Volume 2, Issue 10 October 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal Of Science, Engineering and Information Technology

Chief Editor
Dr. Bilal Ahmad Malik

Publisher

Dr. Bilal Ahmad Malik

Associate Editor

Dr.Nagendra Mani Trapathi

NAIRJC JOURNAL PUBLICATION

North Asian
International
Research Journal Consortium

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi. All research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

M.C.P. Singh Head Information Technology Dr C.V. Rama University	S.P. Singh Department of Botany B.H.U. Varanasi.	A. K. M. Abdul Hakim Dept. of Materials and Metallurgical Engineering, BUET, Dhaka
Abdullah Khan Department of Chemical Engineering & Technology University of the Punjab	Vinay Kumar Department of Physics Shri Mata Vaishno Devi University Jammu	Rajpal Choudhary Dept. Govt. Engg. College Bikaner Rajasthan
Zia ur Rehman Department of Pharmacy PCTE Institute of Pharmacy Ludhiana, Punjab	Rani Devi Department of Physics University of Jammu	Moinuddin Khan Dept. of Botany SinghaniyaUniversity Rajasthan.
Manish Mishra Dept. of Engg, United College Ald.UPTU Lucknow	Ishfaq Hussain Dept. of Computer Science IUST, Kashmir	Ravi Kumar Pandey Director, H.I.M.T, Allahabad
Tihar Pandit Dept. of Environmental Science, University of Kashmir.	Abd El-Aleem Saad Soliman Desoky Dept of Plant Protection, Faculty of Agriculture, Sohag University, Egypt	M.N. Singh Director School of Science UPRTOU Allahabad
Mushtaq Ahmad Dept.of Mathematics Central University of Kashmir	Nisar Hussain Dept. of Medicine A.I. Medical College (U.P) Kanpur University	M.Abdur Razzak Dept. of Electrical & Electronic Engg. I.U Bangladesh

Address: -North Asian International Research Journal Consortium (NAIRJC) 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

IRJIF IMPACT FACTOR: 3.821

EFFECTIVE SUMMARIZATION OF TWEETS USING TWEET SEGMENTATION

P JAYASRI [1] & DR. A P SIVA KUMAR [2]

M.Tech, Department of CSE, JNTUA College of Engineering, Anantapuramu, A.P, India[1]
Assistant Professor, Department of CSE, JNTUA College of Engineering, Anantapuramu, A.P, India[2]

ABSTRACT:

Twitter is one of the most popular platforms with millions of users where people share their opinions on a wide variety of topics. This will result in the production of huge amounts of instant messages every day. Tweets are noisy and short in nature, this makes it difficult to analyze and understand what is being said about numerous topics in the twitter. Summarization is required for data analysis, exploration, opinion mining etc. summarization is necessary before any data analysis can be done otherwise it is highly impossible to read through millions of tweets in the stream. This paper will present the segment based representation for the task of summarization, which includes the method of tweet segmentation. Hybridseg is a framework which is used to split the tweets into meaning full Named Entities; this can be done by maximizing the stickiness score of NEs. It will preserve the semantic and context information of the tweets. Stickiness score will be calculated with the help of local and global contexts. By using these tweet segments we can be able to summarize data to better understand the tweets.

Index Terms—Summarization, Tweet Segmentation, Named Entities

I. INTRODUCTION:

In present globalized network, people are very interested to know trending of various topics like a gadget, a celebrity, a movie etc., and twitter is one such platform where we can get this information. 500 million tweets sent per day where the majority of tweets are not formal or not particularly meaningful. Twitter is credited with providing most trending information about many events before the traditional media and also many organizations monitors their targeted twitter streams to collect the opinions of users majorly for marketing and advertising. Tweet summarization can be useful in exploration of trending information. For example, a note of summary can be attached to any trending topic. Therefore providing a short summary to a user about anytopic related Tweets for understanding all the trending topics instead of reading thousands of random posts.

