IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Comment Analysis in OSN Framework Using Machine Learning Technique

Dr.Kalaimani Shanmugam[1], Haritha G[2], Kannan M[3]

Professor & Head, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India[1]

Student, Department of Computer Science and Engineering, Arasu Engineering College, Kumbakonam, India[2]

Assistant Professor, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India[3]

Abstract: Today's modern life is completely based on Internet. Now a day's people cannot imagine life without Internet. From last few years people share their views, ideas, information with each other using social networking sites. Such interchanges might include diverse sorts of substance such as text, image, audio and video data. One fundamental issue in today On-line Social Networks (OSNs) is to give users the ability to control the messages posted on their own private space to avoid that unwanted content is displayed. Up to now OSNs provide little support to this requirement. Hence Online Social Networks should be extremely secure and should protect the individual's privacy. The Online Social Network provides the security measures but they were limited. While Socializing the user can access the profile of other members which are involved in social sites and even share data such as images, text, videos etc. One critical issue in user wall is to give users the capability to control the messages posted on their own personal space in order to avoid unwanted content to be displayed on their wall. To overcome this problem, we propose a system allowing OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows users to customize the filtering criteria to be matter-of-fact to their walls, and a Machine Learning based soft classifier automatically labelling messages in content-based filtering.

Keywords: On-line Social Networks (OSNs), Machine learning, short Text Classifier, content-based-filtering

I. INTRODUCTION

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. Documents processed in content-based filtering are mostly textual in nature and this makes content-based filtering close to text classification. The activity of filtering can be modeled, in fact, as a case of single label, binary classification, partitioning incoming documents into relevant and non-relevant categories. More complex filtering systems include multi-label text categorization automatically labeling messages into partial thematic categories. Content-based filtering is mainly based on the use of the ML paradigm according to which a classifier is automatically induced by learning from a set of pre-classified examples. The feature extraction procedure maps text into a compact representation of its content and is uniformly applied to training and generalization phases.

Several experiments prove that Bag of Words (BoW) approaches yield good performance and prevail in general over more sophisticated text representation that may have superior semantics but lower statistical quality. The Short texts are classified based on the STC (Short Text Classifier) approach. The system is intended to that where malicious users can be blocked. The System uses machine learning classifier to enforce customizable content dependent rules.

II. BACKGROUND KNOWLEDGE

In this section, most important terms involved and techniques used are discussed to make the clear background knowledge before entering into the main work.

A. Data Mining

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD, a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database

and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data Mining is the process of posing queries to large amounts of data sources and extracting patterns and trends using statistical and machine learning techniques. Data mining tasks include classification, clustering, making associations and anomaly detection.

B. Process In Data Mining

Different data mining processes can be classified into two types: data preparation or data pre-processing and data mining. In fact, the first four processes, that are data cleaning, data integration, data selection and data transformation, are considered as data preparation processes. The last three processes including data mining, pattern evaluation and knowledge representation are integrated into one process called data mining.

Data Cleaning: Data cleaning is the process where the data gets cleaned. Data in the real world is normally incomplete, noisy and inconsistent. The output of data cleaning process is adequately cleaned data.
Data Integration: Data integration is the process where data from different data sources are integrated into one. Data lies in different formats in different locations. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.
Data Selection: Data mining process requires large volumes of historical data for analysis. So, usually the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored. Data selection is the process where the data relevant to the analysis is retrieved from the database.
Data Transformation: Data transformation is the process of transforming and consolidating the data into different forms that are suitable for mining. After data integration, the available data is ready for data mining.
Data Mining: Data mining is the core process where a number of complex and intelligent methods are applied to extract patterns from data. Data mining process includes a number of tasks such as association, classification, prediction, clustering, and time series analysis and so on.
Pattern Evaluation: The pattern evaluation identifies the truly interesting patterns representing knowledge based on different types of interestingness measures. A pattern is considered to be interesting if it is potentially useful, easily understandable by humans, validates some hypothesis that someone wants to confirm or valid on new data with some degree of certainty.
Knowledge Representation: The information mined from the data needs to be presented to the user in an appealing way. Different knowledge representation and visualization techniques are applied to provide the output of data mining to the users.

