FLATM: A Fuzzy Logic Approach Topic Model for Medical Documents

Amir Karami, Aryya Gangopadhyay, Bin Zhou
Information Systems Department
University of Maryland Baltimore County
Baltimore, Maryland 21250
Email: amir.karami, gangopad, bzhou@umbc.edu

Hadi Kharrazi
Bloomberg School of Public Health
Johns Hopkins University
Baltimore, Maryland 21205
Email: kharrazi@jhu.edu

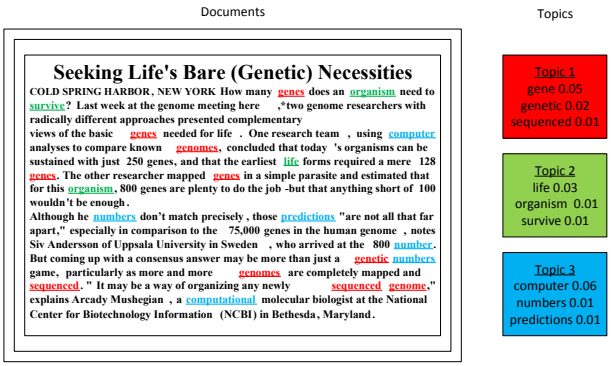
***Abstract*—One of the challenges for text analysis in medical domains is analyzing large-scale medical documents. As a consequence, finding relevant documents has become more difficult. One of the popular methods to retrieve information based on discovering the themes in the documents is topic modeling. The themes in the documents help to retrieve documents on the same topic with and without a query. In this paper, we present a novel approach to topic modeling using fuzzy clustering. To evaluate our model, we experiment with two text datasets of medical documents. The evaluation metrics carried out through document classification and document modeling show that our model produces better performance than LDA, indicating that fuzzy set theory can improve the performance of topic models in medical domains.**

I. Introduction

In the past several years, the medical data have been growing explosively. For example, PubMed[1] is one of the biggest databases for medical research articles. The statistics of PubMed show that the number of papers published in PubMed was increased from 112,177 in 1960 to 2,019,238 in 2013 and the growth rate of publication between 2010 and 2013 is more than 200%. As another example, hospital documents are one major type of medical data. Based on the statistics of U.S. Department of Health and Human Services[2], the annual average number of discharges between 2007 and 2010 is around 35 million records. Analyzing such large-scale medical data is of great importance to enhance health care for millions of people. As reported in [1], more than 44,000 patients died in the hospital as a result of medical errors. In addition, the healthcare industry could save $450 billion a year using advanced data analytical approaches[3].

However, as the majority of medical data are in unstructured free-text format, there is a big challenge to develop methods to analyze large-scale unstructured medical data. Recently, various text mining techniques have been introduced into the medical domain. One fundamental objective of those techniques is to process the unstructured medical data into a proper format for better utilization to recognize explicit facts. Due to the natural probabilistic reasoning of unstructured text data, *topic model* such as Latent Dirichlet Allocation (LDA) [2] has attracted much attention for analyzing medical data. Topic model is one type of statistical models for discovering the latent "topics" that occur in a document collection. It is able to provide a representation of free-text documents in terms of latent features discovered from the collection to generalize an algorithm to unseen documents (Figure 1). Several recent research studies have applied topic models on medical data for different purposes, such as medical document categorization [3], [4], medical document retrieval [5], [6], [7], medical document analysis [8], [9], etc.

Figure1: The Intuition Behind LDA



Despite the usefulness of topic models for medical data analysis [8], [10], existing topic models such as LDA still suffer from several critical issues. One issue of those existing topic models is their computational complexity. Almost all uses of topic models require probabilistic inference, which is arguably hard to achieve without approximate inference algorithms such as Gibbs sampling. Another issue of those existing topic models is their expressive power of representing medical documents.

