International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Natural Language Processing for Categorization and Adult Content Analysis of Social Post into Text and URL Level

Ms. Geetanjali Salunke[1], Mrs. S.S. Patil[2], Mrs. S. S Dhotre[3]
[1, 2, 3]*Department of Computer Engineering , Bharati Vidyapeeth (Deemed to be University),College of Engineering, Pune.*

Abstract**: Use of internet is growing very rapidly in daily life. Different sites are used by internet user to get in touch with each other such as twitter, facebook etc. Among these sites, facebook is the greatest rising site. Internet users expend most of their time on social networks like facebook, twitter etc. Online social network greatly depend on users for getting content and sharing. Information is spread on the social networks in fast and efficient way. Facebook user use facebook to update their status, find new friends and share opinions with friends. Though, at the same time social media networks turn into vulnerable to various types of unwanted adult content such as text messages and URL. Facebook is the most popular online social network for sharing information and content through internet. Most of the facebook user shares different content such as images, videos, messages, opinions and URLs on his wall with their facebook friends. At the same time facebook user post adult messages and URLs on their wall and share with their facebook friends. Social Networking Sites has become an important part of our daily life. Peoples used different OSN for long time to share their opinions, comments, ideas or thoughts and giving any speech about current position. Facebook social site has a many characteristics through which people attracted more towards it, but impact of facebook on our social life may be positive or negative. Negative impact of facebook is spreading adult content on facebook. So there is need to detect adult content on facebook for improving facebook user experience. It's also important to extract opinion in post about different topics for avoiding bad impact on social network users. In this paper we have proposed a method for detecting adult content on facebook at text and URL level. It also extracts opinion from post to decide category of each post.
Keywords: Facebook, adult conent, URL analysis, machine learning etc.

I. INTRODUCTION

Different online social network sites like facebook, twitter, and google+ are experiencing unbelievable development in users. More than millions of users are now active on these sites. In addition immediately creating a profile and connecting with friends, the social networks are now building platforms to run their website. These social sites are presently becoming an example of online communication which makes the use of user's personal information and actions in social links for different services. The social networks are trendy way of communication of the internet users. People are greatly depends on online communications. The internet is giving special options to user to make and keep contacts, relations with other online users. Another way of attack by cybercriminals is the abuse of videos, images and links presented by the user. Cyber attacks mainly happen on social networks. Many of sites such as facebook and twitter presently have millions of active users. Due to increasing use of social media network cybercriminal misuse social network. By clicking upon post it will take the user to some different pages created by malicious user. Cybercriminals builds attractive posts that are really attracts some users. Attacker posts mean post at the time of particular actions and events. Criminals upload adult posts which are connected to different events and misguide users to click those links. Another type of attack is to like the facebook page of user profiles without their information. Detecting malicious URLs is now an important task in network security intelligence. To continue effectiveness of web safety, these malicious URLs have to be detected, identified as well as their corresponding links should be found out. The content contains mean data posted in different form on other user wall. These can be adult content or post and free downloading sites [1].

Opinion targets are nothing but topics on which opinions are expressed. Opinion targets extraction are significant because without identifying the targets and opinions expressed in a sentence are of incomplete use. For example, "I am not happy with the battery life of iPhone". In this example "battery life" is the target of the opinion. Opinion target (i.e. topic) extraction is a complex task in opinion mining. On twitter user expresses the opinions and interests of people in different topics and areas [2]. Facebook has become a part of the daily life understanding for a growing number of people. Facebook is a web supported services which is allowing individual users to build a public profile in a restricted system. Social networking site is useful to individual user to share

