Enhancing Accessibility of Messages Using Clustering and Labeling In Micro blogging's

Syed Thouheed Ahmed, Charan KV, Sudha V,Suma H,Sindhu M, Shree Ranjani J, Prerana K Jayaprakash

Dept. of Computer Science and Engineering,
Dr. T Thimmaiah Institute of Technology, Karnataka, India
Syed@drttit.edu.in

Abstract. In recent years social media micro blogs has gaining popularity due to increasing availability, immediacy and new way of communication medium. The increasing quantity of micro blogging messages creates many challenges for its proper adoption. One such problem of difficulty in accessibility of interested micro blog messages is addressed in our proposed System using NLP technique. The traditional method of message accessing is slow and doesn't establish any semantic and structural relationship between the words in a sentence. The proposed system overcomes this using clustering and labeling of messages having similar semantic and verbal association. The result shows 50% improvement in the message accessibility compared to manual method of searching the particular message in a large micro blogs of 100 pages.

Keywords: Micro blogging, NLP, Social Media, Clustering, Labeling**.**

1 Introduction

Huge amount of data is being generated every second. It causes great difficulty to fulfill user's information needs due to information overload. In recent years micro blogging sites are increasingly used for communicating breaking news, participating events and the information sharing. In recent digital revolution it has become one of the medium for fast communication. According to recent statistics, Twitter has over 500 million posts per day hence the information overload may occur which gives great obscurity to fulfill the information that is need for the users.

The recent messages are more frequently accessible by many users instead of accessing old messages. The short and noisy nature of micro blogging's handles the extract meaning of messages because it is not easily understandable by different kind of people. The micro blogging posts of some people are irrelevant to the discussing topic. These are the few difficult faced by micro blogging's site users because of that they are unable to extract needful information from posts. So NLP techniques are used to minimize these difficulties. Micro blogging is an online platform, a micro-blog is

ICASISET 2020, May 16-17, Chennai, India

DOI 10.4108/eai.16-5-2020.2304025

nothing but a traditional blog which enables the users to exchange the information such as sentences, images or video links. These small messages sometimes called micro posts. It also offers some privacy settings and also the users can control readers who read their micro-blogs. The first micro-blogs were known as tumble logs in April 12, 2005. These Micro blogging services produced a platform called Twitter. Micro blogging services have become the platforms for marketing and public relations. It has become an important source for news-update where we can post news quickly where people share lots of information. Micro blogs are important because, it allows us to connect with the people around the world. It gives a type of community to interact. We can also use this to learn about current events. The main disadvantages of micro blogging are that the content is limited. Since micro blog consists of 140 characters so quite difficult to achieve task.

Natural language processing is a subsection artificial intelligence which is used to deal with human and computer interaction. Speech recognition, natural language understanding and generation are the main challenges of NLP. In 1950, the history begins by Alan Turing by an article called "Computing Machinery and Intelligence". In recent research mainly focused on unsupervised and semi-supervised learning algorithms. These algorithms is difficult than supervised learning and gave less accurate results. In 2010s, representation learning and deep neural network methods widespread in NLP due to its results.

2 Literature Survey

NLP frameworks and techniques are widely used in many domains including education, healthcare, social media and banking. In education it helps in E-learning, automatic evaluation system and improving connectivity between universities and researches to motivate innovation activities. Healthcare system advancements including electronic healthcare, telemedicine, automated tablet dispensers also make use of NLP framework.

There are many applications, challenges and limitations of NLP framework which includes overcoming challenges of managing ever-growing social media microblogging data and data challenges so no.Xia Hu et.al [1] Listed tweets and re tweets of followers in reverse chronological order for better readability, more number of messages are displayed in the interface. The experiments made by authors presents conclusions that novel frameworks in the proposed system enhances the openness of microblog messages by using semantic knowledge. Hemavathi et.al [2] Found that the microblogs in social media are arranged in reverse order and user cannot read all the messages, using clustering and labelling methods the comments are categorized hence the user can quickly access the interested area .The proposed idea works on how the clustering is done for quick access of user, It works only for an Unsupervised messages. The future work was finding technique for semi and supervised texts. Mohammed Sameer et.al [3] Found a problem that requires low cost clustering techniques and cannot add more predefined words for labelling. However, with the help of NLP techniques i.e., analysing the input from Facebook and twitter the authors categorized the

messages using Cluster techniques and Named entity Recognition technique for labelling where most frequently used words in microblog act as label. Hence clustering and labelling technique is done for original short noisy messages.

