International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Latent Dirichlet Analysis Based Opinion Mining of Healthcare Report

Silambarasi. P, Kiran L.N.Eranki

Abstract*: *Intelligent abstraction of knowledge from Social blogs has gained immense attention from Biomedical and online health communities. Healthcare re- views through public opinion platforms not only helped in improving the health- care but also reducing the costs. Advancement in digital healthcare systems enables patients access to Online Health Communities (OHC) through their views, opinions and remedial information. In the current study reviews provided by patients on OHC related to diabetes has been chosen to understand the community effort to address the health care issues of these patients. Reviews shared by patients at various levels of diabetic conditions have been selected and analyzed using LDA text mining techniques. In the current study we have also analyzed the gender specific differences among the diabetic community based on the opinions shared by them on public OHC platforms. Sample for the study has been gathered from a prominent online diabetic community in UK. Community members wrote on different topics from food habits, lifestyle, physical exercise and dis- ease symptoms to predict and possible remedial measures to be taken has been discussed in online health communities. Results of our study show that male users are information-centric than the female users, while female users are more emotionally attached as compared to males. Study also reveals several other findings related to community support and state of healthcare sector with reference to treatment, medication and facilities available to the community.

Keywords: LDA, Sentimental analysis, Online Health Care Forums, Gender differences.

I. INTRODUCTION

Over the last few years, the Internet has become the most important rostrum for most of the patients to interact about some health information with the public to get their opinions. People also paying more attention to health-related information. Through this media, we can interact with others more conveniently. Here the statistical analysis says that 75% of the people or users search the information about the diseases and the clinics where the diseases are get cured, and also it includes 35% of the people genuinely share their personal experience which they faced in their real life via social media such as blog related to healthy communities and 25% of the procure only the information through online counselling [14]. Hence each healthcare ecosystem is taking huge benefits of social network to improve its quality of healthcare communities, Group of community people with common symptoms, interests or goals can participate in some other healthcare community forum to exchange information, to express about their opinion regarding that particular disease in the same way we can establish the relationship with other community members. People are taking initiative and creating their active roles for handling their fitness outside of the clinical surroundings [16]. Doctors have efficient knowledge about the treatment and fostering, even though they have enough knowledge in their fields but sometime they may miss the actual needs of the patients [3]. Hence the online health communities which provide great knowledge about the medical concepts and easily we can obscure the information through these blogs. In general, social media makes Data collection difficult. Several other methods have been implemented, some of them are LDA, Sentimental analysis, classification through links, Prediction based on topics, groups and sub-group detection, overall mining the data. Contributions from the current study are - (1) A brief study about the LDA (2) Next section reviews studies about differences in gender in OHCs. Then we intend our research framework and methodologies. Concluding finally with the research methodologies and contributions of this study analysis. Latent Dirichlet allocation (LDA) is an unsupervised learning model developed by [2] using topic modelling. This approach creates a word cluster of pre-specified topics from the huge collection of reviews [9], [12]. LDA uses feature representation in unification with many Machine Learning (ML) algorithm in order to classify the text brochures. Most of this work uses nonstatistical information retrieval model along with the unstructured text in software repositories such as Vector space model (VSM)/Latent semantic indexing (LSI) [15]. Among all these methods LDA had shown best results as compared to VSM and LSI in terms of rephrasing synonyms, un- certainty and creating more intelligible topics. Onan et. al (2016) [9] have used LDA in several text mining, sentiment classification approaches using classification algorithm such as Support vector machine, Nave Bayes, logistic regression, K-nearest neighbor and Radial basis function network. In addition to these the accuracy of the classifier is improved with ensemble methods such as Bagging, stacking, voting, Random subspace and Ada Boost. Accuracy of each algorithm is evaluated using classification accuracy and f-measure. Hence, as a result, the reviews dataset yields the highest predictive performance whereas the ensemble learning method yields better f-measure values and accuracy.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
Silambarasi. P*, School of Computing, SASTRA Deemed University, Thanjavur, Tamilnadu, India.
Dr.Kiran L.N.Eranki, School of Computing, SASTRA Deemed University, Thanjavur, Tamilnadu, India.(*Corresponding author:)

Wang and Zhang (2016) proposed a hybrid method which assimilates Word2vec and LDA together to acquire a relationship between documents and topics in addition to that contextual relationship from Word2vec [7]. Because of this integration, we are not only spawning the relationship between documents and topics besides the contextual relationship among words. Investigational results showed that document feature which is generated by the
Hybrid method is very useful to improve the classification performance by associating global and local interactions. In this context two prominent research methods include,

1) Applying the new document features to supervised models such as Nave Bayes Model, Neural Network (NN)
2) Clustering topics of similar interest in training and segregating in testing.

