RESEARCH ARTICLE OPEN ACCESS

LATENT DIRICHILET ALLOCATION AND NAIVE BAYES CLASIFICATION BASED TWITTER DATA'S HIERARCHICAL TOPIC MODELING

Mr. C. Mani, M C A, M.E, M Phil*, Mr. P.Harish, M C A **

*(Associate Professor, Department of Computer Science and Engineering,
Nandha Engineering College (Autonomous),
Erode, Tamil Nadu, India
Email: cmanimca@gmail.com)

** (Final MCA, Department of Computer Applications,
Nandha Engineering College (Autonomous),
Erode, Tamil Nadu, India
Email: hariannan98@gmail.com)

----------**********----------

Abstract:

Twitter data accumulated so far make it in all likelihood to find out the distribution and go with the flow of mass tastes and opinions, which greatly help in product recommendation, target advertising and marketing and so on. This challenge proposes a subject model known **as twitter hierarchical Latent Dirichlet Allocation (thLDA)**. Based on hierarchical latent Dirichlet allocation, thLDA aims to automatically mine the hierarchical measurement of tweets' subjects, which may be further employed for text Here, the phrases present in every of the topics with extra importance are extracted out and their beta value is found out. Furthermore, thLDA analyzes the relationships of words in tweets to get a more powerful measurement. Extensive experiments are carried out on Twitter information and the effectiveness of thLDA is evaluated. The consequences show that it outperforms properly amongst other current topic fashions in mining. This venture improves the topics mining amongst two subjects as well as three subjects. In addition, conditional chance is finished for Naïve Bayes Classification of important terms in the given statistics set so that the whole words possibilities in all the categories are found out and displayed. The task is designed the usage of R Studio 1.0. The coding language used is R 3.4.4.

Keywords — LDA, Navive Bayes, OLAP, Data Mining

----------**********----------

I. INTRODUCTION

Data mining or expertise discovery is the computer-assisted manner of digging thru and studying large units of facts and then extracting which means of the information. Data mining device count on behaviors and destiny trends, allowing corporations to make proactive, information-pushed decisions. Data mining system can solution commercial business enterprise questions that traditionally had been too time eating to resolve. They scour databases for hidden patterns, locating predictive facts that experts can also miss because it lies out of doors their expectations.

Data mining derives its name from the similarities between seeking out valuable statistics in a big database and mining a mountain for a vein

of precious ore. Both methods require both sifting via an immense amount of material, and intelligently probing it to locate in which the value resides. Although statistics mining is still in its infancy, groups in a wide variety of industries - which include retail, finance, health care, manufacturing transportation, and aerospace - are already using records mining equipment and techniques to take gain of historical data.

During the beyond few years, Twitter has emerge as increasingly popular as a rising social platform for messaging and conversation among individuals. The massive portions of Twitter statistics accumulated thus far make it possible to find out the distribution and float of mass tastes and opinions, which greatly assist in product recommendation, target advertising and so on. On the alternative hand, OLAP, or on-line analytical processing, lets in assessment to interactively view statistics from all factors in layered granularities, which has already been hooked up especially useful for enterprise intelligence.

As a fashionable unsupervised topic model, the Latent Dirichlet Allocation (LDA) model is green at statistically analyzing textual statistics for the underlying subjects. This project proposed a LDA-primarily based completely model, called MS-LDA, to stumble on the hidden layered hobbies from the Twitter facts. As the extension of LDA, MS-LDA integrated tweets and the social relationships amongst tweeters. Nevertheless, the primitive LDA version can only mine monolayer subjects, in place of the hierarchical ones which OLAP requires. On the other hand, as an unmanaged hierarchical topic version, hLDA can acquire the sibling-sibling relationships between subjects and can set up the topics proper into a hierarchical tree automatically. In fact, Twitter records include sufficient social behavioral data about tweeters, including mentioning, retweeting and following. In addition, there exist a few semantic relationships some of the terms in tweets, which may moreover have an effect on the effectiveness of the modelling method. In different words, to efficaciously discover the hidden layers of topics from Twitter facts for building the hierarchical size for OLAP, we want to endorse a new topic model which leverages the trends of Twitter in its modelling technique.

Unfortunately, OLAP strategies are successful in handling cube information which might be primarily based and formalized, however face issues in processing textual content together with Twitter records. To effectively follow OLAP techniques to Twitter, it is essential to mine the hidden consultant dimensions from it's extensive.

