RESEARCH ARTICLE OPEN ACCESS

Advanced Mechanism for Monitoring The Suspicious Discussions on Online Forums using Text Data Mining

Priyanka Hulde[1], Ms. Nisha Lalwani[2]
(Dept of CSE, JIT Nagpur)[1],
(Asst Professor, Dept of CSE, JIT Nagpur)[2].

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 forumsfor much online discussion. In online forums, the users produce several and various formats of suspicious posts andexchange them online with other people . Monitoring these discussion forums for possible illegal activities that are in textformats can be further used as evidence for investigation.

Keywords- Online forums, Text mining**,** Levenshtein Distance.

I. INTRODUCTION

Web has become a very convenient and effectivecommunication channels for people to share theirknowledge, express their opinion, promote theirproducts, or even educate each others, by publishingtextual data through a browser interface. Mininguseful information from those plain textual data isimportant for people to uncover the hidden data.Themain aim of data mining is to extract informationfrom large data set and transform it in aunderstandable format.

As Internet technology has been increasing more and more,this technology led to many legal and illegalactivities. It is found that much first-hand news hasbeen discussed in Internet forums well before they are
reported in traditional mass media. Thiscommunication channel provides an effective channel
for illegal activities such as dissemination ofcopyrighted movies, threatening messages and online
gambling etc.

The law enforcement agencies are looking forsolutions to monitor these discussion forums forpossible criminal activities and download suspected postings as evidence for investigation. A way bywhich this problem could be tackled is depicted inthis paper.

Text data mining algorithms are used to detectcriminal activities and illegal postings. This systemmonitors and analysis online plain text sources suchas Internet news, blogs, etc. for security purposes.This is done with the help of text mining concept.Information is typically derived through the devisingof patterns and trends. System will analyze onlineplain text sources from selected discussion forumsand will classify the text into different groups andsystem will decide which post is legal and illegal.This system will help to reduce many illegal activitieswhich are held on internet.

II. RESEARCH METHODOLOGY

Identify the Stop Word:

To understand whether the chat or legal or illegal, it is necessary to find the stop words from them. Stop words are the words declare by the administrator or high authority which should not be in forum chat. These stop words are found by using the algorithms like suffix stripping and affix stripping.

Removal of Stop words:

If the stop words are found in the chat then they are analyzed and if the word is suspicious then these

words are removed and ********** blank space or asterisk is seen in chat.

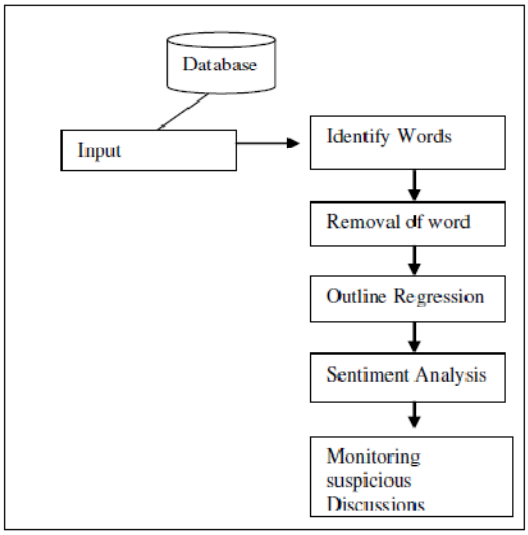

Fig. Flow of the Proposed System

Outline Regression:

In this the words are removed after the removal of these words the system updates the database and finds the most repeated frequent words. The outline regression means the statistical analysis of the words about the frequent occurrence.

Sentiment Analysis:

In this the sentiment analysis of the stop words is done.In this after the statistical analysis of the frequent words, these words are checked and removed with the blank space or asterisks.

Monitoring Suspicious Discussion:

In this the chat is monitored by the removal this unexpected or suspicious words and hence we get the chat free from the suspicious words

ALGORITHMS USED

STOP WORD SELECTION:

Stop words are the words are mostly the pronouns and articles. Pronouns like “I, he, she” and articles like “a, an, the”. Following algorithms are used to detect stop words which are suspicious in system

Stemming Algorithm:

• Stemming is the process of removing the derived words from the stem words, base or root form – generally a written in any of the word form. The process of stemming is also called conflation. These are the programs which commonly referred to as stemming algorithms or stemmers .

