Abstract: With the increasing years, the internet has changed the lives of so many people for better or worse. As internet technology is progressing, many illegal activities have also increased exponentially. The Internet is an unacknowledged path for illegal activities such as hacking, trafficking, betting, fraud and scams etc. The cyber-crime branches are looking for provisions to detect these forums for illegal feedbacks, comments or reviews and download questionable postings as verification for their investigation. Our proposed system will monitor for suspicious postings, collect it from few discussion forums, implement techniques of data mining and extract meaningful data. In this concern, we focus on Data Mining and Sentimental Analysis to enhance the techniques and to extract the features of the text to represent them.

Keywords: Illegal Activities, Discussion forums, Sentimental Analysis.

1. Introduction

Accelerating crimes on digital mediums alert the law implementation bodies to continuously monitor online activities. To achieve the above we need to build a system which detects suspicious postings on online forums. A lot of surveys and facts have proved that it is difficult to manage information which constantly keeps changing on internet thus data mining is the optimal choice to analyse and gather data. Using various data mining techniques, raw data is extracted from a large text corpus and this raw/unstructured data is transformed into structured data in pre-processing. This paper highlights the datamining techniques and sentimental algorithm which is prototyped and implemented using python which is functional in natural language using Natural Language Toolkit (NLTK) library.

A. Existing systems

The digital technology has been impacting human behaviour for a very long time. The existing system analyse text sources from social media and classify the text into different groups. The system distinguishes between legal and illegal data using Stop-Words Selection, Stemmer algorithm and Levenshtein algorithm. In Stop-word Selection, the commonly used words in English language, such as, “we”, “he”, “they” are removed. More such words can be removed using this algorithm.

The Porter Stemmer algorithm removes the suffixes from English words and transforms into its root word, for example:

- The words “Stemmed”, “Stemmer”, “Stemming” have suffixes “ED”, “ER” and “ING” will be removed during the information retrieval and will leave one word “STEM” as the root word.

- Levenshtein distance is a measure of similarity between two words.

B. Some Limitations of current systems

Although the existing system works fine, but there is still scope for improvement. The performances of data retrieval analysis real from online forums remains debatable due to lack of tools. The system is to monitor suspicious discussions automatically on an online forum but the system does not take large amount of data as input. The system is difficult to Moderate. It only limits to Spam reviews and feedbacks. Security vulnerability is also one of the major disadvantage.
2. Literature Survey

A framework has been developed by [4], which is used to detect emotion on online media. EmoTxt finds the emotions and categorize based on the input data provided in a comma separated value (CSV) file format. The output is in the form of CSV file. The file contains text id and predicted label for each input data set. The model classifies the emotions as, joy, sad, and anger etc. According to researchers [4], the model follows a tree structured hierarchical classification of emotions, where latter layers provides an understanding of emotions of the previous layers. The model includes six basic emotions, namely love, happy, anger, sad, fear, and surprise. The data is tested and trained on gold standard dataset using linear Support Vector Machine (SVM).

A research paper published [1], suggests various techniques and algorithms which can be employed. The paper elaborates about Stop-word Selection, Stemming algorithm, Brute-force algorithm, Learning Based algorithm and Matching algorithm. Matching algorithms use two constraints Stemmer Strength and Index Compression. Using these two constraints, stem words in database are compared and their value is calculated. Learning based algorithms include machine learning theories like SVM and conditional random field.

Another paper [2], describes the system will analyse data from few discussion forums and will classify the data into different groups i.e. legal and illegal data using Levenshtein algorithm. Levenshtein is used to measure similarity between two words.

In a paper [5], research is being carried out using web mining. Using Web mining, the data set is collected by crawling large number of web pages. It requires a user interactive query interface intended for predicting crime hotspots from various web pages. The main techniques used are classification, sequential pattern mining, association analysis, outlier analysis and cluster analysis. Clustering and classification techniques, identify the similar items and group them in classes. The association rule mining and sequential pattern mining techniques are similar. They both identify frequently occurring sets and extract a pattern. Using all these techniques in web mining make it more complex. Along with the techniques, a conceptual network i.e. a dynamic structure of nodes connecting in a functional way is required for better visualisation of criminal networks and to reveal the vulnerabilities inside the network. The biggest challenge faced by the researchers was collecting the data from the web pages which consists of hyperlinks, navigation links, advertisements, privacy policies etc. Theses noises should be remove d from the data before processing. Another challenge was that on web the information is never constant. The model intends to concentrate on efficiency by using multiple processes, threads and asynchronous resources.

