Filtration of Unwanted Messages from Osn User Wall Using Machine Learning

M.S.Sivapriya, Aditi Tiwari, Ritesh Kumar Singh

Abstract-The platform to make friends and pass message among each other are becoming a powerful source and tool of communication. Social Networking sites serves as the best platform for entertainment of upcoming generation. OSNs helps the users to connect online to others in order to communicate and share their various experiences in the forms of posts and status. But now-a-days Online Social Networking are facing many problems of posting annoying content on someone else's profile which make others humiliated after seeing this. In arrangement to eliminate these foul words, machine learning is used that will filter the unbearable word from the present content. The content of social media are amorphous Since the data (textual content) on online media is mainly unstructured and often in casual style, the existing research on message-level offensive language detection cannot detect the accurate offensiveness of the content. In comparison with message-level offensiveness detection, the identification at user level will be more viable but this is under analysis stage. In disposition for removal ofobjectionable words from an OSN user's wall, a new system will be offered which will have LSF(Lexical Syntactic Feature), the objectionable content will be filtered based on LSF. Different approaches like Bag of Words(Bow) and n-gram will be used through which filtration of bad words will occur. Thus, the focus of the ongoing work is to offer a mode that will filter the unrelated messages and propose a Filtered Wall (FW).

Keywords: Online Social Network(OSN); Offensive words; Lexical Syntactic Feature(LSF); Bag of Words (BoW); n-gram algorithms; data filtration; short text classification.

I. INTRODUCTION

The online platform to make friends and pass message is a medium to communicate among each other and pass information. These online platforms can be used to share various types of content, which can be texts, images, audio and videos. this provide a platform to make social networks, relations among people over the internet. The online platform to make friends is a service which consist of each users having their own profile. It is an online service which a user can use to make a profile in public mode, or can make a user's list among which they can share connections. A user can make friends and post content on social media. WhatsApp, Instagram, Snapchat are widely used social networking sites. Due to the sudden growing of online platforms for social interaction, the youth are using these platforms for long hours continuously to interact among their friends and family and also to interact with new people over the net, by posting images, status and many more thing on the OSN wall. In a survey conducted it was found that in 2010, almost 60-70% of youngsters are using online social media platform on their daily basis[1] whereas it is increased to 88-90% in 2019.

Revised Manuscript Received on May 10 ,2019

M.S.SIVAPRIYA Assistant Professor in the Dept. of CSE at SRMIST, Chennai India.

Aditi Tiwari, Student in the Dept. of CSE from SRMIST, Chennai India.

Ritesh Kumar Singh, Student in the Dept. of CSE from SRMIST, Chennai India.

Nearly one in four people uses their online platform sites more than ten times in a day [3]. While adolescent's benefits by using online platforms to interact by day-today happening around the globe and to learn from each other. But as everything comes with pros and cons, so is the online platform as it exposes the user to come across foul word and objectionable content online. The survey of Scan Safe global threat, reported that out of every 10 blogs at least 7-8 blogs contain objectionable words and out of every 4 image, video at least 2 of them will have objectionable content [2]. The objectionable message posted on online platforms leads to cyber-bullying. In a survey it was reported that every 2 out of 10 user faces a problem as someone posted something that embarrasses them on online platforms [1]. As this kind of content negatively affect the youngsters so to find objectionable content and to have a solution for this require immediate action.

To mark the objectionable content on the basis of user access, the content is manually reviewed and if came across any objectionable content then it is deleted by the admin of the online platform.

The manual review and identification of objectionable content require more effort to do and it also require more time to identify the foul words, so it is not that much effective in long run.

So there is need for some unmanned filtering application. There are some developed application like Appen and internet security suite which identify and filter out the objectionable content, most of these application will do so by blocking the user or the sentence which contain foul words.

These approach are not that effective when it comes to operation of websites. So in the sentence "Hi xyz the cry baby" will not be termed as objectionable content as this doesn't contain any foul words used in general terms.

As there are certain words which have various meaning so there is a high change for the approach to fail.