EafaNazeer Ahmed Jatti et.al [4] The amount of messages or information in social media is growing and accessing those information it become increasingly important and value of NLP applications. The document is divided into sentences and carry out clustering with help for Named Entity Recognizer and each cluster are given a label which is most frequent in that cluster. Data clustering and Cluster labelling are the two jobs to overcome the disadvantage of access of messages, Here by the use of NLP, by using Bag of words(BOW) the unstructured words are structured which is also a method of NER for enhancing the quality of text representation. SoumiDutta et.al [5] In the world of technology, online social media is one of the most popular platform of exchanging of user and needed information exchange. Due to noisy and precise nature of messages, it is complex work/task to classify data. Here work focuses on comparative study [9] [10]on different clustering methods and performance of each algorithm, so that it will be helpful to the researchers/developers, which of the algorithm suits for clustering. Three clustering algorithms are compared and calculated the performance on similar data sets, i.e., Graph Based Clustering Algorithm, Genetic Algorithm based [6] [7] Tweet Clustering; Feature Selection- based Clustering Algorithm. This study clearly tell that classification of micro loggings or social media data is one of the difficult task and another complex task is to categorizing the social media data is increase with data set [8][11].

3 Proposed System

In the proposed system, we attempted to build an efficient method of extracting messages containing the themes/subtopics of clients intrigue. In the same way to approach this method we use Semantic knowledge by NLP technique, that is Natural Language Process, these should trace the human understandable language because many of them are not aware of present technologies. The various types of messages are accessible in microbloggings, yet they are short and unstructured makes issues to investigate microblogging messages. By utilizing our proposed framework it easy for the clients to experience his/her intrigued messages from microbloggings.

The phases of the proposed system are:

Syntactic decomposition: The process of collecting the information/data in the form of a document and then dividing those documents into sentences is called syntactic decomposition.

Pre-processing: Changing unstructured word to organized word with the help of pre-characterized word reference and removing additional spaces and dots then further it is given to the clustering phase.

Clustering: The process of gathering/grouping correlated sentences is called clustering. Here the input sentences are isolated into gatherings of NER(Named Entity Recognizer) in view of specific conditions.

Label: Based on the clusters that are been made the labels are been assigned.

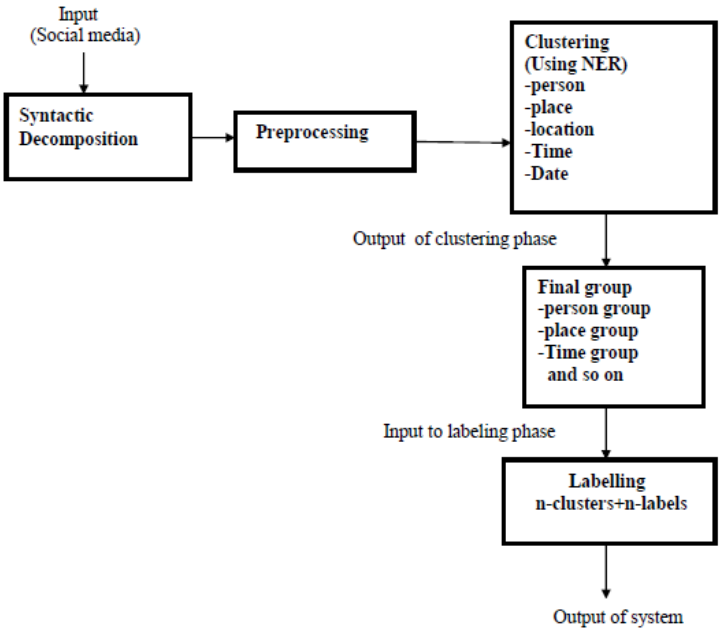

Fig.1. Block Diagram of Proposed System

The Fig.1 speaks the detailed design of clustering and labeling in micro blogging which changes over n input number of micro blogging messages into five diverse gathering of NER (Named substance recognizer, for example, Person, Location, Date, place, Time. By this technique clients can easily search his/her followers' microblogging messages. Few messages are unstructured, so syntactic decomposition and preprocessing phases is added to increase the quality of text representation. And later data is clear and sentenced it will be given to clustering phase for further process as mentioned above by using NER we categorize and grouped in final group phase so that labeling can be done easily, at last phase in the architecture i.e., labeling is done based on groups provided in previous phase.

4 Results

It portrays trial results that were accumulated from execution of venture. Task is created utilizing python, PHP/HTML, The graph considered here depends on the yield of undertaking, by utilizing NLP instrument, effectively performs out the grouping and marking of gathered microblogging. Essential advance to run the undertaking is to run wamp server in online mode.

