IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems
ICIIECS'15

An Approach for Offensive Text Detection and Prevention in Social Networks

Mr. Shashank H. Yadav
Department of Computer Technology
Yeshwantrao Chavan College of Engineering
Nagpur, India
shashank.y89@gmail.com

Mr. Pratik M. Manwatkar
Department of Computer Technology
Yeshwantrao Chavan College of Engineering
Nagpur, India
pratikmm@ymail.com

***Abstract*— Social Network has become a place where people from every corner of the world has established a virtual civilization. In this virtual community, people used to share their views, express their feelings, photos, videos, blogs, etc. Social Networking Sites like Facebook, Twitters, etc. has given a platform to share innumerable contents with just a click of a button. However, there is no restriction applied by them for the uploaded content. These uploaded content may contains abusive words, explicit images which may be unsuitable for social platforms. As such there is no defined mechanism for restricting offensive contents from publishing on social sites. To solve this problem we have used our proposed approach. In our approach we are developing a social network prototype for implementing our approach for automatic filtering of offensive content in social network. Many popular social networking sites today don't have proper mechanism for restricting offensive contents. They use reporting methods in which user report if the content is abuse. This requires substantial human efforts and time. In this paper, we applied pattern matching algorithm for offensive keyword detection from social networking comments and prevent it from publishing on social platform. Apart from conventional method of reporting abusive contents by users our approach does not requires any human intervention and thus restrict offensive words by detecting and preventing it automatically.**

***Index Terms*—Component Social Network, Offensive Words, Pattern Matching, Filtration, Automation**

I. INTRODUCTION

Social network has become an integral part of every individual today. It has change the way we live in 21st century. At the starting of the last decade of the twentieth century, globalization had played an important role in connecting different people around the world. Peoples started understanding each other by communicating with each other. They started working together to achieve a common goal. These scenarios have changed drastically at the beginning of the 21st Century.

With the advancement of the Information Technology a new era has begun. This era shrinks the world with a button click. With the new network enabled technology the information started to share at a lightning speed. After the launch of the Facebook and Twitters people get a common platform where they can share their views with others. However this common platform i.e. Social Networking Sites (SNSs) has given a free access to the resources available with us. With the advancement of the 3G and 4G technologies this information reaches to the users around the world in a blink of an eye. People started sharing their thoughts, views about any topics without any restrictions. These SNSs does not filter any of the comments and status which is uploaded in public platforms. However they have implemented reporting systems where user can report the contents as abuse and the content is them removed from social platform.

But these SNSs do not specify that which no. of reports will make the contents as abusive. For examples, Facebook has many employees working on the contents which being uploaded daily on users walls, profiles, etc. This employee's manually checks the contents which are reported as abusive. Twitter also requests their users to not follow the peoples if they found the content of that user as offensive. None of the above sites has provided any security mechanism at their levels i.e. Server side. So this gave rise to the idea of our proposed approach. In our approach we have developed a prototype of social network such as Facebook to implement our automatic filtering based mechanism.

In our approach we mainly focus on the status and users comments to detect and prevent offensive and abusive words.

II. RELATED WORK

In the following section, we review on work done by different researchers in various text categorization techniques.

Guang Xiang, Bin Fan, Ling Wang, Jason I. Hong and Carolyn P. Rose [3] propose a novel semi-supervised approach for detecting profanity-related offensive content in Twitter. His approach exploits linguistic regularities in profane languages via statistical topic modeling on a huge Twitter corpus, and detects offensive tweets using these automatically generated features.

Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu and Steve Maybank [4] propose a novel framework in which Web pages are divided, using the C4.5 decision tree algorithm. The skin detection algorithm combines multidimensional histograms with the EM algorithm to speed up Gaussian mixture skin modeling without compromising the accuracy of the resulting model [4]. From this paper, we get a known different methods & algorithms for text & image detection.

Félix Gómez Mármol, Manuel Gil Pérez and Gregorio Martínez Pérez [5] propose that SNSs could adopt reputation-based mechanisms to assess accuser's behavior when reporting any content [5]. From this paper, we get to know about how to restrict users from reporting harmless content.

W. H. Ho and P. A. Watters [6] use the Bayes classifier to recognize pornographic texts. They not only consider the influence of different words on the weights of the Bayes network but also assign different weights to the same words when they appear in different Web page components such as title, metadata and body [6].

P. Y. Lee, S.C. Hui and A.C.M. Fong [7], [8] count the frequencies with which keywords appear in a text. The frequencies, together with the relevant Web page features, are used as the input to the Kohonen self-organizing neural network (KSOM). After the learning stage is completed, the KSOM is used to determine whether a text can be classified as pornographic [7].

R. Du, R. Safavi-Naini and W. Susilo [9] extract feature vectors from pornographic and normal texts and save them in a database. To test a new text, its feature vector is extracted and matched to each of the saved feature vectors.

R. Du, R. Safavi-Naini and W. Susilo [9] uses decision making algorithm for classifying the offensive words from the texts.

