ISSN:0975-9646

Decision making using Multi Inference-LDA Algorithm

Kanchana J
Research Scholar: School of Computer Science
Mahatma Gandhi University
Kottayam
Kerala, India

Gladston Raj S
Head, Department of Computer Science
Government College, Nedumangad
Thiruvanathapuram
Kerala, India

***Abstract*—Need of data mining is getting increased day by day, when large organizations rely on data to make business decisions. The core concept of mining rely with a digging conceptual terms from big data. Financial decision mainly based on such concepts can be carried out by modern algorithms like LDA (Latent Dirichlet Allocation). The LDA based clustering models easily classifies business data into business decisions. The various factors considered for digging data using LDA may lead to the development of a new MI-LDA algorithm. In the area of financial decision making, an important factor is feeling insecure in relation to financial risks. This paper presents a model for financial decision making using Multi-Inference LDA (MI-LDA). To evaluate this model, a number of simulation experiments have been performed, which explains the model's ability to make financial decisions of different types of financial situations. The evaluation of new MI-LDA model with the existing LDA model is also made.**

Keywords—LDA, MI LDA, Corpus

I. INTRODUCTION

Data mining based decision making are always been an attraction for researchers and experts to bring better output to the relevant fields. Reading through long text and quotes are difficult, when the number of users in data exchange scenario is increased. Learning big data, clustering information etc. play an important role in decision making based on data mining. The decision making process using the LDA approach consists of two stages. The first one is to make a model using LDA and second one is changing the model into a multi inference LDA model. In previous studies, text clustering[3] was used in the process of partitioning a particular collection of texts into subgroups including content based similar ones. An effective text clustering for sorting large or massive documents help the users to access and organize text documents. Document clustering is one of the most important techniques for organizing documents in an unsupervised manner. The power of clustering may be replaced with the probabilistic model using variation inference.

A. LDA(Latent Dirichlet Allocation)

LDA represents documents as mixtures of topics that split out words with certain probabilities. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. In other words LDA is a generative probabilistic model for discrete data such as bags of words from text documents.

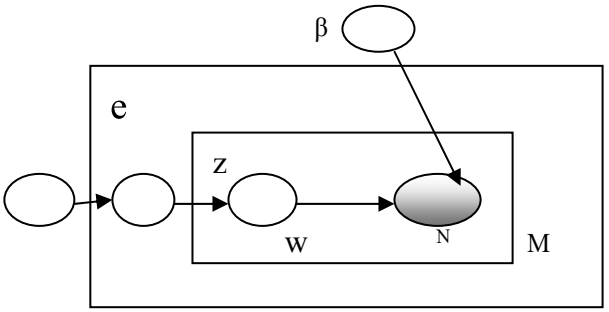

Figure1: Graphical model representation of LDA

The boxes are "plates". The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. The LDA model is represented as a probabilistic graphical model. There are three levels to the LDA representation. The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variable Θ*d is* document-level variable, sampled once per document. And the variables *zdn* and *wdn* are word-level variables and they are sampled once for each word in each document. A classical clustering model would involve a two-level model where as LDA involves three levels. In the case of classical clustering model a document can be associated only with one topic where as in LDA model, documents can be associated with multiple topics.

B. ML-LDA

The general idea behind classical LDA is to generate topics based on variation inference. LDA does not consider the conventional methods involving word frequencies. The TF-IDF (Term Frequency - Inverse Document Frequency) based inference just generates output topics based on text similarity. Other factors like entropy, semantic strength and word polarity etc are not concerned in present LDA based system. Hence a new model is proposed based on the above dimensions. This model may generate more precise results that may suit for decision making.

The objective for MI-LDA is to find a boundary or boundaries that generate topic clusters of data. MI-LDA does this by reading from a basic corpus file and a parameters set. Rather than SVM classifiers, LDA calculates probability distributions and marginal distribution of the document terms.

II. RELATED WORK

The L*atent Dirichlet allocation (LDA)* model is important to emphasize that an assumption of exchangeability is not equivalent to an assumption that, the random variables are independent and identically distributed. But they are conditionally independent and identically distributed. Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. An empirical evaluation of LDA is used in problem domains like document modeling, document classification, and collaborative filtering [1]. The narrow concept of data mining is one critical step of knowledge discovery, an important procedure of drawing useful schemas or building model [7]. Financial analysis is to find out the economy meaning of accounting data in order to understand the running performance and financial position of one company, which helps investors and creditors with their decision making [4]. An analysis model of financial statements based on data mining methods, such as clustering, association rules and decision making tree work together to step by step go into deeper analysis of existing financial statements, during which a annual assets structure statement is worked out. In order to go deep and completely into analyzing the financial data of companies, They use OLAP tool of data mining, build super-cubes according to their own needs, examine or analyze data from multiple dimensions[4].

