Study of Relation between the Proportion of opinionated Tweets and Retweets

K. Sunitha¹, K. Syama Sundara Rao ²

¹M.Tech Student, CSE Department, PVPSIT/ JNTUK University, Vijayawada, A.P, India
²Assistant Professor, CSE Department, PVPSIT/ JNTUK University, Vijayawada, A.P, India

ABSTRACT – Face book and Twitter websites are not just platforms to maintain relationships. They can also be more robust information sources. But, The people are known little about socially generated content. In this project we discover behaviour of relevant item searching on the popular blogging sites like facebook and twitter. By analyzing various sources available in twitter to find relevant information and information related to people and their interests and their opinions. Moreover Twitter and Facebook queries are very short and more popular less likely to progress as part of gathering than Web Queries. More apparently people treat Twitter queries to monitor the search results while changing and developing queries to learn a relevant topic. In our paper We used Latent Dirichlet Allocation technique to compare various documents available in twitter as well as New York Time News to find the relevant and by presumptuous this generative model for a large documents as well as tweets, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection and also In this paper we study the relation among the quantities of prejudiced tweets and topic categories and its types.

Keywords: Tweets, Testing, Training, Microblogging, LDA.

I. INTRODUCTION

Microblogging web pages have evolved to become a source of varied kind of information. This is due to nature of microblogs on which users to communicate with each other. On the twitter users post real time messages about their sentiment on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in our day to day life. In fact, companies manufacturing such products have started to poll these microblogs to get a sense of general opinion of their product. Many times these companies study user reactions and reply to users on microblogs. In particular, it is not clear whether as an information source Twitter can be simply regarded as rapid news feed that covers mostly the equivalent information as conventional news media[1]. In this we empirically compare the content of Twitter with a classical news medium, New York Times, using unsupervised topic modelling. We use a Twitter-LDA model to discover topics from a symbolic sample of the entire Twitter. LDA is a way of automatically discovering topics that these sentences contain. In more detail, LDA represents documents as assortment of topics that give out words with certain probabilities. It assume that documents are produced in the following method when writing each manuscript, you. Decide on the number of words M the document will have. Choose a topic assortment for the document (according to a Dirichlet distribution over a fixed set of N topics). For example, assuming that we have the two entertainment and education topics above, you might choose the document to consist of 1/3 entertainment and 2/3 education [3].

Produce each word in the document by: First we will pick a topic (according to the multinomial distribution that you sampled above; for example, you might pick the entertainment topic with 1/3 probability and the education topic with 2/3 probability). Use the topic to produce the word itself (according to the topic’s multinomial distribution). For example, if we selected the entertainment topic, we might generate the word “movie” with 30% probability, “sports” with 15% probability, and so on. Assuming this generative model for a set of documents, LDA then tries to back down from the documents to find a set of topics that are likely to have generate the collected works

II. LITERATURE SURVEY

A. Empirical study of topic modelling in Twitter by Hong, L., Daviss [1]

Over the past few years, Twitter, microblogging web pages, has become a progressively popular platform for the users to communicate with each other. Because tweets are fast, Twitter has become widely used to spread and
share breaking news, personal updates and spontaneous ideas. The status of this new form of social media has also started to attract the interest of researchers. Several modern studies examined Twitter from different perspectives, including the topological characteristics of Twitter, tweets as social sensors of real-time events, the predict of box-office revenues for movies, etc. However, the explorations are still in an early stage and our understanding of Twitter, especially its large textual content, still remains narrow. Due to the natural history of microblogging, the large amount of text in Twitter may probably contain useful information that can hardly be found in conventional information sources.

B. Earthquake shakes Twitter users: real-time event discovery by social sensors by Sakaki, T., Okazaki, M., Matsuo [2]

Twitter is categorized as a micro-blogging web page. Microblogging is a structure of blogging that allows users to send short text micro media such as images or audio clips. Microblogging services other than Twitter consist of Tumblr, Plurk, Emote.in, Squeelr, Jaiku, identi.ca, and so on. They have their personal individuality. Some examples are the following: Squeelr add geolocation and images to microblogging, and Plurk has a timeline view integrate video and image sharing. Although our study is useful to other microblogging services, in this study, we specifically study Twitter because of its demand and data volume. An important common characteristic among microblogging services is its actual-time environment[2]. Although site users typically bring up to date their sites once every several days, Twitter users write down tweets many times in a single day. Users can know what other users are doing and often what is their opinion about now, users often return to the site and review to see what other people are doing. The large number of update results in few reports relevant to events. They include social events such as parties, baseball sports event, and presidential campaign. They also include terrible events such as storm, fire, traffic jam, riot, heavy rain, and earthquakes.