The performance of various tasks such as document classification and modeling using topic models is still not satisfactory. In this paper, we propose to model medical documents using fuzzy set theory. Fuzzy set theory models membership of objects using a possibility distribution. Most of the studies using fuzzy set theory in the medical domain are related to image processing [11], [12]. A few work have been done in medical text mining using fuzzy clustering [13], [14].

[1]http://www.ncbi.nlm.nih.gov/pubmed
[2]http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6222/version/1
[3]http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care

The main difference between our method and other document fuzzy clustering methods such as [15] is that our method uses fuzzy clustering and word weighting as a preprocessing step for feature transformation before implementing any supervised or unsupervised algorithms; however, other methods use fuzzy clustering as a final step to cluster or classify the documents.

To the best of our knowledge, this is the first study in the medical domain that has been done to use fuzzy set theory to express semantic properties of words and documents in terms of topics. Ideally, if we model words in the documents as objects and a group of relevant words as a latent "topic," the fuzzy set theory provides a natural probabilistic view of free-text documents. Compared with existing topic models such as LDA, the fuzzy set theory is computationally efficient to achieve. We develop several efficient strategies to model medical documents using fuzzy set theory.

Regarding the expressive power, we adopt real medical document collections and compare the performance of our proposed method with LDA by considering different application scenarios. The experimental results show major improvements.

The remainder of this paper is organized as follows. In the next section, we review the related work. In Section III, we present our fuzzy set theory based model in detail. An empirical study was conducted to verify the effectiveness of our method and the results are provided in Section IV. Finally, we present a summary, limitations, and future directions in Section V.

II. Related Work

Medical documents including clinical notes and research papers contain valuable information created by clinicians and researchers. The documents provide rich data waiting to be analyzed. The information in medical documents can be found in the form of narrative and semi-structured format. Some research have been done to extract information from medical documents [16], [17], [18]. There are two major research area in mining medical documents. The first one tracks concepts by looking for frequency of words [19]. The second area categorizes the concepts to find latent variables in medical documents [20]. The first approach leads to high sparse dimensionality data [21]; therefore, researchers have been motivated to use the second approach such as topic modeling. Topic models can help to cluster terms representative of a particular situation such as symptoms or drugs. The goal of topic modeling is to find common topics of dicussion in a corpus. Among topic models, LDA [2] is a popular unsupervised topic model. LDA groups words with similar semantic. Two major outputs of LDA are the probability of each topic for each document, $P(T|D)$, and the probability of each word for each topic, $P(W|T)$. This method is the most effective representation model among supervised and unsupervised topic models [22].

In medical domain, LDA has been leveraged in a wide range of applications. For example, Arnold et al. (2010) used LDA for comparing the topics of patient notes [8] and Bisgin et al. (2011) used LDA in FDA drug side effects labels to cluster drugs [10]. Some other researches propose new variant of LDA to improve its performance for example Cohen et al. (2014) propose a topic model based on LDA to take into account the problem of redundancy in clinical notes [23].

One of methods that has not been fully considered in medical text mining is fuzzy clustering. Since Bellman and Zadeh [24] described the decision-making method in fuzzy environments, an increasing number of studies have dealt with uncertain fuzzy problems by applying fuzzy set theory [25], [26]. Fuzzy Clustering has been used more for image analysis in medical literature [27], [11], [12]. A few work have been done in medical text mining using fuzzy clustering [13], [14]. The main difference between our method and other document fuzzy clustering methods such as [15] is that our method uses fuzzy clustering and word weighting as a preprocessing step for feature transformation before implementing any classification or clustering algorithms; however, other methods use fuzzy clustering as a final step to cluster or classify the documents. Among fuzzy clustering methods, Fuzzy C-means [28] is the most popular one [29]. In addition, we recently used fuzzy clustering as a feature transformation (dimension reduction) approach [30] and also as a method for topic modeling [31] which uses fuzzy clustering for documents in the third step. In this research, we propose a novel method that combines local term weighting and global term weighting with fuzzy clustering to extract latent semantic features from medical documents. In this paper, we extract latent semantic themes from medical documents.

