ISSN:0975-9646

A Framework to Refine Awful Tweets from e-Networks

Ratna Kailashnadh Singamsetty[1], Sravya Tata[1], Vijay Varma Vatsavai[1], Indira Sattiraju[1], Joshua Botcha[2]

[1]B.tech Student, [2]Assistant Professor
Dept. of CSE LIET,
Vizianagaram, India

Abstract—**E-Network enables its users to keep in touch with friends by exchanging several type of content including text, audio, and video data. Users of these sites do not have much control to avoid awful content to be displayed on their own private space generally called as wall. Therefore a major task of today's on stream e-network is information refining. Using machine learning approach and a rule based system, text classification and customization of refining criteria to be applied on user's wall is to be achieved .From this survey, we will be able to see the challenges in short text classification and refining criteria that should be considered while publishing tweets on user wall.**

Keywords- E-Network; Information refining; short text classification

I. INTRODUCTION

E-networks are the latest on stream trend of the last few years to meet people and share information with them. Users of these on stream networking sites forms an e-network, which provides a powerful means of organizing and finding useful information. This communication involves exchange of several types of content including text, image, audio, and video data. Therefore in e-networks, there is a chance of posting awful content on particular public or private areas, generally referred as walls. Information refining has been greatly explored for which it concerns the textual documents and web content. It can be used to give users the ability to automatically control the tweets written on their own walls, by refining out awful tweets. In this paper, our main aim is to survey the classification technique and to study the design of system to refine the awful tweets from e-network user walls. The aim of the present work is therefore to propose and experimentally evaluate an automated system, called refined wall, able to refine awful tweets from e-network. We exploit machine learning text categorization techniques to automatically assign with each short tweet, a set of categories based on its content.

II. RELATED WORK

The main goal of information refining system is to refine awful content from input data before it is presented to the end user. It takes into account the user's profile and compares it with referred characteristics or properties. Supported systems have become popular in recent years. It is a type of information refining system that predicts the preference that user might give to an item or to the social element. It takes into account the user interest and recommends an item. Supported systems works in one of the following ways:

- Content based refining
- Collaborative refining
- Policy- based actualization

Content-based refining

Content based refining selects an item based on the user interest. It uses items previously preferred by the user and then suggests the best matched item. Each user acts independently in content based system. This kind of system chooses item depending on relation between item content and user recommendations against collaborative system that selects item based on relation between people with similar preferences [7]. The content based system creates a content based profile of a user, based on rated items of a user. Items features are weighted based on features preferred by the user and recommendations are given by the system accordingly. In content based refining, the main issue is whether the system is able to learn from user's actions related to a particular content source and use them for other content types. Script classification is similar to content based refining as documents processed in such type of systems are mostly textual. In e-networks, user's social profile has to be taken into account and this makes content based refining system difficult to apply in e-network domain as a standalone system. 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 approaches yield good performance over more sophisticated text representation that may have superior semantics but lower statistical quality.

Collaborative refining

Collaborative refining system selects information item based on user's preferences, actions, and predicts what users will like based on his similarities to other users. Items are rated on the basis of user likes and dislikes [4] .Collaborative refining involves collaboration of multiple agents while refining information. Collaborative refining system often requires large data set. Amazon uses item to item collaborative refining for its recommendation system. The collaborative approach is suitable for popular items but effective content information is not much gained as it is opposed to content based approach which is more suitable for unpopular items and effective content information is easily available.

Policy-based actualization

Policy based actualization is applicable in many different contexts. It adapts a service in specific context according

user defined policies. In on stream social networking sites, user oriented policies can define how communication between two parties or more can be handled. The policy based actualization system in [3] focuses on twitter. It assigns a category to each tweet and shows only that tweets to the user which are of his interest. In this scenario, policy based actualization represent the ability of the user to refine wall tweets according to refining criteria suggested by him. Finally, our policy language has some relationships with policy framework that have been so far proposed to support the specification and enforcement of policies which are expressed in terms of constraints on the machine understandable resource descriptions provided by semantic web languages.

III. GADGET LEARNING TECHNIQUES

A gadget learning approach is from training data and creates classifiers for the classification of new data set. The main task of text classification is, to assign a predefined category with each text. Text classification is accomplished on the basis of endogenous collection of data. The gadget learning based classifier learns how to classify the categories of incoming data on the basis of features extracted from the set of training data. The key methods which are commonly used for a text classification are:

- Neural network classifiers
- Support vector gadgets
- Naive Bayer classifier
- Decision tree

Neural network classifier

Neural network classifiers consist of neurons which are arranged in layers to convert an input vector into an output. The commonly used neural network is a multilayered feed forward network in which a unit feeds its output to all the units of the next layer but, there is no feedback to the previous layer. Radial basis function network is an artificial neural network which uses the radial basis function as an activation function. The output of this network is a linear combination of radial basis function of the inputs and neuron parameters. It is robust to out-lie [5] and therefore more suitable in this context.

