

Analysis of latent Dirichlet allocation and non-negative matrix factorization using latent semantic indexing



Sheikh Muhammad Saqib ^{1,*}, Shakeel Ahmad ², Asif Hassan Syed ², Tariq Naeem ¹, Fahad Mazaed Alotaibi ²

¹Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan

²Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdul Aziz University (KAU), Jeddah, Saudi Arabia

ARTICLE INFO

Article history:

Received 8 May 2019

Received in revised form

15 August 2019

Accepted 16 August 2019

Keywords:

Sentiment analysis

Topic modelling

Latent Dirichlet allocation

Non-negative matrix factorization

Latent semantic indexing

ABSTRACT

A word is a major attribute in the field of opinion/text mining. Based on this attribute, it is decided that whether it is a keyword, aspect, feature, entity, title, or topic? Lots of work has been done to detect such targets using both supervised and unsupervised approaches. These targets can be used in further processing such as text analytics, sentiment analysis, information retrieval, and searches, etc. Latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF) are the major models used for detecting topics. Understanding the depth and details of them algorithms are necessary for those who want to extend these models. The research community of opinion/text mining uses them as a black box. However, there is a question about which model is the most accurate for detecting topics. Latent semantic indexing (LSI) is the best approach for detecting the best match for document in a given query. In this study, we analyzed the LDA and NMF models using LSI to determine the best model for opinion/text mining and found that both are very good, but NMF is slightly better than LDA.

© 2019 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

In the field of opinion mining, topic molding plays a vital role in aspect extraction, key word extraction, and entity extraction (Zhao et al., 2010). Researchers model these using multiple techniques with different accuracy rates. The most commonly used models in topic detection are latent Dirichlet analysis (LDA) and non-negative matrix factorization (NMF). There is lot of confusion, however, regarding which model is most suitable for topic detection. Agrawal et al. (2018) said that LDA methods are more suited in domains where data is in semantic units like words, and NMF methods are more suited to domains where data has the so-called semantic gap." Stevens et al. (2012) stated that "when a descriptive topic is required, LDA is the best choice", and Xue et al. (2014) stated that "NMF is more appropriate when dealing with visual ambiguities". While Chen et al. (2017) and Taniguchi et al. (2018) specified that LDA also shows better adaptability and robustness with clustered visual data. But


opinion/text mining researchers have only textual data and are uncertain with respect to the above answers. In LDA and NMF blogs, the most frequently asked questions show confusion about which model is best (What is a good way to perform topic modeling on short text?). That said, validation of topic detection in both models can be very challenging (Suri and Roy, 2017).

LDA and MNF algorithm details are necessary for the statistical research community, as well as those researchers who want to 1 extend these algorithms or make changes to existing algorithms, such as supervised 2 latent Dirichlet allocation (SLDA) (Blei and McAuliffe, 2010), LDA for multiple languages (MLSLDA) (Boyd-Graber and Resnik, 2010), Constrained-LDA (Zhai et al., 2011), constrained symmetric nonnegative matrix factorization (CSNMF) (Peng and Park, 2011), constrained NMF (Liu and Wu, 2010), and semi-supervised 4 NMF (Chen et al., 2008). The opinion/text mining research community uses LDA and NMF as a black box, where an output is produced based on given inputs. There is no need to understand the depth of each model. But they are concerned with the question of which model is best for topic detection, because they use the detected topic for further processing, i.e., aspect extraction, keyword extraction, grouping aspects into categories, spam detection, opinion topics and finding a common semantic space (He et al., 2011; Gao and Li, 2011; Li et al., 2010). The proposed

* Corresponding Author.

Email Address: saqibsheikh4@gu.edu.pk (S. M. Saqib)

<https://doi.org/10.21833/ijaas.2019.10.015>

 Corresponding author's ORCID profile:

<https://orcid.org/0000-0002-4647-1698>

2313-626X/© 2019 The Authors. Published by IASE.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

methodology is based on the issues related to say community. The aims of this study are as follows:

- We propose to develop a strategy which can determine the best model for opinion/text mining (LDA or NMF) using Latent Semantic Indexing (LSI).
- We propose a method to find out which document is closest to or farthest from the detected topic (LDA- topic and NMF-topic).
- We propose a method to find out which detected topic (LDA-topic and NMF-topic) is closest to the documents.
- We propose to generate an order of documents from the detected topic (LDA-topic and NMF-topic)
- Using LSI, and make a final decision based on these orders.

We used documents from three domains- medicine, politics, and sports to compose the proposed methodology to find which Topic-model is close to the documents.

2. Related work

Topic modeling is an unsupervised learning approach to clustering documents in order to discover topics based on their contents. It is very similar to how the k-means and expectation-maximization algorithms work. Because we are clustering documents; we have to process the individual words in each document to discover topics and assign values to each based on the distribution of these words. This increases the amount of data we are working with, so to handle the large amount of processing required for clustering documents, we have to utilize efficient sparse data structures.

Topic modeling is concerned with aspect extraction, entity extraction, keyword extraction, etc. Keyword.

Extraction has been used in a variety of natural language processing applications, such as information retrieval systems, digital library searching, web content management, document clustering, and text summarization (Rose et al., 2010). Topic detection enables the automatic identification of semantic content and the assignment of a topic label to a given document. Although these approaches are highly useful for a large spectrum of applications, only a limited number of documents with keywords are available online (El-Fishawy, 2014). Keyword extraction is also a process of identifying a short list of words or noun phrases that capture the most important ideas or topics covered in a document (Awajan, 2014). For (Rammal et al., 2015), the aim was to apply local grammar (LG) to develop an indexing system that automatically extracts keywords from titles of Lebanese official journals. Topic modeling offers a computational tool to find relevant topics by capturing meaningful structure among the collections of documents (Wang et al., 2016). For

entity extraction (Pantel et al., 2009) have used a method distribution similarity by comparing the similarity of the surround words of each candidate entity with those of the seed entities; and then ranking the candidate entities based on the similarity values.