III. FLATM

In this section, we detail our ***Fuzzy Logic Approach Topic Model (FLATM)*** and describe the steps. FLATM has seven steps with three main steps including *Local Term Weighting (LTW), Global Term Weighting (GTM), and Fuzzy Clustering (FC)*. In this algorithm (Algorithm 1), the output(s) of each step is the input(s) of the next step(s).

Step 1: The first step is to calculate LTW. Among different LTW methods we use term frequency as a popular method. Symbol f_{ij} defines the number of times term i happens in document j. We have n documents and m words. Let

$$b(f_{ij}) = \begin{cases} 1 & f_{ij} > 0 \\ 0 & f_{ij} = 0 \end{cases} \tag{1}$$

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} \tag{2}$$

The outputs of this step are $b(f_{ij}), f_{ij}$, and p_{ij}. We use them as inputs for the second step.

Step 2: The next step is to calculate GTW. We explore five GTW methods in this paper including *Entropy, Inverse Document Frequency (IDF), Probabilistic Inverse Document Frequency (ProbIDF), Normal*, and *Global Frequency Inverse Document Frequency (GFIDF)* (Table I).

IDF assigns higher weights to rare terms and lower weights to common terms [32]. ProbIDF is similar to IDF and assigns very low negative weight for the terms happen in every document [33]. In Entropy, it gives higher weight for the terms happen less in few documents [34]. Normal is used to correct discrepancies in document lengths and also normalize the document vectors. Finally, GFIDF is another scheme of

Algorithm 1 FLATM algorithm

Functions:E():Entropy;I():IDF;PI():ProbIDF; NO():Normal; GFIDF:GI(); FC():Fuzzy Clustering.

Remove stop words

Step 1: Calculate LTW

for$i = 1$ to n**do**

  for$j = 1$ to m**do**

    Calculate f_{ij}, $b(f_{ij})$, p_{ij}

  endfor

endfor

Step 2: Calculate GTW

for$i = 1$ to m**do**

  for$j = 1$ to n**do**

    Execute E(p_{ij},n),I(f_{ij},n),PI($b(f_{ij})$,n), NO(f_{ij},n), GI(f_{ij},$b(f_{ij})$)

  endfor

endfor

Step 3: Perform Fuzzy Clustering to Find each Topic Membership for each Word $P(T_k|W_i)$

Execute FC(E), FC(I), FC(PI), FC(NO), FC(GI), FC(f_{ij})

Step 4: Calculate Each Word Probability $P(W_i)$ for each of GTW methods in step 3

$\frac{\sum_{j=1}^{n} E_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} E_{ij}}$, $\frac{\sum_{j=1}^{n} I_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}}$, $\frac{\sum_{j=1}^{n} PI_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} PI_{ij}}$, $\frac{\sum_{j=1}^{m} NO}{\sum_{i=1}^{m}\sum_{j=1}^{n} NO}$, $\frac{\sum_{j=1}^{m} GI}{\sum_{i=1}^{m}\sum_{j=1}^{n} GI}$, $\frac{\sum_{j=1}^{n} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}$

Step 5: Calculate the Joint Probability of Word and Topic $P(W_i, T_k)$

$P(T_k|W_i) \times P(W_i)$

Then Calculate the Probability of each Word in each Topic $P(W_i|T_k)$

$\frac{P(W_i,T_k)}{\sum_{i=1}^{m} P(W_i,T_k)}$

Step 6: Calculate the Probability of each Word in each Document $P(W_i|D_j)$

$\frac{E}{\sum_{1}^{n} E}$, $\frac{I}{\sum_{1}^{n} I}$, $\frac{PI}{\sum_{1}^{n} PI}$, $\frac{NO}{\sum_{1}^{n} NO}$, $\frac{GI}{\sum_{1}^{n} GI}$, $\frac{f_{ij}}{\sum_{1}^{n} f_{ij}}$

Step 7: Calculate the Probability of each Topic in each Document $P(T_k|D_i)$

$\sum_{i=1}^{m} P(T_k|W_i) \times P(W_i|D_j)$

IDF. By using this method words that appear once in every document or once in one document get the smallest weight. This method gives weight to words based on frequency in one document and in all documents [34]. The outputs of this step are the inputs of the next step(s).