In addition to this these methods often traces each words as liberated occurrence and not traces the objectionable content source.

In the existing system OSN does not provide a genuine aid to intercept undesired messages to be post on another user profile. A very little contribution is done to detect and filter the indescent words,like some online platforms provide the user to make their profile privatemeans they can grant the accessto other user to send message and post content on the wall (i.e friends or a particular group of thier friends) but there is no privacy given on the basis of preference of the content. Therefore, it is not possible to prevent the wall from getting posted any undesired or unbearable message[4].

Disadvantage in the Current System

1. The online platform provide a privacy in form of restricating the user from posting content on the profile but there is no restrictions on what content they are posting. Due to this some people uses bad and indescent words in commenting on the public posts[4].
2. This is not aacurate as there are content written in informal,which can not be distinguished.

To overcome from the obstacles a solution to enhance the accuracy in detecting the objectionable content is developed. An LSF based model is propoundto productively recognise the objectionable language from the online platform and will remove those foul words. This model will provide elevated precision in the recognition of objectionable content, and will also reduce the error. This model will inspect the user as well with his posted message and the design in which he/she usually posts the message. The model can be imposed as a client side approach for users who are worried about their protection over the net.

The proposed system will recognise the users based on their uses of good word and foul words.

II. RELATED WORK

A. Objectionable Content Refining Methods on Online Platform

The Online platform apply various contraption for screening objectionable contents. In some platforms there is a safety mode, which once activated will hide all the comments posted from users on that profile which has foul words in it. But the derogatory comment will be replaced by asterisks.

The online platform uses simple wordbook based approach to recognise and filter out the objectionable contents. So the correctness of these systems is very low and produces false positive alerts. These procedures are not efficacious for youth who don't have knowledge of the risks. So there is need to design a system in such a way that it is more sensible and correct.

B. Text Mining Expertise to Recognise Objectionable Contents.

Recognition of objectionable language on online platform is a laborious task as the contents posted are casual and structure less. The ongoing method used by the platforms for the elimination of foul words are not sufficient, so the research is under process to launch more feasible expertise.

There is considerable challenge for text mining to recognise objectionable language include

1) Accession of data and pre-process
2) Extraction of attribute and
3) Its stratification.

a) Attribute Extraction at Message Level

Objectionable content recognition includes the following feature

The syntactic and lexical attribute.

The lexical attribute serves every word and aspect as single entity. The design of words like the mien of some keyword and their repetition rate are used for the representation of the model language.

In early research bag of word model was used in the recognition of objectionable word[5]. The BoW model serves a posting as an unordered assembly of words and doesn't bother about the semantic and syntactic attributes.

The correctness of the BoW model is very low in the recognition of objectionable language and it has more false rate through some impassioned argument, and in response to some other person objectionable post and also in chatting among each other. The N-gram technique used to recognise objectionable word is said to be as an improved perspective [6]. The N-grams model is implemented in a sequence of N continual words in post, these are in token formation. The mostly used N gram sequence are Bi-gram and Tri-gram. The Bi-gram uses two-word series and Tri gram uses three-word sequence.

But the N-gram has problem while inspecting the used word which divided by lengthy distance in text. By increasing the N sub series, it may cause trouble and the processing speed will also be slowed.

Syntactic Attribute: The Lexical attributes recognizes the offensive word without considering the whole sentences, but itdoesn't distinguishsentences' offensive behaviour which has the same words with different meaning and order. For example,"he is behaving like a dog"and "The dog is running like a horse". here dog is used in both the sentences but with different meaning. Hence to examine the syntactical attribute in a language, the NLP are instigated to parse a language based on its grammar before selection.

b) Objectionable Recognition at User-Level

The analysis on the recognition of objectionable language keep an attention at message and sentence. As there are no methods which are 100 percent correct, in any case if any user stay in contact of foul words in some websites, there is high chance for the user to be vulnerable to objectionable contents over the net. While for user-level recognition it is complicated and research is still carried out to make it easy.