II. Research Gap and Motivation

According to the study conducted earlier most of the adults were online to seek information which are related to medical conditions. A major place where people find such nobles is Online Health Communities (OHCs), such as ”diabetes.co.uk”. When compared with OHCs, conventional health-related websites that only allow users to retrieve information, but OHCs increased members’ ability to interact with nobles who are facing such health-related issues. In the current study work done by earlier researches to address gender differences in online health community forums has been discussed: 1) Textual reviews provided by various health care centres predict the contemporary state of the patient treatment related to disease diagnosis using healthcare events 2) Analyzing the statistical significance for online health communities towards male and female patients 3) Predicting the correlation among the various healthcare subjects expressing their sentiments through online forums. Applying all the above mentioned approaches to predict the association among gender and disease diagnosis using online healthcare forums. We integrated topic modelling analysis-LDA, sentiment analysis to examine gender differences in online health communities on information and emotional aspects.

III. Methodology and Approach

Designed an Over-all structure to investigate the differences in gender on the traditional dataset from OHCs which are as follows:

1. Acquisition of Data: In this step, we are explaining how we captured the data and how we are performing the data cleaning

2. Analysis of Data: We implemented LDA-Latent Dirichlet Allocation, sentiment analysis to classify gender differences based on the customers statistical needs, and sentimental needs which are in OHCs, respectively. Moreover, performed t- tests and test conducted on significance of statistical analysis in order to scrutinize differences in gender.

3. Gender comparisons: In this step, we are predicting the gender assessment.

A. Acquisition of Data

Diabetes is a kind of chronic disease of constant hyperglycemia. Self-governance is much intrinsic for diabetic care. Studies state that patients with diabetics and their well-being is firmly related to their self-governance ability [6]. This self-governance for patient generally depend on proper medications, treatment and the consistent diet control and their food habits and along with the exercise, so there are more suspicions involved, that makes us examine patient events through online communities. In this work, we decided to propose a technique on UK Based Diabetic community forum as a foundation of data which is the leading and supreme dynamic community forum in the UK. A non-profit organization based online community forum to help people diagnosed with diabetes and educate them with various faces of disease literature. Through this platform patients benefited with various remedial methods can share with early face subjects to proactively cure their disease. To acquire the data, we sneaked the basic post and replies in all the pages of that online community website using the Web scraper script. In the community homepage, details pertaining to various disease diagnostic reviews were posted by both the genders. We extracted the reviews posted in that community forum including the identity of the user as well as their replies of their post. During the pre-processing step we removed their ID, unwanted punctuation and also removed the reviews with missing gender details. Followed by all the valid reviews has been recoded based on male and female replies.

