DISCOVERING EMERGING TOPICS IN SOCIAL STREAMS VIA LINK ANOMALY DETECTION

Tintomon.P.A[1] , N.Santhana Krishna[2]
M.Phil Scholar Department of CS, AJK CAS[1], HOD Department of MSc.CS, AJK CAS[2]
tintomonpa@gmail.com[1], seema.80.sk@gmail.com[2]

ABSTRACT:

Discovery of emerging topics is now getting converted interest motivated by the rapid growth of social networks. We focus on surfacing of topics signaled by social aspect of these networks. Specifically, we focus on mentions of users—link among users that are generated energetically through replies, mentions, and retweets. We propose a probability model of the mentioning performance of a social network user, and propose to detect the emergence of a new topic. We demonstrate our method in numerous real data sets we gathered from Twitter. The experiments show that the proposed mention – anomaly - based approaches can also be detected by new topics at least as early as text-anomaly-based approaches, and in some cases greatly before when the topic is badly identified by the textual contents in posts.

Keywords: *Anomaly detection, Social network communication, knowledge mining.*

1. INTRODUCTION:

1.1 Basic Concepts

Announcement over social networks, such as Twitter and Facebook, is gaining its importance in our daily life. Since the data transformed over networks includes texts, URLs, images, and videos, they are challenging test beds for the study of knowledge mining. In exacting, we are engrossed in the problem of detecting emerging topics from social streams, which can be used to create computerized news and discover hidden market needs or underground political movements. Compared to conservative media, social media are able to capture the earliest, unedited voice of ordinary people. One post may have a number of mentions. Some users may include mentions in their posts rarely; and also being mentioned may be in a unusual circumstance. In this wisdom, a language with the number of words equal to the number of users in a social network.

2.MOTIVATION AND OBJECTIVES :

The objective of this paper is to characterize the computerization aspect of Twitter account, and to organize them into three categories, human, and cyber, accordingly.

This will help Twitter handle the society better and help human users identify who they are twittering with. Based on the measurement results, we propose an automated categorization system may consist of 4 major components:

1. The entropy module uses tweeting period as a measure of behavior density, and detects the interrupted and regular timing that is a display of computerization;

2. The spam discovery module uses tweet substance to check whether text patterns contain spam or not;

3. The account properties module employs useful account property, such as tweeting device makeup, URL ration, to detect deviation from normal; and

4. The resolution maker is based on Random Forest, and it uses the grouping of the features generated by the above three mechanism to classify an unknown user as human, or cyber.

3. METHODOLOGY:

3.1 PROPOSED SYSTEM

We suggest and experimentally estimate an automatic system, called Filtered Wall (FW), able to filter unnecessary communication from OSN user walls. We exploit Machine Learning (ML) text classification techniques [4] to mechanically allocate with every short text message a set of category based on its substance. The main efforts in building a tough short text classifier (STC) are determined in the mining and selection of a set of characterizing and distinguish features. The solution investigated in this paper is an addition of those adopted in a preceding work by us since whom we inherit the learning model and the elicitation method for generating reclassified data. As far as the learning model is concerned, we validate in the present paper the use of neural knowledge which is today accepted as one of the most proficient solutions in text cataloging. The general short text classification strategy on Radial Basis Function Networks (RBFN) for their proven capabilities in temporary as soft classifiers, in supervision noisy data and intrinsically vague classes. Furthermore, the speed in performing the knowledge phase creates the premise for an adequate use in OSN domains, and facilitates the experimental evaluation tasks. The data analysis is to measure the accurate of the tweet. This will validate the tweet content and their account details such as Ip address, Credential and Date time stamp of the received tweet. All details are captured in the dataset to validate the account properties for the user.