Ying Chen, Yilu Zhou and Sencun Zhu, Heng Xu [10] propose Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users in social media.

III. Overview of Different Text Categorization Approach

In the following section we give a brief overview of the classification techniques that we have studied to help us for text detection and categorization.

A. Kohonen Self-Organizing Neural Network (KSONN)

It is a technique for text categorization based on the frequencies with which the words appear in the given text. It counts the frequencies of the keywords appear. These frequencies used as a input to the KSONN. After the learning stage, the KSONN is used to detect offensive text.

B. Bayes Classifier

It is used for the discrete text classification. These texts are the combination of the Metadata, titles, etc. The semantic associations in these texts are less as compare to the Continuous text. So Bayes classifier classifies them by using the probability of occurrences of the keywords present in discrete text.

C. Cellular Neural Network (CNN)

It is a technique for text categorization based on the semantic connections between keywords. It is like a Word Net which gives semantic connections between different keywords. This algorithm is developed by W. Hu et al for recognizing continuous text from pornographic content.

D. N-grams

N-Grams is a statistical based approach for classifying text. The N is the number of keywords used for dividing the input text. Based on the number of keywords used the N-grams are called as 2-grams, 3-grams, etc. It is able to classify the unknown text with highest certainty. This approach is useful if we have small text to categories such as comments in social network sites, conversations in a forum, etc. that because each sentence is divided into small parts.

IV. Algorithm

In the following section, we describe about the algorithm used for abusive word detection in Text Module.

We used AHO-Corasick String Pattern Matching algorithm for offensive word detection. As in social network sites the number of offensive words doesn't change often this algorithm is best suited for our approach. This algorithm works better if we have a set of keywords which doesn't change often. As we have a database of the abusive keywords which we have collected from different datasets available over internet.

The algorithm matches the pattern (i.e. Abusive Keywords) from the given text and finds the occurrences of the same keywords at different index numbers. It will stores the found keywords in the array with respective index numbers. This index number will help to detect the keyword at particular location in the given inputted text.

After the word is detected then we simply replaced it with some special characters to prevent it from publishing on user's wall in our prototype.

Aho-Corasick is useful when you have a set of keywords and you want to find all occurrences of keywords in the text or check if any of the keywords is present in the given chunks of text [12].

The algorithm consists of two parts [12]:-

1) The first part is the building of the tree from keywords you want to search for,

2) The second part is searching the text for the keywords using the previously built tree (state machine).

3) Searching for a keyword is very efficient, because it only moves through the states in the state machine.

If a character is matching, it follows goto function otherwise it follows fail function.

Tree Building:-

- In the first phase of the tree building, keywords are added to the tree.
- Links created in this first step represents the goto function, which returns the next state when a character is matching.
- During the second phase, the fail and output functions are found.

Searching:-

For Searching it will use Breadth First Search (BFS) and searches the text for the keywords using the previously built tree (state machine).

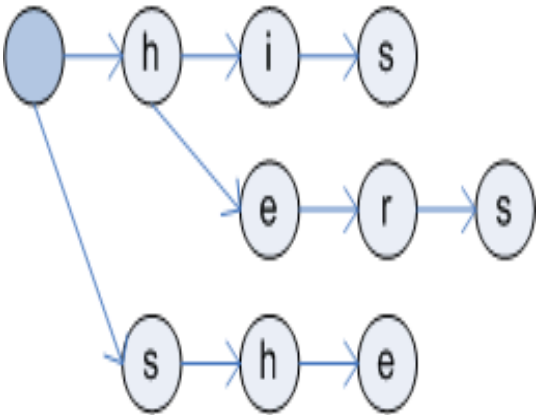

Fig. 1. - after the first step

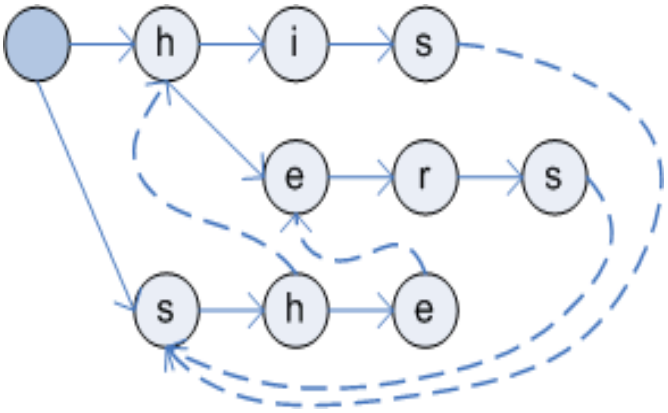

Fig. 2. - tree with the fail function)

V. Our Proposed Text Filtration Module

The Text Filtration Module

A. Offensive Word Detection –
- We represented the experimental results in the form of ListBox Control to show the detected Offensive Keywords.
- We applied the AHO-Corasick Algorithm on the input text for the detection of any offensive word which is present in our database.
- If the matched keyword is found then it will show it in the ListBox with their Index numbers.
- It will find every occurrences of the Offensive Keyword present at different index positions.