Step 3: Fuzzy set theory has been used to model systems that have difficulty assigning an instance to a set [25]. Fuzzy clustering is a soft clustering technique that finds the degree of membership for each data point in each cluster, as opposed to assigning a data point only to one cluster. Fuzzy clustering is a synthesis between clustering and fuzzy logic. Among fuzzy clustering methods, Fuzzy C-means (FCM) [28] is the most popular one [29] and its goal is to minimize an objective function by considering constraints:

$$Min\ J_q(\mu, V, X) = \sum_{k=1}^{c}\sum_{j=1}^{n}(\mu_{kj})^q DIS_{kj}^2 \tag{3}$$

subject to:

$$0 \leq \mu_{kj} \leq 1; \tag{4}$$

Name	Formula
Entropy	$1 + \frac{\sum_j p_{ij}\log_2(p_{ij})}{\log_2 n}$
IDF	$\log_2 \frac{n}{\sum_j f_{ij}}$
ProbIDF	$\log_2 \frac{n - \sum_j b(f_{ij})}{\sum_j b(f_{ij})}$
Normal	$\frac{1}{\sqrt{\sum_j f_{ij}^2}}$
GFIDF	$\frac{\sum_j f_{ij}}{\sum_j b(f_{ij})}$

TableI: GTW Methods

$$\sum_{k=1}^{c} \mu_{kj} = 1 \tag{5}$$

$$0 < \sum_{j=1}^{n} \mu_{kj} < n; \tag{6}$$

Where:

n= number of data
c= number of clusters (topics)
μ_{kj}= membership value
q= fuzzifier, $1 < q \leq \infty$
V= cluster center vector
$DIS_{kj} = d(x_j, v_k)$= distance between x_j and v_k

By optimizing eq.3:

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{c}(\frac{DIS_{kj}}{DIS_{lj}})^{\frac{2}{q-1}}} \tag{7}$$

$$v_i = \frac{\sum_{j=1}^{n}(\mu_{kj})^q x_j}{\sum_{j=1}^{n}(\mu_{kj})^q} \tag{8}$$

The iterations in the clustering algorithms continue till the the maximum change in μ_{ij} becomes less than or equal to a pre-specified threshold. The computational time complexity is $O(n)$. We use μ_{ij} as the degree of clusters' membership for each word or $P(T_k|W_i)$. Topic (T) is the membership degree of a fixed vocabulary in which words with similar semantics have a higher membership degree.

One of the problems in fuzzy clustering is handling a large dimension data. In this paper, we run fuzzy clustering for 9 times with the number of clusters from 10 to 2. Each clustering step is the input for the next one. For example, the output of applying fuzzy clustering on matrix with n words (rows) and m documents (columns) with selecting 10 as the number of cluster is a matrix with n words (rows) and 10 clusters (columns) . Then we again apply fuzzy clustering on matrix with n rows and 10 columns with selecting 9 as the number of cluster. The output of these 9 steps is a matrix with n rows and 2 columns that helps us to reduce the dimension from m to 2. This matrix is the input for fuzzy clustering for the number of topics from 50 to 200 to find $P(T_k|W_i)$.

