A Performance Evaluation of Topic Models based on Fuzzy Latent Semantic Analysis

A Performance Evaluation of Topic Models based on Fuzzy Latent Semantic Analysis

Emil Rijcken[a,b,*], Kalliopi Zervanou[d,e], Pablo Mosteiro[b], Marco Spruit[d,e], Floortje Scheepers[c], Uzay Kaymak[a]

[a]*Jheronimus Academy of Data Science, Eindhoven University of Technology, Sint Janssingel 92, 's-Hertogenbosch, 5211 DA, The Netherlands*
[b]*Department of Information and Computing Sciences, Utrecht University, Heidelberglaan 8, Utrecht, 3584 CS, The Netherlands*
[c]*Psychiatry, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3584 CX, The Netherlands*
[d]*Leiden Institute of Advanced Computer Science, Leiden University, Rapenburg 70, Leiden, 2311 EZ, The Netherlands*
[e]*Public Health & Primary Care, Leiden University Medical Center, Albinusdreef 2, Leiden, 2333 ZA, The Netherlands*

Abstract

Topic models are popular unsupervised statistical methods that extract hidden topics underlying a collection of documents. Several algorithms based on Fuzzy Latent Semantic Analysis have been proposed recently, which use fuzzy clustering. We argue that using fuzzy algorithms to find topics is more intuitive than non-fuzzy-set-based algorithms. The algorithms have shown promising results over existing (non-fuzzy) algorithms. However, they have not been evaluated systematically on open benchmark datasets. This work compares the three algorithms FLSA, FLSA-W and FLSA-V with other state-of-the-art topic modeling algorithms on four open datasets, using three different evaluation metrics. We find that for each evaluation metric, FLSA-W outperforms all the other algorithms in most settings.

Keywords: Topic modeling, NLP, Fuzzy methods, Fuzzy clustering, Unsupervised learning, Information retrieval

1. Introduction

A common and popular task within Natural Language Processing (NLP) is topic modeling, which aims to extract hidden topics from a collection of documents. Some applications of topic modeling include: text categorization [1], finding text similarity [2], text classification [3] and sentiment analysis [4]. In some applications, the topic modeling output is used as input for another downstream task (e.g. text classification) [5], whereas finding latent topics is the main goal for other tasks. Recently, the fuzzy-set-based topic modeling algorithms Fuzzy Latent Semantic Analysis (FLSA)[5], FLSA-W and FLSA-V [6] have been proposed. We argue that using fuzzy algorithms to find topics is more intuitive than non-fuzzy-set-based algorithms. Moreover, these algorithms show promising results compared to the other non-fuzzy-based algorithms. However, a systematic comparison is missing; one that includes various state-of-the-art algorithms is conducted on open datasets and includes various performance metrics. FLSA is compared with Latent Dirichlet Allocation (LDA) in [5]. Yet, only the predictive performance is considered. FLSA, FLSA-W and FLSA-V are compared with LDA in terms of coherence score based on a private dataset in [6]. Lastly, FLSA, FLSA-W and FLSA-V are compared with various other algorithms in terms of inter- and intra-topic performance metrics in [7]. Yet, these experiments are all conducted on a private dataset and are not confirmed with other datasets. This paper compares FLSA, FLSA-W and FLSA-V with existing state-of-art topic models on the four open datasets BBC-news [8], DBLP[1], M10 [9] and 20Newsgroup [10]. We train all the topic models on each dataset with a different number of topics: 10, 20, ... 100. For robustness, each experiment is repeated ten times. We report the average score of the standard evaluation metrics: interpretability-, coherence- and diversity score for each setting; the combination of the dataset, algorithm and number of topics. We find that in most settings, FLSA-W outperforms all the other algorithms for each evaluation metric.
The outline of the paper is as follows. In Section 2, we discuss the different algorithms and performance indicators

*Corresponding author: e.f.g.rijcken@tue.nl. Sint Janssingel 92, 5211 DA 's-Hertogenbosch, The Netherlands
[1]https://github.com/shiruipan/TriDNR/tree/master/data

used for the experiments. In Section 3, we discuss the datasets and the methodology used for the experiments and share the results in Section 4. We discuss the findings and implications in Section 5. Lastly, we conclude the paper in Section 6.

2. Topic Modeling Algorithms

Many different topic modeling algorithms exist [11], but they all return two matrices: **P**(**W**|**T**) and **P**(**T**|**D**), where $p(W_i|T_k)$ gives the probability of word i given topic k, and $p(T_k|D_j)$ gives the probability of topic k given document j. The topics are retrieved from **P**(**W**|**T**) and consist of the n most likely words associated with that topic and their associated probabilities. This section discusses the different topic modeling algorithms, with a detailed analysis of FLSA-W. Then, we end with a discussion of the evaluation metrics used.

2.1. Topic Modeling

This work compares the FLSA-based models [5], [6] with a set of state-of-the-art topic modeling algorithms, as defined in [12]. The most popular topic modeling algorithm is Latent Dirichlet Allocation (LDA) [13]. The algorithms discussed in this section can be roughly divided into LDA-based models and methods based on dimensionality reduction.