C. Technique In Data Mining

Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. The data classification process involves learning and classification. In Learning the training data are analysed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. The algorithm then encodes these parameters into a model called a classifier.
Clustering: Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification.
Predication: Regression technique can be adapted for predication. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. The same model types can often be used for both regression and classification.
Association rule: Association and correlation is usually to find frequent item set findings among large data sets. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. Types of association rule

- Multilevel association rule
- Multidimensional association rule

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Neural networks: Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

III. RELATED WORK

"**DYNAMIC RUMOR INFLUENCE MINIMIZATION WITH USER EXPERIENCE IN SOCIAL NETWORKS**" by Biaowang, the paper investigates the problem of dynamic rumor influence minimization with user experience. First, based on existing works on information diffusion in social networks, we incorporate the rumor popularity dynamics in the diffusion model. We analyze existing investigations on topic propagation dynamics and bursty topic patterns. Then we choose Chi-squared distribution to approximate the global rumor popularity Inspired by the novel energy model proposed by Han et al., we then analyze the individual tendency towards the rumor and present the probability of successful rumor propagation between a pair of nodes. Finally, inspired by the concept of Ising model, we derive the cooperative succeeding probability of rumor propagation that integrates the global rumor popularity with individual tendency. After that, we introduce the concept of user experience utility function and analyze the impact of blocking time of nodes to the rumor propagation process. We then adopt the survival theory to explain the likelihood of nodes getting activated, and propose both greedy and dynamic algorithms based on maximum likelihood principle. We propose a rumor propagation model taking into account the following three elements: First,the global popularity of the rumor over the entire social network, i.e., the general topic dynamics. Second, the attraction dynamics of the rumor to a potential spreader, i.e., the individual tendency to forward the rumor to its neighbors.Third, the acceptance probability of the rumor recipients. In our model, inspired by the Ising model, we combine all three factors together to propose a cooperative rumor propagation probability. In our rumor blocking strategies, we consider the influence of blocking time to user experience in real world social networks.

"**MAXIMIZING ACCEPTANCE PROBABILITY FOR ACTIVE FRIENDING IN ON-LINE SOCIAL NETWORKS**"byDe-Nian Yang**,** thepaper makes a grand suggestion for the social networking service providers to support active friending. To support active friending, the key issue is on the design of the algorithms that select the recommendation candidates. A simple scheme is to provide recommendations by unveiling the shortest path between the initiator and the target in the social network, i.e., recommending one candidate at each step along the path. As such, the initiator can gradually approach the target by acquainting the individuals on the path. However, this shortest-path recommendation approach may fail as soon as a middle-person does not accept the friending invitation (since only one candidate is included in the recommendation list for each step). To address this issue, it is desirable to recommend multiple candidates at each step since the initiator is more likely to share more common friends with the target and thereby more likely to get accepted by the target. Especially, by broadcasting the friending invitations to all neighbors of the initiator's friends, the probability to reach the friending target and get accepted can be effectively maximized as enormous number of paths is flooded with invitations to approach the target. Nevertheless, friending invitations are abused here because the above unidirectional broadcast is aimless and prone to involve many unnecessary neighbors. Moreover, the initiator may not want to handle a large number of tedious invitations. In this paper, we study a new optimization problem, called Acceptance Probability Maximization (APM), for active friending in on-line social networks. The service providers, who eager to explore new monetary tools for revenue increase, may consider charging the users from active friending service.

