ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Tweets Mining: Knowledge from the Social Web

G. Thiyagarajan[1], S.A.K. Jainulabudeen[2]

Systems Engineer, Infosys Limited, Chennai, India[1]

Assistant Professor, Department of CSE, Panimalar Engineering College, Chennai, India[2]

ABSTRACT: In the recent years, Social web mining has gained significant attention with case of it's an interactive platform via which individual of communities creates and shows user generated content and set of social relations that link people through the world wide web. Social Media becoming most powerful tool for information exchange has not only consumed information but also share and discusses information about aspects of their interest. Information retrieval and Text Mining have gained greater momentum in the recent past. Hence there is a need to mine the social media and generate useful knowledge based on identify interesting pattern from the user. The advantage of social media is the freedom of expressing their thoughts in text without following traditional language grammar's eventually this becomes the challenge for mining the social media. Moreover the volume of information is too huge and dynamic. The objective of this proposed work is to mine the social media, in our case twitter. The challenge involved is to understanding the user behavior and generating grammar rules pertaining to the tweet's language we also need to place the proximity of grammar used. The contribution of our work providing information retrieval tools with visual support. By applying the proposed algorithm, a study can be made on user behavior (Tweeters), fact analysis on the context of tweet and to identify the effective tweeters.

KEYWORDS: Text Mining, Topic Model, Latent Dirichlet Allocation.

I. INTRODUCTION

In the current scenario, texts play a vital role in emerging applications in web such as micro-blogging, advertisements in online etc., The searching technique in advertisements includes only few keywords or sentences. Micro-blogging service as Twitter restricts the length of message for a user and less than 140 characters. The techniques for mining short texts are important to application domains. In the past several techniques have been proposed for mining the information from large text in blogs or any other news, when we apply these technique in short texts will lead to poor results. When comparing short text with long texts short text suffered two main difficulties caused by their highly sparse representations: keyword that has not sufficient, content information is not much populated.

Online social streams such as Twitter Streams, Google Buzz, Facebook news feeds, have emerged as important online information. Thousands & Millions of users are reading status, chatting with their friends, sharing information and engage in useful tips. However, some of the conversations in social media are not interesting to read. To avoid this situation, the users with boring conversations are removed and selectively displaying good tweeters pattern to user. The Twitter uses filter correspondence between users (simple rules). Till date many research has been carried out to find the interestingness pattern and prevalence in social streams.

Our research has four high level research questions:

> **RQ1:** How do tweeters may prefer on their conversation? Will it differs? Do their preferences correlate with other words and co-occurrences their usage purposes, i.e., whether they tweeters uses as an information medium else social medium?

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

RQ2: How efficacy can make on different algorithms in selecting interestingness pattern/conversations for tweeters behavior?

RQ3: Do usage purposes of Twitter and preferences of topics affect algorithm performance?

RQ4: Do ensure the latent topics and tweeters behavior in their patterns?

To answer these questions, as explored the design of system that would give potentially interestingness patterns on Twitter, here patterns is also meant conversations. Because of popularity Twitter is chosen and over many other social stream platforms. Specifically over Facebook because Twitter's open APIs will provide us greater flexibility in data gathering, moreover designing the patterns and selection of algorithms along with deploying these algorithms.

Tweeters rated the interestingness of pattern will produced by different algorithms. The study allows by comparing the algorithm which evolves the performances across different Tweeters. The rest of the paper is as follows. First, discuss how existing research relates to our work. Then provide an overview of Twitter and intricacies of Tweeters patterns. Then per-processing techniques is carried out, followed by describe the design of system, and then detail of our studies and the results. Later conclusion as made with discussions of our findings and design implications.

