Topic Modeling: Construct Analysis for Public Issues Using Social Media

Clint Paxton SAMUEL

Mount Zion College of Engineering and Technology

***Abstract*--In recent trends we use social media to share post and tweets and to interact with friends through messages; nowadays it has gone a long way by raising our voice about the issues that we are facing in our day to day life and we often give voice to support others anywhere in the globe. The prime focus is on "topic modeling" of how the tweets are been retrieved and analyse the impact that is been reflected to the issue in that particular period of time and the policy changes done due to these post and tweets by using LDA(Latent Dirichlet Allocation) and Gibbs Method and the impact brought to the very society.**

***Keywords*: Social media, LDA, Gibbs sampling, topic modelling**

I.INTRODUCTION

In order to share our pictures and current mood as post and tweets we use social media such as "facebook" and "twitter" and more often we interact with our friends through messages in these media. Nowadays social media has gone a long way like education, shopping and find our life partner too. The initial reason to create a platform for such social media is to improve communication and get us connected to people who are not near us. Many of us always depend on these social media to gather information, learn stuff that we often don't find in the books and know the culture and taboo of various places around the globe.

Bringing the media accessible in our devices such as our smart phones make us more attached to these devices; instead of spending our time with the real time people who is around us. There is also complaint that students are not spending enough time in studying or performing practical work rather they spend more time in social media and the devices related to it. There are couples who developed relationship and were able to identify their life partners through social media; at the same time there are fake accounts to distract youngster who are not matured enough, misguiding them to get themselves trapped as a victim.

Other than all the criteria given above social media is basically to connect people, and the sites are created in a user friendly way to break our personal fence(privacy) and fly high to see others in the world beyond the sky. More time we are able to identify our own character about what is our role in the drama and find the perfect character that we want our self to perform the way we want and not the way others expect us to do. Before two decades an individual was not able to raise voice for his own rights, he was not identified, he was not heard, he had no means to convey his necessity and he was hidden from the individuals around him. Nevertheless now raising our voice for our rights, for social problem, and to have our cultural heritage and even to save our environment the social media is helpful. These reverberations are not only to be heard by the government and its official instead it is now carefully attended by each and every individual in the social media.

With such voices there had been protest that is been organized for some sort of problems without a strong basement and got success with the help of social media. Such as the cases in "Iran '09", "Egypt '11" and"2017 pro-jallikattu protests". Those who have accounts in any of the social medias are having all rights to create, share and comment any article into the media spreading them around the world. Often government is using social media to spread awareness about the endangering species and many others send facts that is useful for daily life or facts that is been forgotten in our daily life. Even though there is a fun side in sharing all these stuff there is also the other side where the false facts is being published and again creating confused story behind us. It is mandatory for us to identify the truth behind the scene about the article that is been shared, and to find the loyalty and the truth about it.

There are streams that are been reviewed and judged with the help of social media, as there are always positive and negative views for a particular object. The movie review, political parties with more supports, organizes social activities and so on. When it comes to review there is always two faces, sometimes more than that; let us not worry the number of faces instead concentrate on positive and negative faces. In order to find such cases we must run an assessment to analysis and get an output on how the people has shared their views about the subject that we are going to analyse. In this paper with the help of an open source programming language 'R', the post and tweets are extracted from facebook and twitter respectively, analysing and representing them in graphical form and obtain an conclusion. There are number of tweets/post been extracted using R in a several period of time and find the mentality of the public in that time, plotting in graphical structure.

In order to identify the related topics that are been analysed using social media and R we use a statistical method called "Topic Modeling". It is used to group the topics and phrases that is been mainly used by the public and draw a conclusion from the response of the user and plot them as a

graph. The tweets that are being extracted are completely analysed and broken into separate words and categorize them into different set, by the maximum number of usage of these words. The most used words are formed into word cloud with all the topic related to the subject that we are analysing. To complete this statistical method we use some techniques such as LDA (Latent Dirichlet Allocation) and Gibbs method.

Since we look forward to a point that has clear definition of the subject that we are going to analyse, but we extract number of tweets and post which are collected randomly in the period of time. These post and tweets are identified with the help of "metadata tag" which is commonly known as "hash tag (#)"; post are often uses this symbol and type the subject that are being posted or tweeted to categorize and to make sure that they are referring to the particular subject in the posts. So when collect all these post we get a lot of posts referred to the subject and we use topic modelling in order find related and most used topics about that subject; here is where we use LDA to study these post and find how many percentage of this post is referred to the subject that we are searching about. Many times we share many subject into single post which may complicate the analyse by introducing foreign words into our subject, which may affect our topic modelling to avoid such issues we follow LDA method into topic modelling to make it simpler and easier.