When determining the summary of a document or sentiment analysis of an opinion, it is important to find out whether the selected document contains the required key words, aspects, or entities (Chinsha and Joseph, 2015; Qi and Chen, 2011; Thakur and Singh, 2015). Recent studies have proposed a novel, rule-based method for extracting an aspect from reviews of products using an unsupervised approach to uncover the polarity of an aspect in different domains (Gindl et al., 2013; Hu and Liu, 2004). Machine learning and NLP-based rules can also provide better solutions for identifying the aspects, topics, and key words of a paragraph with less effort (Gupta and Ekbal, 2014). The approach uses a classifier trained for each distinct word in a corpus of manually sense-annotated examples as an entirely unsupervised method to cluster the occurrence of words (Raganato et al., 2017). An aspect-based sentiment analysis, which can be carried out by using only particular aspects (Jeyapriya and Selvi, 2015; Gamon et al., 2005; Zhuang et al., 2006; Gojali and Khodra, 2016), requires less effort compared to a sentiment analysis of an object with respect to all aspects. Keyword and topic extraction are not only used in researching the English language, but also in research surrounding other languages, such as Arabic, French, German, Spanish, Chinese, Greek, and Japanese (Pang and Lee, 2008; Tumasjan et al., 2010; Alshammari, 2018). The most important methods used for topic modeling are LDA and MNF (Leek et al., 2000; MacMillan and Wilson, 2017).

In a joint model for sentiment analysis, an aspect-sentiment mixture model was built, based on an aspect (topic) model using LDA and extended LDA (Mei et al., 2007; Lin and He, 2009; Jo and Oh, 2011). A joint model was also proposed in Sauper et al. (2011), which worked only on short snippets already extracted from reviews. Another extension of joint model is semi-supervised joint model, where some topics and aspects are detected by providing some seed aspect terms (Mukherjee and Liu, 2012). A method based on Probabilistic Latent Semantic Analysis PLSA produced a rated aspect summarization of short comments from eBay.com (Lu et al., 2009).

An interdependent LDA (ILDA) has been used to find group aspects and to derive their ratings (Moghaddam and Ester, 2011). The extension of LDA known as ILDA, "it is a type of multilevel latent semantic association, where at the first level, all the words in aspect expressions (each aspect expression can have more than one word) are grouped into a set of concepts or topics using LDA" (Guo et al., 2009). There are also studies in which manifold learning is used for modeling a robot's multimodal information, in these studies, they used manifold learning such as NMF, and multimodal information, which is an

observation of the model represented by low dimensional hidden parameters (Mangin et al., 2015; Chen and Filliat, 2015). Reviews are rated according to an object, so there should be a direct method to determine whether a review is positive or negative. LSI (Latent Semantic Indexing) is better for such a purpose (Saqib et al., 2016). LSI (Huang et al., 2009) has been used for the clustering of documents and for concept representations. An extended method based on LSI can filter unwanted emails in Chinese and English (Yang and Li, 2005). There are many questions about the LDA and NMF models in the opinion mining research community. Various works have been done to test the accuracy of LDA and NMF based on the nature of the data. In text/opinion mining, only the topic of a document which can be determined by LDA or NMF is used for further processing. These researchers use LDA and NMF as a black box tool.

3. Analysis of LDA and NMF using LSI

Using this methodology, we generated a topic from LDA as the LDA-topic and NMF as the NMF-topic. We then used LSI by providing the topic as a query and the document as a list. This method determines the score of each document for the LDA-topic and the NMF-topic. After this, a decision is made by comparing the average LSI score of all documents for the LDA-topic and average LSI score

of all documents for the NMF-topic. Whichever has the greater score is the best model. This method generates two lists of document LSI scores in descending order. The first order is based on the LSI scores of each document for the LDA-topic, and the second order is based on the LSI scores of each document for the NMF-topic. In the first order, the topmost score will be the closest document to the LDA-topic and the last score will be the document which is farthest away from the LDA-topic. In the second order, the topmost score will be the closest document to the NMF-topic, and the last score will be the document which is farthest away from the NMF-topic. The whole process is depicted in Fig. 1.

3.1. LDA

LDA, or Latent Dirichlet Analysis, is a probabilistic model. To obtain cluster assignments, it uses two probability values: P (word-topics) and P (topics-documents). These values are calculated based on an initial random assignment; after which they are repeated for each word in each document to decide their topic assignment. In an iterative procedure, these probabilities are calculated multiple times, until the convergence of the algorithm (Chawla, 2017). Its algorithm is available in course of Advanced Machine Learning at topic "Topic Modeling: Latent Dirichlet Allocation".

