International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr42018 www.irjet.net p-ISSN: 2395-0072

POST SUMMARIZATION AND TEXT CLASSIFICATION IN SOCIAL NETWORKING SITES.

Mr.Vignesh Kumar [1], Ms.Dhivya N [2], Mr.Abishek R K [3], Mr. Dharun B [4]

[1]*Assistant Professor, Department of Computer Science And Engineering, BIT, Tamil Nadu, India.*
[2,3,4] *Final year, Department of Computer Science And Engineering, BIT, Tamil Nadu, India*

Abstract -*Social Networking – It's the way the 21st century communicates now. Social networking is the grouping of individuals into specific groups, like small rural communities or a neighbourhood subdivision. It is that the social networking sites offer all the facility to share information with friends such as posts, comments, private messages etc, with required privacy level but the usage of words is not taken into control in the posts or comments. So we therefore bring out the idea of classifying the words using the text classification process in data mining and also provide post summarization. The user's own post can be summarized based on specific topic and can be collected and kept at database by post summarization technique using data mining and friend matching system. Whenever the user needs he can get the specific topics related post rapidly in his own profile.*

Key Words*: *Social networking site, Privacy level, Post summarization, Text classification, Data-mining.

1. INTRODUCTION

The Social Networking websites which is based on web services allows the individual to get connected to the outside world by the option provided such as friend request, posts, comments, etc .There are social networking sites available such as MySpace, Face book, Twitter which commonly have the profile management module but may vary in some specialized module in their own style. Here in this paper we propose our own style of profile management in our platform namely friend book which help in post summarization and text classification using the concepts of Data Mining. The normal profile management is also associated such as friend request, accepting the friend request ,post and comment on that post etc. Thus the Data mining techniques such as text Classifiers and summarization techniques are used.

2. PROFILE MANAGEMENT SYSTEM

As mentioned above we have come across different types of websites regarding social networks namely Face book , Orkut , MySpace ,Twitter, LinkedIn etc they themselves have their own way of profile management which is common functionality in all networking sites. The Profile Management includes the creation of the account, updating the users information, making new friends around the world etc.

Here the same process have been followed to login in our site too

- Profile Creation is the one in which the users will create their profile to enter the social web world. They can have privacy over their profile that only the persons they like will be able to view or send friend request.
- Updation of information is another necessary thing so that other users will come to know the new users. Updation of the profile means that the location currently they are living, display picture etc .The privacy is maintained in this scope.
- Friend Request is the common module of connecting with the people we like.
- If some one need to get friend then the user can give the interested person the request so that person on the other side would receive the request that some one is interested to make friend with him/her.
- Post is another important module that is available in the social media. Twitter is based mainly on the post. The users can post the image or video or any other files they like on their own profile or their friends profile.
- Comment is nothing but we can comment on the post that out friend post or the friend of us post. It depends upon the privacy that is provided.
- Message Management is also the important module that we can send private messages to our friend we are associated.
- In some social media we have the newsfeed system, making a specific groups within or for any Business needs etc. Thus Profile Management is the vast sector in social networking sites.

3. LITERATURE REVIEW

Data Mining concepts for the post summarization and Text classification involves many techniques such as

- An Enhanced Data Mining for text classification that uses the k-nearest neighbor classification algorithm which is based on the sentence based concept analysis, document based concept analysis and corpus based concept analysis.

- A Novel text classification algorithm based on enhanced associate rule includes Mining frequent item set, generating enhanced association rule.
- Data Mining Techniques for social media analysis for the summarization of word using semantic analysis.
- Enhanced Classification Accuracy on Naïve Baye's Data Mining Model which is primarily based on classification based on similar data set by making the group of similar data using k-means Clustering and training the data set using Naive bayes classification algorithm.

4. EXISTING SYSTEM

The Existing System includes some preliminary techniques such as profile management like friend request, comment, post, profile updation, message management etc., though they provide high level of security through the privacy policy the words that is been used in post and comments are not taken into account in existing system .Thus this problem can be solved by Text classification technique in datamining which will classify the bad text and does not allow them to post in the website. And also in the existing system each user cannot summarize the post under the interested topic the user wish .This can be rectified using the post summarization technique .