II. RELATED WORK

In the recent years, social streams attracted the research community in performing their amount in social web mining. For multitude purpose, social streams have ceeb demonstrated in research. Few papers related to social streams. Java et al [2] focuses on the usage of daily chatting conversations in Twitter as Tweeters, sharing information and reporting events (news). Naaman et al [3] coded twitter messages manually and suggested users of Twitter post for both informational and social purposes. In enterprise social stream site Yammer, Zhang et al [4], analyzed the purposes of web site and found it gives different preferences as a result. This also results in large part of activity on yammer & finds difficulty in relevant content. Conversational aspect of social streams is analyzed in several works, Boyd et al [5] discussed on conversational usage of retweets in Twitter, it reverals different variety in practical Twitter conversations. Honeycutt et al [6] investigated conversations on Twitter, to explore and recommend conversations as an important task in addressing this challenge. In the recent works, they discussed the problem of addressing overload information in social streams by utilizing topic as a key factor. Ramage et al [7] applied LDA (Latent Dirichlet Allocation) to find messages for reading and characterize topics in Twitter. Bernstein et al [8] utilizes topic-based browsing interface of Twitter using search engines. These works gives prior research on Topic modeling and information retrieval, this also includes salton et al [9] and Blei et al [10]. In user preferences, the potential diversity have indicated filtering and recommendation in social streams. Chen et al[11] incorporated news URLs in Twitter by social voting, topic relevance and suggests that single recommender may not satisfy users' needs that differs. The solution may be indicated as personalization. The exploration in interesting conversations recommends the existing body of research in following aspects:

1) While previous research not much concentrate on pre-processing factor, here included spell correction of tweets.

2) As previous research works focuses on single message, the proposed work focuses on conversations in a coherent thread of multiple messages.

3) While previous works as focused on information gathering in social streams and news finding instead the proposed work focuses on tweeters pattern in conversations and facing diversity in usage purposes and preferences. In this work, some factors are explored in tweeters behavior paths. Topic relevance, thread length, tie-strength, latent factors, inference rule on sampling tweets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

III. SYSTEM OVERVIEW

The architecture which shows the method to analysis the tweet's , classify the tweeter's tweets and cluster with the hidden topic through the fact classifier, it also does offer to understanding and tweeter's behavior and classification of tweeter's can be involved further by examining other techniques of twitter communication.

Repository of tweets contains the dataset, which is taken from stanford university and also streamed through Twitter API. Basically dataset contains the raw information which has not applied directly for mining, hence the dataset has to cleanse based on the data structure. Preprocessing task contains the technique which has identified the latent pattern. Through that concept provided by the interestingness and applied in available data. Tweeter's pattern has been modeled. (Restricted by the chosen hash tag). By feeding the data and populated the data extractor and classifier. Through the simple classification, the identification has made on the fact engine.

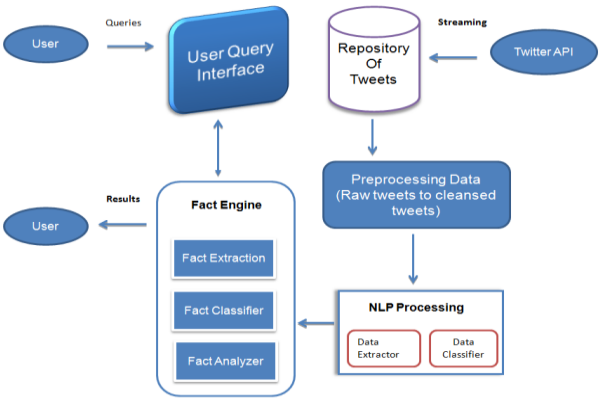

Fig. 1 System Design

As Figure 1, through the fact engine, it's been classify the processed facts and which including the important for tracking the pattern flow within the harvested tweets. As the model, interested in finding the flow of patterns (which in this case, is (retweet flow in this case), the required model involves a method to construct a record of retweets.

A. *PreProcessing:*

Basically Twitter comments are not in form of formal english structure. Since it contains user comments, we couldn't able to apply the algorithm/techniques directly as we apply like in text mining, raw tweets has to cleaned based on the data structure what we like to apply. Here we have done emotion and punctuations handling, which is not necessary to carried out or to analysis the tweeter's nature, stemming to get a root word from a base word, stop word removal, spell correction for expressive words.

B. *Emotions and Punctuations Handling*:

For exacting the pattern by making the pre-processing task much good for effectiveness. We used list of emotions is generally commented in tweet's and also from the Wikipedia list of emotions. Hand tagged into basic five emotion tags. We have replaced the corresponding tagger in the tweet's comments. For example, if emotions was happy or extremely happy then that tagger will replace with 'AhappyA' and sad or extremely sad with 'AsadA'. We append 'A' and prepend 'A' with replacement tags, because the tweet could not mix with such tag replacement for the purpose of prevention. Else we can remove such emotion and punctuations, it don't use such tweeter's behavior. If the sentiment analysis can be made on it, we could able score such tweets like exclamation, question, positive and negative.