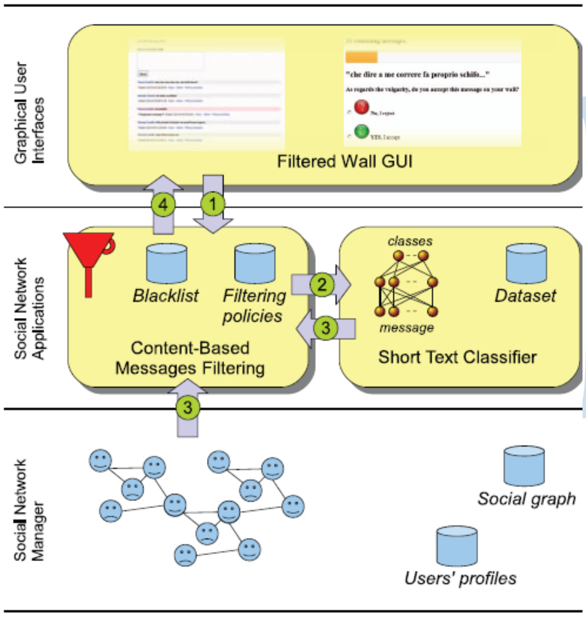

Graphical User Interfaces
"che dire a me correre fa proprio schifo..."
As regards the vulgarity, do you accept this message on your wall?
No, I reject
YES, I accept
Filtered Wall GUI
4
1
Social Network Applications
Blacklist
Filtering policies
Content-Based Messages Filtering
2
3
classes
message
Dataset
Short Text Classifier
3
Social Network Manager
Social graph
Users' profiles

Figure 1 Block diagram of Proposed System

3.2 DATA PREPROCESSING

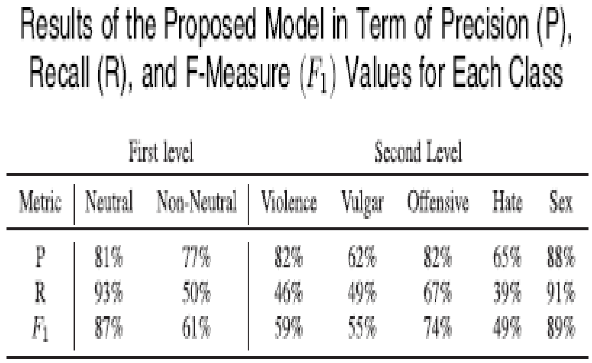
Preprocessing is a very critical and difficult step as its result has a shortest collision on the rules and pattern generated by data mining algorithms. The reason of preprocessing is to transfer various input such as content, construction and usage information into the layout which data mining algorithms can handle straightforwardly. The major task in this phase includes handling missing values, identifying outliers, smooth out noisy data and correct inconsistent data. This section discusses the method used for data preprocessing.

3.2.1 Machine learning-based classification

We address short text classification as a hierarchical two level cataloging process. The first-level classifier performs a binary hard classification that labels messages as Neutral and No neutral. The first-level filter task facilitates the succeeding second-level task in which a finer-grained categorization is performed. The second-level classifier performs a soft-partition of No neutral messages assigning a given message a gradual membership to each of the no neutral classes. The first-level classifier is then ordered as a regular RBFN. In the second level of the organization stage, we introduce a alteration of the standard use of RBFN. Its regular use in organization includes a hard decision on the output values: according to the winner-take-all rule, a known input pattern is assigned to the class corresponding to the winner output neuron which has the highest value. In this approach, we consider all values of the output neurons as a result of the categorization task and we interpreted as gradual estimation of multi membership to classes.

4. PERFORMANCE METRICS:

In order to provide an general estimation of how efficiently the system applies a FR, we look again at Table 1. This table allow us to guess the Precision and Recall of our FRs, and values reported in Table 2 have been compute for FRs with content specification component set to (C, 0:5), where $C \in \Omega$. Let us suppose that the system applies a given ruling on a convinced message.



Results of the Proposed Model in Term of Precision (P), Recall (R), and F-Measure (F_1) Values for Each Class

	First level		Second Level				
Metric	Neutral	Non-Neutral	Violence	Vulgar	Offensive	Hate	Sex
P	81%	77%	82%	62%	82%	65%	88%
R	93%	50%	46%	49%	67%	39%	91%
F_1	87%	61%	59%	55%	74%	49%	89%

. As such, Precision reported in Table 1 is the probability that the verdict taken on the measured message (that is, blocking it or not) is really the exact one. In difference, recall has to be interpreted as the probability that, given a rule that must be applied in a firm message, the rule is really enforced. Let us discuss, with some example, the results presented in Table 1, which reports Precision and Recall values.

The second column of table 1 represents the Precision and the Recall value computed for FRs with content constraint. In difference, the fifth column supplies the Precision and the Recall value computes for FRs with (V ulgar, 0.5) constraint. Results obtained by the content-based condition component, on the first-level arrangement, can be considered with those obtained by well-known information filtering techniques. However, the analysis of the features reported in Table 1 shows that the introduction of contextual significantly improves the ability of the classifier to correctly distinguish between non neutral classes.

