A Survey on Semantically Data Classification Analysis Algorithm for Social Media

Aarti Pandey, Prabhat Pandey

***Abstract:** Now in these days a number of users are participating in the social media and they are actively participating in conversation with their friends and community. Due to this sometimes the youth and teen agers are participating in non-social communities. Thus a new kind of data model is required to design by which the user communication and their patterns are accurately classified according to their semantics meaning. Thus a text content analysis technique is designed using the available automatic text classification technique. Using this technique the correlation between different words and their utilization in different semantics sentences are analyzed and based on the effects of these words a rule based classification technique is developed.*

***Keywords:** sentiment, opinion, semantic, Data Processing*

I. INTRODUCTION

Data mining is an application of data processing by which knowledgeable patterns and information is extracted. This extracted information is consumed using applications and real time systems for making decisions. The web is a rich domain of data, knowledge and entertainment. Large amount of user's access internet, in between they are always still connected through their friends using these services. Sometimes small human greedy nature over internet invites the hidden dangers. Bulling, phishing and abusing is a part of social crime that is not recoverable using technology. But cyber infrastructure can be prevented using small efforts of techniques. The main aim of this study is to overcome the social networking sites abasement and unsocial activities. Therefore a content classification scheme is required to develop for classifying th[i]ese contents. Children and youth worldwide have adopted social networking sites actively, partly because of the loss of children's freedoms in physical world. Additionally technical designs of social networking sites slow down easy management of settings and transparency concern the commercial use of personal information. Therefore social networking filters are required to develop by which user communications are navigated and their semantically meanings are recognized. This work is intended to develop a social networking filter for text based contents. To develop content based filter some methodologies and previous similar works are required to be analyzed.

Manuscript published on 30 June 2016.
***** Correspondence Author (s)

Aarti Pandey, Department of Computer Science, Awadhesh Pratap Singh University, Rewa (Madhya Pradesh). India.

Dr. Prabhat Pandey, Professor, Department of Physics, OSD, Office of Additional Director, Higher Education Rewa (Madhya Pradesh). India.

II. LITERATURE SURVEY

Xia Hu et al [1] investigate whether social relations can help sentiment analysis by proposing a Sociological Approach to handling Noisy and short Texts (SANT) for sentiment classification. In particular, Author presents a mathematical optimization formulation that incorporates the sentiment consistency and emotional contagion theories into the supervised learning process; and utilizes sparse learning to tackle noisy texts in Microblogging. An empirical study of two real-worlds. Twitter datasets shows the superior performance of given framework in handling noisy and short tweets.

Fei Jiang et al [2] emoticons in Chinese microblog messages are used as annotations to automatically label noisy corpora and construct sentiment lexicons. Features including micro blog-specific and sentiment related ones are introduced for sentiment classification. These sentiment signals are useful for Chinese microblogs sentiment analysis. Evaluations ona balanced dataset are conducted, showing an accuracy of 63.9% in a three class sentiment classification of positive, negative and neutral. The features mined from the Chinese microblogs also increase the performances.

Eric Baucom et al [3] seek to investigate how closely Twitter mirrors the real world. Specifically, they wish to characterize the relationship between the language used on Twitter and the results of the 2011NBA Playoff games. Author hypothesizes that the language used by Twitter users will be useful in classifying the users' locations combined with the current status of which team is in the lead during the game. This is based on the common assumption that "fans" of a team have more positive sentiment and will accordingly use different language when their team is doing well. They investigate this hypothesis by labeling each tweet according the location of the user along with the team that is in the lead at the time of the tweet.

Min Wang et al [4] propose a novel Cross-media Bag-of-words Model (CBM) for Microblog sentiment analysis. In this model, we represent the text and image of a Weibo tweet as a unified Bag-of-words representation. Based on this model, we use Logistic Regression to classify the Microblog sentiment. It performs well in the sentiment classification task since it doesn't require the conditional dependence assumption. They also use SVM and Naïve Bayes to make a comparison. Experiments on 5,000 Microblog messages demonstrate that our CBM model performs better than text-based methods. The sentiment classification accuracy on Microblog messages of our model is 80%, improved by 4% than the text-based methods.

III. PROPOSED SYSTEM

In order to resolve the identified issues in the social networking data analysis a new model is proposed in this system. The proposed data model is based on the hybrid concept of graph theory and data mining techniques. The proposed data model can be understood using the given figure1.

