Monitoring of Suspicious and Fraudulent Activities on Online Forums

Megha K B
Dept. of Computer Science & Engineering
Sahyadri College of Engineering, Mangalore

Nishan B
Dept. of Computer Science & Engineering
Sahyadri College of Engineering, Mangalore

Navya Prabhu M
Dept. of Computer Science & Engineering
Sahyadri College of Engineering, Mangalore

Nithin Thomas
Dept. of Computer Science & Engineering
Sahyadri College of Engineering, Mangalore

Shailesh Shetty S
Asst. professor, Dept. of Computer Science & Engineering
Sahyadri College of Engineering, Mangalore

***Abstract*- 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—Datamining, sentimental analysis, forums, Stop word Selection, naive bayes

I. INTRODUCTION

World Wide Web has become a very effective channel for many end users to share their knowledge and express their views. And also, by publishing data through a browser interface, stimulate their products or even educate each other. It was found that a great deal of first-hand news was discussed in online forums before it was reported in traditional mass media. As internet technology has been spreading its hands, this technology led to many such illegal activity.

Data mining interest has also increased miraculously. To expose the hidden data, it is important to extract useful data from the plain text form data. The main objective of data mining is to extract information from large data set and change it in a human knowable format. Data mining, which is the extraction of secluded predictive information from huge databases, is an influential fresh technology with mighty potential to aid organizations focus on the most vital information in their data repository. Organizations are now using data mining techniques to assess their databases for

ongoing trends, relationships, and are now using data mining techniques to assess their functionality.

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.

The rest of the paper is organized as follows: Section-2 presents the Background & Related Work, Section-3 presents the Methods, Section-4 presents the Results and Section-5 presents the Conclusion.

II. BACKGROUND AND RELATED WORK

A. Sentimental Analysis

Sentimental analysis is provisional text mining that searches and obtains intuitive information in source document and helps a company understand its product or service social sentiment while analysing online conversation. However, analyzing streams of social media is usually limited to analyzing basic feelings and counting based metrics.

B. Related Work

In [1] various techniques like stop-word selection,stemming algorithm,Brute-force algorithm are explained in detail .In [2], the system will examine the data from different discussion forum and separates data into different groups i.e. legal and illegal data using Levenshtein algorithm which measures similarity between two words. In [3] ,with the help of web mining large number of web pages are crawled to obtain the dataset. It makes use of user interactive query interface to determine crime hotspots. Also,involves techniques like clustering and classification to identify similar item and association rule mining,sequential pattern mining to obtain frequently occured set . In [4] the developed framework detects emotion using emotext and gives the output in the form of csv file. Each input data set has its own text id and predicted label. In [5], paper presents a method which makes use of different classification methods and hybrid classification on multiple classifiers which increase performance and effectiveness.

III. 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

A. 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.

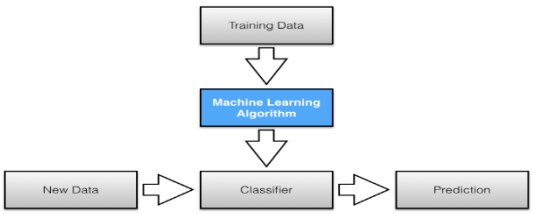

Fig.2. Naïve Bayes classifier

B. WORKING

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.

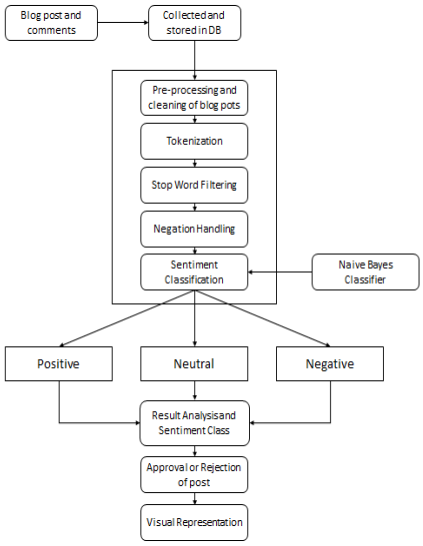

Fig. 3. System Architecture

IV. RESULTS

The system we developed will categorize the discussions in a typical forum like the one we used in our experimentation into spam or ham. This categorization of user posts and comments into spam and ham is performed through various text mining techniques we discussed in the paper. A graph showing the time taken to process a given text with respect to the length of the data is given in Fig. 4

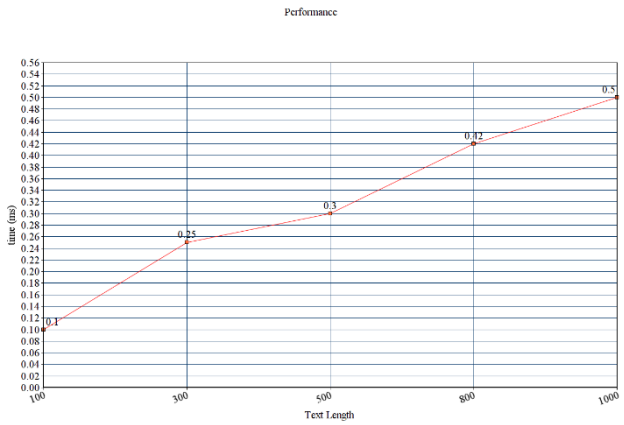

Fig.4. System performance

CONCLUSION

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

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

[3] 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.

[4] 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.