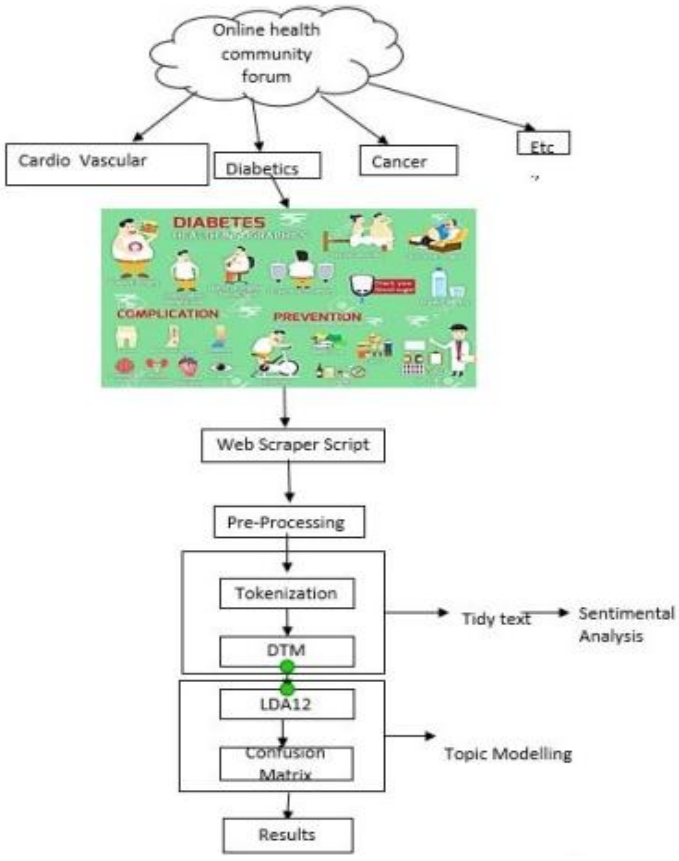

Fig. 1. Research Flow Diagram

B. Analysis of Data

In the remaining sections of the chapter we elaborately discuss about LDA and sentiment analysis. We carried out an analysis of the reviews posted in the diabetes forum, we consolidated the topic which is discussed by the various user with the help of LDA. This method of extraction of topics can help to solve conflicts of thought posted by various users in the community forums.

C. LDA-Text mining

LDA is an unsupervised ML technique which is also known to be a three-layer Bayesian probability model. It assumes that a quantity of text has some probability distribution over topics, and every single topic is related with a dissemination over words. In general, LDA facsimiles a document which has the combination of issues or topics and each topic are allied with the keywords. When fitting a quantity of documents with the LDA, the topics you find often reveal a great deal of information about the relationship and the shared structure between documents [13]. LDA is a propagative statistical model which extends Probabilistic Latent Semantic Analysis (PLSA), is a procedure from the family of topic models, with the purpose of identifying the unseen semantic structure of the data [12]. LDA is a standard natural language processing tool which automatically discovers the bottom-line topics from an amorphous dataset. Mathematically, LDA can be generated as a matrix categorization technique. LDA characterizes a collection of document applying document-term matrix (DTM) in the vector space which is further distributed into two different categories document-topic matrix and topic- term matrix. Generative process of LDA characterizes the joint distribution over random variables. The probability density function of a k-dimensional dirichlet random variable is computed using equation 1 whereas the joint distribution of mixture of topics is solved using the equation 2. Finally, the corpus metric is estimated using the equation 3[1]. For every document M in the dataset, LDA reiterates through each word W" throughout the document and fix the current topic word with a replacement assignment.

$$p\ (\theta/\alpha) = \frac{\Gamma\left(\Sigma_{i=1}^{k}\alpha_i\right)}{\pi_{i=1}^{k}\Gamma(\alpha_i)}\ \theta_1^{\alpha_1-1}\ldots.\ \theta_k^{\alpha_k-1}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$p(\theta,B,w/\alpha,\gamma) = p(\theta/\alpha)\ \pi_{n=1}^{N} p\ (z_n/\theta)\ p(w_n/B_n,\gamma)\cdots\cdots(2)$$

$$\pi_{d=1}^{M}\int p(\theta_d/\alpha)\left(\pi_{n=1}^{N_d}\sum_{B_{dn}} p\left(\frac{B_{dn}}{\theta_d}\right)p(w_{dn}/B_{dn},\gamma)\right)d\theta_d\ \cdots\cdots\cdots(3)$$

For each word W" a new topic B is assigned a probability P where P= product (P1, P2) where P1 and P2 are the two probabilities assigned to each topic in M. organization of words is surveyed for additional processing. LDA objective to categorize the basic underlying topic structure based on the perceived data. In LDA, the perceived data is nothing but words of every single document. For every document in the corpus, the words are created in two staged procedure.1) Distribution over topics are randomly chosen. Based on this selected distribution, a topics is randomly chosen for each word of overall topic for each word of the particular document. In LDA word is a distinct data from the widely chosen vocabulary which are indexed by 1..V , Classification of N words W=[W1, W2.Wn)] and a corpus is a group of M documents which is represented by D=W1,W2, W_M.

For each single document W in corpus D:

1) Choose a multinomial Poisson distribution where N ε
2) Choose a Dirichlet distribution for the document D=W1, W2, W_M with the parameter α
3) For each of the N words Wn a) Choose a topic Zn Multinomial (θ). Choose a word Wn from (Wn—Zn,), a multinomial probability conditioned on the topic Zn.

The progression of LDA can be exhibited by a three-layer Bayesian graphical model, in which nodes are represented by random variables and the edges are represented by the possible dependencies between the variables which are illustrated in information box.