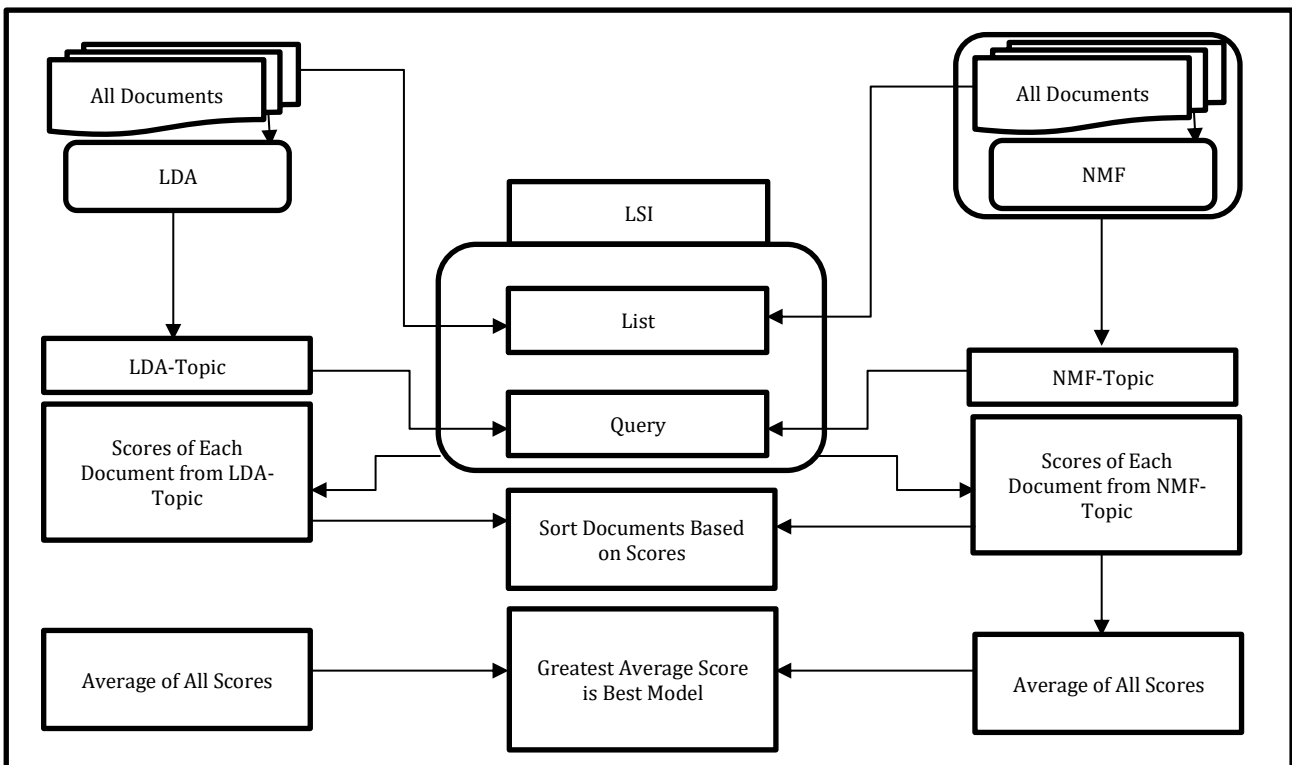


Fig. 1: Proposed framework

We can describe LDA more formally with the following notation (Blei et al., 2010).

"The topics are $\beta_{1:k}$, where each β_k is a distribution over the vocabulary. The topic proportions for the d_{th} document are θ_d , where $\theta_{d,k}$

is the topic proportion for topic k in document d . The topic assignments for the d_{th} document are Z_d , where $Z_{d,n}$ is the topic assignment for the n_{th} word in document d . Finally, the observed words for document d are W_d , where $W_{d,n}$ is the n_{th} word in

document d , which is an element from the fixed vocabulary". With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables in Eq. 1 (Blei et al., 2010):

$$p(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^k p(\beta)_i \prod_{d=1}^D P(\theta_d) \left(\prod_{n=1}^N P(Z_{d,n} | \theta_d) p(W_{d,n} | \beta_{1:k}, Z_{d,n}) \right) \quad (1)$$

The algorithm based on above equation was implemented using TfidfVectorizer, Count Vectorizer classes of package sklearn: Feature extraction; text and Latent Dirichlet Allocation of package sklearn; decomposition in Python.

3.2. NMF

Non-negative matrix factorization is a Linear-algebraic model, that factors high-dimensional vectors into a low-dimensionality representation. Like Principal component analysis (PCA), NMF takes advantage of the fact that the vectors are non-negative. By factoring them into the lower-dimensional form, NMF forces the coefficients to also be non-negative, its algorithm is also implemented in "Topic Modelling with LDA and NMF on the ABC News Headlines dataset" by Chawla (2017). NMF is useful in settings where the domain of the data is inherently non-negative and where parts-based decompositions are desired. In general, "NMF seeks a $n \times d$ non-negative matrix W and a $d \times t$ non-negative matrix H so that $V \sim WH$ ". The matrices W and H are estimated by minimizing the following objective function which is calculated from Eq. 2 (MacMillan and Wilson, 2017):

$$D_{NMF}(W, H) = \|V - H\|_F^2, W \geq 0, H \geq 0 \quad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm. In topic modeling, W and H have a special interpretation: W_{ij} quantities the relevance of topic j in document i , and H_{ij} quantities the relevance of term j in topic i .

