A Machine Learning Approach to Riddle Abused Scraps from OSN User Space

Ch. Pavan Kumar
M.Tech.(SWE),Dept.of CSE
Kakatiya Institute of Technology and Science
Warangal-15,Telangana,India

B. Raghu Ram
Assistant Professor, Dept.of CSE Kakatiya Institute of Technology and Science
Warangal-15, Telangana, India

B. Hanmanthu
Assistant Professor,Dept.of CSE, Kakatiya Institute of Technology and Science
Warangal-15,Telangana,India

ABSTRACT
Today's Online Social Networks are to communicate with people and share information with them. In this kind of network process user space is a space where scraps or temporal notes of its users can be displayed. But users have very less control on scraps posted on their space. Abused Scraps which are sending and received on the user private space are riddled by allowing OSN users to have a direct control on the scraps posted on their spaces through Filtering Rules and Context Based Filtering Mechanism. A Flexible Rule System, that allows users to customize the filtering criteria to be applied to their space. The Abused Scraps from the user space can be riddled by Machine Learning based soft classifier. The user has a system with a sophisticated approach to decide which scraps should be inserted into a Blacklist.

Keywords
Content Based Filtering, Machine Learning, Black List, On-line Social Network, Rule Based Information Filtering System.

1. INTRODUCTION
Online Social Networks provide platform to meet people and share information with them[1]. Communication on these sites involves exchange of various content including text as well as multimedia data. For example Face book allows user to state who is allowed to insert scraps in their space by friends, friends of friends are defined groups of friends. A social network include blogs, private scraping, chat facility and file, photo sharing functions and other ways to share text and multimedia data. We exploit Machine learning text categorization techniques to automatically assign with each shortest scrap [2] a set of categories based on its content. The original set of features derived from properties of short texts is enlarged here. In this including exogenous knowledge related to the context from where scraps originate. As far as the learning model is concerned [3], we confirm in the current paper the use of neural learning which is today recognized as one of the most efficient solutions in text classification. The first proposal of a system is to automatically riddle abused scraps from OSN user space on the basis of both scrap content and the scrap characteristics and creator relationships. In this paper extends for what concerns both the classification and rule layer module. The Major differences is different semantics for riddling rules to fit the domain better, OSA to help users in FR specification the extension of the set of features in the classification process, a more deep performance an update and evaluation study of the prototype implementation to reflect the changes made to the classification.

2. PREVIOUS WORK
Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

2.1 Short Text Categorization
In Machine learning approach Categorization text potentially containing a complex and specific terminology requires the use of learning methods. In this process we use prediction by partial matching (PPM), a method that compresses texts to capture text features and creates a language model adapted to a text. This method achieves a high accuracy of text categorization and can be used as an alternative to state-of-art learning algorithms[4]. A general framework for building classifiers that deal with short and sparse text & Web segments by making the most of hidden topics discovered from large-scale data collections. The main motivation of this work is that many tasks working with short segments of text and Web such as search snippets, forum and chat scraps, blog and news feeds, product reviews, and book and movie summaries, in this process fail to achieve high accuracy due to the data sparseness. We the underlying idea of the frame-work is that for each classification task we collect a large-amount of external data collection and then build a classifier on both a small set of labelled training data and a rich set of hidden topics discovered from that data collection. The framework is applied to different data domains and genres ranging to Web search results. We did a careful evaluation on several hundred megabytes of Wikipedia 30M words and MEDLINE 18M words with following tasks:

i. Web search domain disambiguation.
ii. Disease categorization for medical text.
iii. Achieved significant quality enhancement.