Support vector gadgets

The support vector gadget classifiers analyze data and recognize pattern in it. They are based on supervised learning model and are able to perform nonlinear ranking in addition to linear classification. The support vector gadget classifier is suitable for large amount of unlabeled data and small amount of unlabeled data [4] .The high dimensional input space, irrelevant features, sparse document vectors and linearly separable text classification makes support vector gadget classifier suitable for script categorization [4].

Naive Bayer classifier

Naive Bayer classifier is a probabilistic classifier which is based on the Bayer theorem with the independence assumption [4].Prefer a class variable which assumes the presence or absence of specific feature which is unrelated to the presence or absence of any other feature. Bayer classifier considers each of these features independently to the probability regardless of the presence or absence of any other feature. The main advantage of this classifier is that it requires a small amount of training data to estimate the parameters required for classification.

Decision trees

Decision tree classifiers are used for a hierarchical decomposition of the data space. It determines the predicate or a condition depending on an attribute value. Class labels in the leaf node are used for classification. In order to reduce the over fitting, data pruning is required in the decision tree. This classifier requires iterative training procedure and is oversensitive to the training data [5].

IV. SHORT SCRIPT CLASSIFIER

A hierarchical two level classification is advantageous to the short script classification [1].The first level of a classifier labels the tweet into neutral and non-neutral form. In second level, non neutral tweets are estimated into one or more of the conceived categories.

Script representation

Script representation of a given document is important task which strongly affect the performance of classification process. It is done by extracting features for a given document. The investigation from the previous data [7] suggests three types of features important for text representation. They are bag of words, document properties, and contextual features. The first two types of features are entirely derived from the information contained within the text of the tweet [7] whereas, contextual features are exogenous. Script representation is done using an endogenous function. In bag of words, representation terms are identified with in the words. It is also important to use feature which is extracted from outside the tweet content but, related to tweet itself. A contextual feature is been introduced in [8] which characterize the environment where the user is posting. It determines the semantics of tweet [6]. Vector space model is the model of script representation by which a script document is represented as a vector of a binary or real weight. These three features are experimentally evaluated for short script classification in [8] for their appropriateness.

Gadget learning based classification

As short script classification is hierarchical with two level tasks and it should be robust to out-lie and hence a radial basis function model is used for short script classification. A radial basis function model is chosen as per the experimental evaluation in [14] [11] among the other classifier.

V. REFINED WALL ARCHITECTRE

The architecture of e-network services is a three-tier structure of three layers as shown in the figure 1[7].The three layers are:

- Social Network Manager
- Social Network Application
- Graphical User Interface

The main task of social network manager is the profile and relationship management. It maintains the data related to

user profile and provides the data to the second layer for applying refining rules and blacklists .Second layer composed of content base tweet refining and a short text classifier which is the most important layer. The classifier categorizes each tweet according to its content and content base tweet refining which refines the tweet according to the refining criteria and blacklist provided by the user. Third layer consists of graphical user interface by which user provide his input and is able to see published wall tweets. Additionally, graphical user interface provides user the facility to apply refining rules for his wall tweets, and helps to provide list of black list users who are temporally prevented to publish tweets on user's wall. The graphical user interface also consists of a refined wall where the user is able to see his desired tweets. As per the refined wall architecture, when the user tries to post a tweet on a private wall of his or her contact, it is intercepted by the refined wall. Then, a short script classifier categorizes a tweet according to its content and then content base tweet refining applies refined wall and black list as per the data provided by the third layer. Based on the result of above step, the tweet is published or refined by refined wall.

VI. VII. REFINING RULES AND BLACK LIST MANAGEMENT

Refining rules

User can state what contents should be blocked or displayed on refined wall by means of refining rules. Refining rules are specified on the basis of user profile as well as user social relationship. Refining rules is dependent on following factors:

