International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

Using Natural Language Processing for Detection of Events and Spam Control from user Data Stream in Social Sites

Roshani M. Shete
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Wardha, India

Prof. S. W. Mohod
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Wardha, India

***Abstract* - Detecting trending topics is perfect to summarize information getting from social media. To extract what topic is becoming hot on online media is one of the challenges. As we considering social media so social services are opportunity for spamming which greatly affect on value of real time search. Therefore the next task is to control spamming from social networking sites. For completing these challenges different concepts of data mining will be used. For now whatever work has been done is narrated below like spam control using natural language processing for preprocessing and clustering. One account has been created for making it real.**

Keywords – Event detection, Control spamming, Text mining, information filtering, Social networking site

I. INTRODUCTION

As we know day by day popularity of online media is increasing. Communications and interactions using texts, posts, comments and chats always reflect on dynamics and real time event, which means the importance of information, is increasing dramatically. Now a day's exact and accurate sensor of real time events is none other than social networking sites content. The pervasiveness is expanding of online media. Public base on online media gets more active in producing stuff on real world events.

More than 170 million people are using online media to be connected to their friends, co-worker and family members. There are so many topics, subjects, events, jokes, news discussed by people on it. To know which topic is hot on the social media and why, there is need to invent a system. As few events get more attention whether some get less.

So for producing this in real time, there is need of collection of user generated data on social networking sites. There are two approaches in the proposed work, identifying current and control spamming.

The first step is preprocessing which is important for mining the data or filtering the data. The work of preprocessing has been done. Then the spam control has been done. Spam control is the part of feature extraction. Here used the bisecting K-means clustering algorithm, because clustering is an important step for quality results. So nothing but natural language processing (NLP) technique has been used for preprocessing, clustering etc. For applying all the implementation, one account has been created.

Access of structured information which is seen in databases and unstructured information which is seen in documents or unstructured fields of texts both benefits to information processing applications. So, access of such texts information, also having the benefits of linguistic analysis of text, as contrary of shallower "word basis" analysis. As there are lots of methods, techniques that can be tested on natural language texts, its effect in the amount of search in the natural language processing fields [2].

Next work is identification of current events which is in the processing, it will be completed soon. For that also NLP and machine learning (ML) concepts will be used.

II. RELATED WORK

Topic Detection and Tracking extract event from public generated data on social sources and identify the trend in term of time [4]. In this the public generated data means posts uploaded by them.

In clustering of all the data streams on social media there are two type of methods one cluster by document or cluster by feature. The first is document pivot and the other is feature pivot method. Both approaches are presented by authors differently in previous work. So these two used methods and their work explained below.

Feature pivot method means the term or keyword will be considered while clustering but the drawback of that is it capture misleading term. For example if anyone wants to search like "definition of class", it will shows extra result like subclass, superclass etc. So the accuracy of result is very less, also redundancy and ambiguity gets formed in this.

M. Cataldi, C. Schifanella and L. Di Caro [5] proposed two measures, term frequency to calculate nutrition for each word and a page rank measure. After that Bursty keywords are obtained using nutrition trend. Then by using graph based approach for bursty keywords generates the topic boundary. Sayyadi, Maykov and Hurst [6] used graph approach in which clustering of keywords is done by matching pairs. They used community detection algorithm in which made a graph whose nodes are clustered. Also the topic extraction is carried out by identifying document with

similar term. Lehmann, Kleinberg and Backstorm [7] have used the graph for short phrases. Phrases are connected by edges.

One of the method modeled called Latent Dirichlet Allocation (LDA) [8], the idea of knowing the most breaking news by calculating the bursty terms in document [9]. This avoids the other topics by capturing the high peak [10]. So first find bursty term then cluster them for event detection. In some graph based approach, the first step is to tag the terms, then group it and then find the interest in social media [11].

III. BLOCK DIAGRAM

In proposed work, address the task of detecting topics in real time from social media streams. To keep our approach general, consider the stream is made of pieces of text generated by social media users i.e. posts, messages, or tweets of social media. The flow of proposed work will be phase-1 for user data stream, phase-2 for preprocessing using NLP, phase-3 for spam control and phase-4 for identification of events.