Starting with dimensionality-reduction-based models, the earliest proposed model used for our experiments is 'Non-negative Matrix Factorization' (NMF) [14]. NMF is an unsupervised method in which a corpus is represented as a document-term matrix. Using the document-term-matrix **A**, NMF returns two matrices: **W** (topics × words) & **H** (documents × topics). **W** contains the found topics and **H** contains the coefficients. Then, NMF modifies **W** and **H**'s initial values so that its product approaches **A**. NMF tends to produce high-quality topics on short texts [15]. Another foundational work on topic modeling is Latent Semantic Analysis (LSA) (also referred to as Latent Semantic Indexing), which uses Singular Value Decomposition (SVD) for dimensionality reduction of the document-term matrix [16]. The **V** (words × singular values) and **U** (singular values × documents) matrices produced by SVD are then used to represent topics. Similarly to LSA, FLSA starts with a document-term representation and uses SVD for dimensionality reduction [5]. Yet, after representing the corpus in a document-term matrix, a global term weighting mechanism is used before dimensionality reduction. Then, FLSA takes SVD's **U** matrix (singular values × documents) and performs fuzzy c-means clustering [17] to find different topics. Lastly, it uses Bayes' theorem and linear algebra to find the two output matrices.

Dimensionality reduction methods implicitly assume that the order of words in a document, and the order of documents in a corpus, can be neglected. De Finetti's representation theorem [18] establishes that any collection of exchangeable random variables has a representation as a mixture distribution, the probability distribution of a random variable derived from a collection of other random variables. Because of this, mixture models that capture the exchangeability of words and documents should be used to consider exchangeable representations for both. This line of thought paves the way to LDA [13], the best-known topic modeling algorithm on which multiple other topic models are based. LDA posits that each document can be seen as a probability distribution over topics and that each topic can be seen as a probability distribution over words. A random sample is drawn to represent the topic distribution from a Dirichlet distribution, a multivariate generalization of the beta distribution. Then, a random sample is selected from another Dirichlet distribution to represent the word distribution. Although the posterior distribution is intractable for exact inference, many approximate inference algorithms can be considered for LDA. Popular algorithms are mean-field methods and collapsed Gibbs sampling [19]. The drawback of both algorithms is that applying them to new data requires the inference model to be re-derived, which can be time-consuming. Black-box inference models have been created to overcome this drawback, require minimal- and easy-to-compute information from the model, and can be automatically applied to new models [20]. An inference network [21] is trained through autoencoding variational Bayes (AEVB) [21]; a neural network that directly maps the document-term representation of a document onto a continuous latent representation. A decoder network then reconstructs the document-term matrix by generating its words from the latent document representation [22]. ProdLDA and NeuralLDA [20] are the first topic modeling algorithms that use AEVB inference methods. ProdLDA models the probability distribution over individual words by combining the output from several simpler distributions, whereas NeuralLDA uses mixture models. ProdLDA is reported to produce topics with higher coherence than LDA and trains faster [20].

The embedded topic model (ETM) is proposed to overcome LDA's difficulty dealing with large vocabularies [23]. ETM is a generative model of documents that marries traditional topic models with word embeddings. The ETM models each word as a categorical distribution of which the natural parameter is the inner product between the word's embedding and an embedding of its assigned topic. The produced topics by ETM are reported to have higher coherence- and interpretability scores than LDA and ProdLDA [23].

2.2. *FLSA-W & FLSA-V*

FLSA uses singular value decomposition to decompose document representations into three matrices $\mathbf{U\Sigma V}^T$ [5]. Then, FLSA clusters on the $\mathbf{U}$ matrix, which means in clusters documents. Yet, clustering words might make more sense as topics consist of words. This idea has been the basis for FLSA-W and FLSA-V, discussed below. The following quantities are defined:

M number of unique words in the data set,
N number of documents in the data set,
C number of topics,
S number of SVD dimensions,
i word index $i \in \{1, 2, 3, ..., M\}$,
j document index $j \in \{1, 2, 3, ..., N\}$,
k topic index $k \in \{1, 2, 3, ..., C\}$,
tf_{ij} the number of times that word i occurs in document j,
$b(tf_{ij})$ a binary value indicating whether term i occurs in document j, calculated as: $\begin{cases} 1, & t_{ij} > 0 \\ 0, & t_{ij} = 0, \end{cases}$
p_{ij} the probability that word i appears in document j, calculated as: $t_{ij} / \sum_j t_{ij}$.

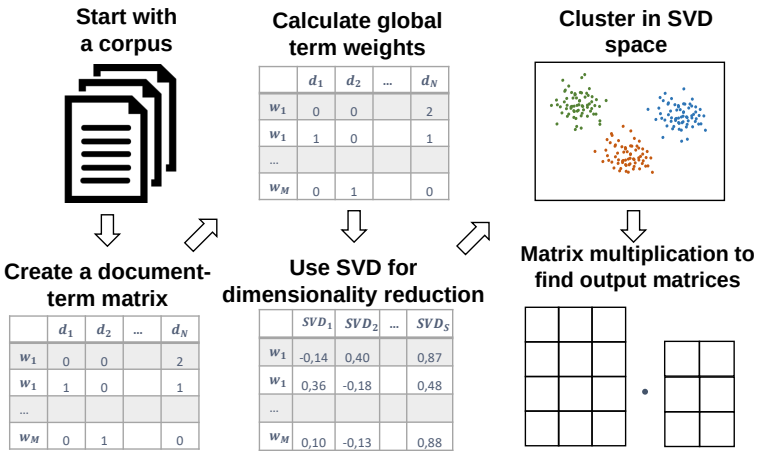

Figure 1: Visual representation of FLSA and FLSA-W's steps. The difference between both methods is that FLSA clusters in SVD's $\mathbf{U}$ matrix and FLSA-W clusters in $\mathbf{V^T}$

.