User level require very little effort. There are some model to track and distinguish between online pillage, some of these are to distinguish victims from pillage and a model to recognise antagonistic discussion.

III. PROPOSED SYTSEM

In the proposed work a system called Machine learning will be used that will filter the unbearable and indecent words from the given content. Algorithms like Bag of Word(BoW)s and n-gram will be used with Lexical Syntactic Feature model in order to filter the offensive words.

Advantages of Proposed System

1. The proposed system will filter out the undesired vulgar content by checking it through blacklist file from the message contentand therelationship and characteristic of message creator.

2. Major difference includes, a different semantics and approach at message-level offensiveness detection that is more feasible and accurate.

IV. SYSTEM ARCHITECTURE

The basic architecture of the online platform for connecting multiple user comprised of Manager, Application and User interface schema.

Manager: This comprised of a management body who will allocate the basic serviceability and manages association.

Application: it allocates extraneous approach to connect.

User interface: the platform which gives access to user to communicate with the Online platform.

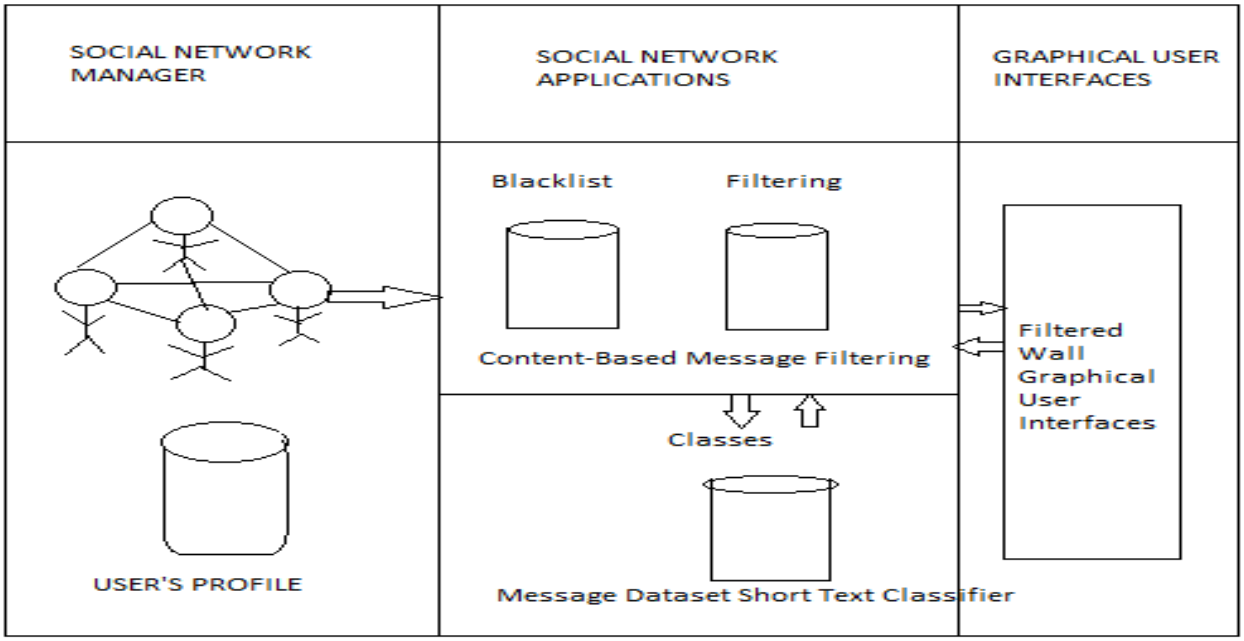

Fig.1: OSN Architecture

V. DESIGN FRAMEWORK

To conquer the recently ongoing dispute, we provide a LSF system which is to fabricate a structure to recognise objectionable content and recognise objectionable users who posts vulgar content over the internet.

The LSF attribute consist of following objectionable detection phase.

Phase 1: The motive of this phase is to recognise vulgar content at sentence extent and

Phase 2: It obtain the objectionable language at user extent.

