messages or documents. Since it is not very easy to find out the code they are using, we can find out the probabilities for hidden data. Criminal replaces harmful words by some innocuous words so that it looks normal to others. In this paper we applied some measures on different types of documents to detect such word substitution. Sentence oddity, Enhance sentence oddity and k grams are used in this research paper. We applied measures on two types of data, General data and Google news data and compared the performance of these measures. Substituted and original sentences are used to classify the data for substitution.

Keywords:- Enhance Sentence Oddity, K gram, Page Count, Random Forest, Sentence Oddity

I. INTRODUCTION

Internet is today‟s need for easy and fast communication. Since late 1980s, it has proven to be a highly dynamic means of communication, reaching an ever-growing audience worldwide. Though it can be used in many applications which reduce work and time of human being, it can also used to do the illegal things by some criminals. It includes sending text messages via email or SMS to the group members either using fake identification or by hacking/stealing the device or network link. Such mail can be separated by scanning every message for the occurrence of sensitive words and then processing it using another level of data mining algorithms.

Internet can also used by the terrorist by various means. Deceptive writing may be a problem related to crime. It is really a challenge to tackle such problems since authorship of document is hidden. [1] *et al* used information gain ratio to detect such problems. But many times criminals use fake accounts to hide their identification and communicate with their group members. One of the primary uses of the Internet by terrorists is for the dissemination of propaganda. Propaganda generally takes the form of multimedia communications providing ideological or practical instruction, explanations, justifications or promotion of terrorist activities. These may include virtual messages, presentations, magazines and treatises, audio and video files developed by terrorist organizations or sympathizers.

Initially terrorist groups like Al Qaida were also using encryption in their communications. They developed their own software like “Asrar el Mojahedeen” or “Mujahedeen Secrets” for encrypt the data [2]. But the problem with data encryption is it draws attention to user. So they started to use some special code in their communications and substitute some harmful words like attack, bomb etc by normal words so that it cannot easily recognize by the others.

Apart from the messages, the terrorist groups are using sites to publish objectionable material like method to prepare Bomb etc. However, the data uploaded on the website is obfuscated such that it looks normal to the users.

Substitutions can also do by the people interested in bribes where they have to communicate at public places. Human being may detect such substitution with the help of contextual information and general sense. However, automatic detection of such obfuscated messages is quite difficult. At the same time, it is not possible to manually scan every message. This paper proposes classification of such messages depends on replaced and non replaced words. Sentence Oddity, Enhance Sentence Oddity and K gram are used for the classification of the data. We compared data from Google news and general data considering the performance of the measures.

II. BACKGROUND

Malicious email detection problem was discussed by Peng Hong *et al* [3]. The email filter can automatically filter the email and host receives when an email server is operating. But in case of substitution of text this email filter could not filter out the data from the document since all words in the documents looks normal. In substitution of text harmful words which can come under the process of filtering is replaced by innocuous words. Recently used word substitution in the sentence by terrorist group Indian Mujahideen(IM) is „H‟ instead of „Hydrabad‟. For e.g. a sentence “work needs to be done in H” instead of “work needs to be done in Hydrabad” was used. Problem of detection of substitution of the word is first discussed by SzeWang Fong *et al* [4]. They used Enron corpus and Brown corpus and applied different measures on it. In their experiment individual measure performed poor so they combined measures and got performance for both corpuses [5]. Word obfuscation detection is one of the many natural language processing tasks that can benefit from characterizing the contexts a word or a phrase typically used in. Sanaz Jabbari *et al* presented a probabilistic model which applied for problem of textual defuscation [6]. They developed this model to check whether certain words are used in or out of context. Some extended measures are discussed by Mrs. Shilpa Mehta to highlight the issues of security over computer communications and legal implications [7]. They presented technical issues and limitations of earlier surveillance techniques. Turney *et al.* 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 [8]. 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. In this research we used some previously suggested measures and applied different datasets for it. Google News data and general dataset were considered to get the page count and used for measures. Comparative study of both data based on performance was done in this experiment.

III. MEASURES USED

3.1 Sentence Oddity:

This measure considers a sentence as a whole and the relationship between the entire sentence and sentence with particular word of interest deleted. SO is based on the observation that if we remove contextually appropriate word from the sentence then it should not change the frequencies of resulting bag of words in comparison with frequency of entire sentence because it co-occurs frequently. But if remove contextually inappropriate word from the sentence it may produce large frequency of remaining bag of words because it co-occurs rarely. SO is given by

$$SO = \frac{Frequency\ of\ bag\ of\ words, target\ word\ removed}{Frequency\ of\ entire\ bag\ of\ words} \quad \textbf{(1)}$$

Here SO is sentence oddity.