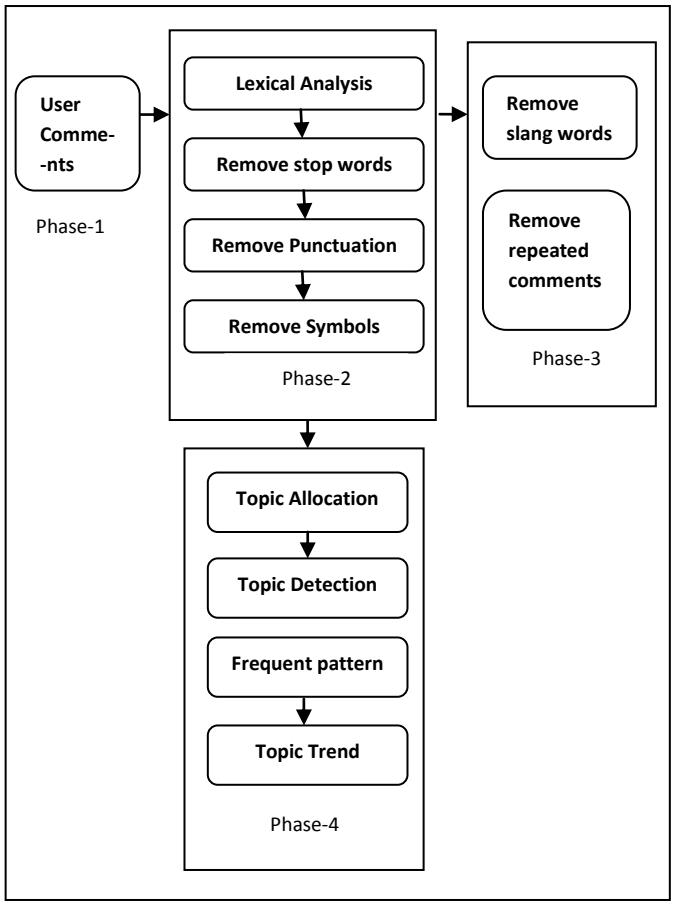

Fig. 1. Flow of Work

IV. IMPLEMENTATION

Analysis of unstructured textual data and automatic processing with the natural language processing always deal with. One of the sides of natural language processing based on statistical techniques and methods, typically contains the processing of words appeared in texts. Leveraging knowledge resources like taxonomies, linguistic rule bases and ontologism, and then rule based methods get used other approaches. Bag of training materials are required for statistical human language processing, by which desirable and undesirable dependencies and relationships are exemplifies. Subsequent changes then needs some extent of system retraining. Rule based methods need knowledge from online dictionaries and linguistic theories instead of getting training material. May be natural language processing make use of these methods and to decide which method or technique to use always depends on availability of external resources, training materials [12].

A. Preprocessing

Preprocessing contains filtering of data. Natural language processing concepts are used for preprocessing of data.

1) Lexical analysis: The lexical analyzer covert sentences into words then words convert into characters.

2) *Elimination of punctuations:* Remove punctuations like comma, full stop etc.

3) Elimination of symbols: Remove symbols like @, # etc.

4) Elimination of stop words: Remove words like in, of, the, is, and, for etc.

B. Account

For showing the results an account is designed. In which master page as Login Panel is created with options like Registration Form, Login, Email id, Password etc.

First to access the account user need to create id using registration form where there is need to fill the information like city, country, mobile number, id, password etc. This information will save into the database. After that whatever entries available in the database, only to them user can add as friend. Now user can post, comment in it. Everything will be saved into database.

C. Feature Extraction

Feature extraction is used for reduction of dimensionality. Before classification there is need of reduction of feature space. Now spam control is also nothing but a feature reduction task. Therefore, slang word reduction is done for the spam control.

For spam control, dictionary of slang words is created. So, whenever user use any slang word in the post or comment that word matches with the words available in the dictionary and it replaces with the stars (****).

For example if user posted something and his/her friend commented “you dog”, so this comment will be replaced by “you ****”. Because the word dog is slang word and it is defined in the dictionary.

So, some of the feature extraction is done with reduction of slang words.

In extracting the data from document there are problems like ambiguity in result, redundancy so the process get annoying and time consuming.

D. Categorization of events

This is the first part of event detection where event will be detected by field wise. It means whether the given comment related to bollywood, politics, sports, education or business.

If comment does not exist in anyone of it, then it will be shown in 'other'. First the dictionaries of bollywood related words, business related words, and politics related words are created. So the process is that, each comment/post will be split word by word. Then each word will be compared with the dictionary words. Then if any word of comment/post is match with one of dictionary after that comment\post will be shown in the respective field.