B. Offensive Word Prevention -
- Content Upload Module
- Filtration Technique Implementation
- Abusive Keyword Detection
- Replacement with special character.

VI. Our Experimental Outputs

A. Text Uploading Module

Fig. 3. Text Uploading Module

B. Offensive Word Detection

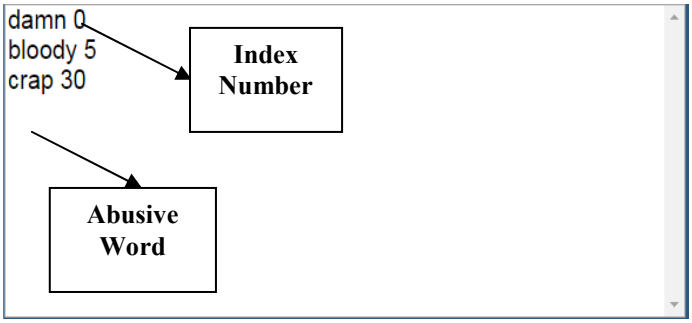

Fig. 4. Offensive Word Detection

C. Offensive Word Prevention

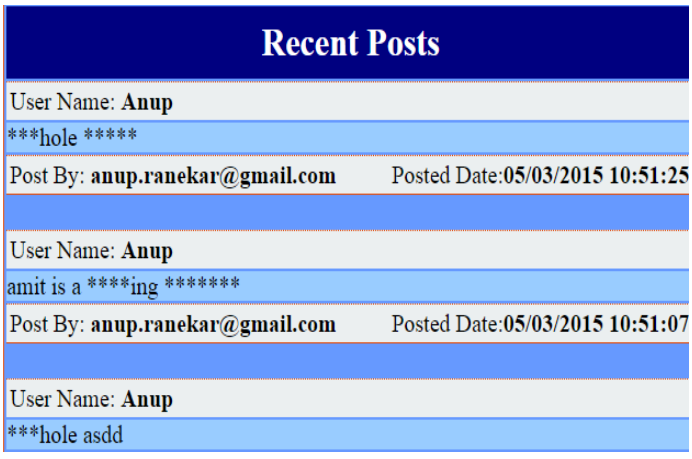

Fig. 5. Offensive Word Prevention

VII. CONCLUSION

In this paper, we have proposed our text filtration approach using Aho-Corasick string pattern matching algorithm for offensive word detection. Also we applied our preventive measures to restrict offensive words from publishing in our social network prototype. We have used dictionary table to match the keywords from given input text. Also, semantic relations between words are ignored because we assume that in social network users generally use slang languages for communication. This assumption is also beneficial for real time processing of our text module.

REFERENCES

[1] Cavnar, W. B. and J. M. Trenkle, "N-Gram-Based Text Categorization," 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161-175, 1994.

[2] ÖLVECKÝ, Tomáš. "N-Gram Based Statistics Aimed at Language Identification." IIT. SRC 2005: Student Research Conference, pp. 1-7, 2005.

[3] Guang Xiang, Bin Fan, Ling Wang, Jason Hong and Carolyn Rose, "Detecting offensive tweets via tropical feature discovery over a large-scale twitter corpus," 21st ACM International Conference on Information Knowledge Management, pp. 1980-1984, 2012.

[4] 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, pp. 1019-1034, 2007.

[5] Felix Gomez Marmol, Manuel Gil Perez, Gregorio Martinez Perez, "Reporting offensive content in social networks toward a reputation-based assessment approach," IEEE Journal on Internet Computing, Vol. 18, Issue 2, 2014.

[6] W. H. Ho and P. A. Watters, "Statistical and structural approaches to filtering internet pornography," IEEE International Conference System, Man and Cybernetics, Vol. 5, pp. 4792-4798, 2004.

[7] P. Y. Lee, S.C. Hui and A.C.M. Fong, "Neural networks for web content filtering," IEEE Journal on Intelligent Systems, Vol. 17, Issue 5, pp. 48-57, 2002.

[8] P. Y. Lee, S. C. Hui and A.C.M. Fong, "An intelligent categorization engine for bilingual web content filtering," IEEE Transactions on multimedia, Vol. 7, Issue 6, pp. 1183-1190, 2005.

[9] R. Du, R. Safavi-Naini and W. Susilo, "Web filtering using text classification," IEEE International Conference on Networks, pp. 325-330, 2003.

[10] Ying Chen, Yilu Zhou and Sencun Zhu, Heng Xu, "Detecting Offensive language in social media to protect adolescent online safety," IEEE International Conference on Social Computing (SocialCom), pp. 71-80, 2012.

[11] http://en.wikipedia.org/wiki/Bag-of-words_model

[12] http://www.codeproject.com/Articles/12383/Aho-Corasick-string-matching-in-C.