• Stemming algorithms mostly differs in terms of accuracy and performance. A stemmer for ENGLISH word, for example, it recognize the STRING "rewarding" is derived from the root "reward", and "expressing" is derived from the root "express". A stemming algorithm identifies the root word example "enforcing", "enforced" to the root word, "enforce".

There are several approaches to stemming. One way to do stemming is to store a table of all index terms and their stems. For example:

Term	Word	Removed
Enginnering	Enginneer	-ing
Engineered	Engineer	-ed

Suffix Stripping Algorithm

This algorithm do not rely on a lookup table has inflected forms and root form relations. Instead, smaller lists of “rules” are stored which provide an input word form, to find root

form. Some examples of the rules include: if the word ends in 'ful', remove the 'ful

if the word ends in 'ing', remove the 'ing if the word ends in 'ly', remove the 'ly'

Removing suffixes by automatic means is an operation which is especially useful in the field of information retrieval. In a typical IR environment the one has a collection of alldocuments and each is described by the words in the document title and possibly by words in the document abstract. Ignoring the issue of precisely where the words originate, we can say that a document is represented by a vector of words, or terms. Terms with all common-stem willnormally have the equal meanings, for eg:

CONNECT CONNECTED CONNECTING CONNECTION CONNECTIONS

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, IONS to leave the single term CONNECT. In addition to this the suffix stripping willdecrease the count of terms from the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous. The nature of the task will vary considerably depending on whether a stem dictionary

is being used, whether a suffix list is being used, and of course on the purpose for which the suffix stripping is being done. Assuming that, that the purpose of the task is to improve IR performance and one is not using of a stem dictionary, thesuffix stripping program will usually be given an explicit list of suffixes, and, with each suffix, the criterion under which it may be removed from a word to leave a valid stem. This is theapproach adopted here. In any suffix stripping program for IR work, two points must be borne in mind. Firstly, the suffixes are being removed simply to improve IR performance, andnot as a linguistic exercise. This means that it would not be at all obvious under what circumstances a suffix should be removed, even if we could exactly determine the suffixes of aword by automatic means. Perhaps the best criterion for removing suffixes from two words W1 and W2 to produce a single stems, is to say that we do so if there appears to be nodifference between the two statements `a document is about W1' and `a document is about W2'. So if W1=`CONNECTION' and W2=`CONNECTIONS' it seems very reasonable to conflate them to a single stem. But if W1=`RELATE' and W2=`RELATIVITY' it seems perhapsunreasonable, especially if the document collection is concerned with theoretical physics. (It should perhaps be added that RELATE and RELATIVITY are conflated together in the algorithm described here.)

Between these two extremes there is a continuum of different cases, and given two terms W1 and W2, there will be some variation in opinion as to whether they should be conflated, just as there is with deciding the relevance of some document to a query. The evaluation of the worth of a suffix stripping system is correspondingly difficult. The second point is that with the approach adopted here, i.e. the use of a suffix list with various rules, the success rate for the suffix

stripping will be significantly less than 100% irrespective of how the process is evaluated. For example, if SAND and SANDER get conflated, so most probably will WAND andWANDER. The error here is that the -ER of WANDER has been treated as a suffix when in fact it is part of the stem. Equally, a suffix may completely alter the meaning of a word, in which case its removal is unhelpful. PROBE and PROBATE for example, have quite distinct meanings inmodern English. (In fact these would not be conflated in our present algorithm.) There comes a stage in the development of a suffix stripping program where the addition of more rules

to increase the performance in one area of the vocabulary causes an equal degradation of performance elsewhere. Unless this phenomenon is noticed in time, it is very easy forthe program to become much more complex than is really necessary. It is also easy to give undue emphasis to cases which appear to be important, but which turn out to be rather rare. For example, cases in which the root of a word changes with the addition of a suffix, as in DECEIVE/DECEPTION, RESUME/RESUMPTION, INDEX/INDICES occur muchmore rarely in real vocabularies than one might at first suppose. In view of the error rate that must in any case be expected, it did not seem worthwhile to try and cope withthese cases.