II. LITERATURE REVIEW

Dongjin Y u et al [1] exploring the distributions and correlations from Twitter records helps accurate customized recommendations. Online Analytical Processing, or OLAP, presents an intuitive form that is suitable for exploring Twitter records. Unfortunately, the conventional OLAP tactics can only cope with structured facts, not unstructured textual information like tweets. The key to applying OLAP to twitter records is to mine and construct a size hierarchy of tweeter interests. However, the modern strategies can extract tweeter hobbies from Twitter facts on a single level, but fail to gain a hierarchy of tweeter hobbies with special granularities. To cope with this problem, they proposed a LDA-based model, referred to as MS-LDA, which mixes tweeters' social relationships and tweets to extract and construct the tweeters' hobby dimension Such a size hierarchy can be further employed to use OLAP strategies to Twitter information. In addition, we rent Word2vec to acquire the linguistic similarity of phrases in tweets, to enhance its effectiveness. The large experiments demonstrate that our method can effectively extract the size hierarchy of tweeters' hobbies for multidimensional analysis.

Maha Azabou, Kaïs Khrouf et al [2] describe the Semantic measurement inside the Diamond Model is represented and integrated as follows:

i) Connection of the Version parameter with the Semantic useful resource parameter of the semantic dimension.

ii) Linking the parameters of the same old dimensions, wealthy in semantic text, with the Concept parameter of the semantic dimension. For example, the Title attribute of the D-Movie

measurement might be linked to the Concept parameter.

Ying Quanzhi Li et al [3] present numerous tweet subject matter class techniques by way of exploiting different varieties of data: tweet text, tweet text plus entity know-how base, word embeddings derived from tweet text, distributed representations of tweets, and topical word. The word embedding, topical word embedding and sentence representation models are generated from billions of phrases from tweets without supervision. To the excellent of their know-how, that is the first look at of applying allotted language representations to tweet topic category task.

Xiong Liu, Kaizhi Tang [4] talk a text cube method to reading different sorts of human, social and cultural behavior (HSCB) embedded within the Twitter stream. Text dice is a new manner to prepare information (e.g., Twitter text) in a couple of dimensions and multiple hierarchies for efficient records question and visualization. With the HSCB measures described in a dice, users are capable of view statistical reports and carry out on line analytical processing. Along with viewing and analyzing Twitter text the use of cubes and charts, they have got also added the functionality to reveal the contents of the cube on a warmness map. The degree of opacity is right away proportional to the charge of the behavioral, social or cultural measure. This shape of map allows the analyst to focus interest on hotspots of difficulty in a vicinity of interest. In addition, the text dice architecture enables the development of information mining models the use of the information taken from cubes. They supplied sever a case studies to demonstrate the text dice technique, together with public sentiment in a U.S. town and political sentiment within the Arab Spring.

HNafees Ur Rehman, et al [5] proposes an evaluation platform for the massive records generated from social sports online. It advocated using the mature information warehousing technology coupled with data mining to enable efficient and multidimensional analysis of social network information from the newly found perspectives. These located perspectives are derived out of the underlying dataset using lots of traditional and expertise discovery methods. The dataset is semantically enriched via extraction and identification of entities, events, language, sentiment, subjects etc from the user messages.

III. MODULES

PREPROCESSING

In the pre-processing stage, the system first queries all tweets from the database that fall inside date d1 and date d2. For the tweets facts source, the set of terms aren't the tweets' keywords, but all precise and relevant terms. First, the language of every queried tweet is identified, dismissing any tweet that isn't always in English. From the closing tweets, all phrases that seem in a forestall phrase list or which can be much less than 3 characters in duration are eliminated. To do away with terms that are not applicable, Unicode characters and punctuators are removed. The terms are then brought as words used for locating beta value.

SYNONYM WORD REPLACEMENT

In this module synonym phrases are taken from a text document and loaded within the vector. Then all the tweets posts are taken checked for words present within the synonym words listing and replaced with the ones terms. This assists in semantic similarity fulfilment for two tweets those containing two phrases with same meaning.

LATENT DIRICHLET ALLOCATION MODEL

In this module, topic models library is used, LDA feature is called, to discover a) Per-Document-Per-Word Probabilities and b) Per-Document-Per-Topic Probabilities. The feature tidy (ap_lda, matrix = "beta") is used to get the values for a). Ap_documents.

NAÏVE BAYES CLASSIFICATION MODEL

In this module, dataset is selected as paragraphs as first column with class as second. Then conditional probabilities which are the basis of Naïve Bayes category are observed out for all the phrases for every class.

IV. EXISTING WORKS

The existing device consists of Data acquisition: Obtain tweeters' profiles, tweets and social relationships via the Twitter package in R Data pre-processing: Remove the fast phrases (the maximum common, short function words along with the, is, at, which, and on) and the web hyperlinks and leave simplest nouns and verbs in the unstructured tweets. Then making use of LDA feature to the ones twitters records using 'topic models' bundle. Then the significance of phrases in each subject matter is located out and displayed as beta values. This beta cost which is 'per record per word probability' is graphically displayed as bar plot.