connections, views, opinions, thoughts with unidentified friends or allow to sharing their view points with observable well known friends. Many of users uses facebook social site for innumerable activities. Amongst the mainly common used are connecting with existing networks, building and rising friendships, generate an online presence for their users, observing content, finding information, building and modifying profiles and so on. Social networking site like facebook that allow persons to create a public of semi-public profile within an enclosed system, make a list of other users with whom they share a connection and observe, pass through their connections list and those made by others inside the system. Using facebook, user gets a platform through which they can share various information, chat, opinions, videos, pictures, upload their profile photo and posted comment on facebook wall. Actually facebook has great attractive characteristics which contain celebrity followers, tollywood, bollywood or hollywood supporters. This social networking site makes different type of cybercrimes which guide to demotivate the people and sometime they may be involved into various illegal or unprofessional works which are considered to be a bad feedback of facebook site. Some bad impacts have been founded at the time of survey and from different literature reviews. The majority of the facebook users are uses facebook just only to verify their profile or a few new update posted on their wall. So most of the users are very busy through this site and they are busy in chatting with their friends for a long period of time. They waste their time during day without doing anything. Using facebook site photos, views, videos or political share by user might distract attention and it creates a bad impact on users carrier life. Spending extra time on facebook it makes people more selfish and most of the people they are more interested in finding and reading information of their non-friends users. They used this site to spread misinformation and perform various types of cybercrimes such as doing sexual and mental crime with unknown friends through chatting. For a time giving a lot of information in facebook profile it increase the risk of detecting theft and it may be unsafe for user status [3].

In the present day people are so greatly comfortable to social networks. Because of this, it is very simple to spread spam contents through them. Anyone can get the detail information of any person very simply through these sites. Not everyone is secure within the social media. The attacks on sites are also increasing very rapidly. Using social sites adult content propagators spreads links which containing adult contents. Clicking on this link by user, the user will be forwarding to malicious sites [4].

Process of extracting opinion targets and opinion words is mostly used for bad comments removal from facebook account. The main use of the facebook is to share communicate with the whole friends at any time. But now facebook account is used to extend the bad opinions and opinion in other topics. It may direct major problems and affect the privacy of people. Social media is one of the largest mediums to state opinions. Sentiment analysis is the process from which information is extracted from the opinions, reviews and emotions of people in related to things, actions and their features. Opinion Mining is to analyse and classify the data generated by user like reviews, comments, blogs, articles etc. [5]. Opinion target is known as an object on which user express their opinions, generally nouns or noun phrases [6]. It is frequently difficult to understand what trending topics are about. Various efforts are being prepared to classify the topics include within comments, opinions and views into general categories with more accuracy for improved information retrieval [7].

It is very exciting to identify what topics are trending and what people in other parts of the world are attracted in. However, a very large proportion of trending topics are denoted by hashtags (#), a name of an individual, or other languages words and it is regularly hard to know what the trending topics are about. So it is very essential to classify topics into general categories for easier understanding of topics and enhance information retrieval. The names of trending topic may or may not be representative of the type of information people are tweeting regarding unless one reads the tendency text associated with it. For example, #happyvalentinesday shows that people are telling about Valentines Day [8]. Web site is responsible for performing online surveys with one million users in a one-hour timeframe [12]. With the fast development of Internet, more people obtain useful information through the Internet. The internet is an open system, a vast number of information is produced and update onto it every day [14].

II. LITERATURE SURVEY

Combination of natural language processing and machine learning techniques is used to identify opinion targets in tweets and to decide category based on topic to which they expressed. It gives better result of analysis of tweet posts [2]. Natural Language Processing (NLP) is a technique which allows a machine to process a natural language similar to english. NLP carry out information from unstructured information [9]. Using NLP malicious tweets are detected. Imperfect sentences in a tweet are identified. Removal of stop words and stemming concepts of NLP was used for spam and malicious post detection. In machine learning techniques machine can be trained on its own. In this techniques training set was used for analysis of tweet data [4]. It is difficult to finding trending topics from information or opinion discussed in tweet. Author proposed solution of this problem using different predefined categories like sports, politics education with the help of Bag-of-Words text classification and information of social network i.e. twitter [8]. Author proposed solution to detect adult account on twitter using URL blacklisting techniques i.e. comparing extracted

URL with set of blacklisted URLs. Adult content detected on text based using natural language processing and machine learning techniques. It is useful for detection of adult web pages [1].

