A Comparative Study of Fuzzy Topic Models and LDA in terms of Interpretability

Emil Rijcken
Jheronimus Academy of Data Science
Eindhoven University of Technology
Eindhoven, The Netherlands
e.f.g.rijcken@tue.nl

Floortje Scheepers
Psychiatry
University Medical Centre Utrecht
Utrecht, The Netherlands
f.e.scheepers-2@umcutrecht.nl

Pablo Mosteiro
Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands
p.mosteiro@uu.nl

Kalliopi Zervanou
Industrial Engineering & Information Sciences
Eindhoven University of Technology
Eindhoven, The Netherlands
k.zervanou@tue.nl

Marco Spruit
Public Health & Primary Care
Leiden University Medical Centre
Leiden, The Netherlands
M.R.Spruit@lumc.nl

Uzay Kaymak
Jheronimus Academy of Data Science
Eindhoven University of Technology
Eindhoven, The Netherlands
u.kaymak@ieee.org

***Abstract*—In many domains that employ machine learning models, both high performing and interpretable models are needed. A typical machine learning task is text classification, where models are hardly interpretable. Topic models, used as topic embeddings, carry the potential to better understand the decisions made by text classification algorithms. With this goal in mind, we propose two new fuzzy topic models; FLSA-W and FLSA-V. Both models are derived from the topic model Fuzzy Latent Semantic Analysis (FLSA). After training each model ten times, we use the mean coherence score to compare the different models with the benchmark models Latent Dirichlet Allocation (LDA) and FLSA. Our proposed models generally lead to higher coherence scores and lower standard deviations than the benchmark models. These proposed models are specifically useful as topic embeddings in text classification, since the coherence scores do not drop for a high number of topics, as opposed to the decay that occurs with LDA and FLSA.**

***Index Terms*—Topic Models, Text Classification, Fuzzy Modelling, Explainable AI, NLP**

I. Introduction

In the last decade, machine learning systems (ML) have achieved (super)human performance in a wide variety of computational tasks [9] and have been used extensively in mining biological data [23]. A common ML computational task is text classification, a sub-domain of *Natural Language Processing* (NLP). With text classification, an algorithm labels each input text from a fixed set of labels. Applications of ML text classification have a wide range; from sentiment analysis [15], [16], [17] to news article classification for stock price prediction [20], [41] to clinical note classification for finding patient cohorts [10], [40]. Although the classification performance is high in state-of-the-art models, these models are hardly interpretable [42]. In this study, we use clinical notes from the psychiatry department of the University Medical Center Utrecht (UMCU) to predict which patient will become violent. In earlier work, we have slightly outperformed the benchmark classifications [28], [29]. Yet, the decisions made by these algorithms were hardly interpretable.

Topic models carry the potential to make text classification more interpretable. They are a group of unsupervised natural language processing algorithms that extract latent topics in texts. Based on a trained topic model, a topic embedding indicates to what extent each topic is represented in texts. This topic embedding is then fed into a classification algorithm, indicating the most important topics for its decisions.

A problem with current topic models is they do not always give interpretable results. Fuzzy Latent Semantic Analysis (FLSA) [19] is a topic modeling algorithm applied to medical data and showed superior performance to the most popular model Latent Dirichlet Allocation (LDA). FLSA is based on grouping similar documents in terms of the words they contain. The membership of each document to a topic is determined by using fuzzy clustering. However, documents might contain different number of topics and may contain different subsets of words in a topic, which influences the quality of the clustering results. This leads typically to large overlap between topics, which reduces the interpretability of the topic model obtained. In this paper, we propose two topic modeling methods that can reduce the overlap between the topics. The first model, FLSA-W, clusters on words rather than on documents. The second model, FLSA-V, applies fuzzy clustering on a 2D word-mapping, based on the open-source software tool VOSviewer [44]. In contrast to the existing topic models LDA and FLSA, our results indicate that the models' performances (as measured by the coherence score) either remain constant or grow as the number of topics increase. This property is specifically useful for text classification, as more topics typically lead to better classification performance.

The contributions of this paper are:

1) we propose two new topic modeling algorithms,
2) we compare four different topic models in terms of interpretability,
3) we study the relation between various parameters of the topic models and their properties that can could be useful

978-1-7281-9048-8/21/$31.00 ©2021 IEEE

for text classification.

In this study, we use clinical notes from the psychiatry department of the University Medical Center Utrecht (UMCU) to predict which patient will become violent. We build upon earlier work in which we have slightly outperformed the benchmark classifications [28], [29], but the decisions made by these algorithms were hardly interpretable. Note that although we trained and tested these techniques on a medical data set, the techniques presented are not limited to the medical domain. Instead, they may be used for all topic modeling and text classification applications with texts of moderate to high length.

The outline of the paper is as follows. In Section II, we describe how text classification methods work, why topic embeddings could make classification more interpretable and briefly discuss the topic modeling literature. In Section III, we describe the details of the FLSA model, and we motivate our decision for the selected coherence metric. In Sections IV-A and IV-B we provide the mathematical details of the two proposed models. In Section V we describe the data set and setup that we used for running our experiments. In Section VI, we describe the experimental results, discuss the implications in Section VII and we conclude our paper in Section VIII.