2.2 Parsing
Parsing is the process of structuring a linear representation, parsing with a given grammar. This definition has been kept abstract on purpose, to allow as wide an interpretation as possible. The "linear representation" may be a sentence, a computer program, a knitting pattern, a sequence of geological strata, a piece of music, actions in ritual behaviour in short any linear sequence in which the preceding elements in some way restrict the next element. Sentences in these

languages have to be processed automatically that is, by a compiler and it was soon recognized that this is a lot easier if the language has a well-defined formal grammar. The syntaxes of all programming languages in use today are defined through a formal grammar. Most of the authors parsing algorithms require a CF grammar to be monotonic. The only way a CF rule can be non-monotonic is by having an empty right-hand side; such a rule is called an ε-rule and a grammar that contains no Such rules are called ε-free. The requirement of being ε-free is not a real restriction, just a nuisance. Any CF grammar can be made ε-free be systematic substitution of the ε- rules. But this in general does not improve the appearance of the grammar. The basic property of CF grammars is that they describe things that nest an object may contain other objects in various places. When during the production process we have produced one of the objects the right-hand side still remembers what has to come after it in the English grammar after having descended into the depth of the non-terminal Subject to produce something like the wistful cat the right-hand side Subject Verb Object still remembers that a Verb must follow. While we are working on the Subject the Verb and Object remain queued at the right in the sentential form.

2.3 Back Propagation

The set of various Riddle Scraps and spam emails parsing the individual mails to extract the words of interest. Porter method implemented by using the Back propagation Algorithm. The back propagation algorithm looks for the minimum of the error function in weight space using the method of gradient descent. The combination of weights which minimizes the error function is considered to be a solution of the learning problem. Since this method requires computation of the gradient of the error function at each iteration step we must guarantee the continuity and differentiability of the error function. Obviously we have to use a kind of activation function other than the step function. We use Back propagation Algorithm because the composite function produced by interconnected perceptions is discontinuous, and therefore the error function. One of the more popular activation functions for back propagation networks is the sigmoid, a real function SC: IR! (0, 1) defined by the expression. The constant c can be selected arbitrarily and its reciprocal 1/c is called the temperature parameter in stochastic neural networks. The shape of the sigmoid changes according to the value of c, the graph shows the shape of the sigmoid for c = 1, c = 2 and c = 3. Higher values of c bring the shape of the sigmoid closer to that of the step function and in the limit c! 1 the sigmoid converges to a step function at the origin. In order to simplify all expressions derived in this chapter we set c = 1, but after going through this material the reader should be able to generalize all the expressions for a variable c. In the following we call the sigmoid s1(x) just S(X).

2.4 Machine Based Learning

In content based recommendation methods, the utility of item s for user c is estimated based on the utilities, i u c s assigned by user c to items is ∈S that are "similar" to item s. For example, in a movie recommendation application in order to recommend movies to user c, the content-based recommender system tries to understand the commonalities among the movies user c has rated highly in the past. Then, only the movies that have a high degree of similarity to whatever user's preferences are would get recomended.The content-based approach to recommendation has its roots in information retrieval and information filtering. Because of the significant and early advancements made by the information retrieval and filtering communities and because of the importance of 6 several text-based applications, many current content-based systems focus on recommending items containing textual information such as documents, Web sites (URLs), and Usenet news messages. The improvement over the traditional information retrieval approaches comes from the use of user profiles that contain information about users' tastes, preferences, and needs. The profiling information can be elicited from users explicitly. It is usually computed by extracting a set of features from item s and is used to determine appropriateness of the item for recommendation purposes. Since, as mentioned earlier, content-based systems are designed mostly to recommend text-based items the content in these systems is usually described with keywords. For example, a content-based component of the Fab system [5], which recommends Web pages to users represents Web page content with the 100 most important words. Similarly the Syskill and Webert system represents documents with the 128 most informative words. The "importance of word ki in document dj is determined with some weighting measure wij that can be defined in several different ways.

One of the best-known measures for specifying keyword weights in Information Retrieval is the term frequency/inverse document frequency (TF-IDF) measure that is defined as follows. Assume that N is the total number of documents that can be recommended to users and that keyword ki appears in ni of them. Moreover, assume that i, j f is the number of times keyword ki appears in document dj. Then i, j TF , the term frequency (or normalized frequency) of keyword ki in document dj, is defined as where the maximum is computed over the frequencies z, j f of all keywords kz that appear in the document dj. However, keywords that appear in many documents are not useful in distinguishing between a relevant document and a non-relevant one. Therefore, the measure of inverse document frequency (IDFi) is often used in combination with simple term frequency (i, j TF).

