ISSN:0975-9646

Detection and Elimination of Censor Words on Online Social Media

Shubhankar Gupta

Jaypee Institute of Information Technology, Noida, India

***Abstract*— Recently, online communication has increased manifold with the advancement of technology, and there cannot be any question raised regarding to the convenience offered by it. Throughout the world, a large number of people interact with each other through the means of social media which has now become very popular. There has been a significant rise in the number of social media platforms, such as, Facebook, Twitter, Instragram, Google+ etc., which allow people to share their experiences, views and knowledge with others. Sadly enough, with online communication getting embedded into our daily communication, incivility and misbehaviour has taken on many nuances from professional misbehaviour to professional decay. During online communication, exchange of rude messages and comments has generally been observed, which in turn triggers conflict. To prevent online communication from getting downgraded, it is essential to check the hostile users on the communication platforms. This paper presents a probable Detection and Elimination Model which can be used to check and prevent online hostility. It can detect and eliminate the presence of censor words while posting information on Online Social Media (OSM).**

***Keywords*— Information Technology, Censor Word Detection Model, Censor Word Elimination Model, Online Social Media.**

I. Introduction

In the last few years, online communication has increased manifold with millions of people discovering the power of Information Technology. Throughout the world, a large number of people interact with each other through the means of social media which has now become very popular among them. Recently, there has been a significant rise in the number of social media platforms such as Facebook, Twitter, and Google+, Instragram etc., which allow people to share their experiences, views, knowledge and most importantly share their daily life activities with others [1].

Undoubtedly, there are several advantages of the mode of online communication, but at the same time there are several disadvantages attached to it which prevails in the form of social media communication. People now have a predilection towards social media to such an extent that it has now become an important part of their life. They pretend as if without social media, their life has no meaning. The posts on social media by these people indicate that how pretentious and self-obsessed they have become. Social media sites make up a sizable portion of all Web Traffic since 75 per cent of all Internet users use social media [2]. According to Facebook, it has about 2 billion monthly active users and more than one billion that log on daily basis. In this paper, Detection and Elimination Models are presented. These models can be used to check and prevent online hostility. It can detect the presence of censor words while posting information on social media.

II. Negative Effects Of Social Media

Social media sites are used to contact colleagues, friends and relatives. However, with the advances in social and technological platforms, people have also started interacting with strangers through social networking sites [1, 3, 4, 5, 6, 7, 8, 9 and several others]. Out of these, strangers' research has also shown that males have a greater tendency to flame than female participants [8]. Even in the male population, it is the young, immature minds, who would account for higher flaming [1].

As mentioned above, OSM sites are in greater use, hence easily available to predators as well and friends. Cyber bullying is one of the pernicious problems of the OSM. The youths are especially vulnerable to the practice of cyber-bullying in which the perpetrators, anonymously or even posing as people their victims trust, terrorize individuals in front of their peers. The devastation of these online attacks can leave deep mental scars. Consequently, victims have been driven to suicide in many cases. According to a 2010 CBS News Report [10], Cyber-bullying has spread widely among youth, with 42% reporting that they have been victims. The use of censor words instigates cyber-bullying. These censor words in communication hampers the relationship among youths and older people. People become intransigent of their opinion when a discussion takes place on OSM. When a person posts something offensive for others, this often ends up in greater form of conflicts. Eventually such forms of communication degrade the sense of camaraderie among friends, business people and others.

It is essential that some kind of integrity is maintained among the individuals. This kind of communication is more common among the youth who seem to be very aggressive in nature. In this society, such exchange of messages is not admissible.

III. Censor Word Detection Model

A Detection Model has been proposed to detect the occurrence of censor words on OSM. This model can be plugged into any of the OSM sites such as Facebook, Twitter, Quora, Google+. Flame Detector Model consists of three components as shown in Figure 1. It consists of Social Networking Sites/OSM, Web Services and Censor Word Detection.

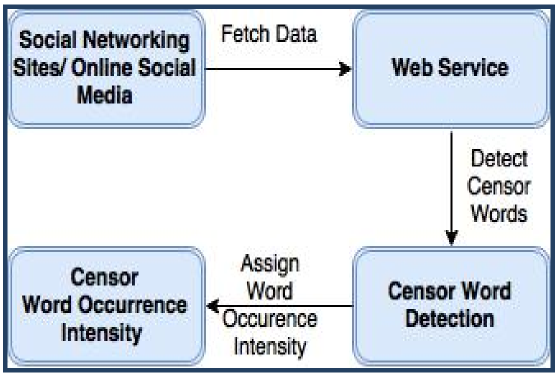
Social Networking Sites/ Online Social Media
Fetch Data
Web Service
Detect Censor Words
Censor Word Detection
Assign Word Occurence Intensity
Censor Word Occurrence Intensity

Fig1 Censor Word Detection Model

As shown in the figure 1, the flame detector model has three components. The first component is social networking sites or the OSM where formal as well as informal discussion takes place and from where the model gets its input. The second component is the Web Service which is used to fetch the data from social networking sites i.e. Facebook, Twitter which is then analysed by the Censor Word Detection Component to assign censor word occurrence intensity to the user. A predefined thesaurus (Censor_word.txt file) has been developed, which contains words that are considered as inappropriate for communication on OSM sites. In addition to this, to avoid the hostile users to use censor words in further communication, there should be some elimination model that expunges these words on the social media.

