AN MACHINE LEARNING AND DATA ANALYTICS APPROACH TO FILTER TEXT CONTENT FROM ONLINE SOCIAL NETWORKING SITES

Mr. Parshuram N Shingote

Ph.D. Research Scholar, Computer Science & Engineering . Dr. A. P. J. Abdul Kalam University, Indore, Indi

Abstract

Now a day's Social networking on social media websites involves the use of the internet to connect users with their friends, family and associates. The most well-known social media sites are Facebook, twitter, Instagram and LinkedIn. These sites allowed sharing photos, videos, and information, the main drawback of this On-line Social Networks (OSNs) is the lack of privacy for the user's own private walls. This research represents system enforcing content-based message filtering consider as a key service for On-line Social Networks (OSNs). Our system provides ability to OSN users to possess a straight control on the messages denoted on their walls. Up to now, OSNs offer very little support to stop undesirable messages on user private walls, there is no content-based preferences are supported and so it's impossible to stop undesirable messages, like political or vulgar ones, regardless of the user who posts them. One basic issue during this system is blocking of user for lifetime. we tend to overcome this drawback by using proposed System; during this paper we tend to proposed a system that performs blocking of undesirable message user and additionally sends back a notification to it who has send undesirable message on user' s wall using threshold frequency worth. To filter messages I tend to proposed an increased filtering system by using machine learning based categorization techniques.

Index Terms—*Online Social Networks, Filtered Wall, Blacklists, Machine Learning text categorization.*

I. INTRODUCTION

On-line Social Networks (OSNs) are platforms that allow people to publish information about them and to connect to other users of the network through links. Now days, the popularity of OSNs is increasing significantly. Twitter, Facebook, LinkedIn have more than a hundred million active users. Today the most interactive medium to communicate with others is online social networks (OSN). This online social network is useful for spreading information, pictures and videos and generally staying in touch with people you wouldn't normally get to interact with all the time. Therefore in OSN there is chance of posting undesirable contents/message on particular public/private area, called user walls. In this paper we generally focus on text based messages. The existence of OSNs that include person- specific information creates both interesting opportunities and challenges. For example, data

The aim of present work is thus to propose and through an experiment assess an automatic system, known as Filtered Wall (FW), able to filter undesirable messages from OSN user walls. We have a tendency to exploit Machine Learning (ML) text categorization techniques to mechanically assign with every short text message a group of classes supported its content [1]. Our focus during this work is on on-line identification of real-world event content. We have a tendency to determine every event and its associated Twitter messages victimization an internet clump technique that teams along locally similar tweets. We have a tendency to then work out revealing options for every cluster to assist confirm that clusters correspond to events. significantly, we have a tendency to style options to tell apart between real-world events and a special family of non-events, namely, Twitter-centric or trending topics that carry very little which means outside the

Twitter system. These Twitter-centric activities usually share similar temporal distribution characteristics with real-world events

II. LITERATURE SURVEY

In the literature survey proposed work will discuss recent methods over the Content-based Filtering in On-line Social Networks. Below in literature we are discussing some of them. There are many research papers on message filtration. Planned work consists of the system which may filter undesirable messages from OSN user walls. So, we have to concentrate on content-based message filtering approach on on-line social Networks (OSNs).

M. Vanetti et al. [1] this paper proposes a system implementing content-based message filtering formed as a key service for On-line Social Networks (OSNs). The system permits OSN users to possess an immediate control on the messages posted on their walls. This can be achieved through a versatile rule-based system, that enables a user to customize the filtering criteria to be applied to their walls, and a Machine Learning based soft classifier automatically producing membership labels in support of content-based filtering. In proposed system initially we tend to secure by permitting solely those users who are within the user's friend list can post or write messages on user's wall. Beside that there is Email notification sending to the user who has posted unwanted message on wall when explicit threshold value (ex. If user send unwanted messages three times then send a notification and if subsequently same user send once more unwanted messages then when he cross the threshold frequency of five messages block user).User receives the undesirable content on his/her personal walls. So, activity of classifying such unauthorized message/data and displaying only that message to user that is consistent with his wish is known as information filtering [8].