Then, FLSA-W is calculated with the following steps [6]:

1. Calculate local term weighting (LTW), a $(N \times M)$ matrix that indicates how much each word occurs in each document.

2. Calculate global term weights (GTW) by multiplying the global weighting vectors element-wise with the LTW matrix, to obtain $p(W_i, D_j)$. Previously defined methods for calculating the global weighting vector include:

$$Entropy = 1 + \frac{\sum_j p_{ij} log_2(p_{ij})}{log_2 n}, \quad (1)$$

$$IDF = log_2 \frac{n}{\sum_j b(tf_{ij})}, \quad (2)$$

$$Normal = \frac{1}{\sqrt{\sum_j tf_{ij}^2}}, \quad (3)$$

$$ProbIDF = log_2 \frac{n - \sum_j b(tf_{ij})}{\sum_j b(tf_{ij})}. \quad (4)$$

3. For dimensionality reduction, perform singular value decomposition (SVD) on the GTW matrix. SVD's output is the decomposed matrix $\mathbf{U\Sigma V}^T$, from which we select the largest S singular values.

4. Perform fuzzy clustering on $\mathbf{V^T}$ to obtain $\mathbf{P(T|W)^T}$ $(M \times S)$.

5. Calculate probability vectors:

$$p(D_j) = \frac{\sum_{i=1}^{M} p(W_i, D_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} p(W_i, D_j)}, \quad (5)$$

$$p(W_i) = \frac{\sum_{j=1}^{N} p(W_i, D_j)}{\sum_{i=1}^{m} \sum_{j=1}^{N} p(W_i, D_j)}, \quad (6)$$

$$p(T_k) = p(T_k|W_i)p(W_i). \quad (7)$$

6. Calculate the probability of word i, given topic k

$$p(W_i, T_k) = (p(T_k|W_i) \otimes p(W_i))^T, \quad (8)$$

$$p(W_i|T_k) = \frac{p(W_i, T_k)}{\sum_{i=1}^{M} p(W_i, T_k)}. \quad (9)$$

   $\otimes$ represents element-wise multiplication.

7. Calculate the probability of topic k, given document j

$$p(W_i|D_j) = \frac{p(W_i, D_j)}{\sum_{i=1}^{M} p(W_i, D_j)}, \quad (10)$$

$$p(D_j|W_i) = \frac{(p(W_i|D_j) \otimes p(D_j))^T}{p(W_i)}, \quad (11)$$

$$p(D_j|T_k) = \sum_{i=1}^{M} p(D_j|W_i)p(W_i|T_k), \quad (12)$$

$$p(T_k|D_j) = \frac{(p(D_j|T_k) \otimes p(T_k))^T}{p(D_j)}. \quad (13)$$

The implicit assumption with FLSA and FSA-W is that semantically related words are located near each other in the lower-dimensional projection. Yet, there is no step explicitly guarantying this meaningfulness. Instead, FLSA-V uses the software tool VOSviewer [24] to project words into a low-dimensional manifold and then performs the same steps as FLSA-W. However, FLSA-V can only be trained on small datasets as VOSviewer runs into memory issues with large corpora [2].

[2]For this reason, FLSA-V cannot be trained on the BBC-news dataset.

2.3. Evaluation Metrics

We use quantitative evaluation metrics instead of qualitative measures for two reasons. Firstly, our goal is to maintain a high level of objectivity. Using qualitative measures would add a degree of subjectivity. Secondly, the time required for qualitatively assessing all topics is infeasible, given that we train 3500 topic models. A topic model's output consists of various topics containing a collection of words. Because the output consists of a set of sets, a topic model's evaluation metric should focus on the quality of words within each topic (intra-topic quality) and the quality between different topics (inter-topic quality). For example, it is not very useful if a topic model contains many topics with the same words. Similarly, topics with a different focus but unrelated words are not useful either. We use the coherence-, diversity- and interpretability score in this work to evaluate inter-, intra-topic- and overall quality. We will discuss each of these metrics in detail in the following subsections.

2.3.1. Coherence Score

The coherence score measures how well each topic's words support each other. We use c_v for calculating the coherence- and interpretability scores as it correlates highest with human topic ranking data [25]. With c_v, the Normalized Pointwise Mutual Information (15) is calculated for the combination of all the top-n words in a topic. Then, the arithmetic mean is calculated based on all these scores. To calculate the probabilities in Normalized Pointwise Mutual Information (NPMI), a sliding window of 110 words is used. NPMI is based on the assumption that semantically related words occur nearby each other. A downside of NPMI is that synonyms are not captured, typically.

$$PMI(w_i, w_j) = log \frac{p(w_i, w_j) + \epsilon}{p(w_i) \cdot p(w_j)} \tag{14}$$

$$NPMI(w_i, w_j)^{\gamma} = \frac{PMI}{\sum_{i=1}^{M} \sum_{j=1}^{N} p(W_i, D_j)}^{\gamma} \tag{15}$$

The coherence score ranges between zero and one, where one means perfect coherence and zero means no coherence.

2.3.2. Diversity Score

A measure for inter-topic quality is *topic diversity*, which measures the unique words in a topic model as a proportion to the total number of words [23]. If the topic diversity equals one, different topics do not share common words, whereas a zero value indicates that all topics contain the same n words. The following quantities are defined:

W_{unique}	the number of unique words in the top-n words of all the topics,
W_{all}	the total number of words in all the topics ($n \times C$),
n	the number of words per topic,
C	the number of topics.

