ISSN 2395-1621

Deducing and Conclusive Analysis of Declassified Obfuscates chat in Terrorism using Word Substitution

[#1]Vaibhav Shinde, [#2]Komal Salunke, [#3]Punam Teli

[1]vaibhavshinde.1066@gmail.com
[2]komalsalunke167@gmail.com
[3]telipunam96@gmail.com

[#123]Computer Engineering Department

JSPM's Rajashri Shahu College of Engineering, Pune, India

ABSTRACT

Online Chatting Web Applications make use of substituted words with another word in sentences. Strategies behind using substituted words by incognito person or terrorist is to convey their messages without getting warned by security and intelligence. Agencies do interpreting of scanning messages, emails and telephone conversations. System is being designed in such a way that there will be admin and two user roles, Admin is able to view the live chat and monitor the same. Terrorist is able to do the chat with any person without monitoring it. While chatting if there is any occurrence of Substituted words then the algorithm is being able to detect the same by converting the substituted words with a normal one. Hadoop database is being used for analysis of sentences, where all the chat will be stored in a text file. The text file will be stored into hadoop as a storage purpose further with the help of Map Reduce Analytics percentage of the substituted words, which is being generated with the output of key value pairs.

Keywords: Counterterrorism, Data Mining, Document Analysis, Map Reduce, Sentence Substitution, Natural Language Processing.

ARTICLE INFO

Article History

Received: 11th March 2017

Received in revised form :

11th March 2017

Accepted: 13th March 2017

Published online :

19th March 2017

I. INTRODUCTION

System is being designed in a way that there will be admin and user will be two user roles, Admin is able to view the live chat and monitor the same. User is able to do the chat with any person without monitoring it. While chatting if there is any occurrence of substituted word then the algorithm is being able to detect the same and is able convert the substituted word with a normal one. Hadoop database is being used for analysis of the same where all the chat will be stored in a text file, the same text file will be stored into hadoop as a storage purpose further with the help of Map Reduce Analytics percentage of the substituted words is being generated with the output of key value pairs. Word substitution is used in communication by terrorist or criminals for conveying their messages without getting warned by security and intelligence. Agencies are interpreting in scanning messages, emails and telephone conversations. In the case of communications analysis, this might involve manually searching for a scintilla of intelligence amongst vast amounts of data. The pervasiveness of electronic modes of communication through web-logs (blogs), chat rooms and e-mail, has inevitably resulted in the exchange of information and messages between criminals and terrorists through these means. This has recently led to the UK Government to put forward plans to track all communications traffic in the spirit of law-enforcement. If this is realized, the amount of data that needs to be analysed will continue to increase.

II. LITERATURE SURVEY

The increase in the amount of data that intelligence and security analysts are required to review has led to a great deal of interest in information management and exploitation solutions to help tackle the information deluge problem. The proposed research, therefore, is concerned with delivering a practical solution to identifying such constructions automatically in order to support the analyst in their role. By developing Natural Language systems which can expose phrases or sentences in which a substitution has occurred, the amount of text that an analyst has to review can be dramatically reduced. The final algorithms must be both accurate in detecting substituted utterances, whilst not

suffering from false positives which could bloat the results returned, nor false negatives in which vital intelligence can be missed.

There is a vast development in communication media, especially in India, in last fifteen years. This includes use of telephones, mobile phones, internet, email etc. This facility is proved beneficial for the illicit acts in terrorisms and crimes too. 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.

Apart from email communication, user groups are using websites to publish objectionable material for example, publishing detailed procedure to manufacture bomb. However, in order to hide the actual meaning of the published material, the data uploaded on the website is substituted such that it looks normal to the users. As substituted words are selected without logic in word selection and they are selected such that the substituted message looks like normal. Another strategy to conceal the content of messages is to replace significant words with other words or locations that are judged less likely to attract attention. This work considered, not individual sentences, but large collections of messages. Speech recognition differs from the problem addressed here because it is limited to the left context, whereas we are able to access both left and right contextual information. Detecting misspellings uses an alteration model that incorporates common keystroke errors, themselves derived from visual, aural, and grammatical error patterns. Spam detection is closer to our problem in the sense that the alteration model assumes human-directed transformations with the intent to evade detection by software.