3.2 Enhance Sentence Oddity:

The numerator in the sentence oddity measure includes some sentences that contain the word being considered; that is the numerator counts some sentences that are also counted in the denominator. It is useful to define enhanced sentence oddity in which the numerator explicitly excludes the word being considered. Hence we define the enhanced sentence oddity of a sentence with respect to a particular target word as:

$$ESO = \frac{Frequency\ of\ bag\ of\ words\ with\ target\ word\ excluded}{Frequency\ of\ entire\ bag\ of\ words} \qquad \textbf{(2)}$$

ESO is Enhance Sentence Oddity.

3.3 K gram Frequencies:

An *n*-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n-1) order Markov model. N-gram models are now widely used in probability, communication theory and computational linguistics. N-grams can also used for efficient approximate matching. By converting a sequence of items to a set of *n*-grams, it can be embedded in a vector space, thus allowing the sequence to be compared to other sequences in an efficient manner. N-gram-based searching can also be used for plagiarism detection. We can also compute n gram statistics in distributed file processing [9]. We can consider 1 gram, 2 gram, 3 gram string and so on. But it has been observed that more than 3 gram or 4 gram string does not occur on search engine with some frequency [10]. However, as calculation of n-gram may increase the time complexity, a more general form of n-gram, k-gram is proposed to be used [11]. The k-gram of a word is the string containing that word and its context up to and including the first non-stopword to its left, and the first non-stopword to its right. Left part is called left K gram and right part is called right K gram. For e.g. in a sentence "Life in metro cities is always busy", if we consider a word metro then left K gram of this sentence is „life in‟ and right K gram is „cities is always busy‟. Left and right K gram can be helpful for calculating various measures. While calculating K gram for detecting substituted word, we can consider left and right K gram of the target word.

IV. EXPERIMENTS

In the experimentation, we used general dataset and Google News dataset and calculated oddity of the sentences. Here dataset comprise of pair of original and substituted sentences. Page count of Google search engine is used to calculate the values of the measures. We selected news having text size less than or equal to 10 words. Specific words were substituted for testing the data. We classified the data in each dataset for probability of replacing and non replacing words in the sentences. In first experiment we calculated Sentence Oddity of general and Google News sentences. We searched all sentences along with individual words of the sentences in the searched engine since almost all sentences were giving page count zero. So we calculated Sentence Oddity by using Google search engine. In a sentence "The bomb is in position", frequency of bag of words was 175000000 and frequency of bag of words without target word bomb was 1240000000. Hence we got SO of this sentence as 7.085. If we assume substituted word as „flower‟ instead of „bomb‟, then substituted sentence is "The flower is in position". Here frequency of bag of words was 227000000 and SO for substituted sentence was 5.463. Table 1, Table 2 and Table 3 shows performance of SO for General Google and Google News for J48, J48Graft and Random Forest algorithms. Here detection rate, false positive rate and area under ROC curve have three values showing performance for cross validation, training set and percentage split respectively.

Table 1. Sentence oddity for J48

Sentence Oddity(Weighted Avg)J48			
Corpus	Detection Rate	False Positive Rate	Area under ROC Curve
General Google	0.5, 0.5, 0.429	0.5, 0.5, 0.429	0.5, 0.5, 0.5
Google News (News text only)	0.909, 0.955, 0.857	0.091, 0.045, 0.107	0.872, 0.955, 0.875

Table 2. Sentence Oddity for J48 Graft

Sentence Oddity(Weighted Avg)J48 Graft			
Corpus	Detection Rate	False Positive Rate	Area under ROC Curve
General Google	0.5, 0.5, 0.429	0.5, 0.5, 0.429	0.5, 0.5, 0.5
Google News (News text only)	0.909, 0.955, 0.857	0.091, 0.045, 0.107	0.872, 0.955, 0.875

Table 3. Sentence oddity for Random Forest

Sentence Oddity(Weighted Avg)Random Forest			
Corpus	Detection Rate	False Positive Rate	Area under ROC Curve
General Google	0.5, 1, 0.429	0.5, 0, 0.429	0.5, 0.5, 0.5
Google News (News text only)	0.909, 1, 0.429	0.091, 0, 0.429	0.938, 1, 1

Performance of Sentence Oddity for Random Forest is giving better result than J48 and J48Graft algorithm. Enhance Sentence Oddity is also calculated which is giving almost same result as SO. Another dataset is used by considering the some latest substitutions used by terrorist group Indian Mujahideen. For e.g.

IM was using „H" instead of „Hydrabad" so for a sentence „work needs to be done in Hydrabad" they were using „work needs to be done in H". Apart from this many other substitutions were used by this group. Some sentences with SO and ESO are shown in Table 4.