C. *Stemming:*

Stemming is a process of getting/deriving a root word from a form of base word, its most important feature for gathering the latent topic and supported nowadays for indexing and search systems. The idea behind the stemming, it has to improve recall by automation of handling the specific word ending by reducing the words to the word root, at the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

time performing latent topic identification. We used porter stemmer, will restrict only 7 forms of suffix removal, tense identification made by using WordNet.

D. *Stop Word Removal:*

The tremendous amount of textual information from social media twitter and it requires quite extensive level of data preparation and deep analysis to extract useful interestingness pattern. However, it has made and duce to the assumptions of various technique as well as amount of data to be large, then data quality has to be low. To improve the quality of data, many author been proposed different techniques / to extract an effective stop word list methods. Stop word play an important role and to improve the retrieval of information. Thus adding more ambiguity in the model formation and stop words don't carry any useful information and thus are no use of us for processing. We have created a list of stop words to implement to removal in the tweets such as he, she, a, the etc., and ignore them while parsing. We also discard words which are of length ≤ 2 for the tweets while parsing

E. *Spell Correction:*

Tweets are tweeted in a random manner, which hasn't focus to correct structure/format and spelling. Spell correction is a most important part in behavior analysis of populated content. Tweeters type certain characters which are arbitrary number of times to emphasis on those corresponding tweets. We use WordNet and spell correction algorithm form (Bora, 2012). In the algorithm they replace a word with any co-occurrence characters (repeation) more than twice with two words. For example the word 'looooovvveee' is replaced with 8 words 'lovvveee', 'looooveee', 'loveee', 'loovee', 'love', and so on. Another form of spelling mistakes occur while tweeting is because of skipping some of characters from the spelling like "university" is generally written as "univ". Such types of spelling mistakes are not currently handled by our system.

IV. APPLYING LDA TO OBTAIN LATENT TOPICS FOR TWEETERS BEHAVIOR

In this section, we have briefly described LDA. Because studying how social mining and textual information associated with entities in the tweeter's behavior can be modeled for insight.

A. *Topic Models:*

Human Languages involve hidden topics words topic modeling deals with assuming latent topics in usage of words, when a searcher uses animal as a query and if author used the word mammal in the document, assuming both might involve same concept (topic) lion by topic modeling. This paves topic model to observe words from unobservable words. Probabilistic generative model to topic model is introduced by PLSI. Equation (1) represents its document generation process based on the probabilistic generative model: by [14], it has identify the distribution of P(d,w) further it is computed.

$$P(d,w) = P(d)P(w|d) = P(d) \sum P(w|z)P(z|d). \quad (1)$$

P(d,w) is the probability of observing a word w in a document d and can be decomposed into the multiplication of P(d)P(w|d) the probability distribution of documents, and P(w|d), the probability distribution of words given in a document. This equation is best described for a word selection in a document, where we select a document first then a word in that document. This selection is iterated multiple times so that we can generate a document and eventually a whole document corpus. By assuming that there is a topic latent z, we can rewrite the equation above with the multiplication of P(w|z), the probability distribution of words given a topic, and P(z|d), the probability distribution of topics in a given document. This equation describes an additional topic by adding selection step between the document selection step and the word selection step. As there are various multiple latent topics where a word may come from, we sum up the term for multiplication over a set of all the independent topics Z. PLSI and other probabilistic topic models support multiple memberships using the probabilities P(w|z) and P(z|d).

For example, if $P(w_{mammal}|z_{animal}) > P(w_{lion}|z_{animal})$, the word mammal is more closely related to the topic animal than the word lion, though they are all related to the topic animal. In this way, we can measure the strength of association between a word w and a topic z by the probability P(w|z). Similarly P(z|d) measures the strength of association between a topic z and a document d. By [14], I have drawn the L value based on the probability factor.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Equation (2) represents the log-likelihood function of PLSI:

$$L = \log[P(d,w)^{n(d,w)}] \quad \ldots\ldots\ldots\ldots\ldots \quad (2)$$

where D and W denote a set of all d and w respectively, and n(d|w) denotes the term frequency in a document (i.e., the number of times w occurred in d). By maximizing the log-likelihood function L, we can maximize the probability to observe the entire corpus and accordingly estimate the P(w|z) and P(z|d) that most likely satisfy Equation (1).