Then, Equation (16) illustrates that the diversity score is calculated as the fraction of the unique words in all the topics and the total number of words in all the topics.

$$Diversity = \frac{W_{unique}}{W_{all}}. \tag{16}$$

Since $W_{unique} =< W_{all}$, diversity ranges between zero and one as well. One means that no topic shares words with other topics, and zero means that all the topics contain the same words.

2.3.3. Interpretability Score

The *interpretability* score combines intra- and inter-topic quality by taking the product between the coherence- and diversity score [23]. Equation (17) illustrates that the interpretability score is calculated as the product of the coherence- and the diversity score [23].

$$Interpretability = Coherence \times Diversity \tag{17}$$

The interpretability score also ranges between zero and one since it is the product of two variables with the same range.

3. Data & Experimental Setup

This section provides more information about the four datasets used and discusses the methodology of the experimental setup.

3.1. Datasets

We use the datasets provided by the OCTIS package, a Python package for optimizing and comparing topic modeling algorithms [12]. The four datasets that we train our algorithms on are: BBC-News [8], DBLP [3], M10 [9] and 20NewsGroup [10]. These datasets have been preprocessed by: stopwords-, punctuation- and number filtering, lowercase folding, word lemmatization and filtering of high- and low-frequency words. Table 1 shows descriptive statistics for each preprocessed dataset.

Table 1: Descriptive Statistics Four Open Datasets.

Dataset	**Number of Texts**	**Unique Words**	**Words per Text**
BBC-News	2225	2949	120.1
DBLP	54595	1513	5.4
M10	8355	1696	5.9
20News	16309	1612	48.1

VOSviewer cannot handle the BBC-News dataset size and would require pruning, while the DBLP-, M10 and 20NewsGroup dataset have not been pruned. For this reason, we do not train FLSA-V on the BBC-News dataset.

3.2. Experimental Setup

We test each model on a number of topics: 10, 20, ..., 100, and we calculate the three evaluation scores on the top 15 words per topic. Then, we train each setting (the combination of dataset, algorithm and number of topics) ten times and report the average interpretability-, coherence- and diversity scores per setting. For all models, we use the default parameter settings provided by the packages. For the FLSA-based models, we use 'normal' global term weights, two SVD factors, and fuzzy C-means clustering [17]. The packages used for topic modeling are FuzzyTM [26] and OCTIS [12]; FuzzyTM uses the pyFume [27] for fuzzy clustering.

4. Results

Tables 2, 3 & 4 show the mean scores per evaluation metric per setting. There is no algorithm with more settings with the highest score for all three evaluation metrics than FLSA-W. In the following subsections, we discuss the results per evaluation metric.

[3] https://github.com/shiruipan/TriDNR/tree/master/data

Table 2: Mean interpretability scores for different datasets, algorithms and number of topics.

BBC-news										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	0.388	0.395	0.403	0.413	**0.414**	**0.412**	**0.411**	**0.406**	**0.404**	**0.392**
FLSA-V	-	-	-	-	-	-	-	-	-	-
FLSA	0.35	0.284	0.277	0.282	0.275	0.274	0.281	0.283	0.281	0.282
LDA	0.141	0.117	0.109	0.101	0.09	0.09	0.086	0.087	0.083	0.083
NeuralLDA	0.464	0.414	0.397	0.375	0.365	0.352	0.346	0.321	0.317	0.302
ProdLDA	**0.593**	**0.504**	**0.459**	**0.418**	0.342	0.313	0.256	0.252	0.245	0.224
LSI	0.268	0.183	0.139	0.119	0.105	0.091	0.084	0.075	0.071	0.068
ETM	0.265	0.11	0.121	0.143	0.122	0.09	0.081	0.04	0.031	0.053
NMF	0.28	0.239	0.211	0.191	0.182	0.169	0.165	0.157	0.151	0.144
DBLP										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**0.462**	**0.451**	**0.447**	**0.429**	**0.429**	**0.416**	**0.404**	**0.384**	**0.362**	**0.338**
FLSA-V	0.229	0.262	0.258	0.246	0.226	0.205	0.189	0.169	0.149	0.132
FLSA	0.106	0.079	0.063	0.053	0.048	0.045	0.042	0.039	0.038	0.037
LDA	0.25	0.274	0.269	0.268	0.272	0.274	0.275	0.274	0.275	0.269
NeuralLDA	0.28	0.265	0.258	0.231	0.214	0.205	0.191	0.173	0.167	0.159
ProdLDA	0.446	0.403	0.28	0.21	0.177	0.154	0.141	0.126	0.117	0.114
LSI	0.127	0.076	0.066	0.059	0.052	0.046	0.041	0.038	0.037	0.036
ETM	0.019	0.01	0.009	0.009	0.011	0.012	0.012	0.012	0.012	0.012
NMF	0.237	0.231	0.215	0.202	0.189	0.181	0.171	0.165	0.157	0.149
M10										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**0.562**	**0.539**	**0.527**	**0.51**	**0.484**	**0.45**	**0.43**	**0.416**	**0.399**	**0.38**
FLSA-V	0.447	0.366	0.265	0.201	0.159	0.132	0.114	0.099	0.088	0.08
FLSA	0.195	0.182	0.221	0.224	0.245	0.252	0.253	0.249	0.24	0.24
LDA	0.17	0.2	0.221	0.237	0.249	0.255	0.264	0.26	0.257	0.246
NeuralLDA	0.377	0.309	0.282	0.251	0.235	0.212	0.203	0.19	0.181	0.172
ProdLDA	0.406	0.276	0.262	0.247	0.236	0.247	0.245	0.245	0.241	0.244
LSI	0.186	0.127	0.095	0.081	0.076	0.071	0.067	0.065	0.064	0.063
ETM	0.097	0.016	0.009	0.007	0.005	0.005	0.005	0.004	0.004	0.004
NMF	0.236	0.212	0.19	0.182	0.174	0.169	0.166	0.163	0.157	0.158
20NewsGroup										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	0.391	0.38	0.369	0.359	0.339	0.322	0.306	0.281	0.256	0.237
FLSA-V	0.191	0.168	0.148	0.142	0.126	0.111	0.098	0.086	0.078	0.07
FLSA	0.373	0.343	0.286	0.257	0.237	0.215	0.201	0.194	0.183	0.175
LDA	0.353	0.378	0.391	0.372	0.371	0.365	0.362	0.353	0.351	0.348
NeuralLDA	0.399	0.4	0.388	0.361	0.351	0.31	0.32	0.273	0.263	0.268
ProdLDA	**0.509**	**0.552**	**0.516**	**0.48**	**0.43**	**0.424**	**0.381**	**0.341**	**0.331**	**0.321**
LSI	0.3	0.229	0.188	0.15	0.13	0.111	0.102	0.093	0.085	0.079
ETM	0.29	0.197	0.163	0.119	0.06	0.054	0.05	0.038	0.041	0.025
NMF	0.469	0.381	0.357	0.321	0.306	0.282	0.268	0.254	0.238	0.23