Where the maximum is computed over the frequencies z, j f of all keywords kz that appear in the document dj. However, keywords that appear in many documents are not useful in distinguishing between a relevant document and a non-relevant one. Therefore, the measure of inverse document frequency (IDFi) is often used in combination with simple term frequenc Then the TF-IDF weight for keyword ki in document dj is defined as i, j i, j i w = TF × IDF and the content of document dj is defined as Content(do) = (w1j, …wkj). As stated earlier, content-based systems recommend items similar to those that a user liked in the past. In particular, various candidate items are compared with items previously rated by the user, and the best-matching item(s) are recommended.

More formally, let Content Based Profile(c) be the profile of user c containing tastes and preferences of this user. These profiles are obtained by analyzing the content of the items previously seen and rated by the user and are usually constructed using keyword analysis techniques from information retrieval. For example, ContentBasedProfile(c) can be defined as a vector of weights, where each weight wci denotes the importance of keyword ki to user c and can be computed from individually rated content vectors using a variety of techniques.

For example, some averaging approach, such as Rocchio algorithm, can be used to compute Content Based Profile (c) as an "average" vector from an individual content vectors. On the other hand use a Bayesian classifier in order to estimate

the probability of a document. To work well, a ML-based classifier needs to be trained with a set of sufficiently complete and consistent pre classified data. The difficulty of satisfying this constraint is essentially related to the subjective character of the interpretation process with which an expert decides whether to classify a document under a given category. In order to limit the effects of this phenomenon, known in literature under the name of inter indexer inconsistency [6], our strategy contemplates the organization of "tuning sessions" aimed at establishing a consensus among experts through discussion of the most controversial interpretation of scraps.

2.5 Content based Methods

Content-based filtering relies on creating associations between items in a collection. When a user shows a preference for specific items, the system compares those items to others in the collection. Items with a high degree of similarity are presented as recommendations. Pure content-based recommendations ignore the preferences of other users. There are a number of methods that can be used to generate a list of similar items. In the simplest form, this can be thought of as grouping items based upon their genre or subject matter. However, content-based filtering takes this concept further by increasing the number of terms that may be considered to compare items. For example a collection of movies could be compared based on genre, actors, director, subject, parental guidance rating, or review ratings. This allows the filtering system to recommend items based on a much larger range of aspects than searching alone would allow. In the case of article and website recommendations, a slightly different system is used. One method weights articles based on the number of times specific keywords appear in the article compared to the overall rarity of that keyword in the articles indexed. Therefore, items that contain terms that relate well to the searched which are statistically less likely to be common are suggested first. In content-based filtering, each user is assumed to operate independently. As a result, a content-based filtering system selects information items based on the correlation between the content of the items and the user preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences [5], [7]. Content-based filtering techniques suffer from the problems of limited content analysis, overspecialization, and new users. Limited content analysis refers to the set of attributes that any given item in the collection has attached to it being unacceptably small [8].

2.6 Block List

A further component of our system is a BL mechanism to avoid scraps from undesired creators, independent from their contents. BLs is directly managed by the system, which should be able to determine who are the users to be inserted in the BL and decide when users retention in the BL is finished. To enhance flexibility, such information is given to the system through a set of rules, hereafter called BL rules. Such rules are not defined by the SNM, therefore they are not meant as general high level directives to be applied to the whole community. Rather, we decide to let the users themselves, i.e., the space owners to specify BL rules regulating who has to be banned from their space and for how long. Therefore, a user might be banned from a space, by, at the same time, being able to post in other space. Similar to FRs our BL rules make the space owner able to identify users to be blocked according to their profiles as well as their relationships in the OSN. Therefore, by means of a BL rule, space owners are for example able to ban from their space users they do not directly know or users that are friend of a given person as they may have a bad opinion of this person.