The processes involved in topic mining and classification consists of term extraction, dimensionality reduction, and feature-selection and so on. Using these, data sets can be trained and the classification model can be created based on the algorithms. Decision support systems are new computerized applications that act as a support system for supporting the organizational and business decision makers in the activities going on in their business and other industries[5]. A DBMS serves as a data bank for the DSS. It stores large quantities of data that are relevant to the class of problems for which the DSS has been designed and provides logical data structures with which the users interact. Multiple criteria decision making is a sub-discipline of operations research that explicitly considers multiple criteria in decision-making environments [5]. Financial Decision Support System (DSS) is the core of corporate management information system, and is also the ultimate goal of the development of accounting information systems.[8]. In the Financial field, text mining is widely used to create an index of economic policy uncertainty [9]. Financial Decision Support System (FDSS) is a decision support subsystem, which establishes based on computerized accounting, information technology and the network, and with the characteristics of human-computer interaction [8].

The function of Financial forecast system which is a part of FDSS, is divided into two aspects: one is to make use of existing financial data to forecast the company's future financial condition and the second one is to make use of expert experience and expertise to forecast a financial topic[8].

III. PROPOSED SYSTEM

The concept of LDA with Multiple Inference is utilized in modeling a financial decision making system. The major input of the system is the press releases by financial organization as well as their stock related statements. The text matter may be processed by different methods like noise and word removal, stemming etc. The resultant document term matrix vector will be subjected to LDA for generating concepts as financial decision making inputs. The multi-factor data generated after LDA will also undergo ML-LDA. The topics generate decisions with relevance on efficiencies and factors considered. The model will provide a semantically strong inference mining and the resultant topic clusters are used for decision making needs.

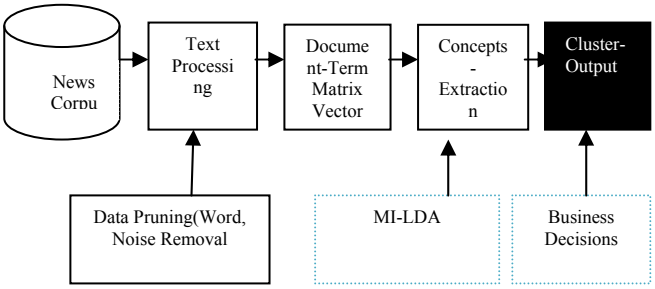

Figure 2 . Research model to study text processing using the MILatent Dirichlet Allocation for concept extraction.

IV RESEARCH METHODOLOGY

This section presents the underlying research methodology as specified in Figure 2. The Section A introduces the data pruning techniques to transform the adhoc announcements into a machine-readable format. We then give an overview on the Latent Dirichlet Allocation method in Section B, which we utilize to extract the concepts from the ad hoc press releases. The section C explains the proposed MI-LDA model for better mining results.

Data pruning

Pruning is implemented by pressing the input sentences extracted from corpus file. It can be done either by word frequency or by the dependency grammar. A phrase structure analysis is carried out first, every word is taken as a node. An oriented edge between two words is it's dependency. It can be done to the semantic, syntactic and morphological representation.

Let S be a sentence and D be the set of dependency tags. A dependency is a triple (w1, w2, d) where w1, w2 € S and d € D. Let Dep(S) denote the set of dependencies of S and root(S) the head of S. Pruning will consist in defining a morph syntactic constraint φ i.e. a condition on

dependencies (and POS tags) of words, the fulfillment of which is necessary for the word to be included in the item set.

The announcements are transformed into a machine-readable format by the following text processing step, which are as follows:

- **Noise removal:** Most announcements contain general information like address, contact information on the releasing company. We remove the underlying event which doesn't provide any relevant information. Some common words and phrases, introduced by the publishing company are also removed from the corpus.
- **Word removal:** The words which do not contribute to the informative value of the announcement (such as the, is), we remove these words from the corpus.
- **Stemming:** In this step, we reduce the inflected words to their stems by using the Porter stemming algorithm.
- **Document-term matrix vector:** At the end of text processing, we create a document-term matrix vector. This matrix vector enables us to numerically represent which words or phrases are present in an article. Then, we apply a tf-idf weighting (i.e. term frequency–inverse document frequency) to identify the most relevant words in the documents (Manning and Schütze, 1999; Salton and McGill, 1986).

