International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

CLUSTERING AND LABELING OF MESSAGES IN SOCIAL MEDIA

Eafa Nazeer Ahmed Jatti[1], Bhavana Chandavar [2],Ranjita Naik [3] , Md Aaqibuddin [4]

[1,2,3,4]*Anjuman Institute of Technology and Management, Bhatkal, Karnataka.*
Under the guidance of
G.Vannur Swamy *(Professor of Information Science and Engineering Department, AITM ,Bhatkal)*

Abstract - *As the amount of social media information available online is growing, the need to access it becomes increasingly important and the value of natural language processing applications becomes clear. It introduces a disadvantage for the user to easily discover his/her follower's opinions when the user interface is overloaded by a large number of messages. So, we applied NLP techniques to manage a huge number of messages into clusters and label them. In this paper, we propose a method that divides the document into sentences and carries out clustering with the help of a Natural Language Processing tool called Named Entity Recogniser and each cluster is given a label which is a most frequent word in the clusters. Generally, with the help of document clusters, a user would prefer to quickly have a look through the collection to easy identify clusters of interest without examining particular documents in detail. This method is commonly used to model and evaluate data that are smaller in size.*

Key Words: Natural Language Processing, Clusters, Labels, Facebook, Twitter.

1.INTRODUCTION

The majority of tasks accomplished by humans are done through language, whether communicated directly or reported using natural language. As technology is increasingly making the methods and platforms on which we communicate ever more accessible, there is an even greater need to understand the languages we use to communicate. By combining the power of artificial intelligence, computational linguistics, and computer science, Natural Language Processing (NLP) helps machines "read" text by simulating the human ability to understand language.

Natural language processing is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. in particular how to program computers to process and analyze a large number of natural language data. Social media is used to send short messages, images, videos, links, etc, which causes a major reason for their popularity and provides difficulty in displaying his/her follower's messages.

Twitter, Facebook, and other social services have become platforms for marketing and public relations, with a sharp growth in the number of social media marketers. The social media sites like Facebook and Twitter which produces a large amount of data on a daily basis. As there is an increase in communicating text, photos, and multimedia of their own devising and share it. By using certain apps users can receive notifications of their friend's activities. Users may join common-interest groups. Facebook and Twitter receive prominent media coverage, including many controversies such as user privacy and psychological effects. This data is a burden and provides difficulties for users to find the information on her/his interest.

Many messages of their interests may be hidden in a large amount of streaming data. To overcome this difficulty of microblogging's site, the concept of clustering and labeling is introduced for random data. With the help of this, users are easily able to find information of their interest.

Clustering is the process of making a group of similar objects. It is done with the NLP technique i.e., Named Entity Recognition and labeling with the most frequent words in the clusters.

2. PROBLEM DEFINITION

With a huge number of messages appearing in the interface, people mostly do not have time to go to each message. It suppresses a user from moving to what he/she is interested in. Here we define two jobs to overcome this:

1) Data Clustering: Let D =d1,d2……, dn be a collection of d microblog's data. Out of these d messages, there are k different subjects. We cluster the n messages into k clusters c1, c2……., ck.

2) Cluster Labeling: For each cluster ci, the label is generated with the most frequent words in the clusters l1,l2……lk.

3. METHODOLOGY

To make a large collection of social media messages accessible to users, we proposed a system which will provide clusters and also labels for each cluster. The users can easily find messages of their interest.

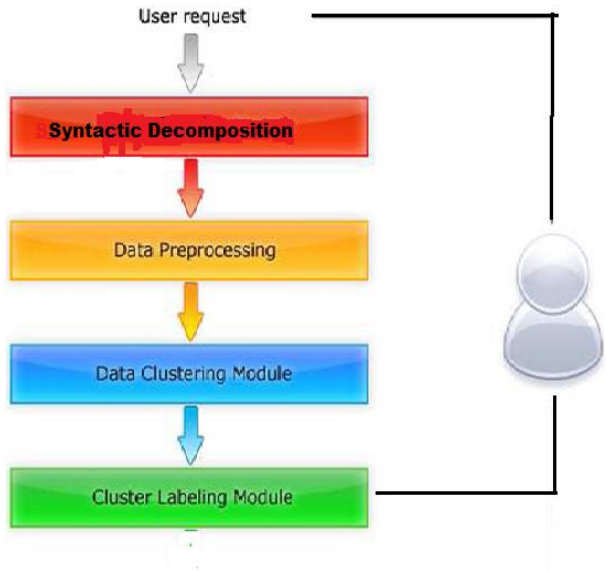

Fig -1: Data Clustering and Labeling Process

3.1 User Request

Data is requested by the user through Facebook and Twitter as input to our system.

