A STUDY OF SURVEILLANCE ON MONITORING SUSPICIOUS DISCUSSIONS ONLINE FORUMS USING DATA MINING

GAURAV KUMAR SUMAN

Research Scholar, Dept. of Computer Application,

Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road, Madhya Pradesh, India,

Dr. Jitendra Seethlani

Research Guide, Dept. of Computer Application,

Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal Indore Road, Madhya Pradesh, India.

ABSTRACT

The Forum is a huge virtual space where to express and share individual opinions, influencing any aspect of life. The exponential advancement in information and communication technology has fostered the creation of new online forums for much online discussion. In online forums, the users produce several and various formats of suspicious posts and exchange them online with other people. Monitoring these discussion forums for possible illegal activities that are in text formats can be further used as evidence for investigation. Recently internet has become a path for online illegal activities such as hacking, tracking, betting, fraud, scams etc. Malicious crowd utilize these online forums for many illegal purpose. Monitoring the suspicious activities is one of the better way to measure clients loyalty and also keeping an account on their sentiments towards the posts. The cybercrime law agencies are searching for solutions to monitor and detect such discussion forums for possible illegal activities and download suspected posting that are in text formats as an evidence. The 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 are focusing on Data mining and Sentiment Analysis to enhance the techniques and to extract the features of the text to represent them.

KEYWORDS—Data mining, sentimental analysis, forums, Stop word Selection, naive bayes.

INTRODUCTION

Web has become a very convenient and effective communication channels for people to share their knowledge, express their opinion, promote their products, or even educate each other's, by publishing textual data through a browser interface. Mining useful information from those plain textual data is important for people to uncover the hidden data. The main aim of data mining is to extract information from large data set and transform it in a

understandable format. As Internet technology has been increasing more and more, this technology led to many legal and illegal activities. It is found that much first-hand news has been discussed in Internet forums well before they are reported in traditional mass media. This communication channel provides an effective channel for illegal activities such as dissemination of copyrighted movies, threatening messages and online gambling etc. The law enforcement agencies are looking for solutions to monitor these discussion forums for possible criminal activities and download suspected postings as evidence for investigation. A way by which this problem could be tackled is depicted in this paper. Text data mining algorithms are used to detect criminal activities and illegal postings.This system monitors and analysis online plain text sources such as Internet news, blogs, etc. for security purposes. This is done with the help of text mining concept. Information is typically derived through the devising of patterns and trends. System will analyze online plain text sources from selected discussion forums and will classify the text into different groups and system will decide which post is legal and illegal. This system will help to reduce many illegal activities which are held on internet. Text algorithms in data mining are used to detect criminal activity and illegal posts. This system analyzes plain text sources for security purposes online, such as net news, blogs, etc. This can be done by using the concept of text mining. Typically, information is obtained from patterns and trends. System monitors plain text sources from chosen online forums online and classifies the text into different groups, and the system decides the whether the post is legal and illegal. 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 data mining techniques and sentimental algorithm which is prototyped and implemented using python which is functional in natural language utilizing Natural Language Toolkit (NLTK) library.

Stop word selection

Stop words are the most used words in the English language which includes the words pronouns such as “I, he, she” or articles such as “a, an, the” or prepositions. Information Retrieval (IR) systems was first introduced the concept of stop-words . For a significant portion of the text size in terms of frequency of appearance small portion of words in the English language accounted. It was noticed that the mentioned pronouns and preposition words were not used as index word to retrieve documents. Thus, it was concluded that such words did not carry significant information about documents. Thus, the same interpretation was given stop words in text mining applications as well .To reducing the size of the feature space the standard practice of removing stop words from the feature space is mainly used. The stop word list that is considered to be removed from the feature space generic stop words list which is application independent. This may have an adverse effect on the text mining application as certain word is dependent on the domain and the application.

Stemming algorithm

Stemming is the process of reducing derived words to stem words, base or root form. The process of stemming is called conflation. They are commonly referred to as stemming algorithms or stemmers. A stemmer for English, for example, it identifies the string "ACCEPTED" based on the root word "ACCEPT".