Step 4: In this step, we use document-term matrices of step 2 with and without GTW methods to find the probability

of words, $P(W_i)$, by:

$$P(W_i) = \frac{\sum_{i=1}^{n}(W_i, D_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}(W_i, D_j)} \quad (9)$$

Step 5: The next step is to find $P(W_i|T_k)$ by first calculate:

$$P(W_i, T_k) = P(T_k|W_i) \times P(W_i) \quad (10)$$

Then we normalize $P(W, T)$ in each topic:

$$P(W_i|T_k) = \frac{P(W_i, T_k)}{\sum_{i=1}^{m} P(W_i, T_k)} \quad (11)$$

Step 6: We do the similar calculation in step 5 to find $P(W_i|D_j)$:

$$P(W_i|D_j) = \frac{P(W_i, D_j)}{\sum_{i=1}^{m} P(W_i, D_j)} \quad (12)$$

Step 7: The final step is to find $P(T_k|D_j)$ by:

$$P(T_k|D_j) = \sum_{i=1}^{m} P(T_k|W_i) \times P(W_i|D_j) \quad (13)$$

IV. Experimental Results

In this section, we discuss our empirical evaluation of FLATM against LDA using three measures: document classification and document modeling. In the experiment, we use Matlab packages for Chib-style estimation[1] and fcm[2] with its default setting for implementing FCM clustering. Moreover, we use Weka[3] for classification evaluation, and MALLET package[4] with its default setting for implementing LDA.

A. Datasets

We leverage two available datasets in this research. The first dataset is a labeled corpus of English scientific medical abstracts from Springer website. It includes 41 medical journals ranging from Neurology to Radiology. In this research, we selected 5 journals including: Arthroscopy, Federal health standard sheet, The anesthetist, The surgeon, and The gynecologist with 1527 documents and 14411 terms.The second dataset called Deidentified Medical Text is an unlabeled corpus of 1607 nursing notes with 11,059 terms (Tables II&III).

#Documents	1527
#Term Tokens	245931
Unique Terms	14411
Avg Term Per Document	96.3

TableII: Basic Statistics for First Dataset

#Documents	1607
#Term Tokens	299449
#Unique Term Words	11059
Avg Term Per Document	124.8

TableIII: Basic Statistics for Second Dataset

[1]http://www.cs.umass.edu/∼wallach/code/etm/

[2]http://www.mathworks.com/help/fuzzy/fcm.html

B. Document Classification

The first evaluation measure is document classification on the first datasest. We use 80% of data for training and 20% for testing, with 5-fold cross validation. We train the models for the five classes and calculate the likelihood for the test data. We present the classification accuracy for the models. Table IV shows that FLATMs with Entropy, IDF, Normal, and ProbIDF have better performances than LDA in almost all different number of topics. The advantage of approach is especially obvious for a large number of topics. In addition, the combination of FLATM using GFIDF and FLATM without using any of GTW methods produces a lower performance in comparison to LDA. We remove the combination of FLATM using GFIDF and FLATM without using any of GTW methods in the rest of the experiments. Finally, the sign test shows that the improvement of FLATMs over LDA is statistically significant with a $p - value < 0.05$.

Method	**50 Topics**	**100 Topics**	**150 Topics**	**200 Topics**
FLATM (Entropy)	71.31	71.91	72.04	72.88
FLATM (ProIDF)	70.13	71.71	71.05	72.76
FLATM (IDF)	71.57	68.96	71.31	70.98
FLATM (Normal)	69.15	69.81	70.81	71.57
LDA	72.29	66.01	65.42	63.91
FLATM	44.66	42.24	53.11	49.91
FLATM (GFIDF)	38.51	38.70	38.83	39.03

TableIV: Document Classification Accuracy (%)