Twitter has limited length of characters and it allows free writing style, tweets may contain misspelling and grammatical mistakes and informal abbreviations, these makes tweets less reliable. It will become very difficult while handling with the twitter data. The tweet segmentation [1] process will split tweets into an order of consecutive n-grams ($n \geq 1$), these are called segments. A segment can be anything like a named entity. Ex:an event like “Olympics 2016” or any semantically meaningful piece of information such as “audio released”, or any other types of phrases which is appearing more than once. For example, a tweet may contain information like “In celebrations of college day with frnds”, here each word can be considered as a keyword or a segment. Some tweets may also contain same keywords like “celebrations of Diwali”, “experienced the first day of college” etc. If we can be able to identify each of these keywords or segments of the tweets it becomes easy to generate a topic and query related summaries.

Extraction based summary creation is related to the identification of the keyword which resides in the original document if we consider the batch of tweets and performed the segmentation process on the tweets then it will become basic for summary creation where we consider each of the valid segment as the keyword for extraction of the summary.

II. RELATED WORK:

Summarization of Microblogging messages is a more general problem compared with other automatic summary generation. Processing of short and informal posts is a new area of research.II. P. Luhn presented that a summary of a document can be formed with the original contents of that document e.g., words, sentences and other components [5]. There are a number of studies presents about the data representation models like Vector Space Model (VSM), term weighting approach [6], [7], n-grams and n-multigrams approach[8], n-gram graph model[9], KNN based classifier used for the word level classification[10]. All these models have several drawbacks regarding with their representations. Other studies presented the entity linking strategy for the identification of named entities from the knowledge bases like Wikipedia [15]. Chenliang Li et.al [1] presented the segmentation technique which is similar like the chines word segmentation (CWS). The external knowledge base will be considered as a valid segment and those will be applied for the segmentation process, it helps in the preserving of semantic meaning of the twitter messages.

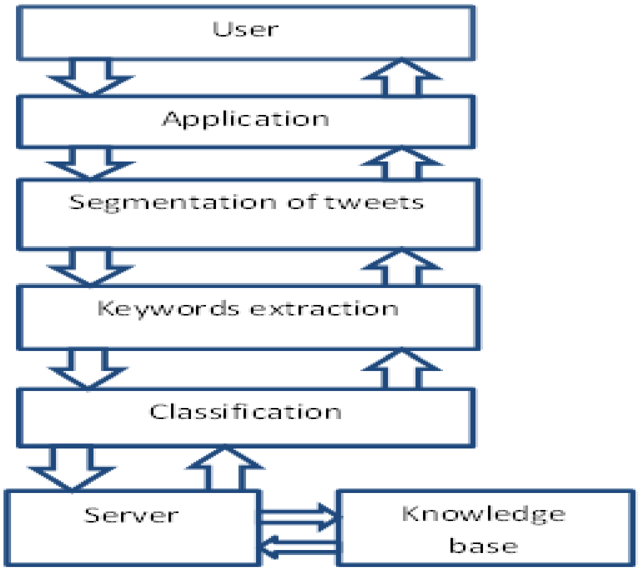

Fig 2: System Architecture

III. PROPOSED WORK:

To perform extraction based summarization the keywords should be identified in the twitter stream where it was very difficult to extract the relevant information of the topic of the summary. The segmentation process will find the keywords of the summarization process based on that the summary will be created. The Segmentation process will consist various steps in which the knowledge base will be considered as the reference while splitting of the messages.

IV. TWEET SEGMENTATION PROCESS:

The segmentation method plays a key role in formation of the summary where the twitter posts are identified as the valid segments of the process; it consists of a knowledge base where we can input relevant information of a particular domain. The reference knowledge base is created by obtaining related sources of the domain such that summary can represent the relevancy of that domain. The segmentation method is carried out by the framework called Hybridseg where stickiness score is taken as the criteria for segmentation, each of the segments is identified as the named entity based on the knowledge base. Fig.1 shows how the segmentation process is continued with the calculation of the stickiness scores of each word (w1, w2,…wl). The segmentation framework involves various preprocessing steps like the length normalization in which meaningful segments created by extracting long segments to preserve specific meanings.