DRAWBACKS OF EXISTING SYSTEM

- Semantic similarity isn't considered, i.e., words which are having identical that means are treated as distinctive words in existing system.
- Beta cost is produced for 2 phrases even with identical which means.
- LDA if implemented with k value 3, then Topic clever, Chapter clever importance among the phrases should be determined out.

V. PROPOSED SYSTEM

Like existing machine, the proposed system also consists of Data acquisition: Obtain tweeters' profiles, tweets and social relationships through the Twitter package deal in R. Data pre-processing: Remove the fast phrases (the most common, short feature phrases along with the, is, at, which, and on) and the internet hyperlinks and leave only nouns and verbs in the unstructured In addition, synonym phrase replacement is likewise finished so that phrases with identical that means even they may be different phrases are replaced as first phrase. Then applying LDA characteristic to those twitter records using 'topic models' bundle but with k cost 3. Then the importance of words in every topic is determined out and displayed as beta values. This beta price which is 'per report per phrase probability' is graphically displayed as bar plot. Conditional probability (Naïve Bayes classification) is also applied.

ADVANTAGES

- Semantic similarity is implemented, i.e., words which are having equal meaning are renamed with first phrase.
- Beta price is produced for replaced word with same meaning.
- LDA is carried out with k cost 3, and so Topic sensible, Chapter wise importance most of the phrases are located out.
- Conditional chance (Naïve Bayes classification) is also carried out which objectives in finding the probability percentage among all categories.

VI. CONCLUSION

Classification is implemented to automatically examine the significance of a word in particular category, in order that its importance amongst all categories is obtained. The absolute price of the text represents the influential electricity and the sign of the text denotes its emotional polarity. This assignment positioned forward a singular hierarchical topic version, i.e., thLDA, which is implemented to mine the measurement hierarchy of tweets' subjects from a massive quantity of unstructured Twitter data. The consequences show that thLDA has a better recognition impact than the other models. When thinking about how social relationships impact on the hierarchical topic version, this assignment attention only topic clever beta price locating so that importance of terms among subjects. In addition, to improve the model effectiveness, conditional chance finding is completed for all given classes to locate the

significance of phrases among all classes. In future, this mission may awareness on how the social effect elements which include hash tags and word semantic similarity affect the experimental outcomes separately, and whether it's far viable to enhance the model using such hash tags.

REFERENCES

[1] D. Yu, J. Sun, Y. Wu, Z. Ni, and Y. Li, ‘‘Discovering hidden interests from Twitter for multidimensional analysis,’’ in Proc. 29th Int. Conf. Softw. Eng. Knowl. Eng., 2017, pp. 329–334.

[2] M. Azabou, K. Khrouf, J. Feki, C. Soulé-Dupuy, and N. Vallès, ‘‘A novel multidimensional model for the OLAP on documents: Modeling, generation and implementation,’’ in Proc. Int. Conf. Model Data Eng. Cham, Switzerland: Springer, 2014, pp. 258–272.

[3] Q. Li, S. Shah, X. Liu, A. Nourbakhsh, and R. Fang, ‘‘Tweet topic classification using distributed language representations,’’ in Proc. IEEE/WIC/ACM Int. Conf. Web In tell. (WI), Oct. 2016, pp. 81–88.

[4] X. Liu et al., ‘‘A text cube approach to human, social and cultural behavior in the Twitter stream,’’ in Proc. Int. Conf. Social Comput., Behav.-Cultural Modeling, Predict. Berlin, Germany: Springer, 2013, pp. 321–330.

[5] N. U. Rehman, A. Weiler, and M. H. Scholl, ‘‘OLAPing social media: The case of Twitter,’’ in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, Aug. 2013, pp. 1139–1146.

[6] X. Pu, M. A. Chatti, H. Thues, and U. Schroeder, ‘‘Wiki-LDA: A mixed method approach for effective interest mining on Twitter data,’’ in Proc. 8th Int. Conf. Comput. Supported Edu. Rome, Italy: ScitePress-Science and Technology Publications, 2016, pp. 426–433.

[7] A. M. Dai and A. J. Storkey, ‘‘the supervised hierarchical Dirichlet process,’’ IEEE Trans. Pattern Anal. Mach. In tell vol. 37, no. 2, pp. 243–255, Feb. 2015.

[8] J.-T. Chien, ‘‘Hierarchical Pitman–Yor–Dirichlet language model,’’ IEEE/ACM Trans. Audio, Speech, Language Process, vol. 23, no. 8, pp. 1259–1272, Aug. 2015.

[9] D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones, ‘‘Word embedding based generalized language model for information retrieval,’’ in Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2015, pp. 795–798.

[10] Haixia Liu, “Sentiment Analysis of Citations Using Word2vec”, Computation and Language (cs.CL), arXiv: 1704.00177v1 [cs.CL], April 2017