"**TOPIC AND ROLE DISCOVERY IN SOCIAL NETWORKS**"by Andrew Mccallum**,**the paper presents the Author-Recipient-Topic (ART) model, a directed graphical model of words in a message generated given their author and a set of recipients. The model is similar to the Author-Topic (AT) model, but with the crucial enhancement that it conditions the per-message topic distribution jointly on both the author and individual recipients, rather than on individual authors. Thus the discovery of topics in the ART model is influenced by the social structure in which messages are sent and received. Each topic consists of a multinomial distribution over words. Each author-recipient pair has a distribution over topics. We can also easily calculate marginal distributions over topics conditioned solely on an author, or solely on a recipient, in order to find the topics on which each person is most likely to send or receive. Most importantly, we can also effectively use these person conditioned topic distributions to measure similarity between people, and thus discover people's roles by clustering using this similarity. For example, people who receive messages containing requests for photocopying, travel bookings, and meeting room arrangements can all be said to have the role "administrative assistant," and can be discovered as such because in the ART model they will all have these topics with high probability in their receiving distribution. Note that we can discover that two people have similar roles even if in the graph they are connected to very different sets of people. Thus, we propose an Author-Recipient-Topic (ART) model for message data. The ART model captures topics and the directed social network of senders and recipients by conditioning the multinomial

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 6, June 2020

distribution over topics distinctly on both the author and one recipient of a message. Unlike the AT, the ART model takes into consideration both author and recipients distinctly, in addition to modeling the email content as a mixture of topics. The ART model is a Bayesian network that simultaneously models message content, as well as the directed social network in which the messages are sent.

IV. PROPOSED SYSTEM

On-line Social Networks (OSNs) are today one of the most popular interactive mediums to communicate, share and disseminate a considerable amount of human life information. One fundamental issue in today On-line Social Networks (OSNs) is to give users the ability to control the messages posted on their own private space to avoid that unwanted content is displayed. Up to now OSNs provide little support to this requirement. To fill the gap, in this paper, we propose a system allowing OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning based soft classifier automatically labeling messages in support of content-based filtering. Machine learning (ML) is used as text categorization techniques to automatically assign each short text message with in a set of categories based on its content.