5. PROPOSED SYSTEM

The Profile Management system in social networking sites have the same process as discussed earlier ,In addition to this we propose two modules which is summarization of post and the text classification using the data mining concepts. The Summarization of the post is that if the particular topic is given the data containing the post for the particular user at each sessions will be summarized at the backend .The main motive of text classification is to avoid the vulgar ,violence word not to be posted or commented so that other users may not feel bad or get hurted .so we go for the text classification using datamining concept .

5.1 SUMMARIZATION

Post Summarization involves the process of summarizing the post under the particular topic with the particular session id. This is done with the help of data mining technique summarization .summarizer is the algorithm which is used to get the needed or required data from the large amount of data thus giving summarized data in a readable and structured way. Automatic summarization is one of the field in natural language processing which makes the system to analyse, predict and understand like human. Summarizer is part of the J4 classifier algorithm which scans the post database and comment database.

When the needed topic is created for the summarization it searches the above mentioned database and separate summarization is created for the user id of that particular session. Thus the summarization is just part of the text classification algorithm.

In Automatic Summarization the particular text from the document is summarized with help of the word frequency .If the word frequency is high then the word will be taken into account for the summarization else it will be left as the normal one. The main need for the summarization is that the business analyst ,researches may need to go through many documents per day to get a particular topic. Thus using this algorithm they can get at an instant .

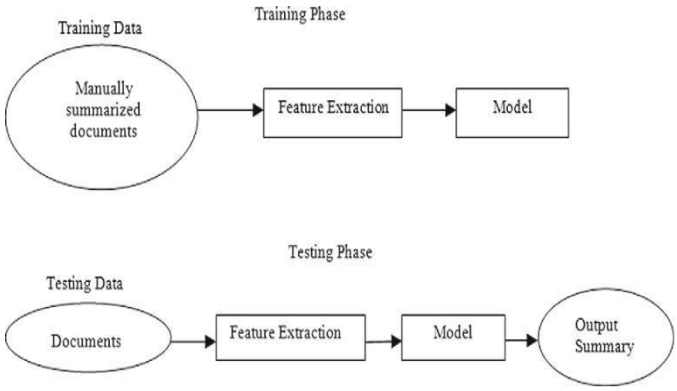

Fig. 1 Summarizing the training data

5.2 TEXT CLASSIFICATION

The Text Classification is the one of the main module that plays the major role to block the vulgar or the violence word in the social networking site. In text classification we classify the words into the soft classification and hard classification. In the Soft classification we classify the word into neutral and the non neutral .

STOP WORD REMOVAL

The process of the stop word removal is that removing the letters like "ing" and the words like article, preposition, and all the unwanted letters around the word. Example for this is the word "beautifully" can be taken as beauty thus removing unwanted stop words.

Four types of stop word removal methods are followed, the methods are used to remove stop words

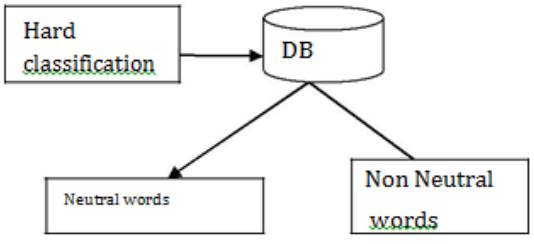

Fig.2 Classificaton of Text

5.2.1 HARD CLASSIFICATION

Hard Classification includes the text that are violence .In the Hard Classification the word which are fed into the spectrum is classified before it is being posted by comparing the words with the training data set .If it matches with the training dataset of the bad words the word will not be pushed inside the spectrum .Thus the Hard Classification of the Text .

5.2.2 SOFT CLASSIFICATION

In Soft Classification, The word is classified under

- Neutral classification
- Non Neutral classification

Neutral Classification

In the neutral classification the words that are classified are normal words that we use normally in the daily life. The words may be like speed ,human etc.

Non Neutral Classification

In the non neutral Classification the words that are classified are of normal usage but feels like scolding or for fun in the soft manner. The words may be like idiot, fool etc.

5.2.3 TEXT PREPROCESSING

Before the classification we use to process the text so that they can be easily classified . The Pre processing of the text includes

- Extraction
- Stopword Removal

EXTRACTION

The Extraction process involves tokenization of the word that splits the sentences into each token by removing the space . from the files .

- Classic Method
- Zipf's law
- Mutual Information method
- Term Based Random sampling

Classic Method

Classic method include the removal of the stop words from pre complied list.

Zipf's law

Zipf's law states that removing the words which occur frequently and the words which occur only once can be removed from the file.