IV. Censor Word Elimination Model

The major task is now the detection of censor words on the online social media. Let the occurrence of the censor words be represented by a counter variable which is initially assigned to zero. Over a period of time as the user exchanges data either in the form of text messages to others or in the form of blogs or in the forms of online social posts, the counter value increases in the step unit of one, whenever certain amount of censored words are encountered on the online social media. The value of the counter variable at the same time can be retrieved by the respective online social media websites which can be very useful for analytics purposes as they can warn a user for avoiding the usage of censor words. An elimination model has been developed which can eventually be very useful in elimination of censor words (Fig 2).

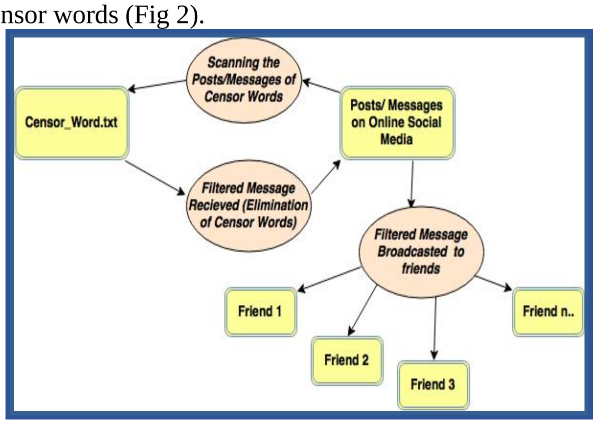
Scanning the Posts/Messages of Censor Words
Censor_Word.txt
Posts/ Messages on Online Social Media
Filtered Message Recieved (Elimination of Censor Words)
Filtered Message Broadcasted to friends
Friend 1
Friend 2
Friend 3
Friend n..

Fig2 Censor Word Elimination Model

The Figure 2 depicts the elimination model which can be injected on the online social media. As shown in the model, whenever a user posts or sends a message on the OSM site, the first task is the scanning the presence of censor word, if any. The list of censor words is stored in a text file (*Censor_Word.txt)* which can be managed by the administrator of the OSM site.

If there is a word that at the same time is present in the list, then the corresponding word is omitted and replaced by asterisk (*). The asterisk (*) thus serves to expurgate the censor words. Subsequently with the omission, the value of the counter variable is also incremented. The filtered message is then received back to the social media which can be broadcasted to one or more friends.

For e.g. if a user sends a message to one of the friends, "*You fucking man, why are you bothering me…*". This message is first scanned of censor words. The substring *"fuck"* of the string *"fucking"* is found in the list. This particular substring is then omitted and replaced by asterisk (*). The filtered message thus sent is:- *"You ****ing man, why are you bothering me.."* . Many such more messages can be encountered on the online social media which can be used to curb the hostile users. In some special cases, if the administrator finds a word which is inappropriate, the particular word can be tagged as censor word and can be added dynamically to the list of censor words for future upgrade.

V. Limitations of Detection and Elimination Models

There are many words which are either not tagged as censored or they are not included in any of the official dictionaries. Hence, these words need to be identified accurately and regularly updated in the thesaurus (*Censor_Word.txt*), which cannot be done manually by an administrator since it is a cumbersome job. Therefore, some kind of mechanism like an *administrator bot* needs to be devised which is able to identify new censor words with the help of machine learning.

VI. Conclusions

The convenience and advantages of online communication is evident to one and all. The extensive rights granted allows a user to share information by posting blogs and exchanging messages with our peers. However, these rights have eventually led to the hostile and aggressive exchange of words. To maintain some civic sense during online communication and to keep it intact, it is essential to prevent it from getting adulterated with aggressive and abusive form of behaviour. This paper presents a probable model for detecting and elimination censor words on the online social media.

References

[1] S Gupta, and N. Chanderwal, "Development of security detection model for the security of social blogs and chatting from hostile users", *The International Association for Computer Investigative Specialists*, *Issues in Information Systems*, Paper ID 213, 2017.

[2] "12 Social Media Facts and Statistics You Should Know in 2016" *[Online](http://www.makeuseof.com/tag/12-social-media-facts-statistics-know-2016/).*

[3] M. L. Markus, (1994). "Finding a happy medium: Explaining the negative effects of electronic communication on social life at work",

ACM Transactions on Information Systems, vol. 12(2), pp. 119-149, 1994.

[4] D.A. Moore, Kurtzberg T.R., Thompson, L.L. and Morris, M.W. “Long and short routes to success in electronically mediated negotiations: Group affiliations and good vibrations”, *Organizational Behavior and Human Decision Processes*, vol. 77(1), pp. 22-43, 1999.

[5] E.M. Landry, “Scrolling around the new organization: The potential for conflict in the on-line environment”, *Negotiation Journal*, vol. 16(2), pp. 133-142, 2000.

[6] R.A. Friedman, and Currall, S.C. “Conflict escalation: Dispute exacerbating elements of e-mail communication conflict”, *Human Relations*, vol. 56(11), pp. 1325-1347, 2003.

[7] P.B. O'Sullivan and Flanagin, A.J. “Reconceptualising "flaming" and other problematic messages”, *New Media & Society*, vol. 5 (1), pp. 69-94, 2003.

[8] M. Zuckerberg, “500 million stories”, *South Atlantic Quarterly*, vol. 92, pp. 559-568, 2010.

[9] R. Verma, and Nitin, “On security negotiation model developed for the security of the social networking sites from the hostile user”, *Issues in Information Systems,* vol. 16, Issue II, pp. 1-15, 2015.

[10] “The Negative Effect of Social Media on Society and Individuals”, *[Online] (http://smallbusiness.chron.com/negative-effect-social-media-society-individuals-27617.html).*