3.3. Latent semantic indexing for LDA and NMF

LSI, which was proposed by Deerwester et al. (1990) and Blei et al. (2003), is an efficient information retrieval algorithm (Phadnis and Gadge, 2014). Basically, in LSI, there is a cosine similarity measurement between the coordinates of a document vector and the coordinates of a query vector. If this value is 1, it means that the document matches the query 100%; if it is 0.5, it means the document matches the query 50%; and if it is 0.9, it means the document matches the query 90%. The important step now is finding the coordinates of each document and query. A singular value decomposition (SVD) can determine the points or coordinates of a document and query. Through the SVD, three values S , V and U which will be used for

further processing can be determined by a matrix. The matrix consists of rows and columns containing integers, where the inputs are different text documents. A feature matrix can be obtained by calculating the frequencies of each word. This means that first, a feature matrix is created from all the documents, and then the SVD is calculated. After this, the supporting variables S , V and U are calculated using NumPy (Numeric Python). The coordinates of all the documents are determined from S , and these coordinates are merged with the query to obtain the query coordinates. Finally, a cosine similarity function is applied to these coordinates to find the documents that best match the query. The whole process can be done using following equations: The Eq. 3 determines the topics of given documents (DOC) with LDA and NMF.

$$Topic_{LDA,NMF} = U_{x=1}^n(LDA(DOC_i), NMF(DOC_i)) \quad (3)$$

The following Eq. 4 determines the LSI scores of each document with the LDA-topic.

$$LSI_{LDA} = U_{x=1}^n(LSI_i(Topic_{LDA}), (DOC_i)) \quad (4)$$

The Eq. 5 determines the LSI scores of each document with the NMF-topic.

$$LSI_{NMF} = U_{x=1}^n(LSI_i(Topic_{NMF}), (DOC_i)) \quad (5)$$

Now coordinates of documents and coordinates of topic will be determined by setting term weights and construct the term-document matrix $DOCS_{mat}$ from all documents and topic matrix $Topic_{mat}$ from Topic (LDA and NMF). Decompose matrix $DOCS_{mat}$ and find the v , S and u matrices using Singular Value Decomposition svd method of Numerical Python numpy package as Eq. 6:

$$v, S, u = numpy.linalg.svd(DOCS_{mat}, full_matrices) = True \quad (6)$$

V_k is a matrix by extracting first two column of V and each row of its inverse. Now transpose of V_k i.e., V_k^t relates to coordinates of document named as $DOCS_{coor}$. Coordinates of Topic can be determined by the product of transpose of $Topic_{mat}$, U_k and S_k^{-1} as depicted in following Eq. 7:

$$Topic_{coor} = (Topic_{mat})^t * U_k * S_k^{-1} \quad (7)$$

where U_k is matrix from extracting first two columns of U and S_k^{-1} is a matrix from extracting inverse of first two column and row of S . Hence decomposing Eq. 7 with respect to LDA and NMF, we can find coordinates of LDA-Topic and NMF-Topic using Eq. 8 and Eq. 9:

$$LDA - Topic_{coor} = (LDA - Topic_{mat})^t * U_k * S_k^{-1} \quad (8)$$

$$NMF - Topic_{coor} = (NMF - Topic_{mat})^t * U_k * S_k^{-1} \quad (9)$$

Now cosine similarities method will find the scores of each document-coordinates $Docs_{coor}$ from

LDA-Topic_{coor} in Eq. 10 and from NMF-Topic_{coor} in Eq. 11:

$$U_{x=1}^m(DOCS(x), LDA - Topic) = \frac{\sum_{i=1}^n DOCS_{coor}(i) * LDA - Topic_{coor}(i)}{\sqrt{\sum_{i=1}^n (DOCS_{coor}(i))^2} \sqrt{\sum_{i=1}^n (LDA - Topic_{coor}(i))^2}} \quad (10)$$

$$U_{x=1}^m(DOCS(x), NMF - Topic) = \frac{\sum_{i=1}^n DOCS_{coor}(i) * NMF - Topic_{coor}(i)}{\sqrt{\sum_{i=1}^n (DOCS_{coor}(i))^2} \sqrt{\sum_{i=1}^n (NMF - Topic_{coor}(i))^2}} \quad (11)$$

4. Results

We took 10 documents each from the political, sports, and medicine domains, each consisting of approximately 50 words. After applying LDA and NMF to these documents, the topics generated (as shown in Table 1, were similar, but there were some differences. LDA and NMF generate more than one

topic, but we only considered the first topic in our analysis. Each topic was no more than eight words.

4.1. Experimental results

To check the accuracy of LDA and NMF, we applied LSI to the documents from each domain.

4.1.1. LSI of medical documents

Table 2 shows the LSI scores of the LDA-topic and NMF-topic for each medical document. The average NMF score (0.725605836) is greater than the average LDA score (0.716558265). This means that the NMF-topic is a better match across all documents than the LDA-topic. The difference between the average scores is 0.0091, which is considered statistically significant.

Table 1: Topics generated by LDA and NMF

Domain	Methods: Topic
Medicine	LDA: drug drugs pharmaceutical effects generic pharmacist NMF: drugs drug effects pharmacology pharmaceutical study sources
Politics	LDA: science politics study power state capitalist said democracy NMF: science politics study society power state behavior theory
Sports	LDA: sports sport physical performance activity exercise NMF: sports physical performance sport exercise psychology activity

Table 2: LDA-topic and NMF-topic scores for medical documents

Documents	LDA	NMF
d[0]	0.434861789692	0.451060322379
d[1]	0.832238695082	0.84211965462
d[2]	0.172602828885	0.190370749154
d[3]	0.623740198795	0.637760328583
d[4]	0.894482221427	0.902414152172
d[5]	0.95325735576	0.947642564682
d[6]	0.998349660891	0.997149140947
d[7]	0.700823132745	0.713596747494
d[8]	0.971180708745	0.975328415051
d[9]	0.58404605613	0.598616286894
Average Score	0.716558265	0.725605836

methods are very close to each other but with little bit difference.