There are four kinds of information filtering i.e. cooperative filtering, content based, hybrid filtering, and social filtering. N.J. Belkin and W.B. croft et al. [8] information filtering systems are designed for unstructured or semi structured data, as opposed to database applications, that use very structured data. Filtering relies on descriptions of individual or cluster data preferences, or profiles that generally represent long interests. Filtering additionally implies removal of information from an incoming stream instead of finding data within the stream; users see solely the information that is extracted in a very projected model the knowledge is broken in to numerous defined categories like hate, offensive, political and so on. For machine learning classification, feature extraction of content is vital, because it is applied to the training stage of machine learning algorithmic rule. Out of the many text illustration techniques Neural networks [13][14] support vector machines [15]are the efficient ways of machine learning based classification over alternative ways like Naïve Bayesian [16]. F. Sebastiani [17] this survey discusses the most approaches to text categorization that fall within the machine learning paradigm. They need to mention thoroughly problems relating 3completely different issues, namely, document representation, classifier construction, and classifier analysis. For text classification we are using Kohonen Self Organizing Map (SOM) with Redial based function (RBF), for receiving higher accuracy over other ways.

III. PROBLEM DEFINITION

The related system has includes following issues:
1)   In this system anyone is able to post the messages on user's wall even though he/she is not within the friend list therefore there could also be chance of overwhelm by useless information.
2)   In the present System Radial Basis function (RBF) network acts as soft classifier so there perhaps chance of problems in Classification in terms of accuracy.
3)   There is direct blocking of user who has sent unwanted message on wall without sending back notification to him/her.

So the proposed system will overcome these problems and provide another way to block the user for specific time periods and also sending the E-mail notification to that user who posed undesirable message on user walls.

IV. SYSTEM DESIGN

Proposed system planning to block the user for particular time limit and also send notification, E-Mail sending to that user who posted on wall. Along with that proposed system using Kohonen Self Organizing Map (SOM) with Redial Based Function (RBF) for classification of text. This system uses the back propagation technique of neural network that takes the proper action by using previous knowledge of user messages.

A.   Architecture

The Filtered wall architecture in support of OSN services is a three-tier structure (see Fig. 1).
Thee three layers are
1.   Social Network Manager (SNM)
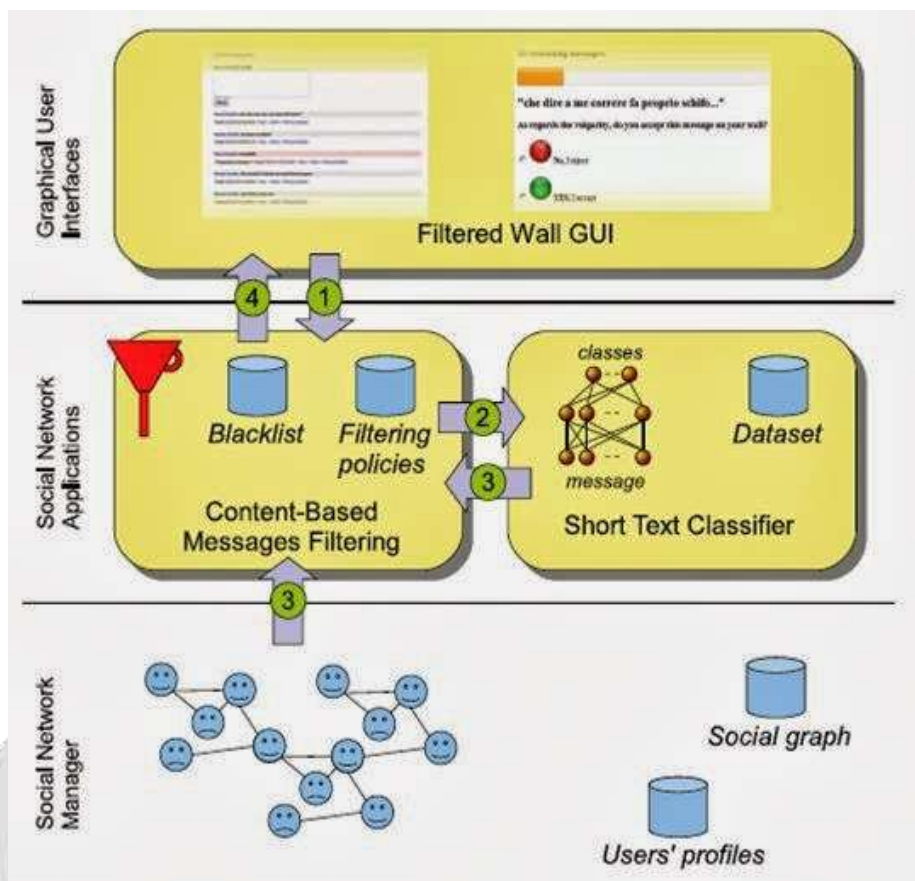2.   Social Network Application (SNA)
3.   Graphical User Interface (GUI)