II. Existing Approach

In a journal titled as "Social media and its effects on individual and social system" completely gives the idea of how the global communication is been improved and make sure that even the civilians in the deserted area are able to express their views and get it exposed to the world. It mainly consider between virtual interaction of global communication and individual communication which has improved our connection with others by making a strong bonding with those who are away from us. [1]

Social media has gone a long way from individual communication to global communication by expressing their views of good and bad of the public cry or an activity performed in the society. The role of each individual they act through social media has their impact to the society like usually the users are mainly fond in streaming news feed and the profile pictures and so on, some focus on blogs and forum to develop their knowledge. Whereas now the number of individuals taking part in activism through social media has increased [2], in order to have a strong movement in the field of political science and social movements the effect will always depend upon the communication between the individuals and the decision maker.

There are number of studies focusing on how a social media can be a tool in shaping social movements for both offline and online at a global level.[3]Social media such as facebook, Twitter, YouTube and the various online blogs have given their voices of support for many individuals otherwise they would have never been heard. The first biggest revolution started with the help of social media is known as 'Arab Spring' which is commonly known as 'Egyptian Revolution', 'Facebook revolution' and Twitter revolution', why does this movement has its name after facebook and twitter, it is only because of the role of social media in this movement. There has been study on the movement of how a social media has played a major role and the changer it has bought to the society. [4] There were blogs written and published in the internet against the Mubarak regime, Egyptian government failed to block or privacy the internet user which lead to share instant message through facebook and the information to form a group in the Tahrir Square to protest against the government. Ultimately in the end they had succeeded their victory.

Social media is not only used for an activism but also to highlight the issues and the problems faced in the society, whereas most of the problems can be rectified instead they can be overcome or find a way to get an aid. Japan is an Earthquake prone area, there is no chance that they could avoid this natural calamity but it can be overcome by investigating the real time interaction of the event using twitter. According to the journal that deals about earthquake and an application is created in order to report the monitor event occurring whereas the Meteorological Agency cannot detect more than three events at a time but we can overcome this issue with the help of social sensors otherwise known as social media.

Previously to get public opinion during election was to walk on the road holding a microphone and recording the opinion of the public in which the high class who move fast in cars and those who are in a hurry to catch a bus are skipped in the opinion poll. There was survey taken to predict the US primary election with the help of twitter. [5] In a particular interval of time the tweets are been extracted to perform the sentiment analysis of the most popularly subject and predict the result, with the predicted result the original outcome is to be compared and find the final output is obtained, this is predicted by the people opinion on each and every candidate and tweets about them. At the end of this journal has not got the result as they expected for they have contradiction between the tweets and the election result and believe that posting tweets will never make any changes to the election result but they can help to capture the public trend in the real world through social media.

A survey in data mining technique is a survey that deals with the different type of mining techniques done in social media for the past years and up to date. [6] With the help of this survey we can understand the techniques that are been followed in the data mining done in social media, it has a significance of how the data is being used in order to share our view critic an event or an

individual. According to the graph theory the major components are nodes and links that is to get the followers and get the link through them. There may be different data mining techniques used in social network analysis which totally depends on the supervision that is been conducted on by retrieving information and contents of the data generated.

How analysis in social media works, is answered by the survey that have been done through some techniques and algorithms. As the journal "Approximate Frequency Counts over Data Streams" [7] deals with an algorithm for finding and calculating the frequently used word count, where the count is declared by the user. It was to give a document full of statements regarding a single subject. The frequently used topic differ for each and every document as their matter of subject is different so make an analyst with a single set of words cannot be used for all the subject, for which a word cloud is being used in order to identify and create a word cloud with which we can run the assessment. Which are clearly states that the frequency count is more than the count given by the user leads to the data stream because the count is being marked approximately which works only for a small document and small memory.