Algorithm1:- Text Pruning

Input – Corpus File C
Threshhold - α
Result – Pruned Document C'
Read(C, α)
S:=Corpus file
D' is the document list in the corpus
For d ∈ D' do
d'=noisremoval (wordremoval(stemming(d)))
c'=c' ∪ d'
end
end

A. LDA based concept extraction.

Latent Dirichlet Alllocation (LDA) is an algorithm that specifically aims to find the topics or concepts from a data collection like Financial statements. Originally proposed in the context of text document modeling, LDA choose the way of summarizing the content of a document quickly is to look at the set of words it uses. Because words carry very strong semantic information, documents that contain similar content will most likely use a similar set of words. As such, mining an entire corpus of text documents can expose sets of words that frequently co-occur within documents. These sets of words may be intuitively interpreted as topics and act as the building blocks of the short descriptions.

LDA is a probabilistic, generative model for discovering latent semantic topics in large collections of text data. Each discovered topic is characterized by its own particular distribution over words. Each Financial document is then characterized as a random mixture of topics indicating the proportion of time the document spends on each topic. This random mixture of topics is essentially our "short description": It not only expresses the semantic content of a document in a concise manner, but also gives us a principled approach for describing documents quantitatively. We can now compare how similar one document is to another by looking at how similar the corresponding topic mixtures are.

Algorithm2. LDA based Topic Extraction

1. LDA assumes the following generative process for each document w in a corpus D and
   Choose N ~ Poisson(s).
2. Choose θ ~ Dir(α). As parameter
3. For each of the N words wn:
   (a) Choose a topic zn ~ Multinomial(θ).
   (b) Choose a word wn from p(wn|zn,β), a multinomial probability conditioned on the topic zn.
4. Parameterize probabilities by a k × V matrix β where β_{ij} = $p(W_j = 1/Z_i = 1)$, which for now we treat as a fixed quantity that is to be estimated.
5. A k-dimensional Dirichlet random variable θ can take values in the (k −1)-simplex (a k-vector θ lies in the (k−1)-simplex if $\theta_i \geq 0$, $\sum_i^k=1\ \theta_i = 1$),
6. Probability density on this simplex: $p(\theta|\alpha) = \Gamma \sum_i^k=1\ \alpha_i\ /(\prod_1^k=1\ \Gamma(\alpha_i))\ \theta_1^{\alpha i-1}\ 1\ \cdots\theta_k^{\alpha k-1}$ where the parameter α is a k-vector with components $\alpha_i > 0$, and where Γ(x) is the Gamma function.
7. Find Joint Distribution $p(\theta, z,w|\alpha,\beta) = p(\theta|\alpha) \prod_{n=1}^N p(zn|\theta)p(wn|zn,\beta)$, where p(zn |θ) is simply θ_i for the unique i such that $z_i\ n = 1$.
8. Calculate marginal distribution of a document: $p(w|\alpha,\beta) = Z\ p(\theta|\alpha) \prod_{n=1}^N \sum zn\ p(zn\ |\theta)p(wn\ |zn,\beta))\ d\theta$.
9. Calculate product of marginal probabilities of, we obtain the probability of a corpus: $p(D\ |\alpha,\beta) = \prod_{d=1}^M Z\ p(\theta d\ |\alpha) \prod_{n=1} Nd \sum zdn\ p(zdn\ |\theta d)p(wdn\ |zdn,\beta)\ )\ d\theta_d$
10. Calculate Z= Calculate p(D |α,β)
11. Output the significant z values greater than a given threshold.

B. Multiple Inference based LDA(MI based LDA)

LDA has been used basically for corpus level concept mining. The basic use of LDA has always been associated with concept mining applications. The basic model does not fit multi-level analysis which is very important in

statistical and financial decision making. LDA is a supervised topic modeling method still it lacks a multi-factor analysis which results poor optimization in implementation.

In multi-factor LDA, we have a set of Z latent concepts in the context of text model and each data point has a lower distribution measurement of Θ over their topics. It can assume K factors modeled with a K dimensional array where each cell of an array has a pointer to word distribution. Conceptually each K tuple 't' functions as a topic in LDA. In our model each word location is associated with K topic rather than a single topic value.

Multi-factor based Latent Dirichlet allocation (LDA) is an enhanced model of LDA. The model is generative probabilistic model of a corpus taken as input. The basic idea is that financial documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

Algorithm3. Multifactor based LDA

LDA assumes the following generative process for each document w in a corpus D:

1. Choose N ∼ Poisson(s)., No.of Factors- n

2. Choose θ ∼ Dir(α). As parameter, Initialise factor f 0.

3. For each of the N words wn:

(a) Choose a topic zn ∼ Multinomial(θ).

(b) Choose a word wn from p(wn | zn ,β), a multinomial probability conditioned on the topic zn.