In another research done on sentimental analysis and opinion mining, [8] Rudy Prabowo1and Mike Thelwall has presented a method which uses different classification methods and hybrid classification on multiple classifiers in their research. They showed that hybrid classification on multiple classifiers can not only improve the performance but can also increase the effectiveness. The paper presents results by comparative study using various automatic classification methods, machine learning classification with hybrid classification.

3. Proposed Method

Data mining can be used to monitor social media as well as discussion forums for suspicious feedbacks or comments. Discussion forums can be used to spread any message to a large population almost instantly. Millions of people share their views and ideas on politics, religion and there are also people who intentionally hurt religious or racial sentiments through malicious posts. Hence it becomes important to monitor the posts on these forums. This application collects the postings and comments from the discussion sites and analyses those comments using data mining techniques and algorithm.

The collected data will be analysed for provoking posts using a set of keywords in the algorithm. Further the set of sensitive keywords are divided into 6 categories: - hacking, sexuality, religious, piracy, gambling, fraud. If the comments in text corpus comes across any of keywords related to any of the 6 categories, then it is classified into that particular category to which the keyword belongs. Our goal is to achieve sentiment analysis for data provided from discussion forums for which we will build Classifiers which consists of different machine learning classifiers.
In this paper, there is a detailed explanation of various existing types of methods to detect suspicious activities of users in online forums. Basically, we found that the techniques developed in this domain, focuses on data mining in order to reduce the suspicious behaviour possessed by the user on web. Study revealed that we need to analyse the user’s behaviour based on feedbacks, comments, reviews shared by them. Also, suspicious behaviours can be categorized under groups such as terrorist activity, financial laundering, hacking, sexual or racial harassment etc. Using this categorization, the corpus of probable suspicious words can be build which will further assist in developing more refined and reliable techniques for detecting these activities.

References


Table 2.1 Literature Survey

<table>
<thead>
<tr>
<th>Work Done</th>
<th>Algorithms Used</th>
<th>Result Obtained</th>
<th>Conclusion</th>
</tr>
</thead>
</table>
| [1] Automated Monitoring of Suspicious Discussions Using Data Mining Statistical Corpus-Based Approach | • Stop word selection  
• Stemming Algorithm  
• Brute-Force Algorithms  
• Suffix-Stripping Algorithms  
• Matching Algorithms  
• Learning-based Methods  
• Emotional Algorithms  
• Keyword Spotting Technique  
• Learning-based Methods | • The purpose of this system is to monitor suspicious discussions automatically on online forum.  
• To detect suspicious posts text analysis is used. This system also focuses to reduce execution time, easier classification to identify more relevant discussions. | This system is to develop an automatic system for detecting illegal discussions on the online forums, through which we can discover illegal activities of all users. |
| [2] Surveillance of Suspicious Forums Using Text Mining | • Stop word selection  
• Stemming algorithm  
• Levenshtein algorithm | • Text mining is used to detect suspicious posts in online forums. Fails to detect the suspicious activity in larger data sets. | This model is presented to detect suspicious activity on smaller data sets using text mining. |
| [3] Algorithm to Monitor Suspicious Activity on Social Networking Sites using Data Mining Techniques | Data Analysis Technique  
• Cleaning and Integration: Selection and transformation  
• Data Mining  
• Pattern Evaluation and Presentation  
• ID3 DECISION TREE | • Using ID3, an accuracy of 86% is achieved which is higher than other n algorithms and neural networks. The 14% error is due to unstructured data which consists of grammar mistakes and abusive comments. | ID3 decision tree algorithm forms a decision tree which will monitor the derogatory comments on social networking sites. The accuracy is 86% using ID3 algorithm. |
| | □ Association rule mining | | Although, the model uses the traditional data mining techniques, the efficiency can increase by using multiple processes and optimal utilization of asynchronous resources. |
| | □ Deviation detection | | |
| | □ Classification | | |
| | □ String comparator techniques | | |
| | □ Sequential Pattern Mining | | |

| [5] Sentiment Analysis and Opinion Mining: A Survey | □ Rule Based Classification General Inquirer Based Classifier (GIBC) | We can use a Sentiment Analysis Tool (SAT), that can apply a semi-automatic, complementary approach, with in each classifier that contributes to other classifiers in achieving effectiveness. |
| | □ Rule-Based Classifier (RBC) | | The hybrid classification on multiple classifiers can not only improve the performance but can also increase the effectiveness. |
| | □ Statistics Based Classifier (SBC) | | |
| | □ Induction Rule Based Classifier (IRBC) | | |
| | □ Support Vector Machines | | |
| | □ Hybrid Classification | | |