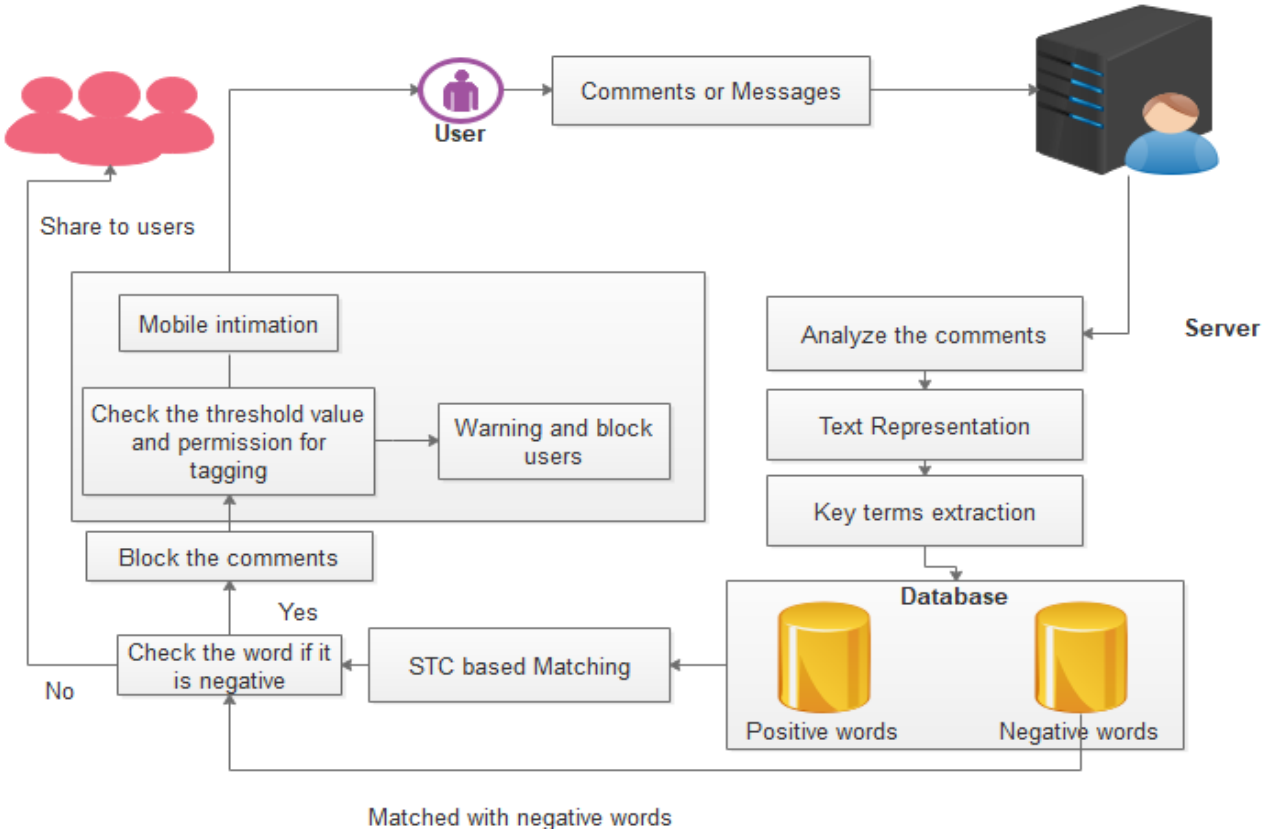

Fig. 1: Workflow of the proposed comment analysis on OSN framework

The major efforts in building a robust Short Text Classifier (STC) concentrate in the extraction and selection of a set characterizing and discriminating features. Here, a database of the categorized words is built and it is used to check the words if it has any indecent words. If the message consists of any vulgar words, then they will be sent to the Blacklists to filter out those words from the message. Finally, the message without the indecent words will be posted in the user's wall on the result of the content-based-filtering technique. A system automatically filters unwanted messages using the blacklists on the basis of both message content and the message creator relationships and characteristics. Major difference includes, a different semantics for filtering rules to better fit the considered domain, to help the users Filtering Rules (FRs) specification, the extension of the set of features considered in the classification process.

A. Framework Construction: A social networking service (also social networking site, SNS or social media) is an online platform that people use to build social networks or social relations with other people who share similar personal or career interests, activities, backgrounds or real-life connections. The variety and evolving range of stand-alone and built-in social networking services in the online space introduces a challenge of definition. Social network refers to interaction among people in which they create, share, and/or exchange information and ideas in virtual

communities and networks. Design the GUI which is the type of user interface that allows users to interact with users through graphical icons and visual indicators. In this module we can create the interface for admin and user. User can login to the system and view the friend request. The user can share the images to friends.

B. ***Post Comments:*** Social media is becoming an integral part of life online as social websites and applications proliferate. Most traditional online media include social components, such as comment fields for users. In business, social media is used to market products, promote brands, and connect to current customers and foster new business. In this module, we can comment in online social network. Comment in the form of text. The text may be uni-gram, bi-gram and multi grams. This module is used to get the input from social users. Comments may be various forms such as links or texts or short texts. Comments are read and send to server page.

C. ***STC Implmentation:*** In this module, we design an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls. The architecture in support of OSN services is a three-tier structure. The first layer commonly aims to provide the basic OSN functionalities (i.e., profile and relationship management). Additionally, some OSNs provide an additional layer allowing the support of external Social Network Applications (SNA). Finally, the supported SNA may require an additional layer for their needed graphical user interfaces (GUIs). The major efforts in building a robust short text classifier (STC) are concentrated in the extraction and selection of a set of characterizing and discriminant features. In order to specify and enforce these constraints, we make use of the text classification. From STC point of view, we approach the task by defining a hierarchical two-level strategy assuming that it is better to identify and eliminate "neutral" sentences, then classify "non-neutral" sentences by the class of interest instead of doing everything in one step.