Steps to extract Facebook Data:

Step 1: Facebook Developer Registration

Go to https://developers.facebook.com and register yourself by clicking on the **Get Started** button at the top right of the page. After this, it would open a form for registration which you need to fill it to get yourself registered.

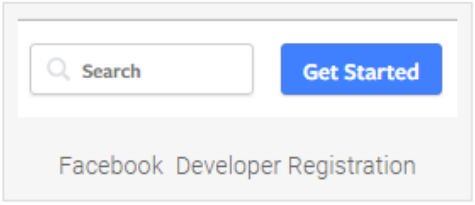

Fig -2: Facebook Developer Registration

Soon after completion of registration as shown in step 1, we need to click on **My Apps** button. Then select Create **New App** from the drop-down list.

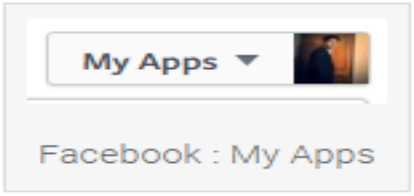

Fig -3: Facebook :My Apps

Then you need to write **Display Name** of App ID and press **Create App ID** button.

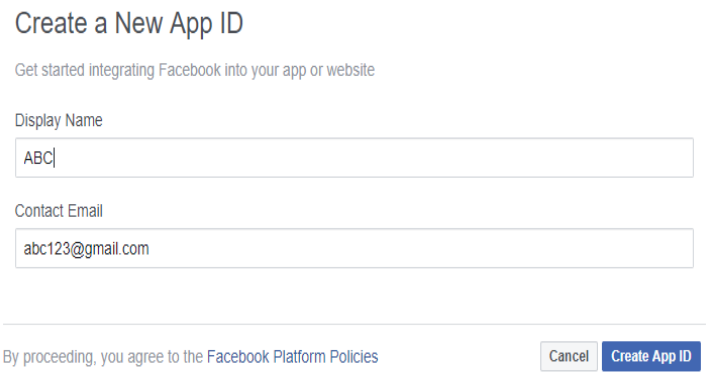

Fig -4: Facebook :Create a New App ID

Step 2: Login with id and password and wait till the binding process is complete.

Step 3: Token Id is generated.

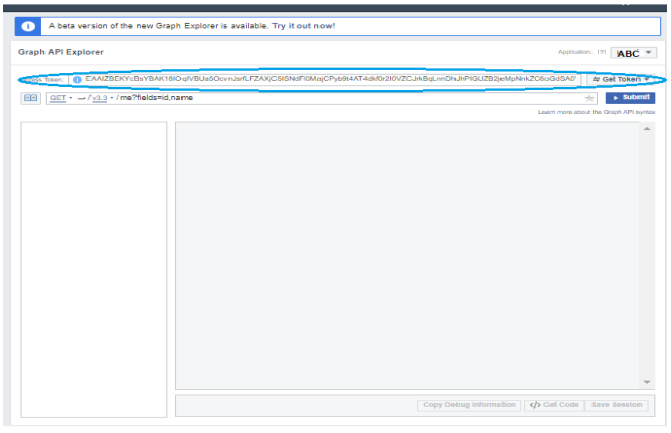

Fig -5: Facebook :Token ID Generated

To bring out Facebook data we need to generate a token from developer.facebook.com. Open **Graph API Explorer** in developers.facebook.com/tools/explorer and click on the "**Get Token**" button.

Step 4: Collect the data of individual.