V. LATENT DIRICHLET ALLOCATION

Though PLSI is equipped with a sound probabilistic generative model and a statistical inference method, it suffers from the over fitting problem and does not cope well with unobserved words. To solve this problem, Blei et al. [10] introduced Dirichlet priors α and β to PLSI, to constrain P(z|d) and P(w|z), respectively, here α is a vector of dimension |Z|, the number of topics examined, and each element in α is a prior for a corresponding element in P(z|d). Thus, a higher α_i implies that the topic z_i appears more frequently than other topics in a corpus. Similarly, β is referred as vector of dimension |W|, the number of words examined, and each element in β is a prior for a corresponding element in P(w|z). Thus, a higher β_j implies that the word w_j appears more frequently than other words in the corpus. As a conjugate indicated prior for the multinomial distribution, the statistical inference can be simplified by Dirichlet Distribution. Dirichlet priors α and β can be placed on the multinomial distributions P(z|d) and P(w|z), those multinomial distributions are smoothed by the amount of α and β and become safe from the over fitting problem of PLSI. It is also known that PLSI emerges as a specific instance of LDA under Dirichlet priors [12, 13].

Topics	College	Social Media	Mobile
Top 3 Latent Topics	Student	Social	Iphone
	Assignment	Media	Features
	Assessment	Internet	Supports

Table 1: LDA Sample Topics with 3 Top Representative Words

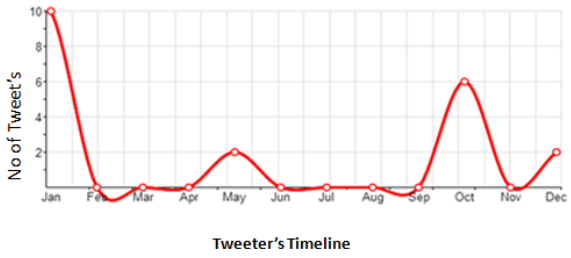

Fig. 2 Tweeter Behavior with Timeline

As Figure 2, represents a simple visualization of tweeter's behavior, it just computes the timeline of individual tweeter's tweet.

VI. EXPERIMENT & RESULTS

As examine the proposed model from four different perspectives: the perplexity of content generation, the performance of predicting retweets and the quality of generated latent topics.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

A. *Perplexity of Content Generation:*

Dataset been streamed based on the timeline, have taken from stanford university. Study on perplexity means the nature of tweeters and tweets which held out on content generation which would based on either timeline or interest of tweeters.

B. *Performance of Tweeting Nature:*

Potentially the behavior of tweets and purpose of topic modeling is to help tweeters find interestingness information from the overwhelming information streams. In the context of Twitter, retweet is the most important signal of tweeters interest, as twitter user are prone to cast their favorite tweets to their followers. Moreover, the mechanism of behavioral nature of tweeters, retweets is a good standard to judge the performance of a topic model.

C. *Performance of Pre-processing Stages:*

By applying these pre-processing methods for improve the pattern extraction , because tweets is user comment simple text mining methods is not fair enough to do. Hence, by performing such pre-processing task, get better result by obtaining the latent topics.

D. *Quality of Latent Topics:*

The method to evaluate the performance of topic models is to obtain top words for the latent topics and evaluate them by experience. By design the content extractor an experiment to compare the latent topics by generating semi-supervides model and the proposed model. Specifically, by setting the number of topics to be 70 then since, amount of topics to be quite low and manually extract the same salient topics for both models. As a result, 11 latent topics are obtained by the content extractor, and the rest of latent topics are either meaningless topics else topics not relevant between two models.

VII. CONCLUSION AND FUTURE WORK

In this paper, a methodology is described that characterizes individual Tweeters by inferring their attributes. In existing methodologies which attempt to infer topics for tweeters with user's tweets or profile, the proposed approach infers by leveraging the freedom of the Tweeters, as focused on Lists of meta-data created by the tweeters. For constructing service for Twitter a inference methodology is proposed for who-is-who service can comprehensively infer an accurate set of attributed over a million Tweeters automatically which includes popular users too.