3. PROPOSED WORK

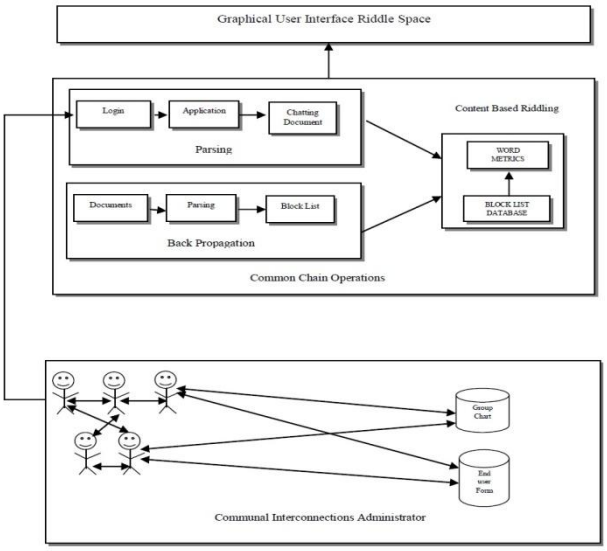

Fig.1: Architecture of Riddling Abused Scrap

The following are steps of Scrap published on the user private space.

1. The scrap which is send through is stored as a chatting documents, this chatting documents is given to parsing for riddling basic verbs & non-verbs.
2. If a probability of occurrence of a particular word exceeds than its threshold then the word is stored in Word Metric.
3. Using Back Propagation the matching pattern of a particular word is stored in the Block List.
4. The matching word which is present in block list matches the word present in word metrics, and then the scrap containing that word is riddled.

The Following is the Procedure for Riddling Abused Scraps:

Online Social Networks had become an essential part in our daily life. Day-by-day communication has being speeded all over the world and being developed very enormously by means of flexible user interface. By the way online social networks are also being emphasized to communicate, share and disseminate a considerable amount of human life information. Daily and continuous communications imply the exchange of several types of content, including free text, image, audio, and video data. In the sharing In OSNs, information filtering can also be used for a different more sensitive purpose. This is due to the fact that in OSNs there is the possibility of posting or commenting other posts on particular public/private areas, called in general space. Information filtering can therefore be used to give users the ability to automatically control the scraps written on their own space, by riddling out abused scraps.

Most of the users when enters into his/her private space, will post the scrap and comment on other photo/post/scrap. Where most of the users will like the other scraps posted by their friends or like comments of the photo/scrap and some of the users will feel discomfort and will not be able to express

themselves about the scraps or comments posted by others. They also feel uncomfortable that the abused scraps posted on their own private space will be viewed by others like friends, friends of friends or group of friends. Online Social Networks had been working on riddling the abused scraps posted on their own user private area by taking the feedback from the users and finding out the solutions to the required problems, but didn't get the perfect and comfort solution to the problem. Till today Online Social Networks has provided a minute support to this requirement.

The OSN Architecture has three layers. The first layer is Communal Interconnections Administrator, handle the basic specifications like form management, user management and establishing the relation. The second layer is Common Chain Operations which takes the external support from Communal Interconnections Administrator (layer-1).The third layer is Graphical User Interface where flexibility is provided to the user to interact with the system to perform and manage the required applications. In this project the proposed system is mainly focused on two layers i.e. Common Chain Operations and Graphical User Interface and some new condition is applied.

The Common Chain Operation provides the user with a Word Metrics, where a particular scrap is published on the user space, if that scrap which is stored in word metrics and doesn't match the scrap stored in the Block Lists (BL).

In Online Social Networks, a user first gets registered by entering all the details like name, address, phone number, email etc. After entering all the details user gets registered and a valid username and password is provided. Then the user enters into his own private space where his/her home page is displayed containing profile photo, update status, searching friends through friends list etc.

Here the user first logins into his/her social account by providing valid username and password, after entering he/she definitely performs some operation like sending friend request, accepting friend request, changing profile picture, updating his/her status. When a user sends a friend request to his/her friends, they will see and accept his/her request, then the communication is established between his/her friends by sending and receiving the scraps together. The scraps communicated between his/her friends are stored in the chatting documents where all the communication history of the scraps is maintained.