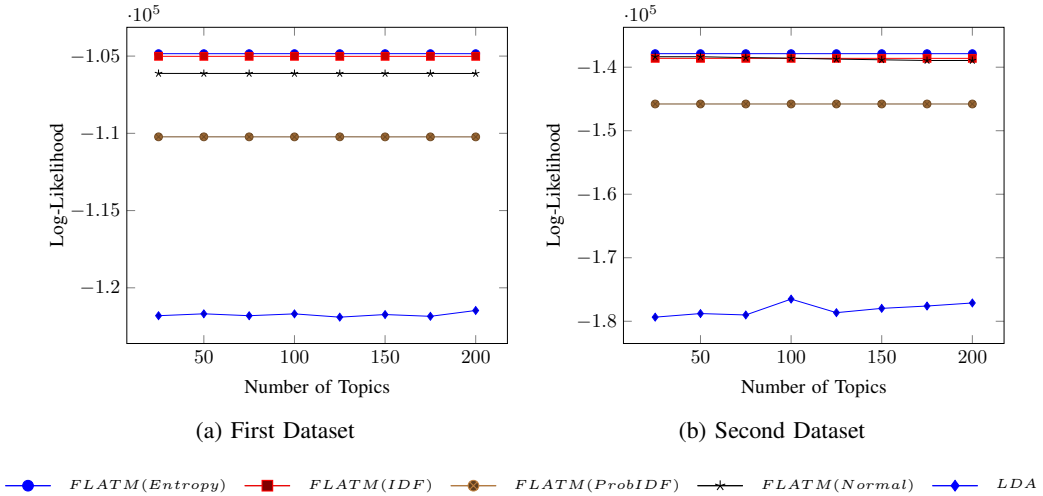
C. Document Modeling

The third evaluation measurement is document modeling using log-likelihood. We trained FLATMs and LDA, on both datasets to compare generalization performance of these models. The documents in the corpora are treated as unlabeled; thus, our goal is density estimation to achieve high likelihood on a held-out test set. We split the first and the second dataset into two sublets with 90% and 10% of the dataset respectively. In preprocessing the data, we removed a standard list of stop words from each corpus. Then we learn topics from the larger set and calculate log-likelihood for the smaller set, $P(D_{test}|T)$.

There are different methods to calculate log-likelihood that among them Chib-style estimation shows better performance [35]. In this evaluation part, we remove FLATM with GFIDF and without any GTW because of their week performance in document classification, and we focus on LDA and the rest of FLATMs. We compare FLATMs with LDA and the result shows that FLATMs have a better performance over LDA with different number of topics and different sets of training data.

[3]http://www.cs.waikato.ac.nz/ml/weka/

[4]http://mallet.cs.umass.edu/



(a) First Dataset (b) Second Dataset

Figure2: Likelihood Comparison

Figures 2.a and 2.b present the log-likelihood for each model on both corpora for different number of topics. FLATMs consistently perform better than LDA.

V. Conclusion

A large volume of medical data has been accumulated in recent years. Analyzing such data is becoming more and more important to advance state-of-the-art healthcare. Due to the unstructured nature of free-text format for the medical data, text mining techniques such as topic modeling are widely adopted to extract latent semantic properties of a medical corpus.

Despite the usefulness of topic models for medical data analysis, existing topic models such as LDA still suffer from several critical issues, such as extremely high computational complexity and unsatisfactory performance for data analytical tasks. In this paper, we proposed the use of fuzzy set theory, the fuzzy clustering technique in particular, for modeling unstructured medical documents. Fuzzy clustering is one of the machine learning techniques that has been used more in medical image processing. To the best of our knowledge, this is the first study that uses fuzzy clustering for topic modeling of medical documents. Our proposed method, FLATM, is a topic model that uses fuzzy clustering with local and global term weighting methods to disclose the latent semantic of medical documents.

Compared to LDA, FLATM has a much lower computational complexity and provides stronger expressive power of medical documents. Our empirical evaluation conducted on several real medical datasets showed that FLATM outperforms LDA in various data analytical tasks including medical document classification and modeling.

There are several interesting directions to explore in the future. For example, prediction of stages of various diseases is very important in healthcare. We are interested in applying our FLATM method on large-scale medical data to provide accurate predictions for patients like using in home healthcare robots [36], [37], [38]. In addition, our approach can be used for other domains such as spam detection in SMS and online reviews [39], [40], [41].

References

[1] L.T. Kohn, J.M. Corrigan, M.S. Donaldson *etal.*, *To err is human: building a safer health system*. National Academies Press, 2000, vol. 627.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol.3, pp. 993–1022, 2003.