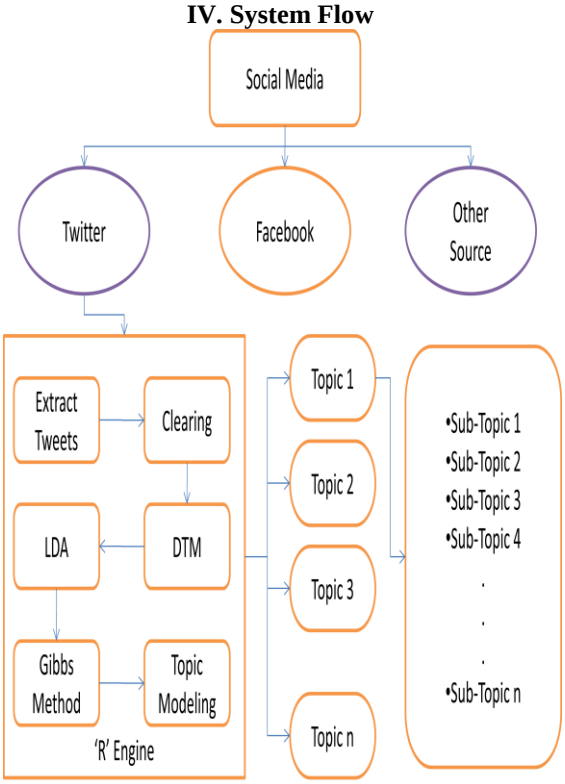
There is another journal that deals with the summarizing data streams using count-min sketch method, "An improved Data Stream Summary: The Count-Min Sketch and its Applications" this journal is fully constructed with the help of mathematical equation and formulae. The survey here is dealing with the large scale document, which has overcome the previously used algorithm for small memory and small document, the time linear has been increased for the update of the data by increasing the speed of the data stream and the sketches are constructed by hash function with strong independence guarantees which may be critical for evaluating in the hardware. With the help of this count-min sketch technique it is now able to sole any sophisticated problems such as big documents or data streams by estimating the fundamental queries.

The algorithm or a statistical method that is been used in the topic modeling is LDA (Latent Dirichlet Allocation) it is a generative statistical model that allows set of observation that is been defined by the group of unobserved and find the similarity between them. The major role is to find the set of description that are collected from the large collection of a document.[8] LDA is fully constructed with the mathematical equation and statistic method, initially LDA was a graphical method for topic discovery that the major topics discussed are found and plotted in a graph. LDA is a collection of discrete data by a flexible generative probabilistic model. It is an easy method for identify the topics that are mainly discussed in a large scale.

III. Proposed Work

Our current focus is on 'Topic modeling' on the current issue based on 'Hydro-Carbon Project' in Neduvasal, Pudukkottai Dt., Tamil Nadu, India. The tweets are been extracted for several interval period of time for my analysis. The high frequently used topics that are been discussed about the Hydro-Carbon Project are been collected in different period are identified and are been compared. The frequency is been set by the user using the DTM(Document Term Matrix) where the topics are tabulated in a matrix form where rows represent topics and columns represent the number of term.

IV. System Flow



The social media is a group of social sites like twitter, facebook and other such sources. Initially my survey is done in twitter, the tweets related #HydroCarbonProject are been extracted and the unwanted characters, punctuations and links are cleared from the extracted tweets. Then the DTM (Document Term Matrix) is done where the frequently used words are been identified with predefined count by the user.

With the use of LDA and Gibbs sampling the topic modelling is carried out by identifying the most used topics in the extracted tweets in different sets and then the each sets are compared with each other and the results are obtained.

V. Methodology

In order to do topic modeling LDA and Gibbs sampling is been used and it is a statistical method that create a set of observed topics from a large document. In natural language processing LDA is a generative statistical model that allows

sets of observation to be explained by unobserved groups that explain why some parts of the data are similar. LDA is constructed with substantial amount of mathematics, but it is important to understand the work of this model and how it is gathering the topics. Given that the mathematical equation to perform LDA is

$$P(\bar{x}|\bar{a}) = \frac{f(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} f(a_k)} \cdot \prod_{k=1}^{K} x_k^{a_{k-1}}$$

Where,

K is the count of the occurrence of the each topic

P is the probability of the occurrence that may occur

$\bar{x}$ is the vector with K component

x_k is a multinomial distribution

$\bar{a}$ is component which contains K parameters

In the right hand side in the equation, the fraction is a normalized constant which is proportional to the Dirichlet parameters. These are expressed as gamma function; the factorial function is generalized as,

$$f(t) = \int_0^{\infty} x^{t-1} e^{-x}\, dx$$

The value of $\bar{a}$ is changed periodically and the changes that are made in the final graph are plotted sequently and their behaviour is being compared.

Gibbs sampling method is used after LDA to get a clear understanding between the topics that are been obtained through LDA method, by comparing the each set of topic with the rest of the sets the final output is obtained. With Gibbs sampling method the topics that are probably expected to predict the relation with all the topics that are been observed using LDA method previously. It is a randomized algorithm that predicts randomly which can be fixed by an approximate count.

