DETECTION OF OFFENSIVE WORDS OF SOCIAL MEDIA POSTS

[1]Pritee Tambat, [2]Vrushali Hajgude, [3]Shilpa Golasangi, [4]Namratha Rupde, [5]Prof. Megha P. Kharche
[01-4] Student, Computer Engineering Department,
K J College of Engineering and Management Research, Pune, India
[05]Assistance Professor, Computer Engineering Department
K J College of Engineering and Management Research, Pune, India

Abstract : This work is about ignoring the adult post on Facebook from the owner of the account. The system is organized on the origin of the post which chooses the account is adult or not, associated with adult words which are stored in the file. Data are obtained from the Facebook and it is distributed on each block of memory and each block has token and it will process according to the token, data is processed for words stop filtering.

IndexTerms - Sentimental Analysis, social media, cyber bullying, adolescent safety, offensive languages**.**

I. INTRODUCTION

Now a day's teenagers are spending their lots of time on social media to share data and to connect with each other. They learn lots of interesting and useful things from this but at another side it having many risks likes social media consist heavy amount of offensive contents so adolescents can get distracted by this.

It is found that 80% of contents includes offensive language, 74% of contents include porn-like images, videos or offensive words, etc. As teenagers are more likely to get affected by bad contents than adults, detecting online offensive contents to protect the adolescents' online safety becomes an urgent task [1].

To solve this problem this project is been proposed "Detection of the offensive language of social media posts". In this project Social media like Facebook is been created in which user can upload photos, can send a friend request, can accept a friend request, can comment on others post and also can chat. If a bad word is detected in message or comments then the word is been replaced with another word by using Core Natural Language Processing and Levenshtein and classification is been done by Naïve Bayes algorithm so this can free offensive word and adult will be free from such contents.

Paper is organized as follows, **Introduction – Related work - Research Method - Results and Discussion – Conclusion**.

II. RELATED WORK

Sadaf Khurshid, Sharifullah Khan, Shariq Bashir [2], In year 2014, proposed a project named "Text-based intelligent content filtering on social platforms" by using Machine Learning algorithm to clean noise from data. Sentiment Analysis is used to classify on based on positive, negative or neutral. Feature engineering is used to make predications on a piece of data. Naïve Bayes classifier and Decision Tree to evaluate but it has limitations which directly impact on the result

Snehal B. Shende, leena Deshpande [3], proposes a "Computational framework for detecting offensive language with support vector machine in social communities" uses Online Communities using Grammatical Relations by Zhi Xu and SencunZhumore focuses on the filtering technique. Offensive word lexicon to detect the actual offensive words. The Naive Bayes is supervised to predict the unique feature which is unrelated to the other feature. N-gram technique As weka or LIBSVM will not understand the text sentence but can understand a set of features in form numerical data say 1,-1

Shuhualiu and Thomas forss[4], proposes "Text Classification Models For Web Content Filtering and And Online Safety" in 2015 uses Topic extraction step takes web textual information as input and generates a set of topic terms. Quality of raw text input and its representation has an immediate and crucial effect on the result of the feature extraction. Naive Bayes models are calibrated to the training distribution, thus changes in the distribution which naturally affects model performances.

Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Steve Maybank[5], propose a "Recognition of Pornographic Web Pages by Classifying Texts and Images" utilizes a novel system for perceiving explicit Web pages. In the context of texts, images, and content, representations of pages are considered jointly by the naive Bayes classification to recognize discrete

pornographic texts. Fusion algorithm, based on Bayes theory. The limitation is that if both detectors find porn content, the web page is classified as porn.

Munish Chopra, Miguel Vargas Martin, Luis Rueda, Patrick C.K[6], Propose "Toward new paradigms to combating internet child pornography" uses Host-based Approaches to The associated complexity of the problem is to be able to retrieve visual files based on their semantics and Combating Child Pornography at the Network Level this uses two approaches, Existing Classification Approaches to propose a network intrusion detection system based on n-grams analysis. And second Promising Network Infrastructure Approaches to which may lead to efficient child pornography detection in routers are, but the drawback is time a consuming process and is an online process

Mingwei Qinghua XueZhi Yong Cui[7], propose "Evading User-Specific Offensive Web Pages via Large-Scale Collaborations" uses OWP prediction algorithm concerns with one-page attribute of one website. Ratings to pages on one website should only be used to predict OWPs on the same website. Though it cannot result in ZERO false positives, it does greatly reduce that especially when large-scale dishonest collaborators exist.