- Author
- Creator Specification
- Content Specification
- Action

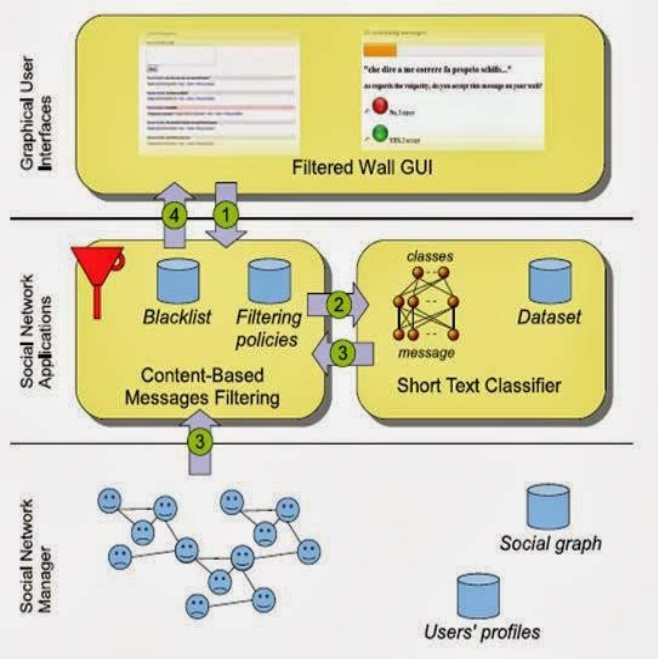

Figure 1: A refined wall conceptual architecture

An author is a person who defines the rules. Creator specification denotes the set of e-network users. Content specification is a Boolean expression defined on the content. Action denotes the action to be performed by the system on the tweet which matches on the content specification and is created by the users identified by Creator specification [8]. In this paper, we do not address the problem of trust computation for indirect relationships since; many algorithms have been proposed in the literature that can be used in our scenario as well. Such algorithms mainly differ on the criteria to select the paths, on which trust computations should be based on many paths which are of the same type, exist between two users.

Black lists

Black listing rules enable the wall owner to determine users to be blocked on the basis of their profiles, and relationship with the wall owner. This boycotting can be done for a specific period or forever according wall owner's desire. Like refining rules, black listing also depends on the author, creator specification, and creator behaviour.

VII. EVALUATION METRICS

First level of classification is evaluated by means of contingency table approach. The overall accuracy index gives the percentage between truth and classification results complimented with Cohen's kappa coefficient. The Cohen's kappa coefficient is more robust measure [13]. The second level of classification is evaluated on the basis of precision and recall. Precision gives the number of false positive values and recall gives the number of false negative values of classification. The overall metric is then computed by the harmonic mean between precision and recall [12].

VIII. INVESTIGATION OUTCOME

Existing technique works well for long text classification but suffers when the text is short. Short text does not have multiple occurrences of words therefore; classification of short text is a challenging task. Short text classification is a hierarchical two level task, consisting of hard and soft classification. Among the variety of machine learning models for text classification, radial basis function model is well suited. Its main advantages are that the classification function is non-linear and the model can generate confidence. It is robust to out-lie and this makes this model well suited.

CONCLUSION

This paper presents an approach of short text classification and design of a system, to refine awful tweets from e-network walls. Additionally, the flexibility of a system can be enhanced through refining rules and blacklist management. In this context, the underlying domain is dynamic n and the collection of pre-classified data provided for training purpose may not be valid for a longer time. A preliminary work in this direction has been done in the context of trust values used for e-network access control purposes. However, we would like to remark that, the system proposed in this paper represents just the core set of functions needed to provide a sophisticated tool for e-network tweet refining. Even if we have complemented our system with an online assistant to set refining rule

thresholds, the development of a complete system easily usable by average e-network users is a wide topic which is out of the scope of the current paper. As such, the face book applications to be meant as a proof-of-concepts of the system core functions, rather than a fully developed system.

REFERENCES

[1] A. Adoma vicius, G .and Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[2] M Chau and H. Chen, "A machine learning approach to web page filtering using content and structure analysis," DecisionSupport Systems, vol. 44, no. 2, pp. 482–494, 2008.

[3] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, 2010, pp. 841–842.

[4] K. Nirmala, S. Satheesh kumar, "A Survey on Text Categorization in Online Social Networks," in Proceedings of International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 9, September 2013**.**

[5] A. K. Jain, R. P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp, 4–37, 2000.

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

[7] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari,"Content-based filtering in on-line social networks," in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning (PSDML 2010), 2010.

[8] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, Moreno Carullo, "A System to Filter unwanted Messages from OSN User Walls," IEEE Transaction on Knowledge and Data Engineering, vol. 25,2013.

[9] S. Zelikovitz and H. Hirsh, "Improving short text classification using unlabeled background knowledge," in Proceedings of 17th International Conference on Machine Learning (ICML-00), P. Langley, Ed.Stanford, US: Morgan Kaufmann Publishers, San Francisco, US,2000, pp. 1183–1190.

[10] S. Pollock, "A rule-based message filtering system," ACM Transactions on Office Information Systems, vol. 6, no. 3, pp. 232–254, 1988.

[11] J. Moody and C. Darken, "Fast learning in networks of locally-tuned processing units," Neural Computation, vol. 1, p: no 281, 1989.

[12] W. B. Frakes and R. A. Baize-Yates, Information Retrieval: Data Structures & Algorithms, Prentice-Hal l, 1992.

[13] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp.159–174, March 1977.

[14] P. E. Baclace, "Competitive agents for information filtering," Communications of the ACM, vol. 35, no. 12, p. 50, 1992.