Table 3: LDA-topic and NMF-topic scores of political documents

Documents	LDA	NMF
d[0]	0.999994942202	0.993473589803
d[1]	0.861276684621	0.912317121888
d[2]	0.966790772781	0.989170117383
d[3]	0.975775271757	0.99401867454
d[4]	0.993861549892	0.999999962808
d[5]	0.358646320534	0.252909815727
d[6]	0.934696892854	0.968350737867
d[7]	0.994473533774	0.999982303212
d[8]	0.860085194999	0.911356355837
d[9]	0.923635874357	0.960443839016
Average Score	0.886923704	0.898202252

4.1.2. LSI on politics documents

Table 3 shows the LSI scores of the LDA-topic and the NMF-topic for each political document. The average NMF score (0.898202252) is greater than the average LDA score (0.886923704). This means that the NMF-topic is a better match across all documents than the LDA-topic. The difference between the average scores is 0.0112, which is considered statistically significant.

4.1.3. LSI on sports documents

Table 4 shows the LSI scores of the LDA-topic and NMF-topic for each sports document. The average NMF score (0.778685937) is lower than the average LDA score (0.779071217). This means that the LDA-topic is a better match across all documents than the NMF-topic. But the difference is only 0.00039, which is not considered statistically significant. Fig. 2 illustrates the comparison of the LDA and the NMF method with documents of different domain. The Fig. 2 clearly shows that the line for the NMF and LDA

Table 4: LDA-topic and NMF-topic scores of sports documents

Documents	LDA	NMF
d[0]	0.629462064395	0.622660949364
d[1]	0.435255235784	0.427386310229
d[2]	0.700007288043	0.693752056845
d[3]	0.727983942163	0.733935922151
d[4]	0.640482025053	0.647155832831
d[5]	0.997249385808	0.996564994993
d[6]	0.88280653158	0.878675965425
d[7]	0.979909541032	0.981611781909
d[8]	0.885040814574	0.889067302946
d[9]	0.91251533836	0.916048248927
Average Score	0.779071217	0.778685937

5. Conclusion and discussion

From the above analysis, it is clear that both methods work well for topic detection, but NMF-generated topics are slightly closer to the documents. After arranging these scores in descending order, all documents are arranged in the same order. In the medical domain, d[6] was very close to both the LDA-topic and the NMF-topic, while

d[2] was far from both, as shown in Table 5. In the political domain, d[0] was very close to the LDA-topic, d[4] was very close to the NMF-topic, and d[5] was far from both, as shown in Table 6. In this domain, only these three documents were in a

different order, rest of the documents were in the same order. In the sports domain, d[5] was very close to both the LDA-topic and the NMF-topic, while d[1] was far from both, as shown in Table 7.

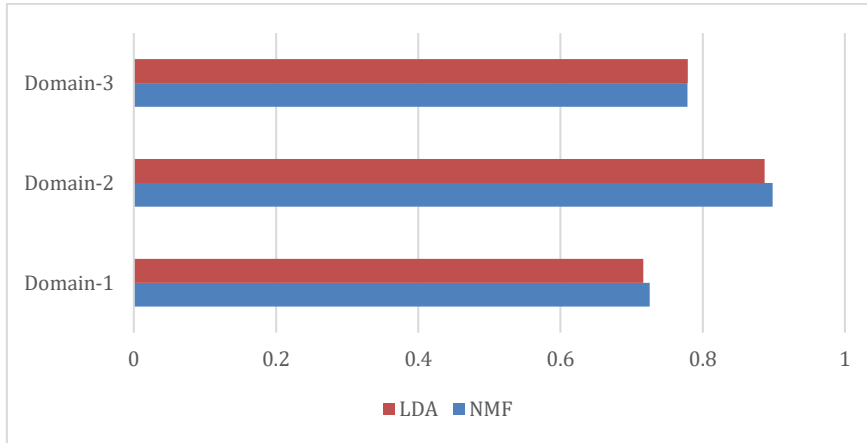


Fig. 2: Comparison of LDA and NMF

Table 5: Medical documents in descending order based on their LDA and NMF Scores

Document	LDA	Document	NMF
d[6]	0.998349661	d[6]	0.997149140947
d[8]	0.971180709	d[8]	0.975328415051
d[5]	0.953257356	d[5]	0.947642564682
d[4]	0.894482221	d[4]	0.902414152172
d[1]	0.832238695	d[1]	0.84211965462
d[7]	0.700823133	d[7]	0.713596747494
d[3]	0.623740199	d[3]	0.637760328583
d[9]	0.584046056	d[9]	0.598616286894
d[0]	0.43486179	d[0]	0.451060322379
d[2]	0.172602829	d[2]	0.190370749154

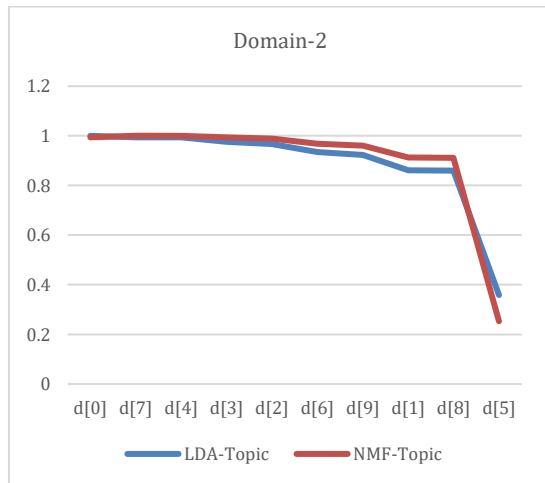
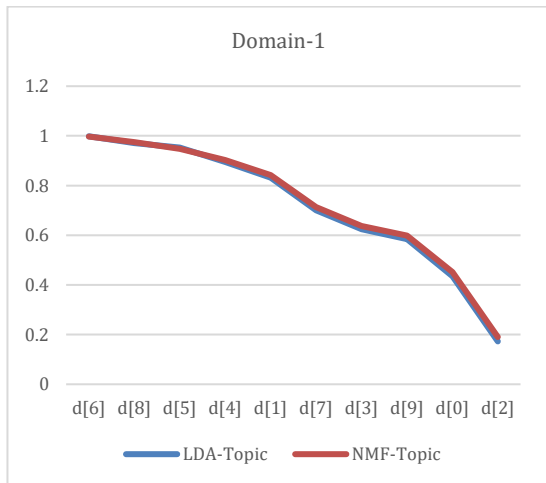
Table 7: Sports documents in descending order based on their LDA and NMF scores