III. Research Method

Methodology

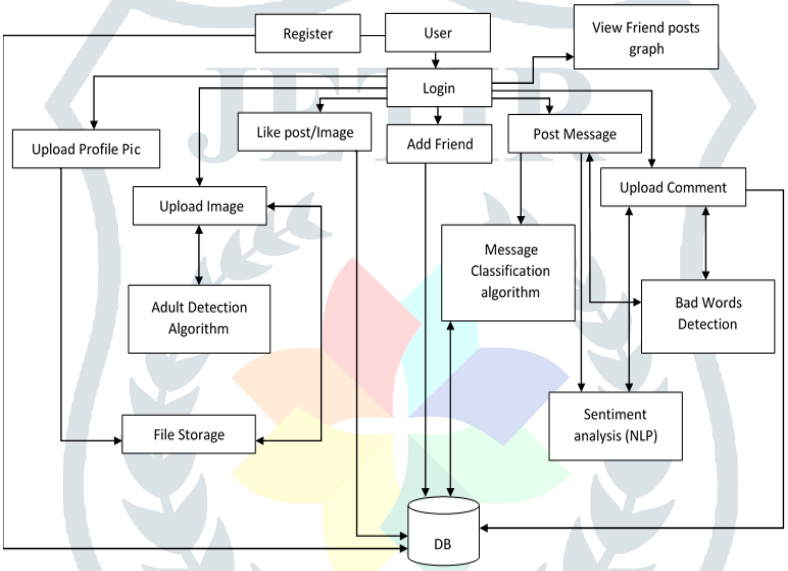

Figure 1: Architecture Diagram

The user had to register their account by providing their personal details like Fist name, Last name, City, valid email id and password on a social media application and next login to their account after login user will able to see their profile. User can also post their comments and also can post images, which is being added in the database. Profile can be uploaded or updated by using file storage, Adult detection algorithm bad words, offensive post, bad (offensive) message can be detected using Natural Language Processing and by using some steps as Sentiment analysis, post comments can be detected be classifier algorithm and classification is based on whether the word is bad or not. Function like login registration performed by the user is been stored at backend od Database.

3.1. Stanford CoreNLP

Stanford CoreNLP [8] is the library which is provided by Stanford University and it is open source to use. Range of technology tools is given by Stanford CoreNLP. Stanford CoreNLP Recognize the statement on which it is been processed. In this project, Stanford CoreNLP performs some steps to detect the offensive word

Stanford CoreNLP is used to detect offensive word in a form on message and comments and then replace it with the symbol or good word. If a statement is having offensive word then by using first step Lexical Analysis, it is used to tokens or break the statement into its individual word and further given as input to Syntactic Analysis, it is used to analysis of word from paragraph or statement for arranging words that express the relation between words. Stop words removal, in this stop word like and, or, the &, etc. is removed to detect the offensive word from the statement. If a bad word is detected then will check into its dataset and remove that word replace it with the symbol [8].

3.2. Levenshtein

Levenshtein distance. Is also called as Edit distance. It is used to find the distance between two words. It is used to clear slang words from messages and comments. This algorithm will detect the slang word and then check into its dataset and then replace it with that word Example "plz msg me" will be converted to "Please message me". In this algorithm insertion-deletion and replacing these three operation is used to convert slang words to meaningful words.

3.2.1. Mathematical Model

S= {S, s, X, Y, T, f main, DD, NDD, f friend, Memory Shared, CPU count}

S(system): IS our planned system which comprises following tuples.

S (initial state at time T): GUI of Facebook post classification, adult content check and categorization using Hadoop. The GUI provides space to enter a query/input for user.

X (input to system): input query. The user has to first enter the query. The query may be ambiguous or not. The query also represents what the user want to search.

Y (output of the system): list of the URLS with snippets. The user has to enter a query into Detection of offensive language of social media posts, then Detection of offensive language of social media posts generates a result which contains relevant and irrelevant URL's and their snippets.

T (No. of steps to be performed): These are the total number of steps required to process a query and generates result.

Fmain (main algorithm): It contains the process p. process p contains the input, output sand subordinate's functions. It shows how the query will be processed into different modules and how the result is generated.