Table 3: Mean coherence scores for different datasets, algorithms and number of topics.

BBC-news										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	0.388	0.406	0.411	0.422	0.422	0.424	0.425	0.428	0.431	0.428
FLSA-V	-	-	-	-	-	-	-	-	-	-
FLSA	0.498	0.472	0.445	0.428	0.404	0.394	0.394	0.397	0.398	0.402
LDA	0.341	0.36	0.363	0.366	0.365	0.369	0.365	0.372	0.368	0.369
NeuralLDA	0.517	0.461	0.453	0.456	0.461	0.464	**0.471**	**0.46**	**0.466**	**0.461**
ProdLDA	**0.652**	**0.647**	**0.645**	**0.636**	**0.575**	**0.504**	0.402	0.422	0.407	0.382
LSI	0.45	0.429	0.393	0.367	0.356	0.342	0.334	0.323	0.321	0.314
ETM	0.457	0.379	0.424	0.479	0.481	0.45	0.434	0.371	0.353	0.422
NMF	0.415	0.42	0.423	0.419	0.413	0.414	0.425	0.419	0.417	0.416
DBLP										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	0.462	0.462	**0.475**	**0.476**	**0.489**	**0.497**	**0.507**	**0.51**	**0.515**	**0.515**
FLSA-V	0.303	0.393	0.437	0.459	0.468	0.469	0.473	0.47	0.46	0.449
FLSA	0.24	0.259	0.267	0.272	0.273	0.277	0.277	0.281	0.281	0.28
LDA	0.317	0.318	0.305	0.304	0.311	0.316	0.32	0.322	0.319	0.314
NeuralLDA	0.282	0.284	0.308	0.308	0.314	0.324	0.326	0.32	0.326	0.325
ProdLDA	**0.467**	**0.475**	0.457	0.428	0.422	0.417	0.415	0.408	0.401	0.396
LSI	0.269	0.258	0.266	0.265	0.258	0.253	0.247	0.245	0.245	0.244
ETM	0.169	0.136	0.108	0.108	0.116	0.126	0.13	0.136	0.138	0.142
NMF	0.332	0.357	0.361	0.357	0.35	0.347	0.342	0.337	0.331	0.325
M10										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**0.585**	0.589	**0.607**	**0.612**	**0.612**	**0.609**	**0.61**	**0.616**	**0.619**	**0.622**
FLSA-V	0.572	**0.601**	0.591	0.59	0.58	0.577	0.58	0.573	0.574	0.575
FLSA	0.351	0.357	0.415	0.425	0.451	0.461	0.467	0.469	0.467	0.476
LDA	0.272	0.317	0.331	0.346	0.359	0.364	0.373	0.372	0.367	0.358
NeuralLDA	0.419	0.412	0.425	0.435	0.441	0.45	0.456	0.455	0.46	0.465
ProdLDA	0.423	0.393	0.406	0.412	0.414	0.434	0.44	0.445	0.452	0.46
LSI	0.331	0.332	0.31	0.309	0.304	0.312	0.319	0.33	0.334	0.34
ETM	0.24	0.162	0.144	0.138	0.134	0.139	0.138	0.136	0.134	0.146
NMF	0.328	0.34	0.336	0.342	0.341	0.342	0.345	0.346	0.348	0.35
20NewsGroup										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	0.391	0.381	0.372	0.37	0.357	0.35	0.348	0.34	0.332	0.327
FLSA-V	0.213	0.202	0.193	0.201	0.202	0.201	0.201	0.201	0.203	0.203
FLSA	0.537	0.534	0.505	0.481	0.467	0.454	0.442	0.435	0.429	0.422
LDA	0.512	0.536	0.542	0.518	0.513	0.499	0.492	0.481	0.473	0.473
NeuralLDA	0.434	0.46	0.479	0.477	0.491	0.475	0.493	0.462	0.466	0.478
ProdLDA	0.563	**0.619**	**0.603**	**0.598**	**0.572**	**0.574**	**0.562**	0.536	0.545	0.545
LSI	0.522	0.465	0.428	0.408	0.386	0.367	0.362	0.357	0.348	0.347
ETM	0.508	0.498	0.509	0.498	0.466	0.461	0.462	0.432	0.442	0.404
NMF	**0.575**	0.565	0.588	0.578	**0.572**	0.569	0.561	**0.555**	**0.551**	**0.549**