4. Follow Steps 3-8 of Algorithm2

5. Find Impact of the factor f as I(f0) and store the score.

6 For all factor fi...j in F(the set of factors)

Z= Calculate p(D |α,β)*k, k as impact of fi

7. Output the significant z value

Decision Factor Processing

For example, the sentence: "AAPL continues its phenomenal run" is a positive sentence as count (positive) = 2 and count (negative) = 0. "Cracks develop in PCLN" is negative heading as count (positive) = 0 and count (negative) =1.

For an entire corpus, we count the positive and negative instances and compute the score as:

Corpus Score = Positive instances / Total instances

Three types of Corpus Scores:
1. Sentences Corpus Score
2. Headlines Corpus Score
3. Short Description Corpus Score

Decision Factor Scoring on Text Corpus PseudoCode

```
# text is from the news, pos and neg are
positiveandnegative  word lists
scoreCorpus < -function(text, pos , neg)
{
corpus <-Corpus( VectorSource (text))
termfreq_control<-list(removePunctuation= TRUE,
stemming=FALSE, stopwords=TRUE,
ordLengths=c(2,100))
dtm <-DocumentTermMatrix(corpus,
control=termfreq_control)
# term frequency matrix
tfidf<-weightTfIdf(dtm)
# identify positive terms
which_pos<-Terms(dtm ) %in% Pos
# identify negative terms
which_neg<-Terms(dtm) %in%Neg
# number of positive terms in each row
score_pos<-row_sums(dtm[, which_pos])
# number of negative terms in each row
score_neg<-row_sums(dtm[, which_neg])
# number of rows having positive score makes up the net
score
net_score<-sum((score_pos –score_neg)>0)
# length is the total number of instances in the corpus
length <-length(score_pos–score_neg)
score <-net_score/length
return(score)
}
```

V EVALUATION

We have compared Decision Score trends of financial news from various websites. Decision scores are able to predict the nature of companies like growing or declining. It helps investors to make investment decisions.

1. Experimental Setup

We have taken corpus from news websites like Yahoo finance, Business news etc.. The site https://in.finance.yahoo.com is taken for collecting yahoo news. The site contains various news clippings regarding ups and downs of companies. The data is applied to text preprocessing methods like Stopword removal and Stemming. We have written algorithms in C Sharp.net programming language. A testing platform is set with Windows Operating system, 2GB memory and Intel core I3 based Processor.

We have used three parameters to generate prediction results. The values of α and β are varied for different data sets to generate different topic distribution. A decision score parameter is verified against three criteria- Positive, Negative and Neutral

2. Empirical Results

Table 1 shows the observations recorded for different datasets to generate positive news count using LDA and MI-LDA algorithms. Table 2 shows the number of negative decision observations using LDA and MI-LDA . It was found that MI-LDA has generated good results.

Datasets	No. of decisions with LDA	
	Positive	Negative
YahooNews	3	27
CNBC News	11	41
MarketWatch	17	55
Bloomberg Business	15	69
Forbes	10	70

Table1-Positive Decision Count Observations using LDA and MI-LDA

Dataset	**No. of decisions with MI-LDA**	
	Postive	Negative
YahooNews	7	30
CNBC News	29	46
MarketWatch	37	65
Bloomberg Business	39	75
Forbes	51	79

Table 2 – Table showing Negative Decision Observations using LDA and Mi-LDA based News Analysis

Dataset	Time Taken-LDA	Time Taken-MI LDA
YahooNews	0.42	0.48
CNBC News	0.57	0.59
MarketWatch	0.64	0.68
Bloomberg Business	0.74	0.76
Forbes	0.75	0.85

Table 3- Table showing the comparison of time taken for LDA and MI-LDA based news processing under different data sets.

A graphical study for analyzing the positive and negative inference on financial news is given below.

