RESEARCH ARTICLE OPEN ACCESS

MONITORING MALICIOUS DISCUSSIONS ON ONLINE FORUMS USING DATA MINING

Priyanka B.Hulde,Prof. Priyanka Dhudhe
Department of computer Science & Eng,
Jhulelal Institute Of Technology, Lonara

Abstract:

The Forum is a big virtual space where to share and express individual opinions, influencing any aspect of life, with implications for marketing and communication alike. Forums are influencing users preferences by shaping their behaviors and attitudes. Monitoring the malicious activities is the best way to calculate users honesty, keeping a track of their sentiment towards their posts. The exponential advancement in information and communication technology has fostered the creation of new online forums for many online discussion and has also reduced a distance between people. Unfortunately, malicious people use these online forums for the illegal purpose. In online forums, the users produce various formats of malicious posts (text, image, video, GIF...) and exchange them online with other people. The law enforcement agencies are looking for solutions to observe these discussion forums for possible illegal activities and download suspected postings that are in text formats as evidence for an investigation. The data in most online forums are stored in text format, so this system will focus only on text posts.

Keywords: **Data Mining; Stop word selection; Stemming Algorithms; Emotional Algorithm; Text Classification.**

1. INTRODUCTION:

The text corpus is the big and structured set of texts posted in the online forums, and different techniques can be hired in this step. We use dataset at this stage. This stage consists to remove stop words and stemming. In computing, stop words are words which are filtered out after, processing of natural language data (text). To simplify the theory we have to delete stop words that contain no useful information, as stop word remove stemming can simplify the processing and decrease errors. Text data mining algorithms are used to find criminal activities and illegal postings. This system analysis and monitors online plain text sources such as Internet blogs, news, etc. For security purposes, this is done with the help of text mining concept. Information is typically derived from the devising of trends and patterns[1].

2. DATA ANALYSIS TECHNIQUE:

Data analysis consists of following steps:

2.1 Cleaning and Integration: The data collected in the data warehouse from multiple databases and is filtered to remove unwanted inconsistent data [2].

2.2. Selection and transformation: In this step, relevant data is retrieved from data Warehouse and transformed into appropriate forms using aggregation operations [2].

2.3. Data Mining: It is a crucial method where different algorithms like C4.5 (decision trees); k-means are applied to extract data patterns [2].

2.4. Pattern Evaluation and Presentation: In this method, various patterns and relationship in between the data set are identified. Then this processed information is represented using bar graphs, another graphical interface [2].

3. EXISTING SYSTEM:

The existing system text corpus is a huge and structured set of texts posted in the online forums, and different techniques can be employed in this step. In this stage we use dataset. This stage consists to remove stop words and stemming. In computing, stop words are words which are filtered out after, processing of natural language data (text). To simplify the study we have to eliminate stop words that contain no useful information, as stop word remove stemming can simplify the processing and reduce errors.[3] 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 have 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[3]. 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.

DISADVANTAGES:

- This may have an adverse effect on the text mining application as the certain word is dependent on the domain and the application.
- Less security and system cost is high.

4. PROPOSED SYSTEM:

The proposed system consider correlations between communication activity in the blogosphere and stock market movement find that postings from stock forums help predicting market volatility and discover that changes in investors' opinions posted to financial forums are closely linked to abnormal returns . There is also some research on predicting the sales volume on the basis of online communication. For instance, and notice that the number of customer reviews is positively associated with Amazon's sales ranking for books. and prove that customer sentiments about movies are good for predicting box office sales. However, an approach which monitors the development of online opinions with regard to marketing campaigns, external events, competing products and sales volume is still missing. Such an approach is crucial for judging the formation of opinions on products. The approach for monitoring opinions on the Internet comprises three steps. First, opinions posted to an online forum are classified as positive, negative or neutral with the aid of methods coming from text mining.

ADVANTAGES:

- Tracking opinions continuously enables an early detection of changes and risks as well as the initiation of appropriate marketing actions.
- Minimize the prize of storage.

5.MODEL& PROBLEM FORMULATION IDENTIFY THE STOPWORD:

To understand whether the chat legal or illegal, it is necessary to find the stop words from them. Stop words are the words declare by the high authority or administrator which should not be present in forum chat. These stop words are found by using the algorithms like affix stripping and suffix stripping. Removal of Stop words: If the stop words are found in the chat when they are observed and if the word is malicious then these words are removed and ********** asterisk or blank space is seen in chat.

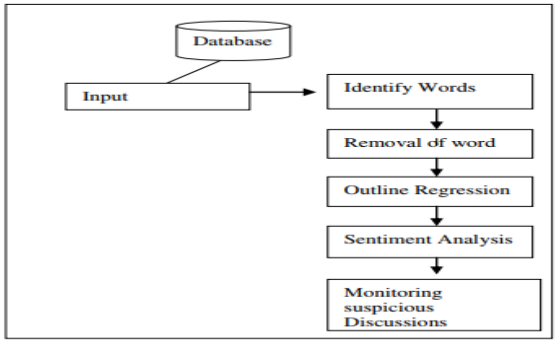

Figure 1. Architecture of the System [4]

Outline Regression: In this, the words are removed after the removal of these words the system updates the database and finds the most repeated used 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 asterisks or blank space. Monitoring Malicious Discussion: In this, the chat is monitored by the removal this unexpected or malicious words and hence we get the chat free from the malicious words.