Table 4: Mean diversity scores for different datasets, algorithms and number of topics.

BBC-news										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**1.0**	**0.975**	**0.981**	**0.98**	**0.981**	**0.972**	**0.966**	**0.949**	**0.937**	**0.914**
FLSA-V	-	-	-	-	-	-	-	-	-	-
FLSA	0.703	0.602	0.622	0.659	0.68	0.695	0.711	0.713	0.706	0.702
LDA	0.414	0.324	0.301	0.277	0.247	0.245	0.237	0.234	0.226	0.225
NeuralLDA	0.898	0.897	0.877	0.822	0.793	0.759	0.734	0.696	0.679	0.654
ProdLDA	0.91	0.778	0.711	0.657	0.606	0.626	0.638	0.601	0.607	0.588
LSI	0.592	0.426	0.354	0.324	0.294	0.267	0.251	0.232	0.222	0.216
ETM	0.529	0.244	0.262	0.299	0.254	0.193	0.173	0.092	0.072	0.114
NMF	0.673	0.57	0.498	0.456	0.44	0.408	0.389	0.376	0.361	0.347
DBLP										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**1.0**	**0.974**	**0.941**	**0.901**	**0.877**	0.837	0.797	0.753	0.703	0.656
FLSA-V	0.756	0.665	0.59	0.534	0.483	0.438	0.399	0.36	0.323	0.294
FLSA	0.44	0.307	0.237	0.195	0.176	0.161	0.15	0.14	0.134	0.13
LDA	0.787	0.863	0.882	0.883	0.875	**0.867**	**0.859**	**0.854**	**0.861**	**0.857**
NeuralLDA	0.994	0.932	0.837	0.751	0.68	0.631	0.587	0.542	0.511	0.49
ProdLDA	0.954	0.848	0.629	0.491	0.418	0.369	0.341	0.308	0.292	0.287
LSI	0.473	0.296	0.248	0.222	0.202	0.18	0.165	0.157	0.153	0.148
ETM	0.113	0.077	0.08	0.085	0.092	0.096	0.094	0.091	0.088	0.086
NMF	0.713	0.645	0.593	0.567	0.541	0.523	0.5	0.488	0.473	0.458
M10										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**0.96**	**0.915**	**0.868**	**0.834**	**0.791**	**0.739**	0.705	0.675	0.644	0.611
FLSA-V	0.782	0.61	0.448	0.34	0.274	0.229	0.197	0.172	0.153	0.138
FLSA	0.558	0.51	0.533	0.528	0.544	0.546	0.543	0.531	0.515	0.505
LDA	0.625	0.633	0.668	0.684	0.694	0.7	**0.707**	**0.699**	**0.7**	**0.686**
NeuralLDA	0.901	0.749	0.665	0.576	0.532	0.472	0.446	0.417	0.392	0.371
ProdLDA	0.959	0.702	0.645	0.6	0.568	0.568	0.556	0.549	0.534	0.53
LSI	0.56	0.383	0.306	0.262	0.248	0.228	0.209	0.196	0.193	0.186
ETM	0.393	0.098	0.062	0.048	0.039	0.035	0.035	0.032	0.027	0.029
NMF	0.72	0.622	0.565	0.532	0.509	0.492	0.481	0.47	0.453	0.451
20NewsGroup										
	10	20	30	40	50	60	70	80	90	100
FLSA-W	**1.0**	**0.997**	**0.992**	**0.972**	**0.949**	**0.921**	**0.879**	**0.827**	**0.771**	0.723
FLSA-V	0.898	0.832	0.768	0.707	0.624	0.551	0.488	0.43	0.385	0.348
FLSA	0.694	0.641	0.567	0.534	0.506	0.475	0.455	0.445	0.426	0.414
LDA	0.689	0.706	0.721	0.718	0.722	0.732	0.735	0.735	0.743	**0.735**
NeuralLDA	0.919	0.87	0.81	0.756	0.714	0.65	0.649	0.583	0.561	0.558
ProdLDA	0.905	0.893	0.854	0.802	0.752	0.739	0.678	0.634	0.606	0.589
LSI	0.575	0.492	0.439	0.368	0.337	0.302	0.283	0.26	0.244	0.227
ETM	0.552	0.384	0.32	0.236	0.127	0.116	0.107	0.087	0.09	0.061
NMF	0.817	0.674	0.607	0.556	0.534	0.495	0.479	0.458	0.432	0.42

