Declassify Obfuscates term in Terrorism using Word Substitution

Komal Morankar[1], Nayna Kapadne[2], Priyanka Nagpure[3], Sharad Doke[4], Prof. Priti Lahane[5]

[1,2,3,4]*Dept. of Information Technology, MET BKC IOE , Nashik, Maharashtra*

Abstract: *Word obfuscation or substitution means replacing one word with another word in sentence to conceal the textual content or Communication. Word Obfuscation is used in adversarial communication by terrorist or criminals for conveying their messages without getting red-flagged by security and intelligence. Agencies interpreting or scanning messages or emails and telephone conversations. System is being designed in such a way that there will be admin and suspect will be two roles, Admin is able to view the live chat and monitor the same. The terrorist 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 to 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 in Hadoop as a storage purpose further with the help of Map-Reduce Analytic percentage of the obfuscated words is being generated with the output of key-value pairs.*

Keywords: Obfuscates term ,word substitution ,Map reduce , Natural Language Processing ,Hadoop

I. INTRODUCTION

Obfuscation is the obscure of intended meaning in communication, making the message puzzling, will fully confusing, or harder to understand. It may be on purpose or unplanned and may result from circumlocution (yielding long-windedness) or from use of language or even speech (yielding economy of words but excluding outsiders from the communicative value). Unintentional obfuscation in expository writing is usually a normal feature of early draft 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 enough reading conception and editing skills. System is being designed in a way that there will be admin and suspect is two users, Admin is able to view the file contents and monitor the same. While monitoring 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[1].

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. 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 weblogs (blogs), chat rooms and email, has inevitably resulted in the exchange of information and messages between criminals and terrorists through these means. If this is realized, the amount of data that needs to be analyzed will continue to increase.

II. LITERATURE SURVEY

In previous research, the substituted words is decoded by human resources and it takes lot of time to get an original text messages. The use of LASSO logic regression for classification . But the weakness of this system is rate of incorrectly classification of word is about 20 percent so cannot used for real time substitution. Detecting word substitution in text but the problem of this system is low percentage of accuracy, time and also the conflict data structure.

Winnow-based approach to context-sensitive spelling correction: Detecting words out of context can also be used to detect (and correct) misspellings. These problem dyers from the problem addressed here because the misspelled words are nonsense, and often nonsense predictably transformed from the properly spelled word, for example by letter reversal. Using common sense knowledge base for detection word Obfuscation using a Hidden Markov Model: Detecting words out of context has also been applied to the problem of spam detection. For example, Spam Assassin uses rules that will detect words such as Viagra. The problem is related to detecting misspellings, apart from that the transformation have properties that maintain certain visual qualities rather than reacting lexical development errors detect word-level manipulations typical of spam using Hidden Markov Models. Using common sense knowledge base for detection word obfuscate in adversarial communication future information security (swati agrawal, ashish sureka): use of concept net to compute conceptual similarity between two term solving problems of word. Conduct empirical analysis on large and real word data set. Conflict occurs in substitution of words due to the complex design [2]. Detecting threats of violence in online discussion using bi- diagrams of important words (Hugo Lewi Hammer): Use of Classification method within text mining, classification performed using LASSO logistic regression. But the drawback of this system is rate of wrongly classification of word is about 20 percent so cannot used for real time substitution [3]. Detecting word substitution in text. IEEE Transaction on knowledge and data (SW. Fong, D. Roussi-novei and D. B. Skillicorn): design the system of measures that applied to sentence, positive detection rate is 90 percent and false rate is around 10 percent. Difficult to understand which design is used and conflict data structure [4].

III. PROPOSED SYSTEM

In our system both user as well as the admin logs into the system. The user can chat with the other users. The admin can supervise the chat between the suspects. The admin can crack the chat between them. In our proposed system we are going to use natural language processing, word substitution, map-reduce like techniques. The input to our system will be in the form of simple text message, or posts from the social Medias or emails. The admin will provide the input for word.