V. RESULTS AND DISCUSSIONS

In implementation one account is created with login and registration form so below snapshot is master page,

This is the master page with registration form and login,

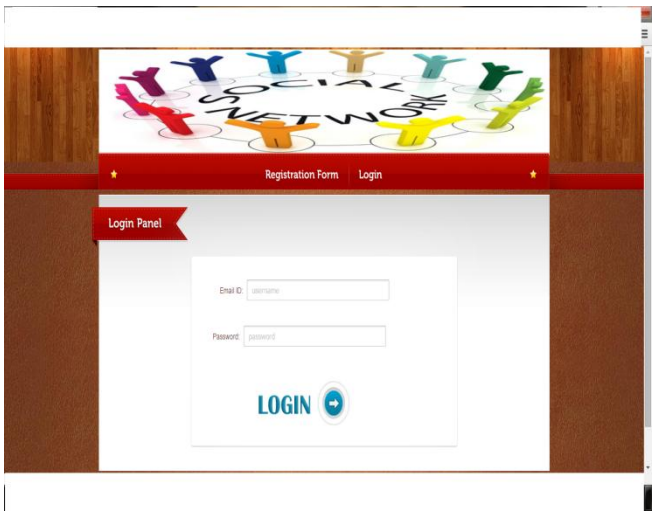

Fig. 2. Master page

This is the registration form where user need to fill their information for profile,