DD (deterministic data): It contains Database data. Here we have considered.
A Facebook post classification which contains a number of ambiguous queries. Such queries are used for showing result. Hence Facebook post classification is our DD.

NDD (non- deterministic data): No. of input queries .IN our system user can enter the number of the queries so that we can not judge how many queries user enters single session. Hence Number of Input queries are our NDD.

Ffriend: WC and IE. In our system, WC and IE are the friend function of the main function. Since we will be using both of the function, both are included in friend function.

Function WC is a web crawler and IE is information extraction which is used for extracting the information on the browser.

Memory Shared: Database. The database will store information like the list of receivers, registration details and number of receivers. Since it is the only memory shared in our system, we have included it in the memory shared.

CPU count: In our system, we required 1 CPU for the server and minimum 1 CPU for the client. Hence CPU count is 2.

Subordinate function:
Identify the process as P.
S= {I, O, P….}
P= {RC, IE}
Where,
WC is the online Web Craw, let technique where post from Facebook are being fetched.
IE is the information extraction.
P is the process.
WC ={U, MAX,CP}.

3.2.2. Algorithm

Step 1: Set n to be the length of s. set m to be the length of t. If n=0 return m and exit. Construct a matrix containing 0..m os and 0..n columns

Step 2: Initialize first row to 0...n and initialize first column to 0…m

Step 3: Examine each character of s(i from 1 to n)

Step 4: Examine each character of t(j from 1 to m)

Step 5: if s[i] equals t[j], the cost is 0
if s[i] doesn't equals t[j] cost is 1

Step 6: set cell d[I,j] of the matrix equal to minimum of:
a) The cell immediately above plus 1: d[i-1,j]+1.
b) The cell immediately l eft plus 1:d[I,j-1]+1
c) The cell diagonally above and to the left to the cost: d[i-1,j-1]+cost

Step 7: After the iteration step(3,4,5,6)are complete, the distance is found in cell d[n,m].

Step 8: Stop.

3.3. Naïve Bayes

Naive Bayes algorithm is based on Bayes theorem and also a collection of the classification algorithm, is used to classify according to the adult offensive word is present or not. It is used to detect offensive word from message and comments and classifies according to it Naïve Bayes uses conditional probability to classify offensive words. Naïve Bayes is also used to classify post on based on sports, Entertainment, news and etc. and shows in pie chart format

P(A|B)=[P(B|A)*P(A)] / P(B)

P(A|B) is the posterior probability of A given predictor B.
P(B|A) is the likelihood which is the probability of predictor given A.
P(A) is the prior probability of A.
P(B) is the prior probability of B.

IV. Results and Analysis

4.1 Data collection

Data get collected from online social media like tweeter and facebook. It can be chat history of two users, comments on posts, or can be a caption of a post. Teenagers can use offensive contents here.

4.2 Data preprocess

To preprocess, firstly the input dataset is taken as input. As the data on the online social network is highly unstructured there is need to pre-process the data before it applied to the actual classification model. The data contain stop words which will get removed in pre-processing.

4.3 Classification

Naive Bayes: The Naive Bayes is used for classification. Naive Bayes based on the probabilistic approach to predict the data. It can predict the unique features which are unrelated to the other feature to get classified easily. To find the accuracy using Naive Bayes firstly the classifier is trained and based on these trained data the prediction can take place.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

p (A|B) – the conditional probability of A given B
p (B|A) – the conditional probability of B given A
p (A) – Probability of A
p (B) – Probability of B

4.4 Identify the offensive sentence in the discussion

Firstly, the messages or comments are classified, based on the classification, After that, each sentence get parsed and detect the offensive content in the discussion of the particular user.

TABLE

User Id	Total Post per User	Offensive word used	Count
User1	1,1,1,-1,-1,1	F**K, Dumb, Ashhole	3

User2	1,1,-1,1	F**k Off,	1
User3	1,-1	Psycho, Basterd	2

In the above table, There are 3 users, first user have total 6 posts, post number 4 and 5 are offensive so it contains -1 and it uses total 3 offensive words in two posts. Similarly user 2 having total 4 posts in which post number 3 in offensive and so on.