III. PROPOSED METHOD

Flow of our proposed system is shown in figure (1). We make system for adult content detection and to decide category of each post on facebook social site. Our main goal is to detect adult content at text and URL level. In this proposed system we extract posts from facebook social site using facebook API [15]; if extracted posts are text messages i.e. simple words or compound sentences then using segmentation method sentences are divided into small sentences. After segmentation process we apply NLP (natural language processing) task i.e tokenization, stemming and stop words removal to make data clean and text filtering. Tasks of NLP are described briefly in following section. After NLP task we trained system for adult text detection using dataset of adult words. If adult text is identified in extracted data then system gives result as adult content. Result shows within graphical notation. If extracted data are URLs then our trained system use set blacklisted URL [16] for adult URL detection. If extracted URL found in dataset then system displays result of URL as adult URL. To decide category of each text post of facebook system performs NLP task and opinion target within extracted posts. We use some set of predefined categories to decide category. Result of adult content detection and categorization are shown by using graphical notation with score of each post.

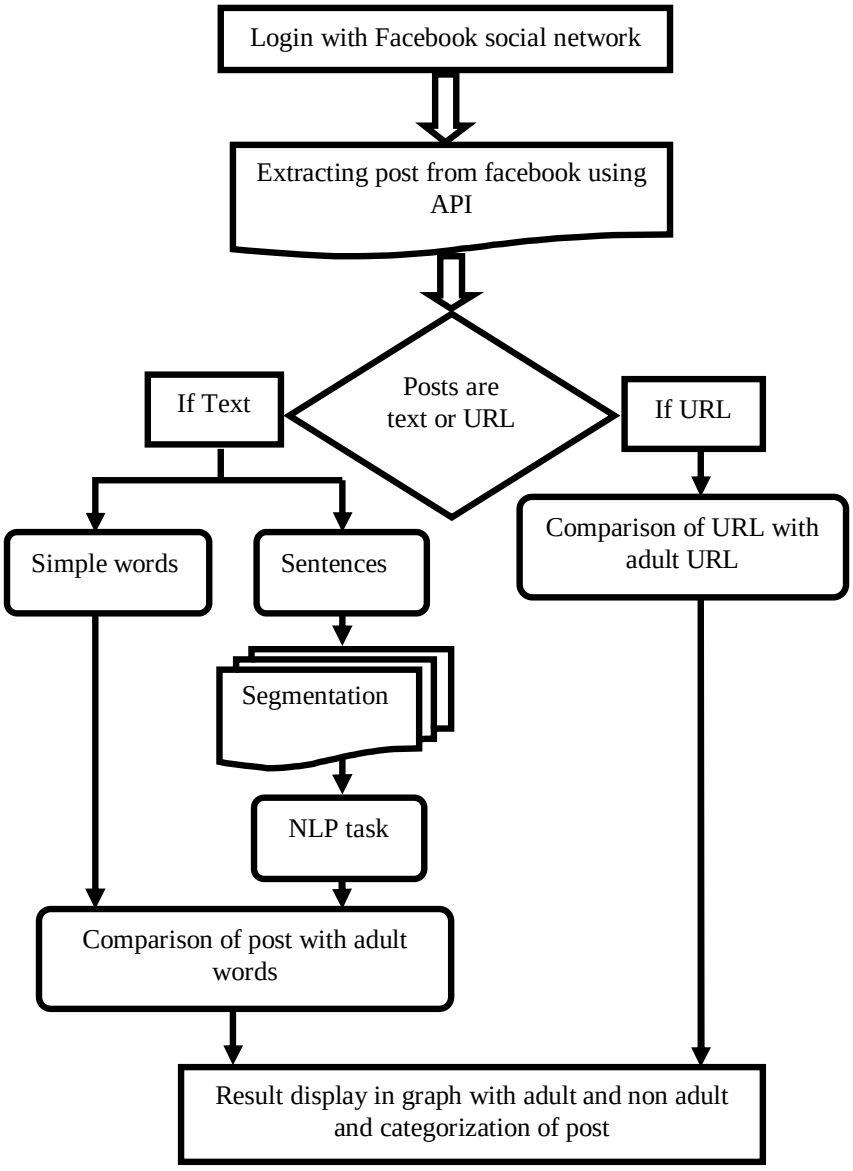

Figure 1. Proposed system

A. Data Pre-processing:

Pre-processing perform very important task in text analysis. It is used to clear the extracted posts (text). It removes unwanted and noisy data[9].