Fig. 1. Architecture of filtered wall

The first layer, called Social Network Manager (SNM),the main task of it provides the basic OSN functionalities (i.e., profile and relationship management), whereas the second layer composed of Content Based Message Filtering (CBMF) and short text classification , it provides the support for external Social Network Applications (SNAs). The supported SNAs may in turn require an additional layer for their needed Graphical User Interfaces (GUIs). According to this reference architecture, the proposed system is placed in the second and third layers. In particular, users interact with the system by means of a GUI to set up and manage their FRs/ BLs. Moreover, the GUI provides users with a FW, that is, a wall where only messages that are authorized according to their FRs/BLs are published [1]

SONN is used with RBFN for classification of message in proposed work. SONN is incremental learning, that is, it can learn new knowledge without destroying the old learned knowledge. Because the neurons in the network are self-organized, it is not necessary to define the network structure and size in advance. SONN maintains previous history of user's friend in user log file based on message that he/she has posted on user's wall and will perform proper action.

In particular, in the proposed work the overall short text classification strategy on Radial Basis Function Networks (RBFN) with Self Organizing Map (SOM) for their proven capabilities in acting as soft classifiers, in managing noisy data and intrinsically vague classes. Moreover, the speed in performing the learning phase creates the premise for an adequate use in OSN domains, as well as facilitates the experimental evaluation tasks. Proposed work inserts the neural model within a hierarchical two level classification strategy. In the first level, the RBFN with Self Organizing Map (SOM) categorizes short messages as Neutral and Non-neutral; in the second stage, Non-neutral messages are classified producing gradual estimates of appropriateness to each of the considered category. Besides classification facilities, the system provides a powerful rule layer exploiting a flexible language to specify Filtering Rules (FRs), by which users can state what contents, should not be displayed on their walls. FRs can support a variety of different filtering criteria that can be combined and customized according to the user needs.

More precisely, FRs exploit user profiles, user relationships as well as the output of the ML categorization process to state the filtering criteria to be enforced. If the friend of user continuously posts the undesirable messages of particular type on users wall then user will send the notification message to that user who posted on wall. In addition, the system provides the support for user-defined Blacklists (BLs), that is, list of users that are temporarily prevented to post any kind of messages on a user wall. The path followed by a message, from its writing to the possible final publication can be summarized as follows:

1   After entering the private wall of one of his/her friends list, the user tries to post a message, which is intercepted by FW.
2   A ML-based text classifier extracts metadata from the content of the message.
3   FW uses metadata provided by the classifier, together with data extracted from the social graph and users' profiles, to enforce the filtering and BL rules.
4   Depending on the result of the previous step, the message will be post or filtered by FW.

V. MACHINE LEARNING BASED CLASSIFICATIONS

A machine learning approach learns from training data and creates classifier for the classification of new data the main task of text categorization is to assign each text as a predefined category. A hierarchical two level classification process is used for short text categorization. The first-level classifier performs a binary hard categorization that labels messages as Neutral and Non-neutral. After first-level filtering task there is subsequent second-level task in which a fine-grained classification is performed. The second-level classifier performs a soft partition of Non-neutral messages assigning a given message a gradual membership to each of the Non-neutral classes. Among the variety of multiclass Machine Learning models well suited for text classification, we choose the RBFN model [8] with Self Organizing Map (SOM) for the experimented competitive behavior with respect to other state-of-the-art classifiers. RFBNs have a single hidden layer of processing units with local, restricted activation domain: a Gaussian function is commonly used, but any other locally tunable function can be used. They were introduced as a neural network evolution of exact interpolation [9], and are demonstrated to have the universal approximation property [10], [11]. As outlined in [12], RBFN main advantages are that classification function is nonlinear, the model may produce confidence values and it may be robust to outliers; drawbacks are the potential sensitivity to input parameters, and potential overtraining sensitivity. The first-level classifier is then structured as a regular RBFN with Self Organizing Map (SOM). In the second level of the classification stage, we introduce a modification of the standard use of RBFN with Self Organizing Map (SOM). Its regular use in classification includes a hard decision on the output values: according to the winner-take-all rule, a given input pattern is assigned with the class corresponding to the winner output neuron which has the highest value. In our approach, we consider all values of the output neurons as a result of the classification task and we interpret them as gradual estimation of multimember ship to classes.