2.1 MOTIVATION:

In our system admin and user roles, Admin is able to view the live chat and monitor the same. User is able to do the chat with any person without monitoring it. While chatting if there is any occurrence of obfuscated word then the algorithm is being able to detect the same and is able convert the obfuscated word with a normal one. Hadoop database is being used for analysis of the same where all the chat will be stored in a text file, the same text file will be stored into hadoop as a storage purpose.

2.2 OBJECTIVE:

In our system word obfuscation means replacing one word with another word in a sentence to conceal the textual content or Communication. Word Obfuscation is used in adversarial communication by user or criminals for communication their messages without getting red-flagged by security and intelligence.

- Admin is able to login with his credentials.
- Admin is able to view the chat and monitor.
- Admin is also able to analyses the chat using Hadoop framework, Along with the storage of the big data.
- Admin is also able to analyses the chat using hadoop framework, Along with the storage of the big data.

III. MATERIAL AND METHOD

- Software Requirement
  1. JAVA 1.7
  2. TOMCAT 7
  3. MySQL 5.3
  4. Hadoop

- Hardware Requirement
  1. 8GB RAM
  2. 500GB HDD

IV. SYSTEM MODEL

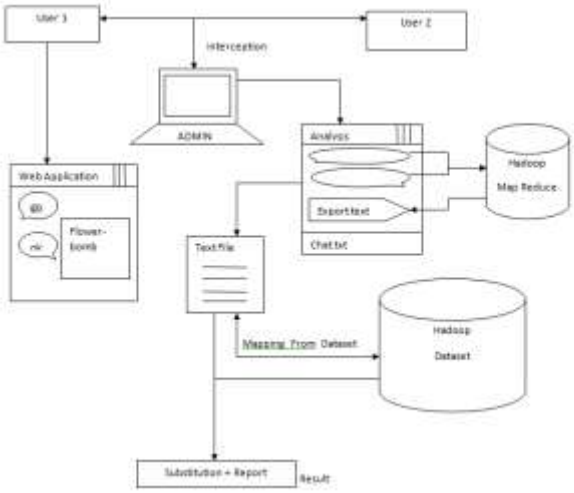

Fig 1. System architecture

V. ALGORITHM

1. Map-reduce:

A large part of the power of Map Reduce comes from its simplicity: in addition to preparing the input data, the programmer needs only to implement the mapper, the reducer, and optionally, the combiner and the partitioned. All other aspects of execution are handled transparently by the execution framework— on clusters ranging from a single node to a few thousand nodes, over datasets ranging from gigabytes to peta bytes. However, this also means that any conceivable algorithm that a programmer wishes to develop must be expressed in terms of a small number of rigidly defined components that must fit together in very specific ways. It may not appear obvious how a multitude of algorithms can be recast into this programming model. The purpose of this chapter is to provide, primarily through examples, a guide to MapReduce algorithm design. These examples illustrate what can be thought of as "design patterns" for MapReduce, which instantiate arrangements of

Components and specific techniques designed to handle frequently-encountered situations across a variety of problem domains. Within a single MapReduce job, there is only one opportunity for cluster-wide synchronization during the shuffle and sort stage where intermediate key-value pairs are copied from the mappers to the reducers and grouped by key. Beyond that, mappers and reducers run in

isolation without any mechanisms for direct communication. Furthermore, the programme has little control over many aspects of execution.

2. Hadoop:

A large part of the power of MapReduce comes from its simplicity: in addition to preparing the input data, the programmer needs only to implement the mapper, the reducer, and optionally, the combiner and the partitioned. All other aspects of execution are handled transparently by the execution framework— on clusters ranging from a single node to a few thousand nodes, over datasets ranging from gigabytes to petabytes. However, this also means that any conceivable algorithm that a programmer wishes to develop must be expressed in terms of a small number of rigidly-defined components that must fit together in very specific ways. It may not appear obvious how a multitude of algorithms can be recast into this programming model. The purpose of this chapter is to provide, primarily through examples, a guide to MapReduce algorithm design. These examples illustrate what can be thought of as "design patterns" for MapReduce, which instantiate arrangements of components and specific techniques designed to handle frequently-encountered situations across a variety of problem domains.Within a single MapReduce job, there is only one opportunity for cluster-wide synchronization— during the shuffle and sort stage where intermediate key-value pairs are copied from the mappers to the reducers and grouped by key. Beyond that, mappers and reducers run in isolation without any mechanisms for direct communication. Furthermore, the programmer has little control over many aspects of execution.