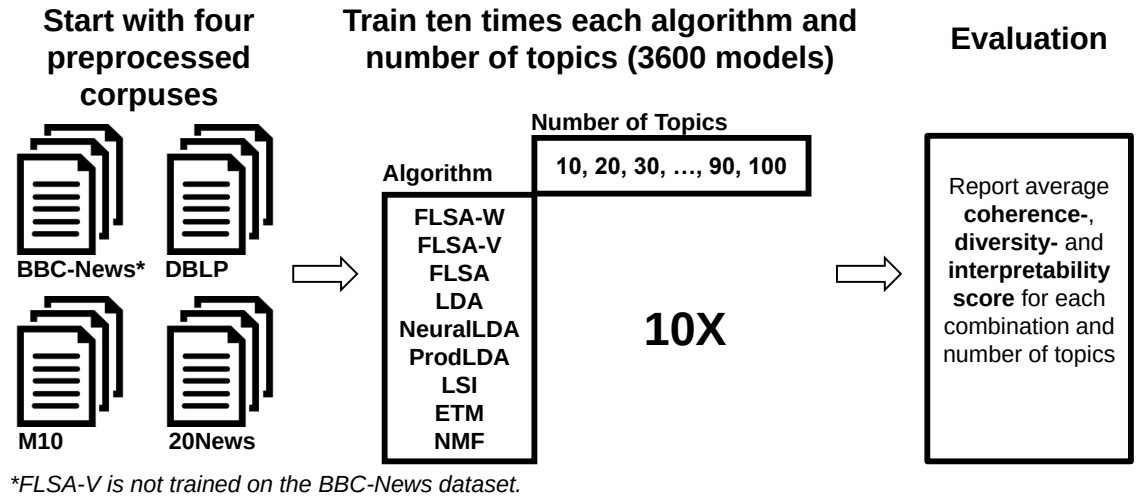

Figure 2: Visual representation of the experimental setup

4.1. Interpretability

Table 2 illustrates that FLSA-W outperforms all the other algorithms on the DBLP and M10 datasets in terms of interpretability scores. For the BBC-News dataset, FLSA-W has outperforms other models in terms of interpretability score for a high number of topics, whereas ProdLDA has the highest interpretability score for a smaller number of topics. Moreover, on the 20NewsGroup dataset, ProdLDA has the highest interpretability score.

4.2. Coherence

FLSA-W has the most number of highest coherence scores on the DBLP and M10 datasets and does not outperform the other models in the 20NewsGroup- and BBC-News datasets. Nevertheless, FLSA-W does not have any number of topics with the highest coherence score on the BBC-News dataset. NeuralLDA and ProdLDA have the highest coherence score on the BBC-News dataset, whereas ProdLDA and NMF have the highest coherence scores on the 20NewsGroup dataset. In contrast to the other algorithms, FLSA-W's coherence score improves as the number of topics increases for the BBC-News, DBLP and M10 datasets. The coherence scores of the other algorithms deteriorate or stay balanced with a higher number of topics.

4.3. Diversity

FLSA-W has perfect diversity scores using ten topics on the BBC-News-. DBLP- and 20NewsGroup dataset. The diversity stays high as the number of topics increases for BBC-News, with a significantly larger vocabulary and length per text than the other datasets. The diversity score deteriorates faster on the DBLP-, M10- and 20NewsGroup datasets as the number of topics increases. Interestingly, as the number of topics increases, LDA scores highest on diversity on the DBLP and M10 datasets, with fewer words per text.

5. Discussion

This work quantitatively tests and compares the FLSA-W algorithm with other state-of-the-art algorithms on four open datasets in terms of interpretability-, coherence-, and diversity scores. We find that FLSA-W has the most number of topics with the highest performance for each evaluation metric amongst all algorithms tested. Most topic models can be seen as being fuzzy, as each word has a membership to various topics, and topics have a membership to documents. For this reason, we argue that using fuzzy algorithms to find topics is more intuitive than non-fuzzy-set-based algorithms. With topic modeling, topics are seen as clusters of words. Since FLSA-W clusters words, it is logical that its performance is higher than FLSA, which clusters documents. FLSA-W and FLSA-V both cluster words but use different projection methods. The results indicate that singular value decomposition captures semantic

relations more effectively than VOSviewer's projections as FLSA-W outperforms FLSA-V for both big and small datasets; FLSA-V does not work for big data sets and requires pruning. Both algorithms project words differently but use the same clustering methods to find topics. Furthermore, our results illustrate that:

- A small number of words per text has both a positive- and negative effect on FLSA-W's interpretability score. FLSA-W outperforms the other models regarding coherence score on the datasets with fewer words per text. However, FLSA-W's diversity score deteriorates more for these small datasets as the number of topics increases compared to the larger datasets.
- In [15], NMF was found to produce high-quality topics on datasets with short texts. Our experimental results cannot confirm this finding. On DBLP and M10, the datasets with short texts, approximately six words per text, NMF has a relatively low performance. Yet, on the 20Newsgroup dataset, with approximately 50 words per text, NMF outperforms the other algorithms in terms of coherence for various number of topics.

5.1. Limitations

The current results are based on four English datasets, all preprocessed similarly. Yet, there is no consensus on the best preprocessing steps. Since we only test datasets preprocessed in one way, we do not know how well the findings generalize to other preprocessing steps. Moreover, the underlying assumption in this work is that default hyperparameter settings are based on optimal results from earlier experiments. Therefore, we have not conducted hyperparameter tuning and used the default values. We do not expect this to impact the performance significantly. Additionally, in these experiments, we take a fixed and arbitrary number of topics that we test. Yet, we do not know if similar numbers of topics reflect our corpora. Lastly, we use quantitative measures to reflect how semantically strong the produced topics are. An advantage of quantitative measures is the objective and deterministic nature of the scores. Yet, no measure correlates perfectly with human interpretation, and therefore the use of quantitative evaluation is limited. However, human interpretation is subjective and non-deterministic, not allowing the results to be reproduced.