[3] E.Sarioglu, H.-A. Choi, and K.Yadav, "Clinical report classification using natural language processing and topic modeling," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol.2. IEEE, 2012, pp. 204–209.

[4] E.Sarioglu, K.Yadav, and H.-A. Choi, "Topic modeling based classification of clinical reports," *ACL 2013*, p.67, 2013.

[5] Q.T. Zeng, D.Redd, T.Rindflesch, and J.Nebeker, "Synonym, topic model and predicate-based query expansion for retrieving clinical documents," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1050.

[6] Z.Huang, W.Dong, L.Ji, C.Gan, X.Lu, and H.Duan, "Discovery of clinical pathway patterns from event logs using probabilistic topic models," *Journal of biomedical informatics*, vol.47, pp. 39–57, 2014.

[7] T.AS01 and K.Eguchi, "Predicting protein-protein relationships from literature using latent topics."

[8] C.W. Arnold, S.M. El-Saden, A.A. Bui, and R.Taira, "Clinical case-based retrieval using latent topic analysis," in *AMIA Annual Symposium Proceedings*, vol. 2010. American Medical Informatics Association, 2010, p.26.

[9] C.Howes, M.Purver, and R.McCabe, "Using conversation topics for predicting therapy outcomes in schizophrenia," *Biomedical informatics insights*, vol.6, no. Suppl 1, p.39, 2013.

[10] H.Bisgin, Z.Liu, R.Kelly, H.Fang, X.Xu, and W.Tong, "Investigating drug repositioning opportunities in fda drug labels through topic modeling," *BMC bioinformatics*, vol.13, no. Suppl 15, p.S6, 2012.

[11] W.Cui, Y.Wang, Y.Fan, Y.Feng, and T.Lei, "Global and local fuzzy clustering with spatial information for medical image segmentation," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*. IEEE, 2013, pp. 533–537.

[12] Z.Beevi and M.Sathik, "A robust segmentation approach for noisy medical images using fuzzy clustering with spatial probability." *International Arab Journal of Information Technology (IAJIT)*, vol.9, no.1, 2012.

[13] D.Ben-Arieh and D.K. Gullipalli, "Data envelopment analysis of clinics with sparse data: Fuzzy clustering approach," *Computers & Industrial Engineering*, vol.63, no.1, pp. 13–21, 2012.

[14] G.Fenza, D.Furno, and V.Loia, "Hybrid approach for context-aware service discovery in healthcare domain," *Journal of Computer and System Sciences*, vol.78, no.4, pp. 1232–1247, 2012.

[15] V.K. Singh, N.Tiwari, and S.Garg, "Document clustering using k-means, heuristic k-means and fuzzy c-means," in *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*. IEEE, 2011, pp. 297–301.

[16] C.Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, and S.B. Johnson, "A general natural-language text processor for clinical radiology," *Journal of the American Medical Informatics Association*, vol.1, no.2, pp. 161–174, 1994.

[17] P.Haug, S.Koehler, L.M. Lau, P.Wang, R.Rocha, and S.Huff, "A natural language understanding system combining syntactic and semantic techniques." in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1994, p. 247.

[18] U.Hahn, M.Romacker, and S.Schulz, "Medsyndikatea natural language system for the extraction of medical information from findings reports," *International journal of medical informatics*, vol.67, no.1, pp. 63–74, 2002.

[19] C.Poulin, B.Shiner, P.Thompson, L.Vepstas, Y.Young-Xu, B.Goertzel, B.Watts, L.Flashman, and T.McAllister, "Predicting the risk of suicide by analyzing the text of clinical notes," *PLOS ONE*, vol.9, no.1, p. e85733, 2014.

[20] J.Lin, D.Karakos, D.Demner-Fushman, and S.Khudanpur, "Generative content models for structural analysis of medical abstracts," in *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*. Association for Computational Linguistics, 2006, pp. 65–72.