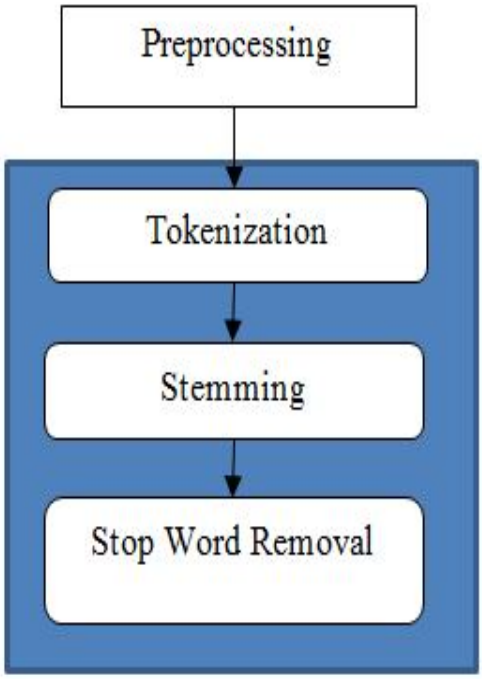

Figure 2. Preprocessing process

B. Tokenization

Tokenization is the process of dividing up sentences into pieces such as words, phrases, keywords, symbols and other elements called tokens. Tokens are may be individual phrases, words or even entire sentences. In this process some characters such punctuation marks are discarded. Tokenization helps to separate the textual information into words. In our system we use tokenization to divide each sentence into small sentences for text analysis and opinion mining [13].

C. Stemming

Stemming is the process of counting words to a few base forms. Stemming finds various terms against one base form and it is then used as a term [10]. This technique is used to find out the root or stem of a word. It converts words into their stems.

D. Stop-word removal

Stop-words are frequent words that hold no information such as pronouns, prepositions and conjunctions. In this process it removes words that are not having meaningful information and not useful in text analysis [11]. In our system we create our own stopword list for categorization and text analysis.

E. Segmentation

We used segmentation process to divide large i.e. compound sentences into small sub sentences to detect adult content. It is also useful to determine opinions and opinion targets from each splited sentence. Using segmentation we can show that category of each (text) post into topics based on opinion target.

F. POS (Part-Of-Speech) Tag:

Part-Of-Speech tagging performs significant role in NLP to process natural language. In text analysis we apply POS tag on each word to identify opinions of users and opinion targets which appear in posts. It used to find out nouns, noun phrase, adverbs, adjectives etc.

G. Experimental setup:

Experimental setup	Description
Web Browser- Mozilla	For online facebook social network
Facebook account	User should have account on facebook social network.
API- www.developers.com	It is Facebook API for extraction of facebook post. User should account on facebook developers.com.
www.urlblacklist.com	Set of blacklisted URL for classification of extracted URL into adult URL.
Categorization Key	Set of different categories are used for topic based categorization of extracted facebook post (text)
Adult words (text)	Set of adult words are required for classification of extracted post into adult and non adult post (text).
Stop words	Different stop words are used for classification and categorization of post (text) like able, about, above, abroad, according etc.

Table 1. Experimental setup of proposed system

IV. RESULTS

In figure 3 we have represents the results of adult post detection at text and URL level. In our result we have create graphical representation by counting number of adult and non adult post of users at text level and counting adult URLs in post. In graphical representation X-axis represents adult and non adult class and Y-axis represents count for the adult and non adult post.