IV. SYSTEM ARCHITECTURE

The above architecture of the proposed system shows, there are two suspicious persons i.e., suspect A and suspect B. Both the suspect communicate with each other. The admin supervise the chat between the two suspicious persons, cracks the chat between them. The admin then provides the chat in the form of normal text message for natural language processing after performing natural language processing on the data the relevant word for the doubtful word is substituted from the Hadoop database. Then the output is provide back to the admin in the form of normal text file. After performing natural language processing on the data the relevant word for the doubtful word is substituted from the Hadoop database. Then the output is provided back to the admin in the form of standard text file.

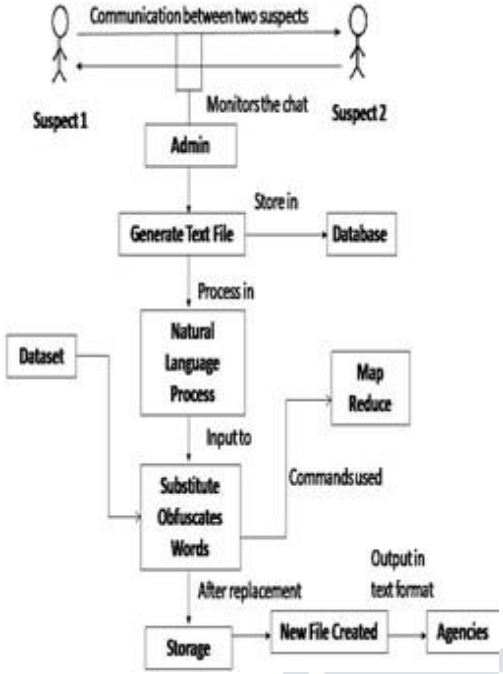

Fig: Architecture of System

V. SYSTEM METHODOLOGIES

1. Map Reduce:

Map-Reduce is a framework for processing parallelizable problems across large datasets using a large number of system, collectively referred to as a cluster (if all nodes are on the same local network and use comparable hardware) or a net (if the nodes are shared across physically and organizationally distributed systems, and use more various hardware). Processing can occur on data stored either in a file system (unstructured) or in a database (structured).

In this proposed system the map- reduce is used for replacing obfuscates words with the combinations of words in the dataset. At the same time the NLP processes the chat messages and classify it in categories like verbs, nouns, etc. In map-reduce it maps the combination of words for the particular word from the chat and find the best choice for the particular word and replace that word with the classified word and then reduce that texts and combine the classified chat and generate the result in another texts file. A huge part of the authority 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 implementation are handled evidently by the execution structure 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 believable algorithm that a programmer wishes to develop must be uttered in terms of a small number of firmly defined components that must fit together in very specific ways. In this proposed system a Map-Reduce framework is usually composed of three operations:

- **Map:** The Obfuscates term find out by the NLP(Natural Language Process). Each obfuscates term applies the map function to the dataset. It will find out the approximate term for the following word.
- **Shuffle :** The Obfuscates words get replaced by the substituted words from data-set.
- **Reduce:** The words will combine as before to form message per key, in parallel.

2. Hadoop:

Hadoop is open source framework that manages data processing and storage for big data applications running in clustered systems. It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications. Hadoop can handle different forms of structured and unstructured information, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide. The core components in the first iteration of Hadoop were Map- Reduce, the

Hadoop Distributed File System (HDFS) and Hadoop Common, a set of shared utilities and libraries. Map-Reduce uses map and reduce functions to split processing jobs into multiple tasks that run at the cluster nodes where data is stored and then to combine what the tasks produce into a consistent set of results. Map-Reduce initially functioned as both Hadoop's processing engine and cluster resource manager, which tied HDFS directly to it and limited users to running Map-Reduce batch applications In this system, the hadoop is used for managing the files which are created the file files which are already created or the newly generated files all files in this system is manages by the HDFS (Hadoop Distributed File System).