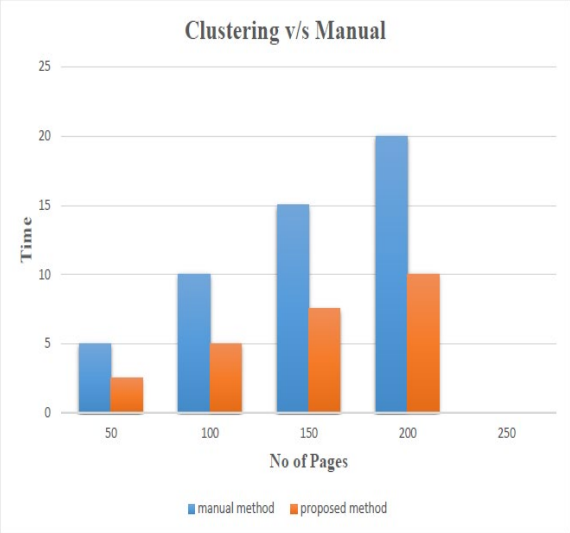

Fig.2. Comparison between Manual and Proposed System

The above graph figure 2 is drawn based on intuition, it shows betterment in accessibility of interested micro blogging messages by decreasing search time approximately 50% compared to manual system of searching micro blogging messages. In proposed system, first syntactic decomposition task is performed where the document is broken down into sentences, the output of syntactic decomposition is given as input to preprocessor. In preprocessor mapping is done by using predefined dictionary for example in sentence if '@' occurs it will be mapped and replaced with 'in' word. Thus in preprocessing unstructured words will be converted into organized one, the output of preprocessing will be given as input to clustering. In clustering phase similar kinds of sentences are grouped together using Named Entity Recognizer (NER) technique and labels are assigned to each group of clusters to represent it.

5 Conclusion

The employments of micro blogging destinations are stacked with incredible measure of micro blogging messages. It is exceptionally hard for clients to experience messages of their advantage, In this paper, the main concentration is on grouping and marking of micro blogging messages utilizing NLP methods, we can separate the messages into bunches(clusters) utilizing Named Entity Recognizer [NER] and appoint name (label) to each bunch to speak to that group. This encourages the client to rapidly explore through their intrigued data. It has additionally changed the unstructured messages to sort out one which has improved the nature of text portrayal. Since the messages produced from microblogging destinations were not in proper configu-

ration, for the gathered microblogging messages bunching and naming is done, with the objective that clients/customers can rapidly get to their intrigued messages.

References

1. Xia Hu, Lie Tang, HuanLiu , “Embracing Information Explosion Without Choking: Clustering and labelling In Microblogging”, IEEE ,vol.1,Issue 1,Jan -Mar2015.
2. Mrs D Hemavathi, Ms M kavitha, Narjiya Ahmed Begum,”Information extracting from social media: Clustering and labeling using microblogging”,2017 International Conferenceon IoT and its application(ICIOT),1-10,2017.
3. H Mohammed Sameer, Smt M Kavitha, MrSrinivasKarur, “Clustering and labeling micro loggings using Natural language processing techniques “ ,International Journal of Innovative Research in Science and Engineering, vol.No.2,Issue 04,April 2016.
4. EafaNazeer Ahmed Jatti1, BhavanaChandavar ,RanjitaNaik , MdAaqibuddin, “Clustering and labeling of messages in social media”, International Research Journal of Engineering and Technology(IRJET), vol.06, Issue 05,May2019.
5. SoumiDutta , Asit Kumar Das, GouravDutta, “A Comparative Study On Cluster Analysis of Micro-Blogging Data”, Emerging Technologies in Data mining and Information Security,pp.873-881,2019.
6. Ahmed, S.T., Sandhya, M. & Sankar, S. TelMED: Dynamic User Clustering Resource Allocation Technique for MooM Datasets Under Optimizing Telemedicine Network. Wireless PersCommun 112, 1061–1077 (2020). https://doi.org/10.1007/s11277-020-07091-x
7. Gunashree, M., Ahmed, S. T., Sindhuja, M., Bhumika, P., Anusha, B., &Ishwarya, B. (2020). A New Approach of Multilevel Unsupervised Clustering for Detecting Replication Level in Large Image Set. Procedia Computer Science, 171, 1624-1633. https://doi.org/10.1016/j.procs.2020.04.174
8. Sreedhar Kumar S, Syed Thouheed Ahmed, NishaBhai, Vinutha B A, “Type of Supervised Text Classification System for Unstructured Text Comments using Probability Theory Technique”, International Journal of Recent Technology and Engineering (IJRTE), Vol.8, No.10, DOI:10.35940/ijrte.B1158.0982S1019, 2019.
9. K. D. Singh and S. T. Ahmed, "Systematic Linear Word String Recognition and Evaluation Technique," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0545-0548, doi: 10.1109/ICCSP48568.2020.9182044
10. M. Anathi, K. Vijayakumar , “An intelligent approach for dynamic network traffic restriction using MAC address verification”, Computer Communications,Elesvier,5 February 2020.
11. K. Pradeep Mohan Kumar, M. Saravanan, M. Thenmozhi ,K. Vijayakumar, “ Intrusion detection system based on GA-fuzzy classifier for detecting malicious attacks”, wiley, Feb 2019.