CONCLUSION:

In this paper, we have obtainable a system to filter undesired messages from OSN walls. The system exploits a ML soft classifier to content-dependent FRs. Furthermore, the suppleness of the system in terms of filtering options is improved through the organization of BLs. In particular, prospect plans consider a deeper examination on two co-dependent tasks. The first concern the withdrawal and/ or selection of contextual features that have been shown to have a high discriminative power. The second task involves the learning phase. The expansion of a GUI and a set of related tools to make easier BL and FR requirement is also a direction we plan to explore, since usability is a key condition for such kind of applications. We do consider that such a tool should propose trust value based on users actions, behaviors, and character in OSN, which might imply to enhance OSN with audit method. However, the intend of these audit-based tools is difficult by several issues, like the implication an audit system might have on users privacy and/or the limitations on what it is possible to audit in current OSNs. However, we like to remark that the system proposed in this paper represents just the center set of functions needed to supply a sophisticated tool for OSN message filtering. Even if we have given our method with an online assistant to set FR thresholds, the expansion of a absolute system simply usable by standard OSN users is a wide topic which is out of the range of the current paper. As such, the developed Facebook purpose is to be doomed as a proof-of-concepts of the system core functionalities, rather than a fully developed system.

5. REFERENCES:

[1] J. Allan et al., “Topic Detection and Tracking Pilot Study: Final Report,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] J. Kleinberg, “Bursty and Hierarchical Structure in Streams,” Data Mining Knowledge Discovery, vol. 7, no. 4, pp. 373-397, 2003.

[3] S. Morinaga and K. Yamanishi, “Tracking Dynamics of Topic Trends Using a Finite Mixture Model,” Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 811-816, 2004.

[4] Q. Mei and C. Zhai, “Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining,” Proc. 11th ACM SIGKDD Int’l Conf. Knowledge Discovery in Data Mining, pp. 198-207, 2005.

[5] A. Krause, J. Leskovec, and C. Guestrin, “Data Association for Topic Intensity Tracking,” Proc. 23rd Int’l Conf. Machine Learning (ICML’ 06), pp. 497-504, 2006.

[6] H. Small, “Visualizing Science by Citation Mapping,” J. Am. Soc. Information Science, vol. 50, no. 9, pp. 799-813, 1999.

[7] D. Aldous, "Exchangeability and Related Topics," _ Ecole d' _ Ete´ de Probabilite´s de Saint-Flour XIII—1983, pp. 1-198, Springer, 1985.

[8] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet Processes," J. Am. Statistical Assoc., vol. 101, no. 476, pp. 1566-1581,2006.

[9] P.E. Baclace, "Competitive Agents for Information Filtering," Comm. ACM, vol. 35, no. 12, p. 50, 1992.

[10] P.J. Hayes, P.M. Andersen, I.B. Nirenburg, and L.M. Schmandt, "Tcs: A Shell for Content-Based Text Categorization," Proc. Sixth IEEE Conf. Artificial Intelligence Applications (CAIA '90), pp. 320- 326, 1990.

[11] G. Amati and F. Crestani, "Probabilistic Learning for Selective Dissemination of Information," Information Processing and Management, vol. 35, no. 5, pp. 633-654, 1999.

[12] M.J. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," Machine Learning, vol. 27, no. 3, pp. 313-331, 1997.

[13] Y. Zhang and J. Callan, "Maximum Likelihood Estimation for Filtering Thresholds," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 294-302, 2001.

[14] C. Apte, F. Damerau, S.M. Weiss, D. Sholom, and M. Weiss, "Automated Learning of Decision Rules for Text Categorization," Trans. Information Systems, vol. 12, no. 3, pp. 233-251, 1994.

[15] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categorization," Proc. Seventh Int'l Conf. Information and Knowledge Management (CIKM '98), pp. 148-155, 1998.

[16] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th ACM Int'l Conf. Research and Development in Information Retrieval (SIGIR '92),N.J. Belkin, P. Ingwersen, and A.M. Pejtersen, eds., pp. 37-50, 1992.

[17] R.E. Schapire and Y. Singer, "Boostexter: A Boosting-Based System for Text Categorization," Machine Learning, vol. 39, nos. 2/3, pp. 135-168, 2000.