D. ***Filtered Rules Implementation:*** The filtering rules should allow users to state constraints on message creators. Thus, creators on which a filtering rule applies should be selected on the basis of several different criteria; one of the most relevant is by imposing conditions on user profile's attributes. In such a way it is, for instance, possible to define rules applying only to young creators, to creators with a given religious/ political view, or to creators that we believe are not expert in a given field (e.g. by posing constraints on the work attribute of user profile). This means filtering rules identifying messages according to constraints on their contents. And block the users who are post the negative comments more than five times and also send mobile intimation to users at the time offline.

E. ***Filtered GUI:*** BL'S are directly managed by the system, which should be able to determine who the users are inserted in the BL and decide when the user retention in the BL is finished. To improve flexibility, this information is in the system by a set of rules; the rules on BL. Rules are generated by server for setting threshold values. Based on threshold values, we can block friends who are providing negative comments. Finally provide mobile intimation to users.

V. PROPOSED METHODOLOGY

In this proposed methodology, the main goal is to make the text classification in sustained way to obtain the predicted results through the following major techniques.

A. ***Text Mining Algorithm:*** In the first step, the comments are collected and forward to admin page.

Document Pre- Processing: In this process, the given input document is processed for removing redundancies, inconsistencies, separate words, stemming and documents are prepared for next step, the stages performed are as follows:
Tokenization - The given document is considered as a string and identifying single word in document.
Removal of Stop Word - In this step the removal of usual words like a, an, but, and, of, the etc. is done.
Stemming - A stem is a natural group of words with equal (or very similar) meaning. This method describes the base of particular word. Inflectional and derivational stemming are two types of method. One of the popular algorithms for stemming is porter's algorithm. e.g. if a document pertains word like resignation, resigned, resigns then it will be considered as resign after applying stemming method.

B. ***Short Text Classification:*** A hierarchical two-level classification is advantageous to short text classification as per the suggestion. The first level of a classifier labels the message into neutral and non-neutral. In second level non neutral messages are estimated into one or more of the conceived categories.

Filtering rule - A filtering rule is a tuple (auth, CreaSpec, ConSpec, action)
1. auth is the user who state the rule.
2. CreaSpec is the Creator specification.
3. ConSpec is a boolean expression.
4. action is the action performed by the system.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Filtering rules will be applied, when a user profile does not hold value for attributes submitted by a FR. This type of situation will be dealt with asking the owner to choose whether to block or notify the messages initiating from the profile which does not match with the wall owners FRs, due to missing of attributes.

C. Blacklist: The main implementation of our paper is to execute the Blacklist Mechanism, which will keep away messages from undesired creators. BL are handled undeviating by the system. This will able to decide the users to be inserted in the blacklist. And it also decides the user preservation in the BL will get over. Set of rules are applied to improve the stiffness, such rules are called BL rules. By applying the BL rule, owner can identify which user should be blocked based on the relationship in OSN and the user's profile. The user may have bad opinion about the users can be banned for an uncertain time period. We have two information's based on bad attitude of user. Two principles are stated. First one is within a given time period user will be inserted in BL for numerous times, he /she must be worthy for staying in BL for another sometime. This principle will be applied to user who inserted in BL at least once. Relative Frequency is used to find out the system, who messages continue to fail the FR. Two measures can be calculated globally and locally, which will consider only the message in local and in global it will consider all the OSN users walls.

VI. RESULTS AND DISCUSSIONS

In this system uses the ML soft classifier is to remove the unwanted messages. BL is used to enhance the flexibility of system for filtering. We will be designing the system which is more sophisticated approach to decide when a user should be inserted into the BL. Moreover, the flexibility of the system in terms of filtering options is enhanced through the management of BLs. This work is the first step of a wider project.