D. Review about OHCs

Community networks or social blogs have become a popular and widely used technology by the societies. Online communities characterize a vital instrument in publics wellbeing support. On the other side the online communities or health care communities are different from public network and other commercial web sites [4]. In online communities or forums usually user share the same opinions, or their experience in common. People who are in the community hardly know each other. Individuals who are in the community portal are not mandatory to interconnect with their friends. People waiting to communicate on particular Areas [4], [10]. The medical management organizations have taking place to turn to public network as devices for joining with the public spectators, as more persons turn to the online network for health-related topic. Several actions that occur in on- line communities such as knowledge dissemination [5], knowledge edifice through team work, and knowledge ingestion are public processes. Various research found that spawning a strong sense of community identification among affiliates of an online community has a progressive impact on their influence to the community. Thus, online community leaders can help accomplish administrative intentions by underlining norms and encouraging people to comply with them. Many Healthcare association have started to build their organization with the social media because most of the people wants to connect to the internet for health-related issues, irrespective of this, public health establishments are previously using social media applications, and some are application are about to launch them.

E. Gender differences in OHCs

Gender expectation towards healthcare are moderately different. In fact, gender differences in healthcare has been investigated in various categories such as outlooks on quality of service, Health seeking behaviour, Resource Operation. With the rise of OHCs empowers the investigators to explore the different gender difficulties. In OHCs huge number of people gain knowledge about the health-related issues and complication towards health in order to accomplish their own fitness. In general, these gender differences are concluded by the human physiological features, public features and social modifications as well as the communications among these factors. Moreover,

the knowledge necessities and Sentimental necessities for both men and women in social health care communities are different. Seale et. al [8] conducted a brief analysis on online health care environments in which the post and replies related to women are emotional and sentimental than men as shown in Figure 2.

Fig. 2. Gender based Opinion Word Cloud

F. Sentimental Analysis with Tidy data

Sentiment analysis is appropriate mining or opinion mining of script which categorizes and mines subjective information in material and helps us to understand the public sentiment of their particular product or facility of service provided by the hospital while watching online conversations. The sentimental analysis which helped to solve the issues with respect to the general and independent classification of texts, sensitive polarity refinement [11] (e.g., positive, negative and neutral emotional sentiments) and sensitive intensity refinement. In this analysis, we made a comparison on both male and female positive sentiment and negative sentiments. An adverse effect most probably impacts a user's psychosomatic state than a neutral or positive effect. Hence the further we measured the expressions on negative sentiments: nervousness, rage, and grief. Here we related male and female emotions among all negative sentiment which are likely to affect the people frame of mind than the positive and neutral sentiment Currently, we have done an analysis test process for the text on post and replies to evaluate all the sentiments which are in the Dataset observation in the research analysis which seem to be in line with studies conducted by other researches.

IV. Results and Discussion

- Statistical analysis of Gender differences: Here we conducted the evaluation between the genders in the OHCs forum. In our sample we noticed nearly 25856 female participants responded actively on the forum comparatively with 12783 male participants Comparative analysis shows that male user are more active than the female users and then we started comparing the average number of replies and post between the female and male users. We have observed 86000 post and replies recorded in the OHC forum. We noticed the pattern of post replied chain carried by male participants are more actively compared with the female participants. Based on these findings, we believe the female users are more thought and emotional when compared to male user. Independent sample t-tests were performed to test differences between the male and female users ($p = 0.6698$) in the number of posts and replies and there is no statistical significance among the gender.

```
tokens from 6 documents.
text1 :
 [1] "elliem"    "saidhaving" "said"      "that"      "i"         "guess"     "the"
 [8] "insulin"   "still"      "gives"     "you"       "more"      "flexibility" "for"
[15] "an"        "occasional" "carb"      "splurge"   "just"      "take"      "lots"
[22] "of"        "insulin"    "for"       "it"        "whereas"   "when"
text2 :
 [1] "also"      "it's"       "worth"     "considering" "that"    "the"
 [7] "latest"    "thinking"   "of"        "pancreatic" """        "burnout"
[13] """         "in"         "type"      "2"         "is"        "that"
[19] "it"        "may"        "actually"  "be"        "more"      "about"

text3 :
 [1] "jim"       "lahey"      "saidno"    "type"      "1"         "is"
 [7] "an"        "autoimmune" "condition" "of"        "insufficient" "insulin"
[13] "type"      "2"          "is"        "a"         "dietary"
```

```
tokens from 6 documents.
text1 :
 [1] "furball64801view" "profileview"   "forum"       "postsprivate"  "messagedd"
 [6] "familygetting"    "much"          "harder"      "to"            "control"
text2 :
 [1] "now"       "its"        "been7hrs"  "since"     "i"         "had"       "novalog"
 [8] "not"       "the"        "first"     "time"      "i"         "have"      "seen"
[15] "weird"     "things"     "but"       "reallytested" "15"     "minutes"   "ago"
[22] "with"      "an"         "86"
text3 :
 [1] "so"        "so"         "happy"     "for"       "you"       "furrafter" "decades"
 [8] "of"        "injecting"  "into"      "your"      "stomach"   "fat"       "it"
[15] "was"       "no"         "longer"    "absorbing" "the"       "insulin"   "i"
[22] "had"       "read"       "about"     "that"      "in"        "the"       "past"
```