Mutual Information method

This works by analyzing the mutual information between the given term and the documentation class. If it is low it means that it has lower discrimination power and vice versa.

Term based random sampling

This method works by iterating over separate chunks of data which are randomly selected. It then ranks terms in each chunk based on their format values using the Kullback-Leibler divergence measure

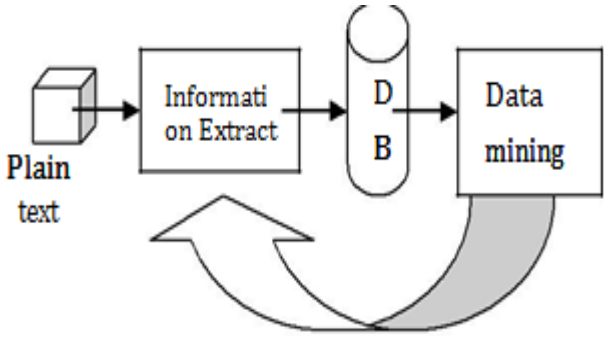

Fig.3 Process of Text Extraction

Thus the preprocessing of the text is the essential thing which is further utilized in the text classification of word making it easier. Here the above process are together performed in the Naïve bayes Classification algorithm.

Thus the Naïve bayes Classification algorithm is the one which is based on the probability theorem. We use the predictive modeling by supervised pattern classification Supervised pattern classification is the task of training the dataset.

6. CONCLUSION AND FUTURE WORK

As of the social networking sites the profile management is the common module in all sites .The post summarization and text classification is the new module which is included to enhance the social networking sites.

The post summarization is used to get the fast retrival of data under a specific topic which is given. The text classification helps in classifying the text into hard and soft so that the hard classified words can be removed from the others post .The Future work of this project include the automatic summarization of the post by taking the term frequency and summarizing under the term .This makes it more efficient than the normal summarization. We can also make use of another module which is profile matching and friend matching system by detecting the life style of the user using the sensor based mobile application which makes its more efficient and user friendly.

REFERENCES

[1] Summarization - Compressing Data into an Informative Representation Varun Chandola and Vipin Kumar Department of Computer Science, University of Minnesota, MN, USA.

[2] Text Summarization in Data Mining Colleen E. Crangle ConverSpeech LLC, 60 Kirby Place, Palo Alto, California 94301, USA crangle@converspeech.com www.converspeech.com

[3] Naïve Bayes, Wikipedia, http://en.wikipedia.org/wiki/Naive_Baves_classifier.

[4] C. M. Bishop, Pattern Recognition and Machine Learning. New York,NY, USA: Springer, 2006.

[5] Mitchell.T.M, McGraw Hill, New York, NY, Machine Learning, 1996..

[6] Le Zhang, Jingbo Zhu, and Tianshun Yao, Natural Language Processing Laboratory Institute of Computer Software & Theory, Northeastern University An Evaluation of Statistical Spam filtering techniques, ACM Transactions on Asian Language Information Processing, Vol. 3, No. 4, December 2004, pages 243-269.

[7] K. A. Hamill, A. Zamora, "The use of titles for automatic document classification", Journal of the American Society for Information Science, pp. 396-402, 1980.

[8] H. Heaps, "A theory of relevance for automatic document classification", Information and Control, pp. 268-278, 1973.

[9] W. Hoyle, "Automatic indexing and generation of classification by algorithm", Information Storage and Retrieval, pp. 233-242, 1973.

[10] An, J. and Chen, Y.: Concept Learning of Text Documents. Web Intelligence 2004: 698-701 .

[11] Clark, P and Niblett, T: The CN2 Induction Algorithm. Machine Learning 3: 261-283 (1989)

[12] Hammouda, K. M., Kamel, M. S.: Phrase-based Document Similarity Based on an Index Graph Model. ICDM 2002: 203-210

[13] Survey Paper on Document Classification and Classifiers Upendra Singh [1], Saqib Hasan [2] UG Students [1] & [2] Department of Computer Science and Engineering Madan Mohan Malaviya University of Technology Gorakhpur - 273010 UP – India

[11] F. Sebastiani, "Text categorization", Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005.

[12] Fetterly, D., Manasse, M. & Najork, M. (2005). Detecting phrase-level duplication on the world wide web. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. pp. 170-177). : ACM Press, Salvador, Brazil