The Common Chain Operations contains Parsing, where it is a mechanism of assembling linear representation with a given principles. The linear representation may be a program, pattern, an action, word, sentence etc. The scrap stored in the chatting documents is given to this parsing where it riddles all the verbs and non-verbs. The Common Chain Operations also contains Content Based Filtering, where it calculates on producing associations between components in a collection. When a user shows a choice for specific components, the system compares those components to others in the collection. Components with a high amount of comparison are presented as suggestions. The Common Chain Operations have Word Metrics where all the words of the scraps are stored and it maintains a Threshold i.e. if the word probability of a particular word has occurred more number of times than the threshold value, then that particular word is stored in the Word Metrics.

Here we use Back Propagation Algorithm, where the set of distinct abused scraps and unsolicited emails which parse the respective mails to abstract the words of relevance. This Back Propagation contains documents where all the scrap history is stored and then these words are given to parsing for checking all the verbs and non-verbs. Back Propagation explains the Level of Classification were in I-Level Classifier the unwanted words are identified and as a result of Back Propagation algorithm, it checks for the II-Level Classifier were the word patterns are searched and if it matches the pattern it is stored in block list. If the matching word which is present in the Block List matches the word present in the Word Metrics, then the scrap containing that word will be riddled and is not published on the user private space.

4. EVALUATION

Frontend, the JAVA language was created by James Gosling in June 1991 for use in a set top box project. The language was initially called Oak, after an oak tree that stood outside Gosling's office - and also went by the name Green - and ended up later being renamed to Java, from a list of random words. Gosling's goals were to implement a virtual machine and a language that had a familiar C/C++ style of notation [9]. The first public implementation was Java 1.0 in 1995. It promised "Write Once, Run anywhere" (WORA), providing no-cost runtimes on popular platforms. It was fairly secure and its security was configurable, allowing network and file access to be restricted. Major web browsers soon incorporated the ability to run secure Java applets within web pages. Java quickly became popular.

Backend Technology is MYSQL which is of Relational Database Management System and is Outcome of Open Source and Free Software. It is Free Widely used – Information Systems/embedded systems, primarily written in C/C++, available for Linux, Solaris, MS Windows and other Operating Systems. It is named after co-founder Monty Widenius's daughter, MymSQL- tweaks and hacks to form MySQL. MySQL AB, now a subsidiary of Sun Microsystems, which holds the copyright to most of the codebase "AB" part of the company name is the acronym for the Swedish "aktiebolag," or "stock company. The name of the MySQL Dolphin (MySQL logo) is "Sakila," which was chosen by the founders of MySQL AB from a huge list of names suggested by users in their "Name the Dolphin" contest.

Online Social Networks are being the most important part of our human life in sharing the information and cultivating the communication around the world [1]. As how the mobile phone generation had been developed to smart phones generation, social networks are also being developed with enormous changes like providing more efficiency in communication, more flexible in nature and in developing easy Graphical User Interface (GUI) by means of user. Day-by-day, number of people using social networks is also being increasing and is spread to more number of countries all over the world.

In this paper, a user firstly fills all his required details like name, address, email etc., in the registration form and gets registered. Then a genuine username and password is given to the registered user. User then types the genuine username and password to enter into his/her own private space, the server checks for authentication that whether the username and password entered by user is valid or not. If it is valid the user enters into his/her private space otherwise the server says to recheck the username and password. After entering the valid username and password user get enters into his/her private

