

Surveillance of Suspicious Analysis on Online Forum using Text Mining

¹Mrs. Valarmathi V, ²Madumathi C P

¹Assistant Professor, ²MSc IT

^{1,2}Department of Information Technology,

^{1,2}Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India

Abstract : Now-a-days people uses social networking sites for communication medium such as Facebook, twitter., where the online forum is a huge space where people can express and share their individual opinions influencing any aspect of life for the purpose of marketing and communication. Surveillance of suspicious Analysis on online forum using data mining is used to mine the information from selected online forum discussion and employ data mining techniques to identify hot topics and cluster authors into different group using word based user profiles. Many malevolent people use these online forums for illegal purposes by posting suspicious chats. High-quality information is typically derived through the origin of patterns and drifts. Mostly the data stored in chatting forums are in the form of text, so the proposed system will focus only text.

Index Terms - Online Forum, Data-mining, Web Crawling, Text Mining, Stop word Selection and Sentimental Analysis.

1. INTRODUCTION

Social media has become a powerful equipment for communication in the modern Era, it is always the first point of spotting out any threat of current discussion, at same time from the other point of view we could consider this as an area where we could get an feedback about the topic under discussion from all the categories of people irrespective of working sectors, age, locality etc. Our System will help all the departments to suspect the illegal words according to their norms and conditions. For Example, in Disaster Management System the word “*red-alert*”, is the suspicious word declares by administrator. Then, if the word flood-affected occurs three times our system will Block the user account for day by maintaining the database of username and a word used by user for investigation purpose. This system will analyse online plain text and will classify the text in different groups and analyse word- by- word whether the words are suspicious or not, whether the entire post is legal or illegal.

2. EXISTING SYSTEM

Forums provide statistic data on each discussion thread, but to help investigation, the system must be able to identify wording, especially those which are important such as name of the people, entities and crime etc. Also, grouping hot discussion threads across forums, so that investigator can have a overall picture of occurring events. There are existing published texts mining applications using data on website. For example: *Blog Pulse, Hot Miner and Google Trends*. They perform text mining on user collaborative content and generate statistics and summary reports. There is an existing publication on how to tackle dark forum in US.

2.1 DOWNSIDES OF EXISTING SYSTEM

The data in the social network can be extracted by so many text mining applications but the survey or the public point of view cannot be determined by these mining applications.

The manual survey on a particular topic of discussion consumes more time and it is also confined to a particular area or locality of where the survey is conducted.

These pitfalls can be overcome, by mapping the online forum discussion to positive and negative comments thereby estimating the amount of for and against, the overall view can be obtained at the point of discussion itself.

3. LITERATURE SURVEY

When a particular fact is searched for, hundreds of thousands of outcomes appear. Users do not have the determination and extends to go through each and every page listed. So search engines might have a massive job role in sorting out the results, in the order of interest to the user within the first page of appearance and a quick encapsulated abstract of the information provided on a page.

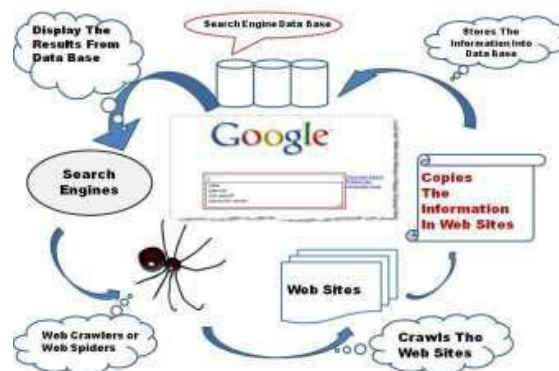


Fig - 1 Components of a Web Search System with Web Crawler.

Recovering effective content from the Web is a crucial task because it heavily influences the recognized effectiveness of a search engine. Users often look at only a few top rated hits, making the precision achieved by the ranking algorithm of supreme importance. Early search engines designate the pages predominantly based on their lexical similarity to the query. The key strategy was to devise and crawl the best weighting description to represent Web pages and query in a vector pointed space, so that intimacy in such a space would be correlated with semantic relevance.

4. PROPOSED SYSTEM

This paper proposes a way to monitor Internet forum, which has the capability to process the English language. It provides a way to let user monitoring a large amount of hot discussion threads by performing Text Mining to simplify the process and get accurate results.

4.1 MODULE ENCRYPTED

1. Crawler

For each forum type, a special crawler is written to crawl. The crawler fetches pages and identifies the page type using URLs. Once the type is identified, only the essential information are extracted and processed. Crawler also has to handle login process to unlock privileged posts. The extracted data are then tagged as title, content, author etc. and are stored into data base.