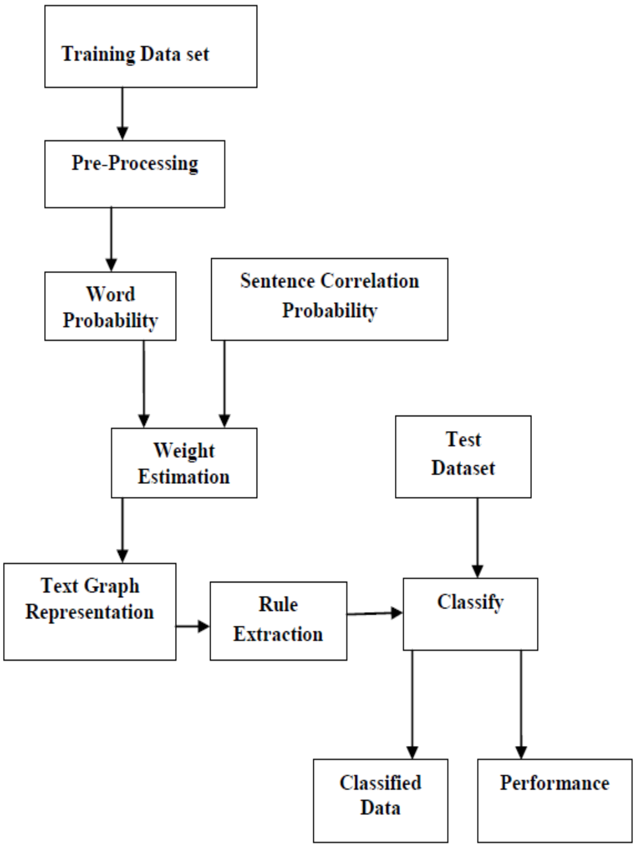

Figure 1: Proposed System

The proposed data model is given in figure 1 in this diagram the model training and the test dataset classification technique is simulated. Thus the entire data model development is performed in two major modules first training and then testing. In training first the system accept the training data samples on which the pre-processing is performed, during pre-processing the stop words are removed from the input text and the significant text remains from the training set. The word probability is majored in this measured from the text remain using the below given formula.

$$Word\ \text{Probability}= \sum_{i=1}^{N} word / N$$

Where the N is number of total words in text documents, in the similar ways the sentence formation probability is estimated from the text using the following formula.

$$\text{Sentence Probability}= \sum_{i=1}^{M} word_i\ / \text{ M}$$

Where the M is the number of sentences in the text document

After that the weights for the weighted graph is computed from both the probability

$$weight = sentence\ probability * Word\ \text{Probability}$$

Using the computed weights and the user feedback the correlation graph is developed in further steps and using these weighted graphs the classification rules are developed. These rules sets are consumed to identify the sentiments of the words involved in the micro-blog text.

IV. APPLICATION

The proposed text analysis technique and the sentiment classification methodology are also applicable in the following domains.

1. Social network analysis
2. Military message encoding
3. Privacy improvement in social networking
4. Detection of unwanted and malicious conversation in electronic messaging systems

V. CONCLUSION

Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. There are likely to be many other applications that is not discussed. It is found that sentiment classifiers are severely dependent on domains or topics. From the above work it is evident that neither classification model consistently outperforms the other, different types of features have distinct distributions. It is also found that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefit from each other‟s merits, and finally enhance the sentiment classification performance.

REFERENCES

1. Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, “Exploiting Social Relations for Sentiment Analysisin Microblogging”, permission and/or a fee.WSDM ’13, February 4–8, 2013, Rome, Italy.Copyright 2013 ACM 978-1-4503-1869-3/13/02
2. Fei Jiang, Anqi Cui, Yiqun Liu, Min Zhang, and Shaoping Ma, “Every Term Has Sentiment:Learning from Emoticon Evidencesfor Chinese Microblog Sentiment Analysis”,c Springer-Verlag Berlin Heidelberg 2013
3. Eric Baucom,AzadeSanjari, Xiaozhong Liu,Miao Chen, “Mirroring the Real World in Social Media: Twitter,Geolocation, and Sentiment Analysis”,Copyright 2013ACM,78-1-4503-2415-1/13/10http://dx.doi.org/10.1145/2513549.2513559
   Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, RongrongJi, “Microblog Sentiment Analysis Based on Cross-mediaBag-of-words Model”,ICIMCS’14, July 10–12, 2014, Xiamen, Fujian, China.Copyright 2014 ACM 978-1-4503-2810-4/14/07
4. Felipe Bravo-Marquez, Marcelo Mendoza,Barbara Poblete, “Combining Strengths, Emotions and Polarities forBoosting Twitter Sentiment Analysis”,WISDOM’13, August 11 2013, Chicago, IL, USACopyright 2013 ACM 978-1-4503-2332-1/13/08.
5. Pedro Calais Guerra, Wagner Meira Jr.,Claire Cardie, “Sentiment Analysis on Evolving Social Streams:How Self-Report Imbalances Can Help”,WSDM’14, February 24–28, 2014, New York, New York, USA.Copyright 2014 ACM 978-1-4503-2351-2/14/02

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)*

Author Profiles

Aarti Pandey, is currently pursuing Doctor of Philosophy (PhD) in Computer Science from Department of Computer Science Awadhesh Pratap Singh University, from Rewa (M.P.) . She has done her Master of Science (M.Phil) from Department of Computer Science Awadhesh Pratap Singh University, Rewa (M.P.) in the year 2009.

Dr. Prabhat Pandey is Professor of physics currently working as OSD Office of the Additional Director, Higher Education, Rewa Division Rewa (M.P.). He has done his Doctor of Philosophy (PhD) in Space Physics from Department of Physics from Awadhesh Pratap Singh University, Rewa(M.P.) in the year1989. He has done his Masters of Science (M.Sc.) in Physics from Awadesh Pratap Singh Univesity Rewa (M.P) in the year1982.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)