A.LOCAL AND GLOBAL CONTEXTS:

Local context represents local knowledge base for the segmentation process it may include official accounts of the twitter or the organizations' pages where the posts are mostly to be in formal language, these will be useful to segment the other user posts. The global context will be referred as an external knowledge base like Wikipedia or any other web corpus can be treated to serve as the external knowledge base. In this paper we externally input domain name, as well as the domain related information which indicates the relevancy of the keywords according to the segmentation process. The segmentation framework includes various methods to deal with the linguistic features of the language where it learns from weak NEs by the voting method it is required because of tweets noisy nature.

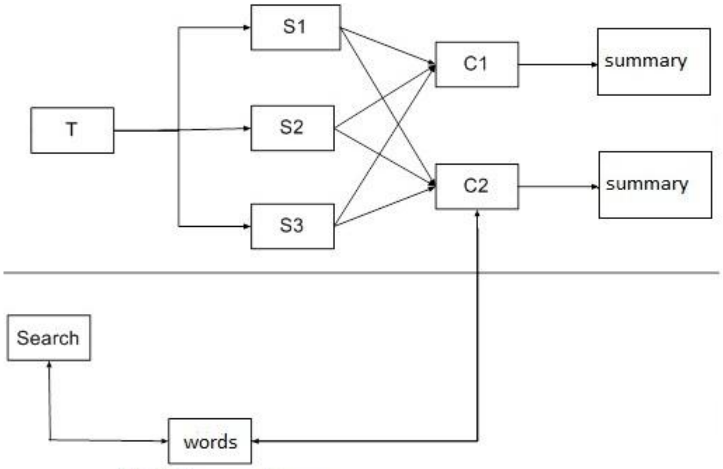

Fig 3. Process diagram

The collocations of the segments are identified from these tweets where the correlated words and sub n-grams of n-grams will be treated as valid segments.

B. ITERATIVE APPROACH:

To make the segmentation in the optimized manner pseudo feedback method is used in which most confidant segments from the previous iteration taken as the reference to process the weak segments. Fig3 shows architecture diagram of the summarization of tweets with the help of segmented partitions of the posts. The NEs can be identified by using random walk method which considers observation in the co-occurrence of NEs with other named entities and POS (Parts of Speech)-tags based algorithm, to identify the noun phrases of the tweets.

V. SUMMARIZATION:

Summarization of tweets will be useful to find responses of users on a specific topic. It will be useful in opinion mining and in other areas. Summarization of tweets carried out with the help of the segmentation process that can be divided into three steps pre-processing of the tweets includes the removal of stop words, hash tags, quotation marks, smiley symbols and other special characters. In the second step, the segmentation process will take place with the help of domain knowledge base which is given as the input for the segmentation process, tweets with informal notations will be considered as the weak segments while the strong keywords is identified related to the domain to forward to the next process that is classification. Finally, third step includes the classification of domain or topic that means when the user wants a specific topic related summary then that will be classified with available segments and knowledge base of that topic. In order to classify the keywords with the related domain, the NEs should be taken as a reference because we are performing the extraction based summarization of twitter posts. The classification of known keywords can be defined as the NEs itself. Fig2 will show how the summarization of tweets application is working. The process shows architecture flow of the proposed system.

- Knowledge base construction is a task of collecting the entire domain and domain related descriptions it works as the input.
- Collection of tweets from the user accounts.
- Segmenting the tweets using tweet segmentation.
- Search for the domain related keywords of the segments.

VI. RESULTS:

The method of segmentation will give accurate keywords in order to produce the summary of the tweets. We have taken a knowledge base compared with the domain topic; in which each domain will consider a specific keyword. The relevancy of the domain has its own keywords. Tweets are taken to find the production of summary with the relevancy of keywords, the results graph represents the no of segments of the domain related keywords with no of tweets which belongs to the keywords. The accuracy of the summary will be highly affected by the segmentation process; it is useful in the extraction based summary generation which helps in terms of identification of specific keywords.