6. Conclusion

In this work, we have compared FLSA-W with other state-of-the-art algorithms on four open datasets in terms of interpretability-, coherence- and diversity scores. We find that FLSA-W outperforms the other algorithms in all evaluation metrics in the most number of topics. Future work will analyze why FLSA-W and the other algorithms perform differently on different datasets and characterize preferred hyperparameter settings per algorithm. Additionally, experimental settings should be broader for more robust findings. Therefore, future work will focus on datasets from different languages, experiment with different kinds of preprocessing, compare optimized models and compare a different number of words per topic.

Acknowledgment

We acknowledge the COVIDA funding provided by the strategic alliance of TU/e, WUR, UU, and UMC Utrecht.

Conflict of Interest Statement

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Y. Haribhakta, A. Malgaonkar, P. Kulkarni, Unsupervised topic detection model and its application in text categorization, in: Proceedings of the CUBE International Information Technology Conference, 2012, pp. 314–319.

[2] D. Spina, J. Gonzalo, E. Amigó, Learning similarity functions for topic detection in online reputation monitoring, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 527–536.

[3] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers, M. Spruit, Machine learning for violence risk assessment using Dutch clinical notes, Journal of Artificial Intelligence for Medical Sciences 2 (1-2) (2021) 44–54.

[4] T. A. Rana, Y.-N. Cheah, S. Letchmunan, Topic modeling in sentiment analysis: A systematic review., Journal of ICT Research & Applications 10 (1) (2016).
[5] A. Karami, A. Gangopadhyay, B. Zhou, H. Kharrazi, Fuzzy approach topic discovery in health and medical corpora, International Journal of Fuzzy Systems 20 (4) (2018) 1334–1345.
[6] E. Rijcken, F. Scheepers, P. Mosteiro, K. Zervanou, M. Spruit, U. Kaymak, A comparative study of fuzzy topic models and LDA in terms of interpretability, in: Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021, p. accepted.
[7] E. Rijcken, U. Kaymak, F. Scheepers, P. Mosteiro, K. Zervanou, M. Spruit, Topic modeling for interpretable text classification from EHRs, Frontiers in Big Data 5 (2022). doi:10.3389/fdata.2022.846930.
URL https://www.frontiersin.org/article/10.3389/fdata.2022.846930
[8] D. Greene, P. Cunningham, Practical solutions to the problem of diagonal dominance in kernel document clustering, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 377–384.
[9] K. W. Lim, W. Buntine, Bibliographic analysis with the citation network topic model, in: Asian conference on machine learning, PMLR, 2015, pp. 142–158.
[10] J. Rennie, 20 newsgroups (Jan 2008).
URL http://qwone.com/ jason/20Newsgroups/
[11] I. Vayansky, S. A. Kumar, A review of topic modeling methods, Information Systems 94 (2020) 101582.
[12] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, A. Candelieri, OCTIS: Comparing and optimizing topic models is simple!, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 263–270. doi:10.18653/v1/2021.eacl-demos.31.
URL https://aclanthology.org/2021.eacl-demos.31
[13] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning research 3 (2003) 993–1022.
[14] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[15] Y. Chen, H. Zhang, R. Liu, Z. Ye, J. Lin, Experimental explorations on short text topic mining between LDA and NMF based schemes, Knowledge-Based Systems 163 (2019) 1–13.
[16] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Processes 25 (2-3) (1998) 259–284.
[17] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Springer Science & Business Media, 2013.
[18] B. De Finetti, Theory of Probability: A Critical Introductory Treatment, Vol. 6, John Wiley & Sons, 2017.
[19] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, M. Welling, Fast collapsed gibbs sampling for latent dirichlet allocation, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 569–577.
[20] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, in: 5th International Conference on Learning Representations,ICLR, 2017.
[21] P. Dayan, G. E. Hinton, R. M. Neal, R. S. Zemel, The helmholtz machine, Neural Computation 7 (5) (1995) 889–904.
[22] D. Kingma, M. Welling, Auto-encoding variational bayes, in: The International Conference on Learning Representations, ICLR, 2014.
[23] A. B. Dieng, F. J. R. Ruiz, D. M. Blei, Topic modeling in embedding spaces, Transactions of the Association for Computational Linguistics 8 (2020) 439–453.
URL https://aclanthology.org/2020.tacl-1.29
[24] N. J. Van Eck, L. Waltman, Software survey: VOSviewer, a computer program for bibliometric mapping, Scientometrics 84 (2) (2010) 523–538.
[25] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM International Conference on Web Search and Data Mining, 2015, pp. 399–408.
[26] E. Rijcken, P. Mosteiro, K. Zervanou, M. Spruit, F. Scheepers, U. Kaymak, FuzzyTM: a software package for fuzzy topic modeling, in: 2022 IEEE international conference on fuzzy systems (FUZZ-IEEE), 2022.
[27] C. Fuchs, S. Spolaor, M. S. Nobile, U. Kaymak, pyFume: a Python package for fuzzy model estimation, in: 2020 IEEE international conference on fuzzy systems (FUZZ-IEEE), IEEE, 2020, pp. 1–8.