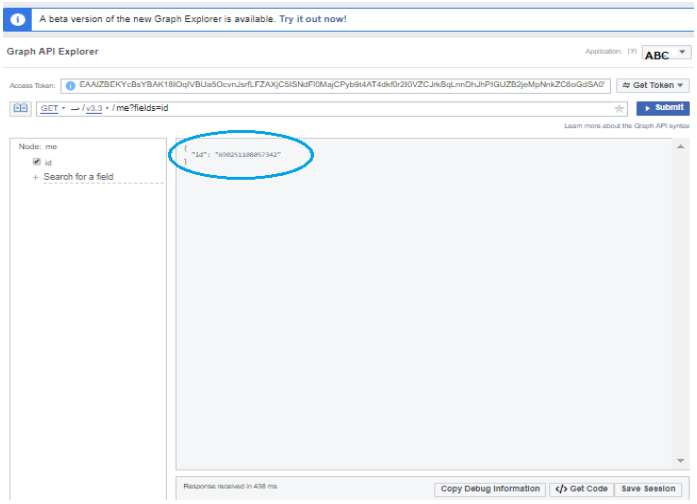

Fig -6: Facebook :Collect data of individual

3.2 Syntactic Decomposition

Data decomposition is a highly effective technique for breaking sentences into small tasks that can be parallelized called tokens(words).

User Request: Hii friends, wish you all the best for the future.

Output: Hii, friends, wish, you, all, the, best, for, the, future.

3.3 Data Pre-Processing

Data preprocessing is a data mining technique that deals with transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and lacks certain behaviors or trends, and is likely to contain many errors. Data preprocessing resolves such issues.

Ex: @=at, grt=great etc.

3.4 Data Clustering Module

As we know NLP, is human-computer interaction. With the help of NLP, technique clustering can be done. Clustering is the process of grouping a set of similar objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters).

Text clustering may be used for different tasks, such as grouping similar documents (news, tweets, etc.)In this paper, clustering concept for grouping the messages is done by NER technique, Named Entity Recognition (NER), information extraction to recognize and divide the named entities and classify or categorize them under various predefined classes such as the person names, organizations, locations, time expressions, quantities, etc.

Ex: Rahul bought 100 shares of ABC Corp. in 2019

[Rahul]*Person* bought 100 shares of [ABC Corp.]*Organization* in [2019]*Time*.

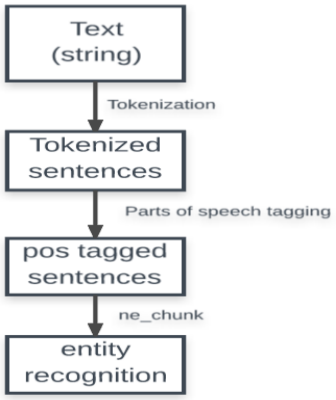

Fig -7:NER Technique

1)Text (String): Both a string and a text field will hold information that you can freely write in. The major difference arises in a number of characters in it. Given a piece of text, find the sentence boundaries. Sentence boundaries are marked by periods or other punctuation marks.

Ex: ABC in Bangalore.

2)Tokenized Sentences: "Tokens" are usually individual words and "tokenization" is taking a text or set of text and breaking it up into its individual words. These tokens are then used as input for other types of analysis or tasks, like parsing. Commonly used tokenization methods include the Bag-of-words model and model. Removing stop words and punctuation.

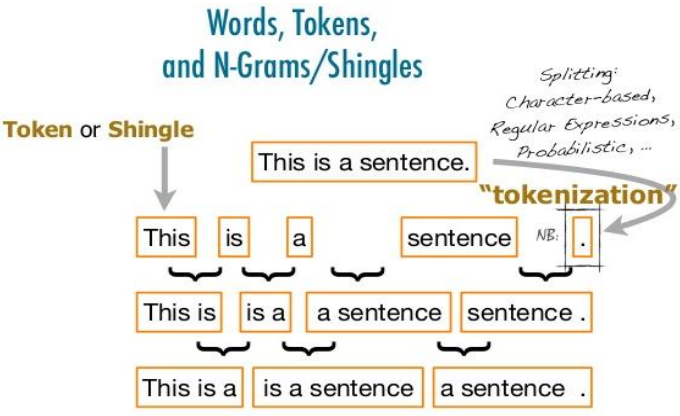

Fig -8: Tokenized sentence

3)Pos -Tagging: It is a process of converting a sentence to forms – a list of words, a list of tuples. Where each tuple is having a form (tag, word). The tag in the case is a part-of-speech tag and signifies whether the word is a noun, adjective, verb, and so on.