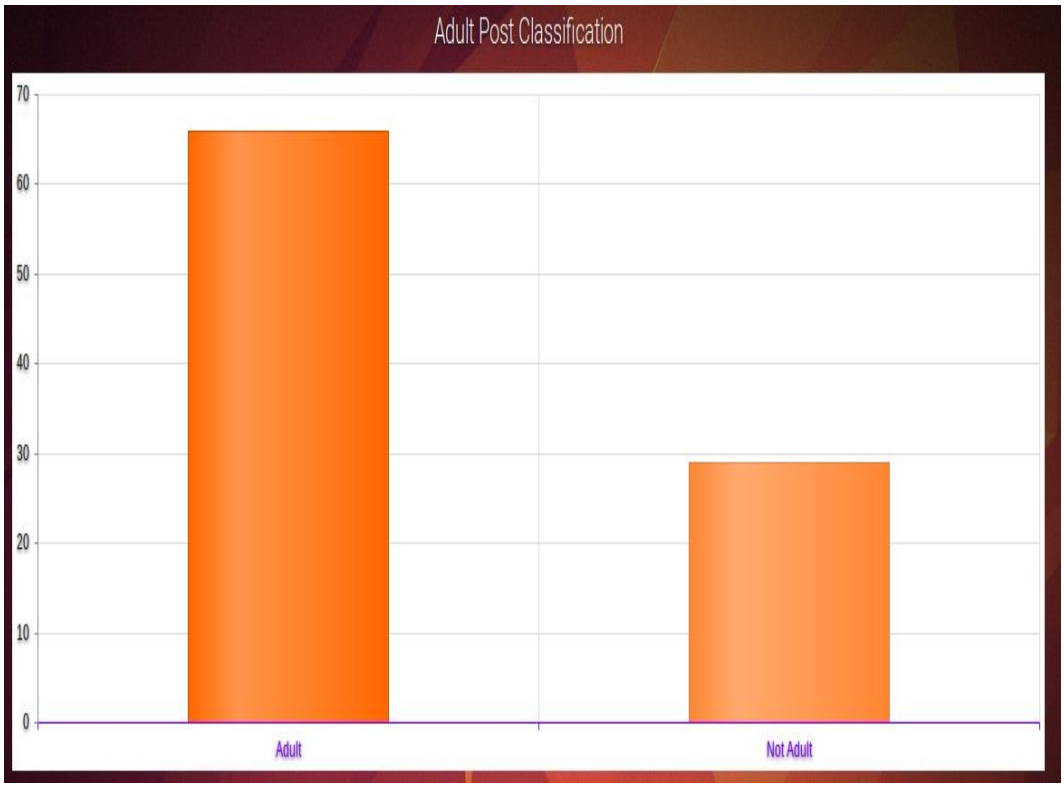

Figure 3 Graphical result of adult and non adult post

Equally figure 4 shows topic based opinion extraction; in this we are representing all topics of that particular text post by considering opinion on that post, here x-axis represents number of categories and y-axis represents count of categories based on topic. We have generated results for all post of different topics.