Hadoop Applications:

Hadoop is primarily geared to analytics uses, and its ability to process and store different types of data makes it a mostly good fit for big data analytics applications. Big data environments typically involve not only large amounts of data, but also a variety of, structured transaction data to semi structured and unstructured forms of information, such as internet click stream records, web server and mobile application logs, social media posts, customer emails and sensor data from the internet of things (IOT). Insurers use Hadoop for applications such as analyzing policy pricing and control safe driver reduce programs. YARN greatly stiff the applications that Hadoop clusters can handle to include stream processing and real- time analytics applications run in cycle with processing engines, like spark and flink. For example, some manufacturers are using real-time data that's streaming into Hadoop in analytical preservation applications to try to detect equipment failures before they occur. Fraud detection, website personalization and customer understanding scoring are other real-time use cases. Because Hadoop can process and store such a extensive collection of data, it enables organizations to set up data lakes as expansive reservoirs for incoming streams of information. Data lakes generally serve dissimilar purposes than conservative data warehouses that hold cleansed sets of transaction data. But, in some cases, companies outlook their Hadoop data lakes as modern- day data warehouses. Either way, the rising role of big data analytics in business decision-making has made effective data supremacy and data security processes a priority in data lake deployments.

3. Natural language

Process: Natural language processing (NLP) is an part of computer science and reproduction intelligence anxious with the relations involving computer and human (natural) languages, in particular how to program computers to process and examine large amounts of natural language data.

Challenges in natural language processing frequently involve speech reorganization, natural language understanding, and natural language generation.

This proposed system the terrorist and criminals use textual or word obfuscation to prevent their messages from getting intercepted by the law enforcement agencies.

.

Textual or word substitution consists of replacing a red-flagged term (which is likely to be present in the watch-list) with an ordinary or an innocuous term. Innocuous terms are those terms which are less likely to attract attention of security agencies.

For example, the word attack being replaced by the phrase birthday function and bomb being replaced by the term milk. This obfuscates words find out with the help of Natural Language Process. It classify the message into different categories like noun, verbs, etc. using this it find out the suspected word. The technology interprets the significant fundamentals of the human language sentence, such as those that might correspond to specific features in a data set, and returns an answer. Natural Language Processing can be used to interpret free text and make it analyzable. There is a incredible amount of information stored in free text

files, this information was unapproachable to computer assisted analysis and could not be analyzed in any kind of systematic way. But Natural Language Processing allows analysts to filter through massive troves of free text to find appropriate information in the files.

• Natural language generation. Translate information from computer databases or semantic intents into understandable human language.

• Natural language understanding Translate chunks of text into more formal representation such as first order logic structures that are easier for computer programs to influence. Natural language understanding involves the classification of the planned semantic from the multiple achievable semantics which can be imitative from a natural language manifestation which usually takes the form of organized notations of natural language conception.

VI. EXPERIMENTAL SETUP

The system is developed using java-jdk1.7. The database is stored in mysql database server. A web application is created using eclipse system with 4 GB ram and i3 processor.

Dataset: A synthetic dataset is generated for 200 to 300 words. Dataset contains the words which are needed to replace or the suitable pairs for those words.

VII. CONCLUSION

The terrorist attacks are the major problem for the society. To know their intentions early and avoid the attacks is very necessary. So the proposed system will declassify the obfuscates chat with the help of word substitutions within textual communication can detect chat and form original messages where the natural language processing technique allows to automatically detect suspicious messages and Map reduce will substitute the word so that the further investigation can be completed.

VIII. ACKNOWLEDGEMENT

Firstly we gladly thanks to our project guide Prof. Priti Lahane, for her valuable guidance for implementation of proposed system. We will remain thankful for excellent as well as polite guidance for preparation of this report. Also we would sincerely like to thank to HOD of Our department Prof. Dr. S. V. Gumaste and other staff for their helpful coordination and support in project work.

IX. REFERENCES

[1] Vaibhav Shinde, Komal Salunke, Punam Teli , "Deducing and conclusive analysis of declassified obfuscate chat in terrorisomusing word substitution", *IERJ, March 2017.*

[2] Swati Agarwal, Ashish Sureka, "using Commonsense Knowledge- base for detecting word obfuscation in adversarial communication",*Future information securi ty workshop,IEEE 2015.*

[3] Hugo Lewi Hammer , "detecting treats of violence in online discussion using bigram of important words", *IEEE Joint Intelligence and Security Informatics Conference. IEEE 2014*SW. Fong, D. Roussinov, and D.B. Skillicorn, "Detecting word substitutions in text", *IEEE Transactions on Knowledge and Data Engineering.*

[4] 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.*