Document	LDA	Document	NMF
d[5]	0.997249385808	d[5]	0.996564994993
d[7]	0.979909541032	d[7]	0.981611781909
d[9]	0.91251533836	d[9]	0.916048248927
d[8]	0.885040814574	d[8]	0.889067302946
d[6]	0.88280653158	d[6]	0.878675965425
d[3]	0.727983942163	d[3]	0.733935922151
d[2]	0.700007288043	d[2]	0.693752056845
d[4]	0.640482025053	d[4]	0.647155832831
d[0]	0.629462064395	d[0]	0.622660949364
d[1]	0.435255235784	d[1]	0.427386310229

Table 6: Political documents in descending order based on their LDA and NMF scores

Document	LDA	Document	NMF
d[0]	0.999994942202	d[4]	0.999999962808
d[7]	0.994473533774	d[7]	0.999982303212
d[4]	0.993861549892	d[3]	0.99401867454
d[3]	0.975775271757	d[0]	0.993473589803
d[2]	0.966790772781	d[2]	0.989170117383
d[6]	0.934696892854	d[6]	0.968350737867
d[9]	0.923635874357	d[9]	0.960443839016
d[1]	0.861276684621	d[1]	0.912317121888
d[8]	0.860085194999	d[8]	0.911356355837
d[5]	0.358646320534	d[5]	0.252909815727

From Table 5, Table 6, and Table 7, it is clear that both the LDA and NMF topics had the same relevancy for each document, but as whole, the average LSI score of NMF was greater than that of LDA. After the analysis of both models based on scores generated using LSI, it is very hard to determine the best model for topic detection in text mining, but NMF can be considered slightly better than LDA as depicted in Fig. 3.