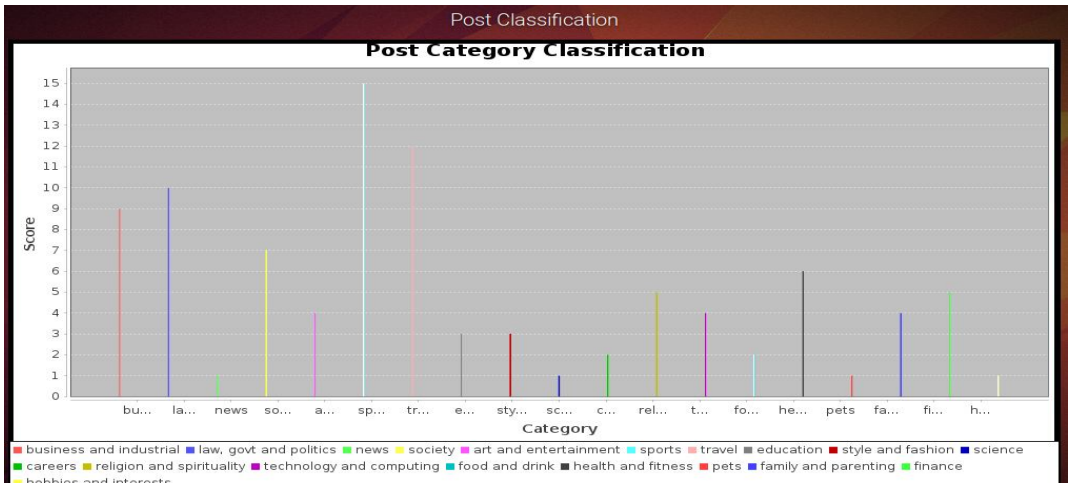

Figure 4 Graphical result of topic based categorization of post

V. CONCLUSION

In this paper, new approach for automatically detecting adult content with natural language processing technique is introduced. This approach categorized post into different topics based on opinion target. In this concept, NLP is used to get the required data and also used text mining, machine learning method and some dataset to calculate the results. The result shows that posts extracted from facebook social site can be adult and improve users experience. It categorized extracted post for avoiding bad impact on social network users.

ACKNOWLEDGMENT

This is to acknowledge and thank all the individuals who played defining role in shaping this paper. Without their constant support, guidance and assistance this paper would not have been completed. Without their coordination, guidance and reviewing this task could not be completed alone. I avail this opportunity to express my deep sense of gratitude and whole thanks to my guide Mrs. S. S. Patil for giving her valuable guidance, inspiration and encouragement to embark this paper.

REFERENCES

[1] Hanqiang Cheng, Xinyu Xing, Xue Liu, and Qin Lv "ISC: An Iterative Social Based Classifier for Adult Account Detection on Twitter", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 4, pp. 1045-1056, April 2015.
[2] Anna Stavrianou, Caroline Brun, Tomi Silander, Claude Roux, "NLP-based feature extraction for automated tweet classification".
[3] Indrajit Roy chowdhury , Biswajeet Saha, Jagadish Chandra Basu, Bhairab Ganguly, "Impact of facebook as a social networking site (SNS) on Youth generations: A case study of kolkata city" International journal of humanities and social science invention, volume 4 Issue 6, PP.28-42, June 2015.
[4] R.Hemavathi, B.Benita, Mrs.D.C.Joy, WinnieWise, RS.J.Doulas, "An Enriched Method for Opinion Target and Word Extraction Using Semantic Labeling" International Conference on Emerging trends in Engineering and Technology, pp.56-63, 2016.
[5] Jyoti S Bhoosanurmath, Gambhir Halse, M.S Sheshgiri KLE, M.S., "Extracting Opinion from Reviews for Better Analysis of Products and Services", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.
[6] Mohit Tare, Indrajit Gohokar, Jayant Sable, Devendra Paratwar, Rakhi Wajgi, "Multi-Class Tweet Categorization Using Map Reduce Paradigm", International Journal of Computer Trends and Technology
[7] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary, "Twitter Trending Topic Classification".
[8] Neeraja M, John Prakash, "Detecting malicious posts in social networks using text analysis", International journal of science and research, volume 5 Issue 6, June 2016.
[9] Thorsten Brants, "Natural Language Processing in Information Retrieval", Google Inc.
[10] Advanced Natural Language Processing Basic Text Process, 2010.
[11] Dharmendra Sharma, Suresh Jain, Mewar University, Chittorgarh, Rajasthan, "Evaluation of Stemming and Stop Word Techniques on Text Classification Problem".
[12] Mrs. S.S.Patil, S.D.Joshi, "Identification of performance improving factors for web application by performance testing", International journal of emerging technology and advanced engineering, vol. 2, pp. 433-436, 2012.
[13] Ms. Geetanjali salunke, Mrs. S.S.Patil, "A research effort to categorize posts applying two phase natural language processing methodology", International journal of advanced science and technology, vol-101, pp. 1-12, 2017.
[14] Dhotre S. S. "Intelligent E-learning system using Web 3.0.," Journal of Engineering Research and Studies 1.2 (2010): 230-232.
[15] http://developers.facebook.com.
[16] URL Blacklist [online]. http://urlblacklist.com.