VI. CONCLUSION

In this paper, a system gives ability to OSN users to have a direct control on the messages posted on their walls. Undesired messages get filtered out and will not get posted on wall. The system exploits a ML soft classifier to enforce customizable content-dependent filtering rules. The flexibility of the system in terms of filtering options is enhanced through the management of BLs. System proposed in this paper represents just the core set of functionalities needed to provide a sophisticated tool for OSN message filtering with temporary blocking of user and also send notification, E-Mail to that who has posted undesirable message on wall. This might enhance services provided by OSN. BL and FR specification are made easier by development of GUI and a set of related tools. Along with it, our proposed system provides a better accuracy for classification of message as compare to previous implemented methods. The miss-classification of message is reduced by using RBF with SONN. By computing matrix (FP, FN, TP, TN, precision and recall) we compare results with other classification techniques.

REFERENCES

[1] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu,J. Freeman, D. Tsai, M. Amde, S. Owen et al., "MLlib: Machine learning in apache spark," J. Mach. Learn. Res., vol. 17, no. 1, pp. 1235–1241, 2016.

[2] E. Inc., "About E-Learning," Available at: http://www.elearningnc.gov/ about elearning/, 2016.

[3] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, " A System to Filter Undesirable Messages from OSN User Walls"," IEEE transactions on knowledge and data engineering, vol. 25, no. 2, pp.285-297, February 2013..

[4] Fang, L., LeFevre, K., " Privacy wizards for social networking sites;" In: WWW ' 10: Proceedings of the 19th international conference on World Wide Web, pp. 351-360. ACM, New York, NY, USA, 2010.

[5] Fong, P.W.L., Anwar, M.M., Zhao, Z., " A privacy preservation model for facebook-style social network systems;" In: Proceedings of 14th European Symposium on Research in Computer Security (ESORICS), pp. 303-320, 2009.

[6] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, " Short Text Classification in Twitter to Improve Information Filtering," Proc. 33rd Int' l ACM SIGIR and Development in Information Retrieval (SIGIR ' 10), pp. 841-842, 2010.

[7] Carminati, B., Ferrari, E, " Access control and privacy in web-based social networks," International Journal of Web Information Systems, pp. 395– 415, 2008.

[8] Churcharoenkrung N., Kim, Y.S., Kang, B.H., " Dynamic web content filtering based on user' s knowledge," International Conference on Information Technology, Coding and Computing 1, pp. 184-188 2005.

[9] F. Sebastiani, " Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1– 47, 2002.

[10] A.K. Jain, R.P.W. Duin, and J. Mao, " Statistical Pattern Recognition: A Review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[11] T. Joachims, " Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. European Conf. Machine Learning, pp. 137-142, 1998.

[12] H. Schu¨ tze, D.A. Hull, and J.O. Pedersen, " A Comparison of Classifiers and Document Representations for the Routing Problem," Proc. 18th Ann. ACM/SIGIR Conf. Research and Development in Information Retrieval , pp. 229-237, 1995.

[13] E.D. Wiener, J.O. Pedersen, and A.S. Weigend, " A Neural Network Approach to Topic Spotting," Proc. Fourth Ann. Symp. Document Analysis and Information Retrieval (SDAIR ' 95), pp. 317- 332, 1995.

[14] J. Park and I.W. Sandberg, " Approximation and Radial-Basis- Function Networks," Neural Computation, vol. 5, pp. 305-316, 1993.

[15] N.J. Belkin and W.B. Croft, " Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Comm. ACM, vol. 35, no. 12, pp. 29-38, 1992.

[16] E.J. Hartman, J.D. Keeler, and J.M. Kowalski, " Layered Neural Networks with Gaussian Hidden Units as Universal Approximations, Neural Computation, vol. 2, pp. 210-215, 1990.

[17] J. Moody and C. Darken, " Fast Learning in Networks of Locally-Tuned Processing Units," Neural Computation, vol. 1, no. 2, pp.

281-294, 1989.

[18] M.J.D. Powell, " Radial Basis Functions for Multivariable Interpolation: A Review," Algorithms for Approximation, pp. 143-167, Clarendon Press, 1987.

[19] S.E. Robertson and K.S. Jones, " Relevance Weighting of Search Terms," J. Am. Soc. for Information Science, vol. 27, no. 3, pp. 129- 146, 1976.