



Monitoring Suspicious Discussions on Social Media Case Study

TridibChakraborty

Assistant Professor

Department of Information Technology

Gurunanak Institute of Technology

Kolkata, India

AdrishGoswami

Department of Information technology

Guru Nanak Institute of Technology

Kolkata, India

Aliva Ghosh

Department of Information technology

Guru Nanak Institute of Technology

Kolkata, India

Jeevan Krishna Ghosh

Department of Information Technology

Guru Nanak Institute of Technology

Kolkata, India

Puja Datta

Department of Information technology

Guru Nanak Institute of Technology

Kolkata, India

Rangoli Das

Department of Information technology

Guru Nanak Institute of Technology

Kolkata, India

Shankha Chatterjee

Department of Information Technology

Guru Nanak Institute of Technology

Kolkata, India

Abstract: The goal of data analysis is to selecting words from the social media like “face book”, “twitter”, “Orkut” etc. The words are analyzed using algorithm like c4.5 or with any other algorithm, the patterns are judged to discover that whether the word or comment can create any negative impact or not. It is the nature of some people making suspicious discussions and sometimes they cross their limits and this may create havoc in the society. Monitoring these staffs very carefully & effectively is necessary. The steps involve collecting data from social media known as Data Mining, analyzing them with some negative & positive words which are affecting peoples is known as Data Analysis, making decision with some effective Algorithms, & lastly Monitoring them carefully. It's become much more helpful now-a -days to watch criminal activities and to deal with cybercrime.

Keywords: Data Mining, Data Analysis, Algorithms, Monitoring *et.*

I. INTRODUCTION

In recent days people are addicted to the social media like anything, it has become the part & parcel of our life. And we have started using it as a live platform to express our feelings, opinions, promotions of the current events on any topic. Fraud or misguided people doesn't leave any space to spread criminal activities & social media is one of the popular medium of them. Data mining & Data analysis is the technique by which we can keep eyes on social media. The process includes

mainly four steps – □ Data Mining

- Data Analysis
- Making Decision Tree
- Monitoring them

II. LITERATURE SURVEY

DATA MINING^{[3][6]} is the first step of data analysis. It is a process of discovering patterns of datasets. Main goal of data

mining is to extract the information from the dataset and to give it a proper structure for further use. So, collecting data and giving it a understandable structure by extracting it's proper meaning, up to this part is done in the first step of data analysis, i.e. data mining. The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history.

Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning. Data mining is a data analysis approach that has been quickly adapted and used in a large number of domains that were already using statistics.

Next step **DATA ANALYTICS**^[5] is a method in which data is collected and organized so that one can derive helpful information from it. In other words, the main purpose of data analysis is to look at what the data is trying to tell us. For example, what does the data show or do? What does the data not show or do? For Cane, will his data show that there are more young hunters out hunting deer each year? Or, will it shock Cane and show that more young hunters are hunting bears? There are many different methods of collecting data. It is the process of cleaning and modifying a data or a dataset to find out useful information. In this step we

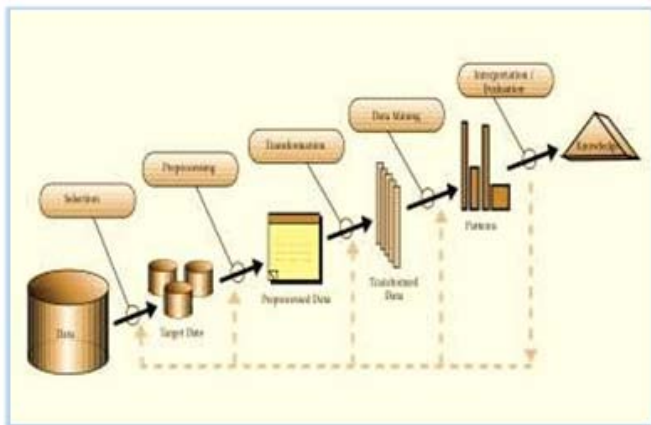
can get a useful meaning of data by processing it after data mining process. The next step is to make a **DECISION TREE**^[2] which is generally used as a *decision support* tool and has a tree like structure contains decisions and their possible consequences. First a table is created which contains comment number versus all 5 categories for a training data set. The goal of constructing ID3^[4] decision tree is attribute selection. The two mathematical terms named Entropy and Information Gain is used for attribute selection. These 2 terms will be used by the ID3 algorithm to determine which attribute will be selected to become a node of decision tree. All the records in the table are referred as the Collection. It is used in decision analysis. It is used as it is simple and easy to interpret and allows addition of new possibilities. It has the ability to determine worst and best values for different possibilities and can be combined with other decision techniques. Entropy is defined as the average amount of the information contained in the event, sample or record.