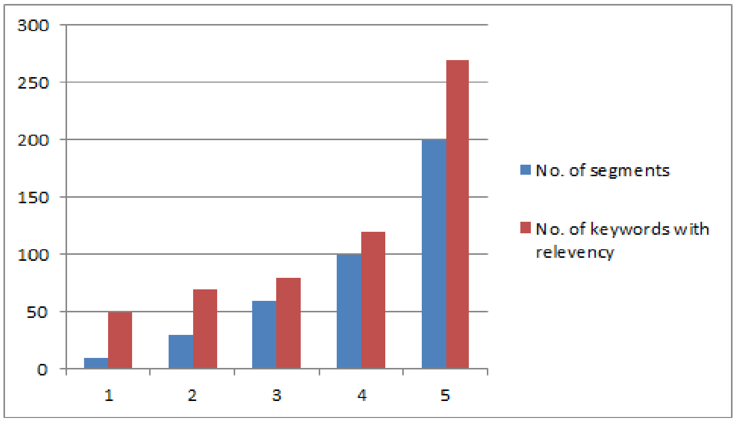

Fig 1. Result of segmentation

VII. CONCLUSION AND FUTURE WORK:

The task of summary helps in reducing the time to know about a particular topic, it is very difficult if we consider social media platforms like twitter to perform such a big task of summarizing of tweets. But if we use the segmentation method to represent the task of summarization it reduces difficulty in the identification of a specific keyword or NE, thus giving an opportunity for understanding the informal texts which are presented in various social media platforms. The segmentation framework will highly evaluate the factors of the representing such extraction based summaries. In future, we can extend this by improving the segmentation method, including the local factors such as local languages.

REFERENCES:

1. Chenliang Li, Aixin Sun, JianshuWeng, Qi He “Tweet Segmentation And Its Application To Named Entity Recognition”, in IEEE transactions on Knowlegde and Datamining,vol.27, NO.2, February 2015.
2. C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, “Twiner: Named entity recognition in targeted twitter stream,” in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 721–730.
3. Kolcz, A., Prabakarmurthi, V., and Kalita, J. (2001) Summarization as Feature Selection for Text Categorization. Proceedings of the Tenth International Conference on Information and Knowledge Management, Atlanta, GA, USA, pp. 365–370.
4. Li Y. H. and Jain A. K., “*Classification of Text Documents*”, The Computer Journal, Vol. 41, No. 8, 1998, IEEE Journal.

5. II. P. Luhn. The automatic creation of literature abstracts. In *IRE* National Convention, pages 60-68, 1958
6. Harish B. S., Guru D. S. and Manjunath S.; "*Representation and Classification of TextDocuments: A Brief Review*", in IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR, 2010.
7. Patra A. and Singh D.: "*A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms*", International Journal of Computer Applications, Volume 75, August 2013.
8. Shen D, Sun J-T, Yang Q, Chen Z: *"Text Classification Improved through Multigram Models*" at ACM Transactions at *CIKM'06,* Nov. 2006, Virginia, USA.
9. Giannakopoulos G, Mavridi P, Paliouras G, Papadakis G, Tserpes K: "*Representation Models for Text Classification: a comparative analysis over three Web document types"*, ACM Transactions at WIMS"12, June 2012 , Romania.
10. Gayathri K, Marimuthu A: "*Text Document Pre-Processing with the KNN for Classification Using the SVM*", Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013) IEEE
11. D. N. Milne and I. H. Witten, "Learning to link with wikipedia," in Proc. 17th ACM Int. Conf. Inf. Knowl.., 2008, pp. 509–518.

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- North Asian International Research Journal Consortium (NAIRJC)
221, Gangoo Pulwama - 192301
Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com