2. Database

Database stores the structural information of internet forum which is processed and sent by the crawler. The database provides data for creating statistic report and text mining. It also provides crawler information about the topics crawler currently processing.

3. Statistical Engine

Statistical Engine is used to create meta-summary information related in a fixed time interval. Using the meta-summary data, user interfaces, alert system and the data mining engines can easily locate popular discussion thread according to their need.

4. Data Mining Engine

Data mining engine performs text-mining on discussion threads which are stored in the database. It extracts the words from the text content recently added. These words and the known words in dictionary are then used to create profiles for users. Clustering algorithm is applied to classify user into different groups by their profiles and creates a summary on recently trends on forums.

5. User Interface

The user interface is used to show user recent hot discussion threads, especially for those which contain some keywords under monitoring. The interface provides user with current trend to give user ideas on the words which are interesting but not appears in the sensitive word list, finally the user interface also provides a way to use the user profiles collected to spot interesting individuals and their activities on the forum.



Fig – 2 System Architecture

4.2 STOP WORD REMOVAL

A) Identify the Stop Word

To understand whether the chat is 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.

B) Removal of Stop words

If the stop words are found in the chat, then they are analysed and if the word is suspicious, then these words are removed and blank space or asterisk is seen in chat.

C) Sentiment Analysis

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

D) Monitoring Suspicious Discussion

The chat is monitored by the removal of unexpected or suspicious words and hence we get the chat, free from the suspicious words.

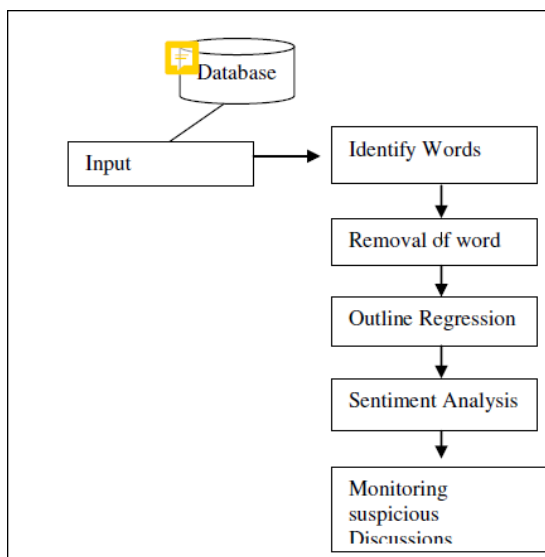


Fig – 3 Stop Word Removal Procedure

4.3 SENTIMENT ANALYSIS

Sentiment Analysis is completely based on feelings such as attitude, emotions and opinions. But it never supports the subjective facts. Generally, a binary opposition in opinion is assumed. For example: for/against, good/bad, like/dislike. Using NLP (Natural Language Processing), statistics, or machine learning methods to extract, identify or otherwise characterize the sentiment content of a text unit. Sometimes referred to as *opinion mining*, although the prominence in this case as withdrawal.

Page No: 65

A) Terms presence and frequency

B) **Parts of speech (POS)** This feature find adjectives, as they are important designators of opinions.

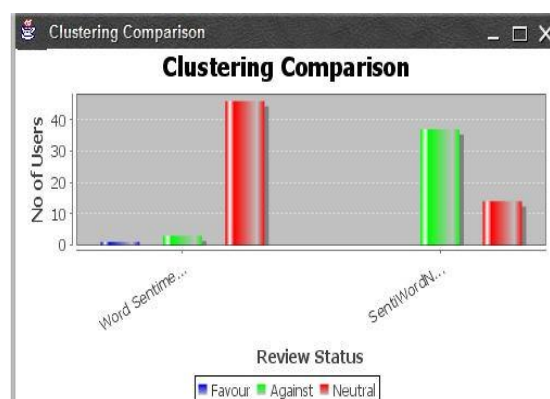
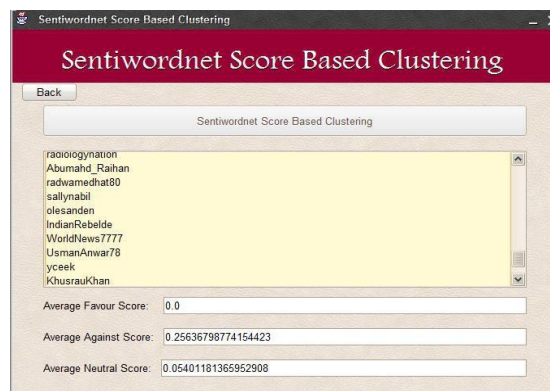
D) Negations The appearance of negative words may change the opinion inclination like not good is equivalent to bad.