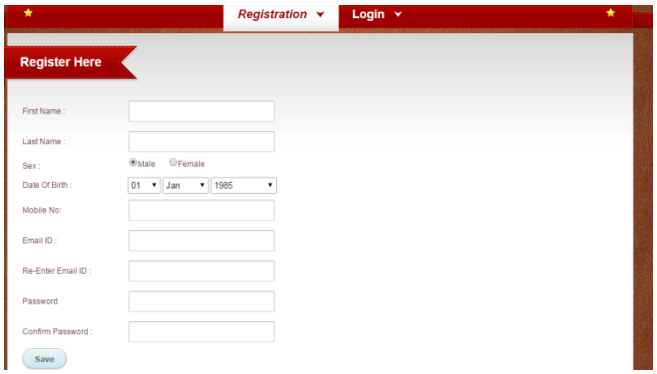

Fig. 3.Registration form

This is dataset of information of users who registered for profile,

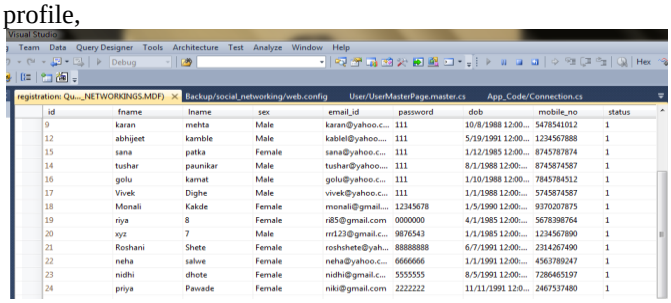

Fig. 4. Registration data

Here, user need to type ID and Password for login,

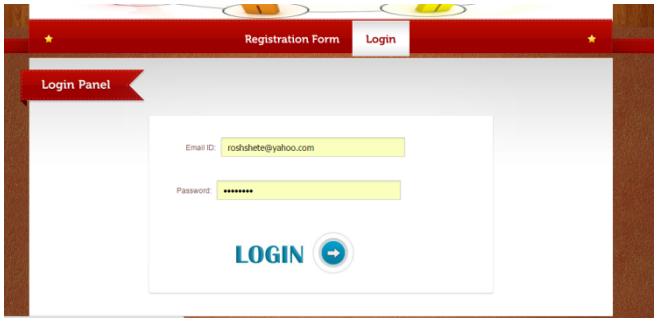

Fig. 5. Login with created ID

This is the created profile page,

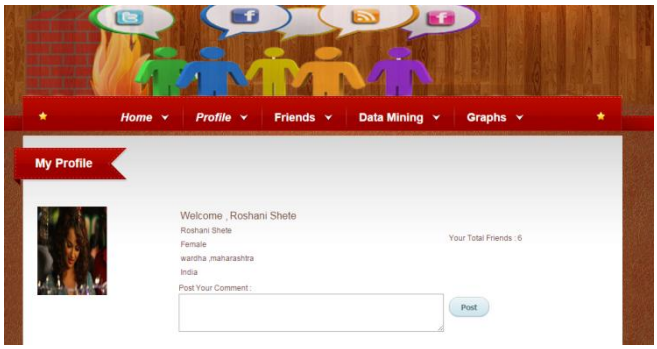

Fig. 6. Profile

These are comments and posts of user and their friends,

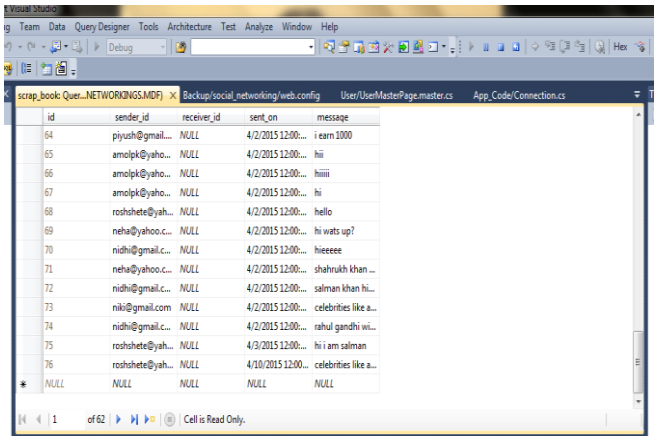

Fig. 7. Comments/Posts

This is example where slang word replaced by ****

Fig. 8. Example

This is slang word dictionary,

Fig. 9. Dictionary

This is the mining of events, field wise detection of event. Here, comments are classified as per the fields.

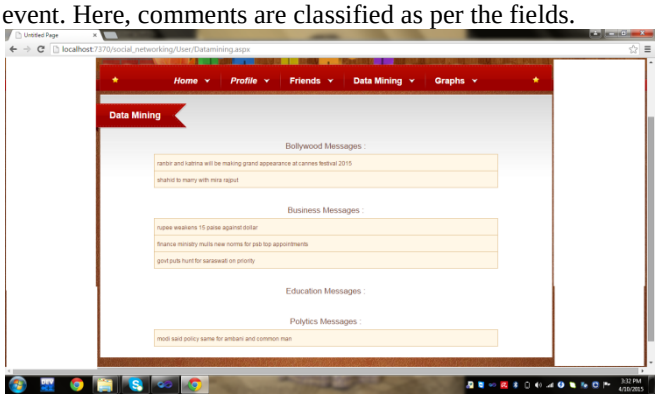

Fig. 10. Categorization of events

VI. CONCLUSION

The main aspects of the proposed work are to detect the current topics of real world and to control the spamming created by spammer. Preprocessing process is done. One account is created for showing results. Also feature extraction is the part of spam control has done. Then, one part of event detection i.e. classify events field wise is done. So the next work is to implement current event detection i.e. second part of event detection in the same created account.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guide Prof. S. W. Mohod, coordinator of M.Tech. in Computer Engineering Department, for his constant guidance. I am extremely grateful to him for his sincere, expert and valuable guidance extended to me.

REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker and Ioannis Kompatsiaris, "Sensing Trending Topics in Twitter", IEEE Transactions on Multimedia, Vol.15, No.6, October 2013.

[2] Jurafsky D. and Martin, J., "Speech and Language Processing", Prentice Hall, Upper Sale River, NJ 2000.

[3] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in Proc. ICSWM: 5th Int. AAAI Conf. Weblogs and Social Media, 2011.

[4] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are strange when you're a stranger: Impact and influence of bots on social networks," in Proc. ICWSM: 6th AAAI Int. Conf. Weblogs and Social Media. AAAI, 2012, pp. 10–17.

[5] M. Cataldi, L. Di Caro and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in Proc. MDMKDD: 10th Int. Workshop Multimedia Data Mining, New York, NY, USA, 2010, pp. 4:1–4:10, ACM.

[6] Sayyadi, M. Hurst and A.Maykov, "Event detection and tracking in social streams," in ICWSM, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds. Palo Alto, CA, USA: AAAI Press, 2009.

[7] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proc. KDD: 15th ACM Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, 2009, pp. 497–506.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003

[9] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and persistence: Modeling the shape of microblog conversations," in Proc. CSCW: ACM Conf. Computer Supported Cooperative Work, New York, NY, USA, 2011, pp. 355–358

[10] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in Proc. WSDM: 4th ACM Int. Conf. Web Search and Data Mining, New York, NY, USA, 2011, pp. 177–186.

[11] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali, "A graph-based clustering scheme for identifying related tags in folksonomies," in Proc. DaWaK: 12th Int. Conf. Data Warehousing and Knowledge Discovery. Berlin, Germany: Springer-Verlag, 2010, pp. 65–76.

[12] Manning C. and Schutze H. "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, MA, 1999.