3. Obfuscated word:

Obfuscation is the obscuring of intended meaning in communication, making the message confusing, will fully ambiguous, or harder to understand. It may be intentional or unintentional (although the former is usually connoted) and may result from circumlocution (yielding wordiness) or from use of jargon or even argot(yielding economy of words but excluding outsiders from the communicative value). Unintended obfuscation in expository writing is usually a natural trait of early drafts in the writing process, when the composition is not yet advanced, and it can be improved with critical thinking and revising, either by the writer or by another person with sufficient reading comprehension and editing skills. Obfuscation may be used for many purposes. Doctors have been accused of using jargon to conceal unpleasant facts from a patient; American author Michael Crichton claimed that medical writing is a "highly skilled, calculated attempt to confuse the reader". B. F. Skinner, noted psychologist, commented on medical notation as a form of multiple audience control, which allows the doctor to communicate to the pharmacist things which might be opposed by the patient if they could understand it.

VI. MATHEMATICAL MODEL

System Specification:

S= {S, s, X, Y, T, fmain, DD, NDD, ffriend, memory shared, CPUcount}

- S (system):- Is our proposed system which includes following tuple.
- s (initial state at time T) :-GUI of search engine. The GUI provides space to enter a query/input for user.
- X (input to system):- Input Query. The user has to first enter the query. The query may be ambiguous or not. The query also represents what user wants to search.
- Y (output of system):- List of URLs with Snippets. User has to enter a query into search engine then search engine generates a result which contains relevant and irrelevant URL's and their snippets.
- T (No. of steps to be performed):- 6. These are the total number of steps required to process a query and generates results.
- fmain(main algorithm) :- It contains Process P. Process P contains Input ,Output and subordinates functions. It shows how the query will be processed into different modules and how the results are generated.
- DD (deterministic data):- It contains Database data. Here we have considered Hadoop database maintained by System itself.
- NDD (non-deterministic data):- No. of input queries. In our system, user can enter numbers of queries so that we cannot judge how many queries user enters into single session. Hence, Number of Input queries are our NDD.
- ffriend :- WC And IE. In our system, WC and IE are the friend functions of the main functions. Since we will be using both the functions, both are included in ffriend function. WC is Web Crawler which is bot and IE is Information Extraction which is used for extracting information on browser.
- Memory shared: - Database. Database will store information like list of receivers, registration details and numbers of receivers. Since it is the only memory shared in our system, we have included it in the memory shared.
- CPU count: - 2. In our system, we require 1 CPU for server and minimum 1 CPU for client. Hence, CPU count is 2.

Subordinate functions:

- Identify the processes as P.
  S= {I, O, P....}
  P= {GCD, DI}
- GCD is Gather Chat data.
- DI for Data Interpretation.
- P is processes.
- GCD= {U, MAX, CV}
- U= Chat data collected from Server File.
- MAX = {1, 2, 3, … , n}
- CV is for sending gathered data to DI.
- TC = {CV, A, Info}
- CV is input which is taken from GCD.

- A is use for analysing chat to find out obfuscate words.

VII. CONCLUSION

This technique allows us to automatically flag suspicious messages, means replacing one word with another word in a sentence of Run-time Application. Admin is able to view the live chat. While chatting if there is any occurrence of Substituted words then the algorithm is being able to detect the same by converting the substituted words with a normal one. Hadoop database is being used for analysis of the same word, and all the chat will be stored in a text file, the same text file will be stored into hadoop as a storage purpose.

REFRENCES

[1] Swati Agarwal,Ashish Sureka Using Commonsense knowledge-base for Detecting Word Obfuscation in adversarial Communication.Future information Security Workshop,IEEE 2015.

[2] Hugo Lewi Hammer, Detecting threats of violence in online discussions using bigrams of important words.2014 IEEE Joint Intelligence and Security Informatics Conference

[3] Sonal N. Deshmukh, Ratnadeep R. Deshmukh and Sachin N. Deshmukh Finding Real Semantic of Replaced Words Using K-gram and NGD. World Congress on Engineering 2013 Vol III, WCE 2013, July 3 - 5, 2013, London, U.K.

[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 WCE2010, June 30 - July 2, 2010, London, U.K.

[5] SW. Fong, D. Roussinov, and D.B. Skillicorn. Detecting word substitutions in text. IEEE Transactions on Knowledge and Data Engineering, 20(8):1067–1076, 2008.