In Phase 1, Mining approach and NLP approach to define syntactic and lexical attributes. On this basis the objectionable values for the content is calculated. In phase 2, user level is also established. It comprises of three elements. Which include pre-process, message vulgar conjecture and user objectionable assessment.

In the pre process juncture, the discussed language is bundled in the post message and then is converted in sentence formation.

In the message vulgar conjecture, the objectionable part is obtained from vulgar words and its context.

Calculation of Objectionable Sentence

To recognise the limitation of foregoing procedure for message vulgarity detection [9], a recent procedure for message-level inspection is done on the basis of offensive words and syntactic attribute. Primarily, two vulgar wordbooks on the basis of the toughness of the vulgar word will be formulated. Secondly, the idea of phonological modifier is established to regulate the words objectionable extent on context basis. And at the endmost step an objectionable value is calculated on the basis of aggregation of vulgar words.

Lexical Attributes: Objectionable wordbook creation

The objectionable post consists of derogatory obscenity or imprecation. Vigorous imprecation are doubtlessly termed as vulgar words when used in context for a user. But in many cases there may be debilitated imprecation and obscenities, like ‘stupid or cry-baby’ which also a little bit vulgar.Here it is focused on to differentiate the vulgar words in two extents on the basis of their robustness.The foul words which will be used here will encompass wordbook used in xu and zhu’s study [10]. Here each and every imprecation is termed as

strong objectionable word.

VI. METHODOLOGY

A. Process of Filtration

In FRs identification, three main issues are considered. Primarily, in online platforms to contact each other in day today life, the precise words may have various meaning in various different context. As a ramification, it allows the user to form a condition for restriction on text producer. Producer may be selected on the basis of different benchmark. This implies to a state on classification, deepness and the value of its relationship. The input message is passed through a categorizing word attribute which categorises the words based on bad words, restricted words and correct words. This will further filter out the bad words and gives the correct word format as an output.

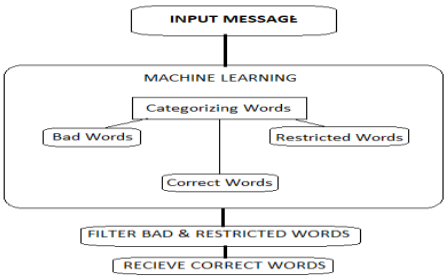

Fig.2: Filtration Process

B. Blacklisting Process

The process of filtration will further be taken to blacklisting process. The additional part of this system will have a BL(blacklisting) mechanism which will result in circumventing texts from unrecognisable producers. The indecent and undesirable word will be added to the blacklist and only the left over message will be displayed.

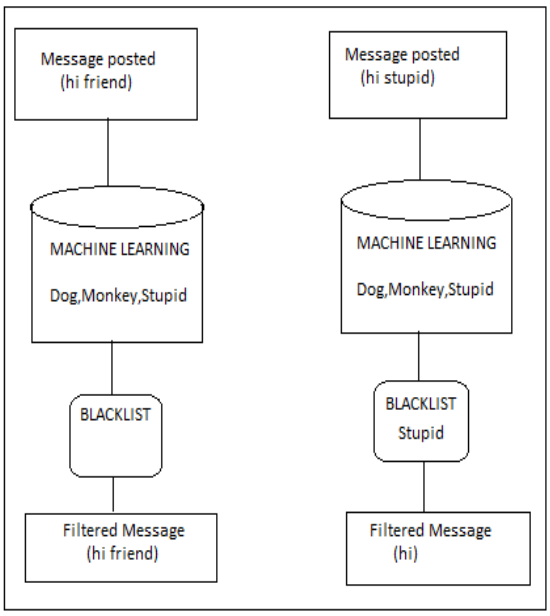

Fig. 3: Blacklist Processing

Similar to Filtering rules, Blacklist rules also make the wall independent of the undesired words.

C. Algorithm Used

Basic algorithm:

Step 1: Start the application

Step 2: A User posts a text, image on an online platform.