II. Related Work

In this work, we propose two new topic models intended as topic embeddings in text classification. Since the intended goal is text classification, we briefly discuss text classification in this Section and discuss topic models afterwards.

A. Text Classification

A typical ML text classification pipeline includes two steps:

- Representation step;
- Classification step.

In the representation step, a text file is transformed into a numeric representation, called a text embedding. The classification algorithm in the classification step calculates the most likely label based on the text embedding. Several classification algorithms, such as random forests, logistic regression or fuzzy systems, can reveal which input variables are most important for determining a label.

1) Representation Techniques: Early ML text classification algorithms used the bag-of-word approach (BOW) to represent each word as a one-hot-encoding [18]. Yet, BOW suffers from two significant limitations: i) it is hard to scale since it is a sparse matrix. ii) it does not consider syntactic information, as it only considers the presence of a word in a text and not the word's location.

Since their appearance in 2013, neural text embeddings such as Word2Vec [26] are widely used since they do not suffer from the limitations of BOW. Neural text embeddings represent words as dense vectors in a high-dimensional space such that semantically similar words are located close to each other in the embedding space. Since 2013, several similar neural text embedding approaches have been applied, such as BERT [7], Doc2Vec [21], Glove [34] and ELMO [35]. These neural models have improved the performance of text classification significantly. However, relatively little is known about the information captured by these embeddings' features. Therefore, there is still little understanding of the classification decisions in the subsequent step.

2) Classification Models: classification models are a set of techniques that map input data (the feature space) to a fixed set of output labels [11]. The choice of technique depends on various aspects such as the number of features, size of the data set and whether a technique should be interpretable. Amongst classification models, the subset of commonly used interpretable models include: linear regression, logistic regression, decision trees, and fuzzy systems [1], [13], [27].

B. Topic Models

While there is little known about the neural embeddings' features, the meaning of topics in a topic model is known. Instead of neural embeddings, a topic embedding' can be used to represent texts numerically. Topic models are a different branch of NLP techniques, and their output consists of two matrices:

1) $P(W_i|T_k)$- The probability of word i given topic k,
2) $P(T_k|D_j)$- The probability of topic k given document j

with:

i	word index $i \in \{1, 2, 3, ..., M\}$,
j	document index $j \in \{1, 2, 3, ..., N\}$,
k	topic index $k \in \{1, 2, 3, ..., C\}$,
M	the number of unique words in the data set,
N	the number of documents in the data set,
C	the number of topics.

A topic model aims to find topics in which the most probable words in each topic are coherent with each other so that the topic is interpretable. Using topic embeddings for text classification, each input document is transformed into a vector of size C, in which each cell indicates the extent to which the document belongs to a topic. After labels are calculated for each input text, interpretable classification algorithms can reveal which topics were most important for performing classifications.

The first step in developing a topic-based text classification algorithm is selecting a suitable topic model. A good topic model should both be interpretable and leading to high classification performance. Typically, classification models use many (at least 50) topics [19], [40]. Therefore, topic models must remain coherent as the number of topics increase.

Many different topic models exist, all having a specific purpose [47]. Applicable topic models for documents with less than 50 words per document are: a mixture of unigrams [31], the pseudo-document approach [49] and the self-aggregation

based topic model [37]. For documents with more words, the Topics over Time model [48] and Dynamic Topic Model [4] are suitable for capturing changes of topics over time. The Pachinko Allocation Model [22] and the Correlated Topic Model [5] are suitable when there is an interest in the correlation between topics.

The most popular topic modelling technique is called *Latent Dirichlet Allocation* (LDA) [3]. LDA performs well on data documents with more than 50 words and no complex topic relationships [47] and has been used before as a topic embedding in the medical domain [29], [40]. LDA is a probabilistic model and assumes that documents are formed by a generative process [3]. Each document is a distribution over C topics in this process, and each topic is a distribution over M words. Thus, if document j has a high probability of containing topic k, then topic k's most probable words are likely to be present in document j. Yet, both the distributions of topics over documents $P(T_k|D_j)$ and words over topics $P(W_i|T_k)$ are latent, thus, cannot be observed from the documents. To infer $P(T_k|D_j)$ and $P(W_i|T_k)$, LDA conditions on the words, the only observable variable. This conditioning can be seen as a reversal of the generative process. Yet, the computations of these distributions are exponentially large and, therefore, intractable [6]. Although $P(T_k|D_j)$ and $P(W_i|T_k)$ cannot be computed exactly, statistical methods such as sampling-based algorithms and variational-based algorithms can approximate them close enough [43]. More recently, a new topic modelling technique called *Fuzzy Latent Semantic Analysis* was proposed, applied to the medical domain, and shows better classification performance than LDA [19]. Whereas LDA uses a reversed generative process to find its topics, FLSA uses fuzzy clustering to find its topics. With FLSA, documents are clustered. However, documents are distributions over topics and topics are distributions over words. Therefore, the clustering is influenced by two interrelated process, making it harder to identify topics. Since topics are assumed to be a set of coherent words, it would be reasonable to identify them in an embedding space where related words are located nearby one another. This is what our proposed methods FLSA-W and FLSA-V aim to achieve by clustering in a space of "word embedding".