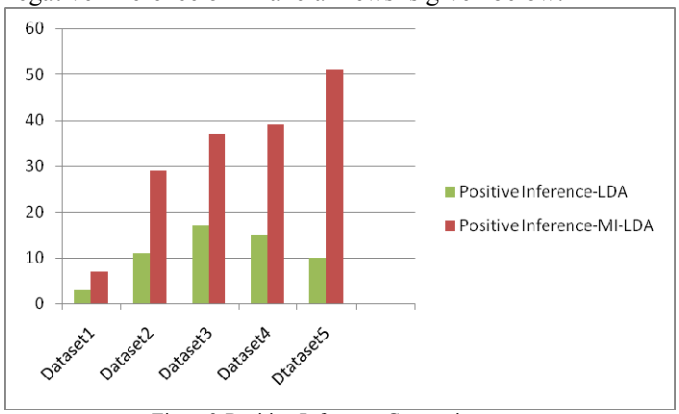

Figure3-Positive Inference Comparison

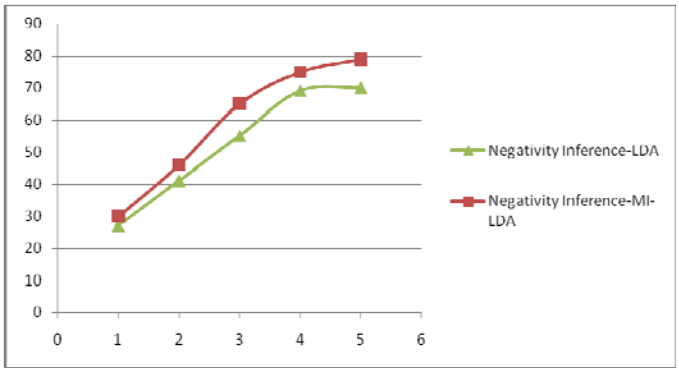

Figure 4- Negative Inference Comparison

3. Time Complexity

Both the algorithms are observed for performancetime. There is no significant change in performance time taken when MI-LDA is implemented. Figure 5 shows the comparison of running time of the new algorithm against the conventional LDA algorithm.

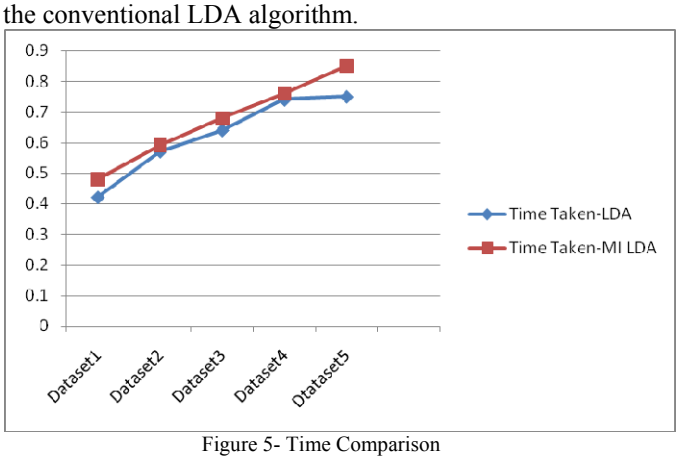

Figure 5- Time Comparison

CONCLUSIONS

From the experimental studies effective prediction results are found based on a decision score. The system is dependable for taking decisions on financial investments. The power of topic distribution based on statistical inference is enhanced with multi-factors like positive, negative and neutral decision factors. The system could optimized the results with a suitable value for α and β. The algorithm can be further enhanced to incorporate new features like analysis of present share values along with the financial news corpus. It is also better to find provisions to reduce information asymmetry and communication style difference in financial news analysis

REFERENCES

[1] David M Bei and Michael Y.Ng, " Latent Dirichlet Allocation – Journel of Machine Learning Research " 3(2003)993-1022
[2] Tam P. Ngo ,"Clustering high dimensional data using SVM" ,2006
[3] M Simmi John and R.Boopathi raj, "High Dimensional Hierarchical Data Clustering using SVM with Kernel region approximation" ,Vol 2(4), 464-467
[4] Li Yanhong, Liu Peng and Qin Zheng, "An Analysis Model of Financial Statements based on Data Mining" ,2006
[5] Shakiba Khademolqorani and Ali Zeinal Hamadani, "An Adjusted Decision Support System through Datamining and Multiple Criteria Decision Making" ,2013
[6] Li Yanhong, Liu Peng and Qin Zheng, "Adaptive Topic Models for mining Text Streams with Application to Topic Detection and Tracking" ,2008
[7] Huang Jiejun,Pan Heping and Wan Youyong," Research on applications of data mining technology, Computer Engineering and Application",No.39,pp.45-48, 2003.
[8] Ma Xiao Hu and Li Gaojin, "Research on Application of Datamining technology in financial Decision Support System" ,IEEE Xplore,2010
[9] Yukari Shirota and Takako Hashimoto and Tamaki Sakura, "Extraction of the financial policy topics by Latent Dirichlet Allocation" ,IEEE Xplore,2014
[10] Anurag Nagar and Michael Hahsler , "News Sentiment Analysis Using R to Predict Stock Market Trends", Southern Methodist University Dallas, TX,2012
[11] Ratku, Antal, "Analysis of how underlying topics in financial news affect stock prices using Latent Dirichlet Allocation" ,University of Freibur.Germany,2014