The algorithm description:

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,j-1] +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]

Keyword Spotting Technique

The keyword pattern matching problem can be detected as the problem of finding occurrences of keywords from a given set as substrings in a described. This problem has been studied in the past and algorithms have been suggested for evaluating it. In the context of emotion detection this method is based on certain predefined keywords. These words are classified as such as disgusted, dull, enjoy, fearness ,exclamied etc. Process of Keyword spotting methods:

1.Tokenization
2.Text
3.Emotion
4.Negation Check
5.Analysis of Intensity
6.Identify Emotion Words

Where a text data is taken as input and output is generated as an emotion class. At the basic step text data is converted into tokens, from these tokens emotion words are detected. Initially this technique will take some text as input and in next step we perform tokenization to the input text data. Words related to emotions will be identified in the next step after analysis of the intensity of these words will be performed and evaluated. Sentence is checked.
an emotion class will be found as the required output.

Learning-based Methods

Learning-based methods are used to evaluate the problem differently. Originally the problem was to identify the emotions from input data but now the problem is to classify the input texts into various emotions. Unlike keyword-based identified methods, learning-based methods try to identify emotions based on a previously trained classifier, which provide various theories of machine learning such as support vector machine and conditional random field, to identify which emotion should the input text data belongs.

Hybrid Methods

Since keyword-based method and the learning-based method could not acquire satisfactory result, some systems use this type of approach by combining both keyword spotting technique and learning based method, which help to improve accuracy outputs. The most significant hybrid based system is the work of Wu, Chuang and Lin , that utilizes a rule-based on this approach is to extract semantics related to specific emotions and Chinese lexicon ontology to get the attributes. These semantics and attributes are associated with emotions . As a result, these emotion association rules, replace the original emotion keywords, serve as the trained features of this learning module based on the seperate mixture models. This method performs earlier approaches, but categories of emotions are still limited in number.

CONCLUSION

This paper presents a way for detecting suspiciousdiscussions on the online forums, through which wecan uncover suspicious activities and interests ofusers. The purpose of this system is to monitorsuspicious discussions on online forum. Text miningis used to detect suspicious posts in online forums.

REFERENCES

[1] M.SuruthiMurugesan, R.Pavitha Devi, S.Deepthi, V.SriLavanya, Dr.AnniePrincy, "Automated MonitoringSuspicious Discussions on Online Forums Using DataMining Statistical Corpus based Approach," In Proceedingsof 2016 IEEE Imperial journal of InterdisciplinaryResearch(IJIR), Volume 2, Issue 5, 2016.

[2] Salim Alami, Omar el Beqqali, "Detecting SuspiciousProfiles using Text Analysis within Social Media," InProceedings of 2015 IEEE Journal of Theoretical andApplied Information Technology, Volume 73, Issue 3, 2015.

[3] Z.Yao, C.Ze-wen "Research on the Construction and FilterMethod of Stop-word List in Text Preprocessing," InProceedings of 2011 IEEE Intelligent ComputationTechnology and Automation(ICICTA), Pp. 217-221, 11-13March 2011.

[4] Y.M.Lai, K.P.Chow, C.K.Hui, S.M.Yiu, "Automatic OnlineMonitoring and Data Mining Internet Forums," InProceedings of 2011 IEEE 7th International Conference OnIntelligent Information Hiding and Multimedia SignalProcessing, Pp.384-387, 7-9 August 2011.

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

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

[7] Yu Shaoqian, "Stemming Algorithm for Text Data andApplication to Data Mining," In Proceedings of 2010 IEEE5th International Conference On Computer Science &Education(ICCSE), pp. 507-510, 24–27 August 2010.

[8] 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.

[9] Zhiyong Zeng, Hui Yang, Tao Feng, "Data Mining Methodsfor Knowledge Discovery," In Proceedings of 2011 IEEEInternational Conference On Data Mining Methods ForExtraction Of Data, Pp. 412-415, 29-31 July 2011.

[10] XinqingGeng, Fengmei Tao, "Automatic Internet Monitoring
and Data Online Forums," In Proceedings of 2012 IEEE 4t International Conference On Intelligent Information HidingAnd Signal Processing, Pp. 492-495, 2-4 November 2012.
[11] Y. Yang, "An evaluation of statistical approaches to textcategorization," In Proceedings of 1999 IEEE Journal OnInformation Retrieval, Volume 1, Issue 1, 1999.