Fig -9: Parts of Speech(POS)

Fig -10: Parts of Speech tags

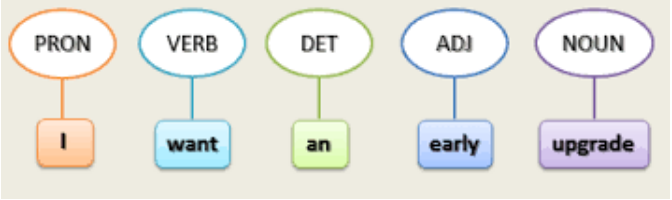

Fig -11 : Example for Parts of Speech tags

4) Named Entity Recognition (NER): Also known as entity extraction classifies named entities that are present in a text into pre-defined categories like "individuals", "companies", "places","organization","cities","dates","product terminologies" etc. It adds meaning to a text which is easily understandable.

With the help of ne_chunk function classifies tokens into different entities like 'PERSON', 'ORGANISATION', 'LOCATION', etc.

ne_chunk makes use of a classifier-based approach and has been trained on parts of speech tagged data. The classifier tries to understand the relationship between the occurrence of various parts of speech and entities in a sentence.

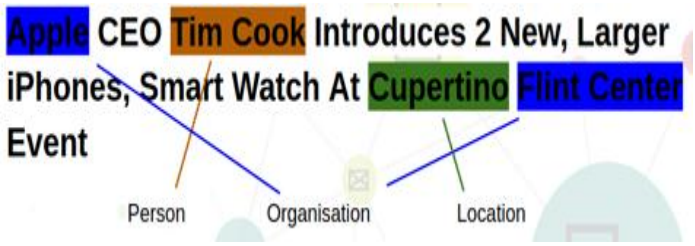

Fig -12 : Example for Entity Recognition

3.5 Cluster Labeling Module

The process of attaching a label to a group of clusters, which is the most frequent word in the clusters. In this way, the user can save a lot of time in finding messages of their interests.

4. CONCLUSIONS

By the use of NLP, the unstructured words of social media messages are structured using Bag-of-Word(BOW), which is a method of NER, to enhance the representation of text messages. Clustering and labeling give the user an efficient way to quickly get messages of their interest.

ACKNOWLEDGEMENT

The successful completion of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along with the completion of our project and also for the comments that greatly improved the manuscript. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

So, we would like to express our first and foremost special thanks of gratitude to our project guide **Prof.G.Vannur Swamy**, Department of Computer Science and Information Science, AITM, Bhatkal.

We also owe deep gratitude to our project coordinator **Prof.S.G. Bhagwath** and our HOD **Prof.Anil Kadle,** Department of Computer Science and Information Science, AITM, Bhatkal, who took a keen interest on our project work and providing all the necessary information for developing a good system.

We respect and thank **Prof. M. A. Bhavikatti**, AITM, Bhatkal, for providing us a golden opportunity to do this task.

REFERENCES

[1] Technical Writer's Xia Hu, Lei Tang, Huan Liu, "Embracing Information Explosion without hoking: Clustering and Labeling in Microblogging", IEEE TRANSACTIONS ON BIG DATA, January 2015.

[2] Logamani .Kand Punitha. S. C , "Density Based Clustering using Enhanced KD Tree", in International Journal of Computer Science Engineering and Technology(IJCSET), November 2014 | Vol 4, Issue 11,314-318.

[3] H. Mohammed Sameer, Smt. M.Kavitha, Mr.Srinivas Karur," clustering and labeling in microblogging using nlp techniques", in International Journal of Innovative Research in Science and Engineering(IJIRSE),April 2016|Vol.No. 2,Issue 4,366-370

[4] Bharat Chaudhari, Manan Parikh, "A Comparative Study of clustering algorithms Using weka tools "International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 1, Issue 2, October 2012.