The image displays four sequential screenshots of a web application interface for text analysis.

HomePage: The interface features a green header with the title "HomePage". Below the header, a message states "Choose any one method for Monitoring Suspicious Detection". A dropdown menu is open, showing two options: "Based on Twitter Reviews" (highlighted in blue) and "Based on Datasets". Below the dropdown are two buttons: "Submit" and "Clear".

Preprocessing: The interface has a green header with the title "Preprocessing". A "Back" button is on the left. A central text area contains a list of words: "stcr", "httpscogdz4mrc5s", "httpscocedcgbf1e", "sharaman", "chal", "jaa", "bacha", "thoda", "bada", "home", "ke", "baad", "baat", "karne", "aaji", "us", "man", "burning", "car", "helped", "save", "httpsc05mftmygm", "mdaya3", "pstco6c6kimbkg", "pstcoyh3feacyl", "bsvrx", "072015", "bomb", "blast", "chithirai", "veedhi", "temple", "2016", "foods", "httpsc08cm3zzrx3", "orgot", "tweet", "terrorism", "religion", "httpsc04thcknjghi", "t", "daynil", "district", "mogadishu", "httpscotc60p0xbes". To the left of the text area is a "Symbols Removal" button, and to the right is a "Stop Words Removal" button. Below the text area is a "User Clustering" button.

Word Sentiment Based Clustering: The interface has a green header with the title "Word Sentiment Based Clustering". A "Back" button is on the left. A central text area contains a list of words: "RadioPakistan", "radvamedhat80", "journoamanish", "sallynabi", "afrika_report", "khadarHared", "GolanSh", "_BeingSaniya", "yahayasa39ad". To the left of the text area is a "Word Sentiment Based Clustering" button. Below the text area are three buttons: "Average Favour Score", "Average Against Score", and "Average Neutral Score".



6. CONCLUSION

In this paper the detailed explanation of how suspicious analysis on online forum can be done using Text Mining was discussed and the system was tested with proper data. The system results in regular timely preparation of the required output. In comparison with existing manual system. It is very easy to obtain the number of supporters for the thread under discussion. The entire system is certified and can be easily understood by the end users. The forms are very user friendly and also it is very easy to handle even by the starters with very little effort and guidance.

REFERENCES

- [1] M. Suruthi Murugesan, R. Pavitha Devi, S. Deepthi, V. Shri Lavanya, and Dr. Annie Princy PhD: "Automated Monitoring Suspicious Discussions on online forum by data mining" Imperial Journal of Interdisciplinary Research Vol-2 ISSN: 2454-1362.
- [2] Priyanka, B. Hulde, Prof. Priyanka Dhudhe. Department of Computer Science & Eng, Jhulelal Institute Of Technology, Lonara "Monitoring malicious discussions on online forums using data mining".
- [3] Jashanjot Kaur, Department of Computer Science and Engineering "Stop words removal and its algorithms based on different method". ISSN No. 0976-5697.
- [4] Zhou, L. and Chaovalit, P. (2008). "Ontology-supported polarity mining" Journal of the American Society for Information Science and Technology, 69:98-110.
- [5] Wiebe J. M., Wilson T., Bruce, R., Bell M and Martin M. (2004) "Learning Subjective Language. Computational Linguistics" Journal of Natural Language Engineering, 4:135-144.
- [6] Benamara F, Cesarano C, Picariello A, Reforgiato D and Subrahmanian V (2007). "Sentiment Analysis: Adjectives and adverbs are better than adjectives alone" In Proceedings of the International Conference in Weblogs and Social Media (ICWSM).
- [7] Das, S. and Chen, M. (2001b). "Yahoo! For Amazon: Sentiment Parsing from small talk on the web" Processing of the 8th annual conference of the Asia Pacific Finance Association (APFA).
- [8] Esuli, A. and Sebastiani, F. (2006). "Sentiwordnet: A publicly available Lexical resource for opinion mining" Proceedings of the 5th Conference of Language Resources and Evaluation (LREC).
- [9] Grimes S (2008). "Sentiment Analysis: A focus on Application".
- [10] Hiroshi K., Tetsuya N., and Hideo W. (2004). "Deeper sentiment analysis using machine translation technology" Proceedings of the International Conference on Computational Linguistics (COLING).