III. PROBLEM DEFINITION

We empirically compare the content of Twitter with a classical news medium, New York Times, using unsupervised topic modelling, but the content analysis on Twitter has not been well studied [1]. So to overcome this problem we use a Twitter-LDA[3] model to discover topics from a representative sample of the entire Twitter.

IV. PROPOSED APPROACH

In the proposed system we use the Twitter-LDA model for the content analysis on the twitter. In Twitter all the tweets of a user as a single document. In fact this treatment can be regarded as an application of the author-topic model to tweets, where each tweet) has a single author[1]. However, this action does not exploit the following important consideration: A single tweet is usually about a single topic. When writing a tweet, a user first chooses topic based on her topic distribution. Then the user chooses some particular words one by one based on the chosen topic or the background model. In this twitter LDA First we identified the topics of information which is related to the user and then we identify the the background words of the user selected particular topic and we search for noisy words and eliminate noisy and background words and later we can able to analysis the tweets of any particular user for selected topic[3].

A. Flow Diagram

![Flow Diagram of testing and Training](image)

**Figure 1:** Flow Diagram of testing and Training

In the above flow diagram First we enter the input text which contains the tweets[1] and retweets of the topics .Then it tokenizes the words in each topic. After the tokenization of the words it identifies the background words in each topic and then identifies the noisy words and eliminates them. Then it analyses the words in each topic. Then it repeats the iterations and finally analyses the each topic.

B. Algorithm

**Step 1:** Draw $p^B \sim \text{Dir}(l)$, $\pi \sim \text{Dir}(o)$

**Step 2:** For each topic $u = 1, \ldots, U$ ,
(a) draw $p^d \sim \text{Dir}(l)$

**Step 3:** For each user $m = 1, \ldots, M$ ,
(a) draw $q^m \sim \text{Dir}(s)$
(b) for each tweet $t = 1, \ldots, k_m$
   i. draw $x_{m,t} \sim \text{Multi}(q^m)$
   ii. for each word $k = 1, \ldots, k_{m,t}$
A. draw $w_{m,t,k} \sim \text{Multi}(r)$
B. draw $z_{m,t,k} \sim \text{Multi}(p^B)$
if $w_{m,t,k} = 0$ and $x_{m,t,k} \sim \text{Multi}(p_m)$
if $w_{m,t,k} = 1$.

V. IMPLEMENTATION

Testing is a process of revealing the errors in the program. It is the major feature measures the employed during software improvement, during testing, the program is execute with a set of test cases and the output of the program for the test cases is evaluate to find out that if the program is performing as it is expected to perform.

VI. RESULT AND DISCUSSION

Figure 2: The processes of Iteration

We open the net beans and upload our project into the net beans. Then after adding library files we run TweeterLDAmain.java file. Then it repeats 100 iterations after completion of iteration it will identify the background words, unique words, the analysis of the particular word is done.

Figure 3: Each word in each topic

The word probability of each word in topics is identified.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

In this project empirically compare the content of Twitter with a typical classical news medium, New York Times, concentrates on the differences between these two[1]. We developed a new Twitter-LDA model that is designed for short tweets and showed its effectiveness compared with existing models. We introduced the concepts of topic categories and topic types to facilitate our study of the topical differences between Twitter and traditional news media[1]. Our empirical relation confirmed some previous observations and also revealed some new findings.

In particular, we find that Twitter[2] can be a good source of entity orient topic that has low coverage in traditional news media. In the future, we will study how to review and visualize Twitter content in a systematic way. Our method of associating tweets with different categories and type may also help out visualization of Twitter content.

B. Future work

In Future we can extend this project with respect of LSA and LDA comparison. Moreover LDA is more powerful than existed LSA technique for Segmentation of Twitter Timeline.
VIII. REFERENCES