A. 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 have 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 data or documents. Thus, the same interpretation was given stop words in text mining applications as well. To minimizing 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 a certain word is dependent on the application and the domain[4].

B. Stemming Algorithm:

Stemming is the process of reducing derived words to stem words, root or base form – generally a written word form. The process of stemming is also called conflation.

• Stemming algorithms have many types which differ with respect to performance and accuracy.

• A stemmer for ENGLISH, for example, it identifies the STRING "rewarding" as based on the root "reward", and "expressing" as based on "express". A stemming algorithm identifies the base word example "enforcing", "enforced" to the root word, "enforce"[4].

C. Brute Force Algorithms:

The stemmers lookup table contains relations between inflected forms and root forms. The table is queried to find a matching inflection to stem a word. The associated root form is returned, if a matching inflection is found Suffix Stripping Algorithms. This algorithm does not rely on a lookup table has inflected forms and base form relations. Instead, a smaller list of "rules" are stored which provide an input word form, to find its base form. Some examples of the rules include: if the word ends in 'ing', remove the 'ing' if the word ends in 'full', remove the 'full' if the word ends in 'ly', remove the 'ly'[4].

D. Affix Stemmers

In linguistics, the term affix refers to either a prefix or suffix. Several approaches also attempt to remove common prefixes in order to handle suffixes. For example, in the word incredible, it identifies that "in" is a prefix and it can be removed. Many of the same approaches mentioned earlier apply, and called as affix stripping[4].

E. Matching Algorithms

These algorithms use a stem database (Example: a set of documents that contains stem words). These stem words are not necessarily valid words themselves. In order to stem a word the algorithm tries to match it with stems that stored in storage or database, having various constraints, on the relative length of the candidate stem within the word (example, the short prefix "inter", which is the stem word of such words as "international", "interpersonal", should not consider as the stem of the word "interest[4].

F. Emotional Algorithms

The Emotional algorithm is used to detect the emotions of the human beings via audio, video, text and so on. In online forums, users are posting their comments or sharing their thoughts mainly in a text format. So, the emotional algorithm is mainly used to detect emotions through text in this system. The following methods are used to detect emotions in the text[4].

1. Keyword Spotting Technique

2. Learning-Based Methods

3. Hybrid Methods

6. UNITS

I • Stemmer Strength

A number of words per conflation class are the average size of the group of words converted to a stem word. Word collection of the given size depends on the number of words processed, a higher value indicates that the stemmer is heavier. The value evaluated using the following formula: MWC = Mean number of words per conflation class, BS = Number of unique words before Stemming, AS = Number of unique stems after Stemming, MWC =BS/AS [5].

II • Index Compression

The Index Compression Factor represents the extent that a collection of unique words is compressed (reduced) by stemming, the idea being that the heavier the Stemmer, greater the Index Compression Factor. This is calculated by; ICF = Index Compression Factor, AS = Number of unique stems after Stemming, BS = Number of unique words before Stemming, ICF =(BS-AS)/BS [5].

7. CONCLUSION

The increasing number of consumer opinions on the Web represents a important source of knowledge for companies. The outlined approach allows the monitoring of opinions by employing text mining methods. Opinions are first identified, then aggregated with the aid of an index and finally observed with respect to their development. The analysis considers the effects of external events and campaigns, the opinions towards competing the influence on the sales volumes and products. This case study shows the economic relevance of the monitoring approach exemplarily. Future work will extend the basic approach. The aim is to build an early warning system which recognizes risks and chances automatically by taking the opinions towards a product and its competing product as well as the information about marketing campaigns, external events and sales volumes into account. A warning will be sent to the marketing manager when risks are detected.

REFERENCES:

[1]HARIKA UPGANLAWAR, 2NILESH SAMBHE 1,2Computer Engineering, Department Of Computer Science, Yeshwantrao Chavan College of Engineering, Nagpur. “SURVEILLANCE OF SUSPICIOUS DISCUSSIONS ON ONLINE FORUMS USING TEXT DATA MINING”.International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-4, Issue-4, Aprl.-2017.

[2]Suhas Pandhe, Sahil Pawar Computer Engineering Department, Pune Institute Of Computer Technology, Savitribai Phule Pune University, India. “The algorithm to Monitor Suspicious Activity on Social Networking Sites using Data Mining Techniques”. International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 12, April 2015.

[3]Salim Alami, Omar el Beqqali, “Detecting Suspicious Profiles using Text Analysis within Social Media,” In proceddings f 2015 IEEE Journal of Theoretical and Applied Information Technology Volume 73, Issue 3, 2015.

[4] Bavane A.B., Ambilwade Priyanka V., Bachhav Mourvika D., Dafal Sumit N, Fulari Priyanka Y. Monitoring Suspicious Discussions on Online Forum by Data Mining. International Journal of Advanced Engineering & Science Research (IJAES) Volume 5, Issue 1, March 2017.

[5] M.Suruthi Murugesan, R.Pavitha Devi, S.Deepthi, V.Sri Lavanya, Dr. Annie Princy “Automated Monitoring Suspicious Discussions on Online Forums Using Data Mining Statistical Corpus Based Approach”.In Proceedings of 2016 IEEE Imperial journal of Interdisciplinary Research(IJIR), Volume 2, Issue 5, 2016.