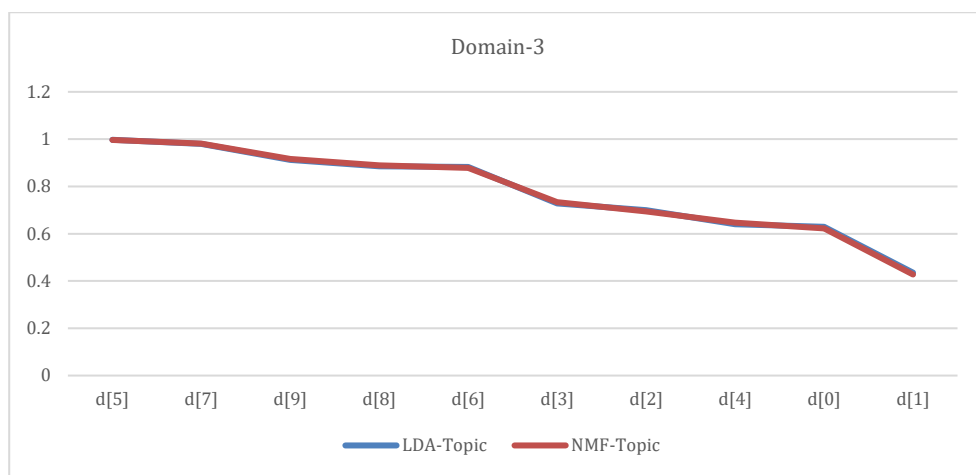


Fig. 3: Sorted documents with respect to LDA-topic and NMF-topic

6. Limitations and future work

We used datasets from only three domains; further domains could also be analyzed. We also limited each document to 50 words and each topic to 8 words, which could be increased or decreased. More than one topic can be generated using both LDA and NMF; in this study, we used only the first topic. Furthermore, the topics themselves could also be considered for analysis. We used well-written documents without noise, whereas if we want to detect topics from user reviews, there is still a need for further study, as user reviews contain mistakes, omitted words, incomplete sentences, and misspellings.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Agrawal A, Fu W, and Menzies T (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, 98: 74-88. <https://doi.org/10.1016/j.infsof.2018.02.005>
- Alshammari R (2018). Arabic text categorization using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, 9(3): 226-230. <https://doi.org/10.14569/IJACSA.2018.090332>
- Awajan AA (2014). Unsupervised approach for automatic keyword extraction from Arabic documents. In the 26th Conference on Computational Linguistics and Speech Processing, The Association for Computational Linguistics and Chinese Language Processing, Jhongli, Taiwan: 175-184.
- Blei D, Carin L, and Dunson D (2010). Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE Signal Processing Magazine*, 27(6): 55-65. <https://doi.org/10.1109/MSP.2010.938079>
PMid:25104898 PMCID:PMC412269
- Blei DM and McAuliffe JD (2010). Supervised topic models. In *Proceedings of NIPS*, Vancouver, Canada. Available online at: <https://bit.ly/2ZAq00A>
- Blei DM, Ng AY, and Jordan MI (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.
- Boyd-Graber J and Resnik P (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Cambridge, USA: 45-55.
- Chawla R (2017). Topic modeling with LDA and NMF on the ABC News headlines dataset. Available online at: <https://bit.ly/2MljpGA>
- Chen Y and Filliat D (2015). Cross-situational noun and adjective learning in an interactive scenario. In the 2015 Joint IEEE International Conference on Development and Learning and Epigenetic, IEEE, Providence, USA: 129-134. <https://doi.org/10.1109/DEVLRN.2015.7346129>
- Chen Y, Bordes JB, and Filliat D (2017). An experimental comparison between NMF and LDA for active cross-situational object-word learning. In the 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), IEEE, Cergy-Pontoise, France: 217-222. <https://doi.org/10.1109/DEVLRN.2016.7846822>
- Chen Y, Rege M, Dong M, and Hua J (2008). Non-negative matrix factorization for semi-supervised data clustering. *Knowledge and Information Systems*, 17(3): 355-379. <https://doi.org/10.1007/s10115-008-0134-6>
- Chinsha TC and Joseph S (2015). A syntactic approach for aspect based opinion mining. In the 2015 IEEE 9th International Conference on Semantic Computing, IEEE, Anaheim, USA: 24-31. <https://doi.org/10.1109/ICOSC.2015.7050774>
- Deerwester S, Dumais ST, Furnas GW, Landauer TK, and Harshman R (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391-407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-AS11>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-AS11>3.0.CO;2-9)
- El-Fishawy N, Hamouda A, Attiya GM, and Atef M (2014). Arabic summarization in twitter social network. *Ain Shams Engineering Journal*, 5(2): 411-420. <https://doi.org/10.1016/j.asej.2013.11.002>
- Gamon M, Aue A, Corston-Oliver S, and Ringger E (2005). Pulse: Mining customer opinions from free text. In the International Symposium on Intelligent Data Analysis, Springer, Hertogenbosch, Netherlands: 121-132. https://doi.org/10.1007/11552253_12
- Gao S and Li H (2011). A cross-domain adaptation method for sentiment classification using probabilistic latent analysis. In the 20th ACM International Conference on Information and Knowledge Management, ACM, Glasgow, Scotland: 1047-1052. <https://doi.org/10.1145/2063576.2063728>

- Gindl S, Weichselbraun A, and Scharl A (2013). Rule-based opinion target and aspect extraction to acquire affective knowledge. In the 22nd International Conference on World Wide Web, ACM, Rio de Janeiro, Brazil: 557-564.
<https://doi.org/10.1145/2487788.2487994>
- Gojali S and Khodra ML (2016). Aspect based sentiment analysis for review rating prediction. In the 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, IEEE, George Town, Malaysia: 1-6.
<https://doi.org/10.1109/ICAICTA.2016.7803110>
- Guo H, Zhu H, Guo Z, Zhang X, and Su Z (2009). Product feature categorization with multilevel latent semantic association. In the 18th ACM Conference on Information and Knowledge Management, ACM, Hong Kong, China: 1087-1096.
<https://doi.org/10.1145/1645953.1646091> **PMid:19757454**
- Gupta DK and Ekbal A (2014). IITP: Supervised machine learning for aspect based sentiment analysis. In the 8th International Workshop on Semantic Evaluation, Dublin, Ireland: 319-323.
<https://doi.org/10.3115/v1/S14-2053>
- He Y, Lin C, and Alani H (2011). Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, USA, 1: 123-131.
- Hu M and Liu B (2004). Mining and summarizing customer reviews. In the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Seattle, USA: 168-177.
<https://doi.org/10.1145/1014052.1014073>
- Huang A, Milne D, Frank E, and Witten IH (2009). Clustering documents using a Wikipedia-based concept representation. In the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Macau, China: 628-636.
https://doi.org/10.1007/978-3-642-01307-2_62
- Jeyapriya A and Selvi CK (2015). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In the 2015 2nd International Conference on Electronics and Communication Systems, IEEE, Coimbatore, India: 548-552.
<https://doi.org/10.1109/ECS.2015.7124967>
- Jo Y and Oh AH (2011). Aspect and sentiment unification model for online review analysis. In the 4th ACM International Conference on Web Search and Data Mining, ACM, Hong Kong, China: 815-824.
<https://doi.org/10.1145/1935826.1935932>
- Leek T, Jin H, Sista S, and Schwartz R (2000). The BBN cross lingual topic detection and tracking system. In The Working Notes of the Third Topic Detection and Tracking Workshop, BBN Technologies, Cambridge, USA.
- Li S, Lee SYM, Chen Y, Huang CR, and Zhou G (2010). Sentiment classification and polarity shifting. In the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, Beijing, China: 635-643.
- Lin C and He Y (2009). Joint sentiment/topic model for sentiment analysis. In the 18th ACM Conference on Information and Knowledge Management, ACM, Hong Kong, China: 375-384.
<https://doi.org/10.1145/1645953.1646003> **PMCID:PMC2779244**
- Liu H and Wu Z (2010). Non-negative matrix factorization with constraints. In the 24th AAAI Conference on Artificial Intelligence, AAAI Press, Atlanta, USA: 506-511.
- Lu Y, Zhai C, and Sundaresan N (2009). Rated aspect summarization of short comments. In the 18th International Conference on World Wide Web, ACM, Madrid, Spain: 131-140.
<https://doi.org/10.1145/1526709.1526728> **PMCID:PMC3280738**
- MacMillan K and Wilson JD (2017). Topic supervised non-negative matrix factorization. Available online at:
<https://bit.ly/30I6WWM>
- Mangin O, Filliat D, Ten Bosch L, and Oudeyer PY (2015). MCA-NMF: Multimodal concept acquisition with non-negative matrix factorization. PloS One, 10(10): e0140732.
<https://doi.org/10.1371/journal.pone.0140732> **PMid:26489021 PMCID:PMC4619362**
- Mei Q, Ling X, Wondra M, Su H, and Zhai C (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In the 16th International Conference on World Wide Web, ACM, Banff, Alberta, Canada: 171-180.
<https://doi.org/10.1145/1242572.1242596> **PMid:17825604**
- Moghaddam S and Ester M (2011). ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Beijing, China: 665-674.
<https://doi.org/10.1145/2009916.2010006>
- Mukherjee A and Liu B (2012). Aspect extraction through semi-supervised modeling. In the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, Association for Computational Linguistics, Jeju Island, South Korea, 1: 339-348.
- Pang B and Lee L (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2): 1-135.
<https://doi.org/10.1561/1500000011>
- Pantel P, Crestan E, Borkovsky A, Popescu AM, and Vyas V (2009). Web-scale distributional similarity and entity set expansion. In the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, Singapore, 2: 938-947.
<https://doi.org/10.3115/1699571.1699635>
- Peng W and Park DH (2011). Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In the 50th International AAAI Conference on Weblogs and Social Media, Barcelona, Spain: 273-280.
- Phadnis N and Gadge J (2014). Framework for document retrieval using latent semantic indexing. International Journal of Computer Applications, 94(14): 37-41.
<https://doi.org/10.5120/16414-6065>
- Qi L and Chen L (2011). Comparison of model-based learning methods for feature-level opinion mining. In the 2011 International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, Washington, USA, 1: 265-273.
<https://doi.org/10.1109/WI-IAT.2011.64>
- Raganato A, Camacho-Collados J, and Navigli R (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Valencia, Spain, 1: 99-110.
<https://doi.org/10.18653/v1/E17-1010>
- Rammal M, Bahsoun Z, and Al Achkar Jabbour M (2015). Keyword extraction from Arabic legal texts. Interactive Technology and Smart Education, 12(1): 62-71.
<https://doi.org/10.1108/ITSE-11-2013-0030>
- Rose S, Engel D, Cramer N, and Cowley W (2010). Automatic keyword extraction from individual documents. In: Berry MW and Kogan J (Eds.), Text mining: Applications and theory: 1-20. John Wiley and Sons, Hoboken, USA.
<https://doi.org/10.1002/9780470689646.ch1>
- Saqib SM, Mahmood K, and Naem T (2016). Comparison of LSI algorithms without and with pre-processing: Using text document based search. ACCENTS Transactions on Information Security, 1(4): 44-51.