Levenshtein algorithm

Levenshtein distance is a measure of similarity between two words. The two words are referred to as source word and target word. The distance between two words are calculated where the distance tells about the number of insertions, deletions or substitutions required. For ex. If source word(s) is "TEST" and target word(t) is "TEST" the distance(s,t) = 0. The greater the Levenshtein distance, the more different the strings are. Levenshtein distance is named after the Russian scientist Vladimir Levenshtein, who devised the algorithm in 1965. The Levenshtein distance algorithm has also been used in

- Spell checking
- Speech recognition
- DNA analysis
- Plagiarism detection

The algorithm decsription:

1. Set n to be the length of s.

Set m o be the length of t.

if n=0,return m and exit.

if m=0,return n and exit.

Construct a matrix containig 0..m rows and 0..n columns.

2. Initialize the first row to 0..n. Initailze the first column to 0..m.

3. Examine each character of s (i from 1 to n). Examine each character of t (i from 1 to m).

4. If s[i] equals t[j],the cost is 0. If s[i] doesnt equals t[j],the cost is 1.

5. Set cell d[i,j] of he matrix equal to the minimum of:

a. The cell immediately above plus 1: d[i-1,j]+1.

b. The cell immediately to the left plus 1: d[i,j1] +1.

c. The cell diagonally above and to the left plus the cost: d[i-1,j-1]+cost.

6. After the iteration steps (3,4,5,6) are complete, the distance i found in cell d[n,m].

METHODOLOGY

The system architecture is composed of four processing phases. Initially, a user will post or comment something on the blog. In the first phase, this data from the blog is saved to the database, and is also subjected to Natural Language Processing. Data tokenizer then converts this unstructured data into structured form along with stemming the words into their root form. In the second phase, a Feature extracting algorithm will extract the required features from the text corpus, sending the data to the Naive Bayes classifier.

Based on the probability that given record or data point belongs to a particular class, the third phase implements a Naive Bayes algorithm which classifies data as positive, negative or neutral. In the final phase of sentiment prediction, this classified dataset is subjected to a sentimental analysis algorithm which will further classify data into particular categories and thus providing an optimal result.

Fig. 1. Phases of sentiment analysis

ALGORITHMS USED

1. Tokenization

A process of breaking down the text corpus into individual elements is called tokenization. This is the first phase of pre-processing where the given textual information is split into individual words.

A swimmer likes swimming, thus he swims.						
↓						
a	swimmer	likes	swimming	thus	he	swims

Tab. 1. Tokenization

2. Stop word removal

This a major form of pre-processing where the data in unstructured form will be converted into structured form by filtering out useless words (stop words) from the given set of information.

A swimmer likes swimming, thus he swims.					
↓					
swimmer	likes	swimming	,	swims	.

Tab. 2. Stop word removal

3. Stemming and Lemmatization

Stemming is the process of converting a word into its root form. Porter stemmer is the algorithm used perform this operation

A swimmer likes swimming, thus he swims.								
↓								
a	swimmer	like	swim	,	thu	he	swim	.

Tab. 3. Stemming

Stemming can result in non-real words. To counter this limitation lemmatization is used. Lemmatization generated canonical form of a stemmed word.

A swimmer likes swimming, thus he swims.							
↓							
A	swimmer	like	swimming	,	thus	he	swim	.

Tab. 4. Lemmatization

4. Naive Bayes classification A classification approach using Bayes theorem where the presence of a particular feature in a class is assumed as unrelated to the presence of the rest of the feature of the class. It has multiple application areas. In our system Naïve Bayes classifier is used to categorize user entered textual data as spam and ham.

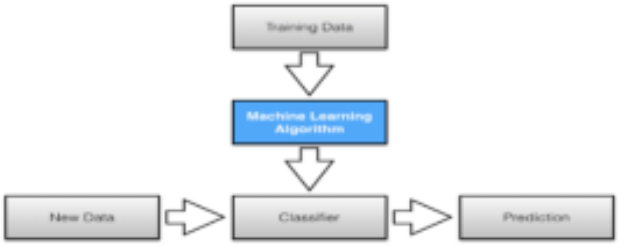

Fig.2. Naïve Bayes classifier

The system will take input from blog which are of comments and post created by the user. Then this data will be stored in the database. The stored data is preprocessed before undergoing various test mining techniques. The preprocessed data will be sent for tokenization, stop word filtering, stemming and lemmatization, and negation handling. By using a Naive Bayes classifier, we will be predicting whether the given data is positive, negative or neutral. Based on the result post will be approved or rejected. If it's negative then the posts are temporarily blocked and waits for admins approval. If admin approves the post it is made public or else the post will be blocked. The system architecture is depicted in Fig. 2.