[21] C.C. Aggarwal and C.Zhai, "An introduction to text mining," in *Mining Text Data*. Springer, 2012, pp. 1–10.

[22] Y.Halpern, S.Horng, L.A. Nathanson, N.I. Shapiro, and D.Sontag, "A comparison of dimensionality reduction techniques for unstructured clinical text," in *ICML 2012 Workshop on Clinical Data Analysis*, 2012.

[23] R.Cohen, I.Aviram, M.Elhadad, and N.Elhadad, "Redundancy-aware topic modeling for patient record notes," *PloS one*, vol.9, no.2, p. e87555, 2014.

[24] R.E. Bellman and L.A. Zadeh, "Decision-making in a fuzzy environment," *Management science*, vol.17, no.4, pp. B–141, 1970.

[25] A.Karami and Z.Guo, "A fuzzy logic multi-criteria decision framework for selecting it service providers," in *45th Hawaii International Conference on System Science (HICSS)*. IEEE, 2012, pp. 1118–1127.

[26] A.Karami, H.Yazdani, H.Beiryaie, and N.Hosseinzadeh, "A risk based model for is outsourcing vendor selection," in *the 2nd IEEE International Conference on Information and Financial Engineering (ICIFE)*, 2010, pp. 250–254.

[27] I.Saha and U.Maulik, "Multiobjective differential evolution-based fuzzy clustering for mr brain image segmentation image segmentation," in *Advanced Computational Approaches to Biomedical Engineering*. Springer, 2014, pp. 71–86.

[28] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[29] K.Bataineh, M.Naji, and M.Saqer, "A comparison study between various fuzzy clustering algorithms." *Jordan Journal of Mechanical & Industrial Engineering*, vol.5, no.4, 2011.

[30] A.Karami and A.Gangopadhyay, "Fftm: A fuzzy feature transformation method for medical documents," in *Proceedings of the BioNLP 2014 Workshop*. Association for Computational Linguistics(ACL), 2014.

[31] A.Karami, A.Gangopadhyay, B.Zhou, and H.Kharrazi, "A fuzzy approach model for uncovering hidden latent semantic structure in medical text collections," *iConference 2015 Proceedings*, 2015.

[32] K.Papineni, "Why inverse document frequency?" in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, 2001, pp. 1–8.

[33] T.G. Kolda, "Limited-memory matrix methods with applications," 1998.

[34] S.Dumais, "Enhancing performance in latent semantic indexing (lsi) retrieval," 1992.

[35] H.M. Wallach, I.Murray, R.Salakhutdinov, and D.Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112.

[36] A.Alaiad and L.Zhou, "The determinants of home healthcare robots adoption: An empirical investigation," *International Journal of Medical Informatics*, vol.83, no.11, pp. 825–840, 2014.

[37] A.Alaiad, L.Zhou, and G.Koru, "An exploratory study of home healthcare robots adoption applying the utaut model," *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, vol.9, no.4, pp. 44–59, 2014.

[38] A.Alaiad and L.Zhou, "Patients behavioral intention toward using healthcare robots," *19th Americas Conference on Information Systems, AMCIS*, 2013.

[39] A.Karami and L.Zhou, "Exploiting latent content based features for the detection of static sms spams," submitted to ASIST 2014: The 77th Annual Meeting of the American Society for Information Science and Technology (ASIST).

[40] A.Karami and B.Zhou, "Online review spam detection by new linguistic features," *iConference 2015 Proceedings*, 2015.

[41] A.Karami and L.Zhou, "Improving static SMS spam detection by using new content-based features," in *20th Americas Conference on Information Systems, AMCIS 2014, Savannah, Georgia, USA, August 7-9, 2014*, 2014.

This figure "tm.png" is available in "png" format from:

http://arxiv.org/ps/1911.10953v1

This figure "tm2.png" is available in "png" format from:

http://arxiv.org/ps/1911.10953v1