The main contributions of the present study – a methodology and a service to accurately infer topics related to Tweeters, mechanism to provide user behavior bath and identification of patterns, have a number of potential applications in building search services on Twitter, by providing better visualizing tool for active tweeters.

REFERENCES

1. X.-H. Phan, L.-M. Nguyen and S. Horiguchi. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections", ACM Proceeding of the 17th international conference on World Wide Web, WWW pp. 91-100, 2008.
2. Akshay Java, Xiadon Song, Tim Finin, and Belle Tseng,"Why we twitter: Understanding Micro blogging usage and communities", ACM Joint 9th WEBKDD and 1st SNA-KDD Workshop, pp. 56-65, 2007.
3. Naaman, M., Boase, J., and Lai, C, "Is it really about me? Message content in social awareness streams", ACM Conference on Computer Supported Cooperative Work (CSCW), 2010.
4. Jun Zhang, Yan Qu, Jane Cody, and Yuling Wu, "A Case Study of Micro-blogging in the Enterprise: Use, Value, and Related Issues", Proceedings of ACM Conference on Human Factors in Computing Systems (CHI), pp. 123-132, 2010.
5. Danah Boyd, Scott Golder, and Gilad Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.' Proceedings of ACM 43rd Hawaii International Conference on System Sciences (HICSS), pp.1-10, 2010.
6. Courtenay Honeycutt and Susan C. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter", Proceedings of ACM 42rd Hawaii International Conference on System Sciences (HICSS), pp.1-10, 2009.
7. Daniel Ramage, Susan Dumais and Dan Liebling, "Characterizing Microblogs with Topic Models", Proceedings of the Fourth International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Weblogs and Social Media (ICWSM), pp.130-137, 2010.
8. Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairan, Ed.H.Chi, "Eddi: Interactive topic-based browsing of social status streams", Proceedings of the 23rd annual ACM Symposium on User Interface Software and Technology (UIST), pp.303-312, 2010.
9. Gerard Salton and Christopher Buckley, "Term-weighting approaches in automatic text retrieval", International Journal on Information Processing and Management, Vol. 24, Issue 5, pp.513-523, 1988.
10. David M. Blei, Andrew Y. Ng, and Micheal I. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research Vol. 3, Issue 1, pp. 993-1022, 2003.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

11. Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, Ed Chi, "Short and tweet: Experiments on recommending content from information streams'. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.1185-1194, 2010.
12. Mark Girolami and Ata Kaban, "On an equivalence between PLSI and LDA", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.433-434, 2003.
13. Matthew D. Hoff man, David M. Blei, and Francis Bach, "Online Learning for Latent Dirichlet Allocation", Advances in Neural Information Processing Systems (NIPS) Vol. 23, pp. 856-864, 2010.
14. Youngchul Cha and Junghoo Cho, "Social-Network Analysis Using Topic Model", In 35th ACM Annual SIGIR Conference USA, pp.565-574, 2012.
15. Mark Guzdial and Jennifer Turns, "Effective Discussion through a Computer-Mediated Anchored Forum.", Journal of the Learning Sciences, Vol. 9, Issue 4, pp. 437–469, 2000
16. Eric Gilbert and Karrie Karrahalios, "Predicting Tie Strength with Social Media", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 211-220, 2009.

BIOGRAPHY

Thiyagarajan Gnanasigamani is a Systems Engineer currently working in Infosys Limited, Mahindra City, Chennai, India. He received his Bachelor of Engineering (B.E) degree in 2011 from Anna University, Chennai, and Master of Technology (M.Tech) degree in 2013 from B.S. Abdur Rahman Institute of Science and Technology, B.S. Abdur Rahman University, Chennai. His research interests are Data Mining, Image Processing, and Big Data Analytics.

Jainulabudeen S A K is Assistant Professor in Computer Science and Engineering Department, Panimalar Engineering College, Chennai, India. He received his Bachelor of Engineering (B.E) degree in 2010 from Anna University, Chennai, and Master of Technology (M.Tech) degree in 2013 from B.S. Abdur Rahman Institute of Science and Technology, B.S. Abdur Rahman University, Chennai. His research interests are Security issues in Android, Image Processing, and Cloud Computing.