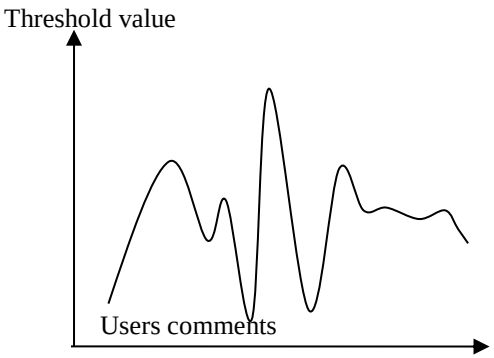

Fig. 2: An illustration of threshold prediction curve

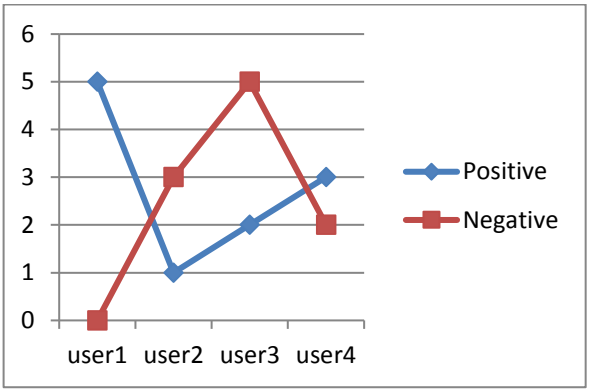

Fig. 3: Prediction of friends to Black List

Threshold is a point beyond which there is a change in the manner a program executes. As illustrated in the figure 2, threshold value is predicted on the basis of negative comments provided by the user.By applying the BL rule,owner can identify which user should be blocked based on the relationship in OSN and the user's profile. The threshold value is showed by using positive and negative comments posted by several users are shown in the figure 3. The user may have bad opinion about the users can be banned for an uncertain time period. The count value of negative post which is posted by same friend is checked and based on the threshold value it automatically blocks the friend. Then send the mobile intimation about blocking comments and users.

VII. CONCLUSION AND THE FUTURE WORK

In this paper, a system is designed to filter undesired messages from OSN walls. The system exploits a ML soft classifier to enforce customizable content dependent FRS. The major efforts in building a robust short text classifier are concentrated in the extraction and selection of a set of characterizing and discriminant features. Moreover, the flexibility of the system in terms of filtering options is enhanced through the management of BLs. This work is the first step of a wider project. The early encouraging results we have obtained on the classification procedure prompt us to continue with other work that will aim to improve the quality of classification. 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. In addition, the system provides the support for user-defined Black Lists (BLs), that is, lists of users that are temporarily prevented to post any kind of messages on a

user wall. In future work, the framework can be intended to exploit similar techniques to infer Black List rules and Filter Rules. Then the system can be implemented with various languages with improved accuracy.

REFERENCES

[1]. B. Wang, G. Chen, L. Fu, L. Song, and X. Wang, "Drimux: Dynamic rumor influence minimization with user experience in social networks," in Proc. 30th AAAI Int. Conf. Artif. Intell., Feb. 2016.
[2]. D. N. Yang, H. J. Hung, W. C. Lee, and W. Chen, "Maximizing acceptance probability for active friending in online social networks," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 713–721.
[3]. A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," in Proc. 19th Int. Joint Conf. Artif. Intell., 2005, pp. 786–791.
[4]. L. Fu, W. Huang, X. Gan, F. Yang, and X. Wang, "Capacity of wireless networks with social characteristics," IEEE Trans. Wireless Commun., vol. 15, pp. 1505–1516, Feb. 2016.
[5]. A. Montanari and A. Saberi, "The spread of innovations in social networks," in Proc. National Academy of Sciences of the United States of America PNAS, Aug. 2010, pp. 20 196–20 201
[6]. X. Rong and Q. Mei, "Diffusion of innovations revisited: From social network to innovation network," in Proc. 22Nd ACM Int. Conf. Inf. Knowl. Manag., 2013, pp. 499–508.
[7]. C. Budak, D. Agrawal, and A. E. Abbadi, "Limiting the spread of misinformation in social networks," in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 665–674.
[8]. A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Viral misinformation: The role of homophily and polarization," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 355–356.
[9]. E. Serrano, C. A. Iglesias, and M. Garijo, "A novel agent-based rumor spreading model in twitter," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 811–814.
[10]. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003, pp. 1175–1180.