Step 3: A person adds a comment on the user post.

Step 4: Each and every part of the user post and the comment on it will be processed using NLP.

Step 4: If the machine after processing found out that the comment or the posts resemble any kind of vulgar attribute then go to step 6.

Step 5: if after processing it is found out that there is no objectionable content then it will be posted as an output on the wall.

Step 6: Using blacklisting process it will filter out each and every words which indulges in any kind of vulgar act.

Step 7: Stops the process.

NLP is an arrangement which learns and formulate on the data it reads and on this basis it will draw conclusion. Like it filters out the spam or non-spam content from the email inbox.

VII. EXPERIMENAL WORK

The investigation is supervised with NLP procedure by using the restricted and correct words format from the objectionable wordbook which already has all the content which are being attached on online platforms. It also distinguishes among the approved and not approved individual who is trying to write something on the online profile.

Correct words = Σ Wi {S} $\neq \Sigma$ BLw
i=1 w=1

Restricted Words = Σ Wi {S} = Σ BLw
i=1 w=1

Where, Wi is the ithname from the sentence S
S Sentence entered.
BLw wth word from the blacklisting process.
Like,
Blacklisted words
BLw = {Stupid, Cry-baby, idiot, Fool}

No. of sentence entered
S = {Stupid, MFC, Kutty, Cry-baby}

SENTENCE ENTERED	
S1	Hi stupid boy
S2	FOOL
S3	Buffalo
S4	Hey there

Output= (BL = = S1) = Hi
Output= (BL = = S2) = Sentence not passed
Output= (BL = = S3) = Sentence not passed
Output= (BL = = S4) = Hey there

VIII. CONCLUSION

Here, we have discussed the ongoing text mining method for the recognitionof the offensiveness of the messages and the problems faced in the present system. To overcome from those present scenario of the existing system, we have already recommended an arrangement which will intercept the impure/indelicate word from the text format from the Online platform using LSF model. By implementing NLP methodology, the output is refined and the result generated has more correctness toward the restricted and correct words and to discriminate among the approved and not approved user on the online platform profiles. So it can be termed as machine leaning model plays and an importance role this paper in arrangement to produce blacklisting of the text and unrecognizable users.

This paper contains the uses of LSF attribute to enhanceorthodox machine learning methods that yields more feasible results with more accuracy.

REFERENCES

1. T. Johnson, R. Shapiro, and R. Tourangeau, “National survey on American attitudes on substance abuse XVI: Teens and parents.,” in The National Center on Addiction and Substance Abuse.
2. J. Cheng, "Report on 80 percent of blogs contain "offensive" content,"inarstechnica.
3. S.O.K Gwenn, C.-P. Kathleen, "Clinical report--the impact of social media on children, adolescents, and families.,"
4. K.Babu, P.Charles,”A System to Filter Words Using Blacklists In Social Networks”.
5. A. Mahmud, Ahmed, KaziZubair, and Khan, Mumit "Detecting flames and insults in text," in Proc. of 6th International Conference on Natural Language Processing.
6. N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," in Proceedings of the First IEEE International Conference on Semantic Computing.
7. A. Kontostathis, L. Edwards, and A. Leatherman, "Chatcoder: Toward the tracking and categorization of internet predators," In Proc. Text Mining workshop 2009 held in conjunction with the Ninth SIAM International conference on Data Mining.
8. M. Pazienza and A. Tudorache, "Interdisciplinary contributions to flame modeling," AI* IA 2011: Artificial Intelligence Around Man and Beyond.
9. E. Spertus, "Smokey: Automatic recognition of hostile messages, "Innovative Applications of Artificial Intelligence.
10. A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," Advances in Artificial Intelligence.

AUTHORS PROFILE

M.S.SIVAPRIYAis an Assistant Professor in the Dept. of CSE at SRMIST, Chennai India.

ADITI TIWARI is a Student in the Dept. of CSE from SRMIST, Chennai India.

RITESH KUMAR SINGH is a Student in the Dept. of CSE from SRMIST, Chennai India.