space, which contains profile photo, status, and friends list to send friend request. The user can wish to do any kind of operation of what he/she wants, like changing the profile image, updating his/her status and sending the friend request to his/her friends to establish and start the communication together. User sends a friend request to his/her friends, after his/her friends accept his/her request, they start communicating each other and a relation is developed between them. In my project, a system is proposed that have a direct control on the scraps posted on their own private area. This was achieved by rule based classifier that allows users which customizes the criteria of filtering posted on their wall and Machine Learning Based soft classifier that filter the messages with support of content based filtering. When a user enters his/her account, he/she views the profile and if they wish to change their status and profile photo they can and they will add friends available in the friends list. Once the request is sent to his/her friend, his/her friend would see and accept the request. Then the communication is established between the two friends by sending messages and commenting on their photos. Suppose a user has posted a comment/message to his his/her friend and so the user didn't liked the message/comment, and then the user felt difficult or bad that he/she didn't liked it and it could be viewed by others. This problem was faced by many of the users. So to overcome this problem, by using short text classifier, content based message filtering, filtered wall and block list, Online Social Network allows users to have a direct control on the messages posted on their own private space i.e. when user-1 tries to post to a message to other user (user-2) and if user-2 didn't liked that message, the Administrator will filter that unwanted word through Filtered Wall(FW) and by adding the filter words into his/her database into some set of categories. Then the posted message will be automatically filtered and will not be published on his/her user private space. Not only the word will be filtered, the entire content of the message will be filtered by the Filtered Wall. So that the user feels comfort and happy that unwanted messages are not being published on their own private space.

Moreover the Filtered Wall (FW) not only filters the word which was filtered by the Administrator by adding the words to filter, we also filter the words of their respective synonyms by using Word Processing Software (RITA Word Processing Software). For example if a user didn't liked the word "bad" and it was filtered and some other time user got a message with a word "ugly", then this word processing software will automatically filters the word "ugly" which was actually not filtered by Administrator. So by this most of the users who are using Online Social Networks feels very happy and enthusiastic with lot of excitement in all of these improvements made easier and flexible to the user. Finally we could say Online Social Networks play a crucial role in human orientation.

Among three systems connected in distributed environment. For both PIMA and Heart datasets we created unique id for each row and split in to three parts where each part row carry same id and stored in three different systems. In second step proposed web service model implemented on training data sets at local sites and generated results are encrypted with senders public key at receivers side it will be decrypted using receivers private key it ensures the security and all the users keys i.e., private and public keys send to trusted third party for this method we adopted RSA algorithm. In third step combining all local data of individual sites gathered at trusted third party and the same will be used to generate global web service.

5. CONCLUSION

In this paper we provide capability to the system to riddle abused scraps from OSN user space. The development of a GUI and a set of related tools to make easier BL and FR specification is also a direction we plan to investigate since usability is a key requirement for such kind of applications. In particular we aim at investigating a tool able to automatically recommend trust values for those contacts user does not personally known. We do believe that such a tool should suggest trust value based on users actions, behaviours, and reputation in OSN, which might imply to enhance OSN with audit mechanisms. Thus this paper provides two levels of riddling capabilities.

6. ACKNOWLEDGMENT

Our thanks to the management members and principal of Kakatiya Institute of Technology and Science-Warangal who have facilitated resources to read and compute in order to develop this model and narrate this article and our sincere thanks to Head of the Department Prof.P.Niranjan who encouraged us research and publish this paper.

7. REFERENCES

[1] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, "A System to Filter Unwanted Messages from OSN User Walls", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 2, February 2013.

[2] M. Chau and H. Chen, "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," Decision Support Systems, vol. 44, no. 2, pp. 482-494, 2008.

[3] A. Adomavicius and G. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June 2005.

[4] S. Zelikovitz and H. Hirsh, "Improving Short Text Classification Using Unlabeled Background Knowledge," Proc. 17th Int'l Conf.Machine Learning (ICML '00), P. Langley, ed., pp. 1183-1190, 2000.

[5] P.W. Foltz and S.T. Dumais, "Personalized Information Delivery:An Analysis of Information Filtering Methods," Comm. ACM,vol. 35, no. 12, pp. 51-60, 1992.

[6] C. Cleverdon, "Optimizing Convenient Online Access to Bibliographic Databases," Information Services and Use, vol. 4, no. 1, pp. 37-47, 1984.

[7] R.J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," Proc. Fifth ACM Conf. Digital Libraries, pp. 195-204, 2000.

[8] M. Chau and H. Chen, "A Machine Learning Approach to Web Webpage Filtering Using Content and Structure Analysis," Decision Support Systems, vol. 44, no. 2, pp. 482-494, 2008.

[9] K. Arnold and J. Gosling, The Java Programming Language, 1996 :Addison Wesley