Fig. 3. Male and Female Opinion Mined Tokens

Latent Dirichlet Allocation (LDA) We now discuss LDA topic based analysis using R applying Gibbs sampling to approximate the limitations of the LDA model. After running the LDA program a number of times, we identified 10 topics which form a bigger cluster of words in each topic. Major topics that are discussed in the OHCs among male and female has accumulated into 10 topics which are listed in the Fig. 3.

- Confusion Matrix Analysis Confusion matrix in machine learning technique tend of the matrix represented as predicted class whereas each column represented as instances in an actual class. There is a superior contingency table which has two dimensions one is actual and another one predicted. Figure 2 we have analyzed set of emotions that have taken from the dataset, in that we have identified specific set of words to understand the pattern of positive and negative opinions among the forums discussions. Our list included combination of word such as anger, sadness, disgust, fear, litigious are showing negative opinion whereas words such as Joy, anticipating, surprise, trust, positive shows positive side of emotions.

REFERENCES

1. AlSumait, L., Barbara, D., Domeniconi, C, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking", Eighth IEEE international conference on data mining. pp. 3–12. IEEE (2008)
2. Blei, D.M., Ng, A.Y., Jordan, M.I., "Latent dirichlet allocation", Journal of machine Learning research, 3(Jan), 993–1022 (2003)
3. Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A., Hołyst, J.A., "Collective emotions online and their influence on community life", PloS one 6(7), e22207 (2011)
4. Johnson, G.J., Ambrose, P.J., "Neo-tribes: The power and potential of online communities in health care", Communications of the ACM 49(1), 107–113 (2006)
5. Lieberman, M.A., "Gender and online cancer support groups: issues facing male cancer patients", Journal of cancer education 23(3), 167–171 (2008)
6. Liu, X., Sun, M., Li, J., "Research on gender differences in online health communities", International journal of medical informatics 111, 172–181 (2018)
7. Lu, Y., Zhang, P., Liu, J., Li, J., Deng, S., "Health-related hot topic detection in online communities using text clustering", Plos one 8(2), e56221 (2013)
8. Mo, P.K., Malik, S.H., Coulson, N.S., "Gender differences in computer-mediated communication: a systematic literature review of online health-related support groups", Patient education and counseling, 75(1), 16–24 (2009)
9. Onan, A., Korukoglu, S., Bulut, H., "Lda-based topic modelling in text sentiment classification: An empirical analysis", International Journal of Comput. Linguistics Appl., 7(1), 101–119 (2016)
10. Park, H., Park, M.S., "Cancer information-seeking behaviors and information needs among korean americans in the online community", Journal of community health, 39(2), 213–220 (2014)
11. Rowley, J., Johnson, F., Sbaffi, L., "Gender as an influencer of online health information seeking and evaluation behavior", Journal of the Association for Information Science and Technology, 68(1), 36–47 (2017)
12. Shah, A.M., Yan, X., Shah, S.J., Khan, S. "Use of sentiment mining and online nmf for topic modeling through the analysis of patients online unstructured comments", International Conference on Smart Health, pp. 191–203. Springer (2018)
13. Silge, J., Robinson, D., "Text mining with R: A tidy approach." O'Reilly Media, Inc." (2017)
14. Yang, H., Yang, C.C., "Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis", ACM Transactions on Intelligent Systems and Technology (TIST), 6(4), 55 (2015)
15. Zhai, K., Boyd-Graber, J., Asadi, N., Alkhouja, M.L., "Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce", Proceedings of the 21st international conference on World Wide Web, pp. 879–888. ACM (2012)
16. Zhang, Y., Dang, Y., Chen, H., "Research note: Examining gender emotional differences in web forum communication", Decision Support Systems, 55(3), 851–860 (2013)