Table 4. SO and ESO for Substituted Sentences used by IM

Sentence	**SO**	**ESO**
works need to be done in Hydrabad	0.19349	10900000
works need to be done in H	0.02333	13.02
you should arrange for a preparation of blast	0.92149	34800
you should arrange for a daawati	44472.04	34800
my friend will come to deliver you a pistol	6.20879	4233.33
my friend will come to deliver you a CD	0.86923	34.6994
collect some people for work from Gujarat	9.23456	4633.33
collect some people for work from Musa	6.93877	3475
you will find some bullets in the bag	83.7662	5.3383
you will find some pen drives in the bag	58.6363	4733.33

come at Delhi for meeting	6.10559	37.470
come at Sham for meeting	72.2794	193.93
send one person to Bangalore	212.546	146.536
send one person to Bagu	95.3642	4140.425
Arrange some riffles for next operation	0.73711	3477777.77
Arrange some DVDs for next operation	0.96621	2.6982
preparation of blast will start in next month	16.9376	11.4606
Daawati work will start in next month	661375.66	3686746.98
find one place at Hydrabad for operation	13.6521	2770.37
find one place at H for operation	0.86980	42.9885

Performance of SO and ESO for dataset where substitutions are done by IM for cross validation and training set is shown in Table 5 and Table 6 respectively. In this dataset, SO with random forest is giving detection rate 1and false positive rate 0 for both general Google and Google news search for this dataset. Also with ESO, random forest is giving detection rate 0.75 and false positive rate 0.25 for both searches. We also calculated K gram for this dataset. We divide each sentence into left and right K gram according to target word. Performance of left and right k gram for both searches are almost same. Performance of left k gram for J48 and Random Forest algorithms is shown in Table 7.

Table 5. Sentence Oddity

Sentence Oddity(Weighted Avg)random forest			
Corpus	Detection Rate	False Positive Rate	Area under ROC Curve
General Google	0.5, 1	0.5, 0	0.5, 1
Google News	0.5, 1	0.5, 0	0.5, 1

Table 7. Left K gram for Random Forest

Left k gram(Weighted Avg)J48, random forest			
Corpus	Detection Rate	False Positive Rate	Area under ROC Curve
General Google	0.5, 1	0.5, 0	0.5, 1
Google News	0.5, 1	0.5, 0	0.5, 1

Table 6. Enhance Sentence Oddity for random Forest

Enhance Sentence Oddity(Weighted Avg)random forest			
Corpus	Detection Rate	False Positive Rate	Area under ROC Curve
General Google	0.6, 0.75	0.4, 0.25	0.595, 0.393
Google News	0.6, 0.75	0.4, 0.25	0.595, 0.393

V. CONCLUSION

Comparing the performance of datasets by using various algorithms we can conclude about the possibility of substitutions of words in the sentences. When we tested dataset for general sentences and Google news sentences in Google search engine, it is showing that performance for Google news search gives better result than general search. Also random forest algorithm works well for data comparing with other algorithms for both SO and ESO. In case of dataset currently revealed by Security Agencies in India and used by IM, detection rate of substitution is very low in case of cross validation for SO, ESO and K grams. This is because less number of news available in this regards.

REFERENCES

[1]. Afroz S. , Brenan M. , Greenstadt R. , Detecting Hoaxes, Frauds, and Deception in Writing Style Online, Security and Privacy(SP), *Symposium on 20-23 2013 San Francisco,* CA, ISSN :1081-6011 pp 461-475.

[2]. 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.

[3]. Peng Hong, Wang Jun, Teifeng Wu, Dongna Zhang, Malicious Email Detection Method Based on Support Vector Machine, *8th International Conference on Control, Automation, Robotics and Vision Kunming,* China, 6-9 Dec 2004.

[4]. SzeWang Fong, Dmitri Roussinov and David B Skillicon, Detecting Word Substitution in Text, *IEEE transaction on Knowledge and Data Engineering,* Vol 20, No. 8, August 2008, pp. 1067-1076.

[5]. Dmitri Roussinov, Szewang Fong, David Skillcorn: Measures to Detect Word Substitution in Intercepted Communication: Proceeding of Intelligence and Security Informatics, *IEEE International Conference on Intelligence and Security Informatics*, ISI (2006)

[6]. Sanaz Jabbari, Ben Allison, Louise Guthrie, Using A Probabilistic Model Of Context To Detect Word Obfuscation, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* 28-30 May 2008, Morocco.

[7]. 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.

[8]. Peter D. Turney, Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, *01 Proceeding of the 12th Europium Conference on Machine Learning* ,pages 491-502,Springer-Verlag, UK,2001

[9]. Klaus Berberich, Srikanta Bedathur, Computing n-Gram Statistics in MapReduce, *EDBT 13 Proceeding of 16th International Conference on Extending Database Technology*, Pages 101-112 ACM, New York, NY, USA, 2013.

[10]. Xiaojin Zhu and Ronald Rosenfeld, Improving Trigram Language Modeling with the World Wide Web, School of Computer Science, Carnegie Mellon University,5000 Forbes Avenue, USA.

[11]. SW. Fong, D.B. Skillicorn and D. Roussinov, Detecting Word Substitution in Adversarial Communication, 6th SIAM International conference on data mining (2006), Bethesda, Maryland.