- Saupér C, Haghghi A, and Barzilay R (2011). Content models with attitude. In the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, USA, 1: 350-358.
- Stevens K, Kegelmeyer P, Andrzejewski D, and Buttler D (2012). Exploring topic coherence over many models and many topics. In the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, Jeju Island, South Korea: 952-961.
- Suri P and Roy NR (2017). Comparison between LDA & NMF for event-detection from large text stream data. In the 2017 3rd International Conference on Computational Intelligence and Communication Technology (CICT), IEEE, Ghaziabad, India: 1-5.
<https://doi.org/10.1109/CICT.2017.7977281>
PMid:30241224
- Taniguchi T, Yoshino R, and Takano T (2018). Multimodal hierarchical Dirichlet process-based active perception by a robot. *Frontiers in Neurorobotics*, 12: 22.
<https://doi.org/10.3389/fnbot.2018.00022>
PMid:29872389 PMCID:PMC5972223
- Thakur D and Singh J (2015). The SAFE miner: A fine grained aspect level approach for resolving the sentiment. In the 2015 3rd International Conference on Computer, Communication, Control and Information Technology, IEEE, Hooghly, India: 1-6.
<https://doi.org/10.1109/C3IT.2015.7060151>
- Tumasjan A, Sprenger TO, Sandner PG, and Welpé IM (2010). Predicting elections with twitter: What 140 characters' reveal about political sentiment. In the 4th International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, Washington, USA: 178-185.
- Wang SH, Ding Y, Zhao W, Huang YH, Perkins R, Zou W, and Chen JJ (2016). Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC Public Health*, 16: 279.
<https://doi.org/10.1186/s12889-016-2932-1>
PMid:26993983 PMCID:PMC4799597
- Xue Y, Tong CS, and Yuan JY (2014). LDA-based non-negative matrix factorization for supervised face recognition. *Journal of Software*, 9(5): 1294-1301.
<https://doi.org/10.4304/jsw.9.5.1294-1301>
- Yang Q and Li FM (2005). Support vector machine for customized email filtering based on improving latent semantic indexing. In the 2005 International Conference on Machine Learning and Cybernetics, IEEE, Guangzhou, China, 6: 3787-3791.
<https://doi.org/10.1109/ICMLC.2005.1527599>
- Zhai Z, Liu B, Xu H, and Jia P (2011). Clustering product features for opinion mining. In the 4th ACM International Conference on Web Search and Data Mining, ACM, Hong Kong, China: 347-354.
<https://doi.org/10.1145/1935826.1935884>
- Zhao WX, Jiang J, Yan H, and Li X (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Cambridge, USA: 56-65.
- Zhuang L, Jing F, and Zhu XY (2006). Movie review mining and summarization. In the 15th ACM International Conference on Information and Knowledge Management, ACM, Arlington, USA: 43-50.
<https://doi.org/10.1145/1183614.1183625>