Fig. 3. System Architecture

CONCLUSION

This paper presents a way for detecting suspicious discussions on the online forums, through which we can uncover suspicious activities and interests of users. The purpose of this system is to monitor suspicious discussions on online forum. Text mining is used to detect suspicious posts in online forums. The Social Networking sites are affecting human life. Hence this system successfully detects the suspicious words from chats and prevents the suspicious activities. This system is applicable to every department where there is need. Not only in social-networking sites is this system applicable in forest department, disaster management system to prevent illegal activities. Text mining technology used to detect suspicious words from discussion forum. Internet is radically changing the way people communicate and share their opinion globally through various online platforms like discussion forums and social network platforms. But as the internet is growing rapidly, various cyber-crimes like scamming, illegal postings and other illicit activities are increasing exponentially. This paper proposes a system which not only identifies and reports illegal activities on online discussion forums but also helps in their reduction by posing certain restriction to the content a user can share publicly. Numerous text mining techniques are implemented in our system to filter illegal and fraudulent posts and ultimately providing a legitimate platform for the users to share their opinions.

REFERENCES

Salim Alami, Omar el Beqqali, "Detecting Suspicious Profiles using Text Analysis within Social Media," In Proceedings of 2015 IEEE Journal of Theoretical and Applied Information Technology, Volume 73, Issue 3, 2015.

Z.Yao, C.Ze-wen "Research on the Construction and Filter Method of Stop-word List in Text Preprocessing," In Proceedings of 2011 IEEE Intelligent Computation Technology and Automation(ICICTA), Pp. 217-221, 11-13 March 2011.

Y.M.Lai, K.P.Chow, C.K.Hui, S.M.Yiu, "Automatic Online Monitoring and Data Mining Internet Forums," In Proceedings of 2011 IEEE 7th International Conference On Intelligent Information Hiding and Multimedia Signal Processing, Pp.384-387, 7-9 August 2011.

T.K.Ho, "Stop Word Location and Identification for Adaptive Text Recognition," In Proceedings of 2000 IEEE International Journal on Document Analysis and Recognition, Volume 3, Issue 1, 2000.

J. Saxe, D. Mentis, and C. Greamo, "Visualization of shared system call sequence relationships in large malware corpora," In Proceedings of 2012 9th International Conference on Visualization for Cyber Security, Pp. 33-40, 2012.

Yu Shaoqian, "Stemming Algorithm for Text Data and Application to Data Mining," In Proceedings of 2010 IEEE 5th International Conference On Computer Science & Education(ICCSE), pp. 507-510, 24–27 August 2010.

T.Bhaskar, "Fast identification of stop words for font learning and keyword spotting," In Proceedings of 1999 IEEE 5th International Conference on Document Analysis and Recognition (ICDAR), Pp. 333-336, September 1999.

Zhiyong Zeng, Hui Yang, Tao Feng, "Data Mining Methods for Knowledge Discovery," In Proceedings of 2011 IEEE International Conference On Data Mining Methods For Extraction Of Data, Pp. 412-415, 29-31 July 2011.

Xinqing Geng, Fengmei Tao, "Automatic Internet Monitoring and Data Online Forums," In Proceedings of 2012 IEEE 4th International Conference On Intelligent Information Hiding And Signal Processing, Pp. 492-495, 2-4 November 2012.

Y. Yang, "An evaluation of statistical approaches to text categorization," In Proceedings of 1999 IEEE Journal On Information Retrieval, Volume 1, Issue 1, 1999.

K.T. Frantzi, S. Ananiadou, and J. Tsujii, "The C-Value/NC-Value Method of Automatic Recognition for Multi-Word Terms," Proc. Second European Conf. Research and Advanced Technology for Digital Libraries (ECDL '98), pp. 585-604, 1998. [5]

T. K. Ho, "Fast identification of stop words for font learning and keyword spotting", In Proc of Document Analysis and Recognition, Fifth International Conference on (ICDAR). IEEE; pp. 333-336 Sep. 1999.

M. F. Porter. An algorithm for suffix stripping .Program, 14(3):130–137, 1980.

B. Connor, R. Balasubramanyan, B. R.Routledge, and N. A. Smith."From tweets to polls: Linking text sentiment to public opinion time series". In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media 2010.