VI. Experimental Results

Topic modeling is used to group the topics and phrases that is been mainly used in the tweets that is been extracted using #HydroCarbonProject and draw a conclusion from the response of the user. The tweets that are being extracted are completely analysed and broken into separate words and categorize them into different sets, by the maximum number of usage of these words. The result obtained is

```
Topic 1
"hydrocarbonproject, stop,
saveneduvas, rt, tamilnadu, farmer,
neduvas"

Topic 2
          "hydrocarbonproject, í,
hydrocarbon, methan, rt, neduvas,
íí"

Topic 3
          "hydrocarbonproject, rt,
project, methan, neduvas, ban,
dont"

Topic 4
              "save,
hydrocarbonproject, pleas, rt,
share, awar, creat"

Topic 5
    "hydrocarbonproject, savefarm,
rt, tamilnadu, support, among,
fire"

Topic 6
    "hydrocarbonproject, rt,
neduvas, let, stand, gvprakash,
skycinema"

Topic 7
     "hydrocarbonproject, rt, get,
readi, youngster, avadhaar,
central"

Topic 8
 "hydrocarbonproject, need, rt,
neduvas, profession, protest,
pmoindia"
```

VII.Conclusion & Further Work

In the end of my survey there are eight sets of topics that are mainly discussed in a large scale, the featured topics are 'farmer',''savefarm','support','tamilnadu' and 'student protest'. By analysing these topics we can come to an conclusion that 'Hydro Carbon Project' must be reconsidered by the government for the welfare of the farming land and the sentiment of farmers, according to the tweets from twitter it is against the 'Hydro Carbon Project' in Neduvasal. So with my survey result the government should reconsider the project that is been initiated in Neduvasal and make necessary changes in the policy for the welfare of the farmers and farming land in Tamil Nadu.

The tweets are extracted in the initial stage of the protest during the month of February,2017, so the tweets in several interval of time must

extracted to find the changes in the topics that are been discussed largely, and find the topic changes by comparing with all the extraction and analysement.

References

1. Natascha Zeitel-Bank. Social Media and Its Effects on Individuals and Social Systems in Management Center Innsbruck, Austria on June, 2014.
2. Bart Cammaerts. Social Media and Activism in Mansell, R., Hwa, P., The International Encyclopedia of Digital Communication and Society. Oxford, UK: Wiley-Blackwell on 2015, pp. 1027-1034.
3. Amandha Rohr Lopes. The Impact of Social Media on Social Movements: The New Opportunity and Mobilizing Structure in Creighton University on 2014.
4. Caroline S. Sheedy. Social Media for Social Change: A Case Study of Social Media Use in the 2011 Egyptian Revolution, Presented to the Faculty of the School of Communication on April, 2011.
5. Graham Cormode & S. Muthukrishnan. An Improved Data Stream Summary: The Count-Min Sketch and its Applications, submitted to Elsevier Science on December, 2003.
6. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber & Frederic Stahl. A Survey of Data Mining Techniques for Social Network Analysis in World Wide Web.
7. Gurmeet Singh Manku & Rajeev Motwani. Approximate Frequency Counts over Data Streams in Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002.
8. David M. Blei, Andrew Y. Ng & Michael I. Jordan. Latent Dirichlet Allocation in Journal of Machine Learning Research 3 (2003) 993-1022.
9. Zeynep Tufekci & Christopher Wilson. Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square in Journal of Communication ISSN 0021-9916 on 2012.
10. C. Wang, D. Blei & D. Heckerman. Continuous Time Dynamic Topic Models in Uncertainty in Artificial Intelligence, Helsinki, Finland, on July 2008.
11. Sebastia´n Valenzuela, Arturo Arriagada & Andre´ s Scherman. The Social Media Basis of Youth Protest Behavior: The Case of Chile in Journal of Communication ISSN 0021-9916 on 2012.
12. Tim Markham. Social Media, Protest Cultures and Political Subjectivities of the Arab Spring in Media, Culture & Society 36(1): 89-104 on 2014
13. Takeshi Sakaki, Makoto Okazaki & Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors in World Wide Web on April,2010
14. Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg & Jacob Spoelstra. Predicting US Primary Elections with Twitter in World Wide Web.
15. Rui Miguel Forte. Mastering Predictive Analytics with R, Pages 317-323.

Biography

Clint Paxton SAMUEL, Studying B.E. in Computer Science and Engineering at Mount Zion College of Engineering and Technology, Lena Vilakku, Pudukkottai Dt., Tamil Nadu 622507, India.