Entropy(S) = $-\sum_{i=1}^c p(I) \log_2 p(I)$ refers to the proportion of S belonging to class \sum is over c i.e. summation of all the classifier items. Last step of this process is MONITORING DATA.

KDD GRAPH^[3] is also used in data analysis. KDD graph is graphical representation of steps of this process. We can properly expand the steps of Data Analysis with this KDD graph. Image of KDD graph is given.

According to this graph in the first step we have to collect all the information and then all the data or data sets are processed and transformed. All the transformed data is then send for data mining at last all the data is evaluated and interpreted using patterns and then the data is stored in the database. So by this KDD graph we can describe all the steps of data analysis.

KDD process



III. Related work

A minor work is done for this case, if a sentence or group of string is given, the string or data is then compared with the words in the database dictionary. As two categories of data is mentioned in the database. The given string or dataset is compared with the words contains in the dictionary and then the data set is judged properly to understand that the data will create positive impact or negative. If it creates any negative

impact which can create havoc or riots in the society then it is printed as SUSPICIOUS DATA and then that data will be modified using different algorithms.

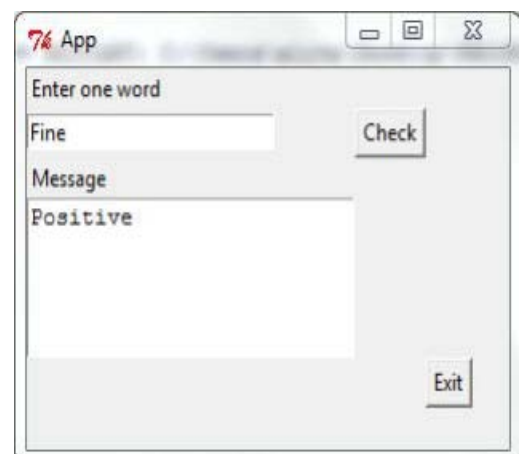
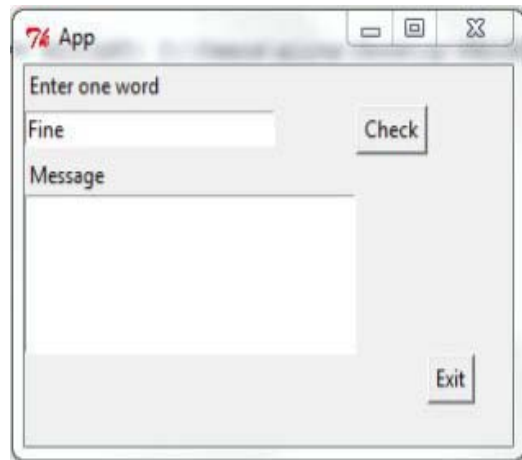
But if the data set is matched with positive words of the database dictionary then the data will remain same and will never be modified.

Algorithm:

1. Start
2. Scan the selected or given data or dataset
3. Datasets are compared with the words in the database dictionary.(There are two types of words in dictionary- positive and negative)
4. If the data is matched with negative words then the data is printed as SUSPICIOUS DATA.
5. Else the data will be printed as POSITIVE
6. END

Implementation

Here an application is created to check the suspicious words used in social media. In this application the following fields are used: 'Enter one word' is used to enter a word for checking. 'Check' button is used to get the probable answer. The 'Message' field is used to print the output to know whether the word is suspicious or not.



IV.CONCLUSION

In this paper we proposed an algorithm to solve the problems of society that occurs in the because of using social networking sites. Sometimes somehow some comments on social can make a situation complicated or can create panic or bad situation as the social networking sites has become the main platform for showing anger, happiness and for expressing feeling and thoughts regarding each and every matter. So sometimes it creates controversy. To avoid these situations this algorithm can be used to find out the negative comments.

V.REFERENCES

- [1] S.Somasundaran and J. Wiebe. Recognizing stances in ideological online debates. In workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 116-124. ACM,2010
- [2] Ana C.E.S Lima and Leandro N.de Castro, “Automatic Sentiment Analysis of Twitter Messages”, IEEE Fourth International Conference on Computational Aspect .of Social Networks (CASoN), p.52-57, 2012.
- [3] Keisuke Mizumoto, HidekazuYanagimoto and Michifumi Yoshioka, “Sentiment Analysis of Stock Market News with Semi-supervised Learning”, IEEE Computer Society,IEEE/ACIS 11th International Conference on Computer and Information Science, p.325-328,2012.
- [4] Sang-Hyun Cho and Hang-Bong Kang, “Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary”, IEEE International Conference on Conference on consumer Electronics (ICCE), p.717-718, 2012.
- [5] Aurangzeb Khan and BaharumBaharudin, “Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms form Blogs”, 2011.