4.5. Accuracy graph

This is graph for accuracy in predicting occurrence of offensive words in social media comments will have upper hand over others (Messages, post's caption, tweets)

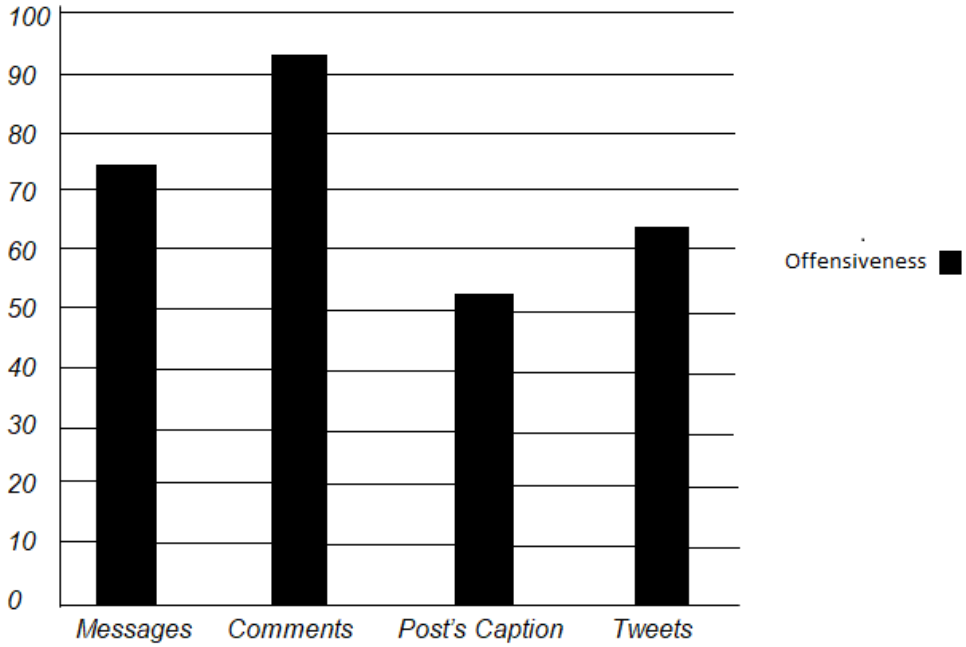

Figure 2:Offensive word occurrence and detection in percentage(%)

V. Conclusion

In this paper, it is proposed the system "Detection of the offensive language of social media posts". Offensive words are detected and the process is been done like tokenize or splitting statement into individual words and remove stop words from the statement and remove the words. Classification is done on based on adult words or not and classify the post on based on classes like sports, Entertainment, and etc and represent it into pie chart for admin. The Twitter post is accessible in this project and also the detection of offensive words can be performed. Concluding output will be based on adult content, replacing offensive word into a symbol by checking offensive word into a dataset which is stored in the backend

References

[1] Ying Chen, Sencun Zhu, Yilu Zhou, Heng Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety", ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk, and Trust, 2012.

[2] Sadaf Khurshid, Sharifullah Khan, Shariq Bashir, a "Text-based intelligent content filtering on social platforms",12th International Conference on Frontiers of Information Technology

[3] Snehal B. Shende, leena Deshpande, "Computational framework fordetecting offensive language with support vector machine in social communities", 2017, IEEE - 40222

[4] Shuhualiu and Thomas forss, "Text Classification Models For Web Content Filtering and And Online Safety",2015 IEEE 15th International Conference on Data Mining Workshops

[5] Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Steve Maybank, "Recognition of Pornographic Web Pages by Classifying Texts and Images",IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 29, NO. 6, JUNE 2007

[6] Munish Chopra, Miguel Vargas Martin, Luis Rueda, Patrick C.K, "Toward new paradigms to combating internet child pornography", IEEE CCECE/CCGEI, Ottawa, May 2006

[7] Mingwei Qinghua XueZhi Yong Cui, "Evading User-Specific Offensive Web Pagesvia Large-Scale Collaborations",This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

[8] StanfordcoreNLPhttps://books.google.co.in/books?id=t1PoSh4uwVcC&printsec=frontcover&dq=STANFORD+CORENLP&hl=en&sa=X&ved=0ahUKEwjhwMegi6ffAhWUfX0KHR77DhkQ6AEIODAC#v=onepage&q&f=false

[9] Levenshtein- en.wikipedia.org/wiki/Levenshtein_distance