In this work, we use both methods to benchmark our proposed models. Therefore, we briefly describe both models in the next section.

III. Methodological background

The models in this work are derived from the FLSA model [19]. In this section, we first describe the steps how FLSA is calculated. Then, based on these steps, we will motivate why we are proposing two new methods, derived from FLSA. Lastly, we discuss the evaluation metric that we use for comparing our topic models.

A. Fuzzy Latent Semantic Analysis

The FLSA approach uses fuzzy clustering to find topics.

We define the following quantities:

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 $k \in \{1, 2, 3, ..., C\}$.

FLSA is calculates with the following steps [19].

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 by multiplying the GTW vectors element-wise with the LTW matrix, to obtain $P(W_i, D_j)$. In the paper four different GTW methods are explored: Entropy, IDF, Normal and ProbIDF.

3) Perform singular value decomposition (SVD) on the GTW matrix, for dimensionality reduction, to get S singular value dimensions.

4) Perform SVD on the GTW matrix, for dimensionality reduction, to get S singular value dimensions.

5) Perform fuzzy clustering on SVD's U^T to obtain $P(T_k|D_j)^T$.

6) Calculate the probability of document j

$$P(D_j) = \frac{\sum_{i=1}^{M} P(W_i, D_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} P(W_i, D_j)}. \tag{1}$$

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

$$P(D_j, T_k) = (P(T_k|D_j) \otimes P(D_j))^T, \tag{2}$$

$$P(D_j|T_k) = \frac{P(D_j, T_k)}{\sum_{j=1}^{N} P(D_j, T_k)}, \tag{3}$$

$\otimes$ represents element-wise multiplication.

8) Calculate the probability of word i, given document j

$$P(W_i|D_j) = \frac{P(W_i, D_j)}{\sum_{i=1}^{M} P(W_i, D_j)}. \tag{4}$$

9) Calculate the probability of word i, given topic k

$$P(W_i|T_k) = \sum_{j=1}^{N} P(W_i|D_j)^T P(D_j|T_k). \tag{5}$$

B. Coherence

Topic coherence can be described as the average semantic relatedness between a topic's words with highest probabilities [30]. The underlying idea of topic coherence is rooted in the

distributional hypothesis of linguistics, which states that words with similar meanings tend to occur in similar contexts [14]. From the available coherence measures, C_v correlates highest with human topic ranking data [39]. Therefore, we will use this measure for our evaluation, similar to [43]. C_v retrieves the co-occurrence counts for the words in the word set using a context window. These counts are then used to compute the normalized point-wise mutual information of every word to every other word, resulting in vectors for each word. The average of cosine similarities between the vector of each word and the sum of all word vectors gives the C_v coherence score [36]. The score ranges between zero and one; one means the topic model is perfectly coherent, and zero means it is not coherent at all.

IV. Methodology

With FLSA, the U^T matrix from SVD is fed to the clustering algorithm to obtain $P(T_k|D_j)^T$. Thus, documents are being clustered with FLSA. However, documents are distributions over topics and topics are distributions over words. Therefore, the clustering is conducted on two mixed distributions, making it harder to distinguish between clusters. Further, documents might contain multiple topics, which make them difficult to cluster. Another possibility is to cluster on words. In that case, clusters are based on the co-occurrence of words only, and such clusters are possibly more meaningful. Based on the clustered output $P(T_k|W_i)^T$, we can use Bayes' Theorem to calculate necessary matrices. Since we cluster words, we refer to this approach as FLSA-W.

Both with FLSA and FLSA-W we cluster on SVD's output. SVD's input is the GTW matrix, which is calculated by weighting words/documents. However, with GTW, there is no guarantee that frequently re-occurring concepts are located nearby each other. Thus, SVD's S-dimensional output may not reflect semantic relations within and between topics. This is our inspiration for the second model that we propose FLSA-V. In this approach we use the tool VOSviewer [44] to obtain a 2D mapping of words. VOSviewer is an open-source software tool for bibliometric mapping. It uses a projection algorithm (similar to multi-dimensional scaling) to locate co-occurring terms nearby each other. With FLSA-V we feed the word coordinates from the 2D mapping ($M \times 2$ into the fuzzy clustering algorithm to obtain $P(T_k|W_i)$. Again, we use Bayes's Theorem to calculate the other necessary matrices. Section IV-B contains the mathematical details of FLSA-V.

A. FLSA-W

In this section, we describe the mathematical details of FLSA-W.

We define the following quantities:

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 $k \in \{1, 2, 3, ..., C\}$.

In this approach, steps one and 2a are the same as FLSA. Yet, in step 2b, we use SVD's V matrix ($M \times S$), instead of the U matrix [46].

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 by multiplying the GTW vectors element-wise with the LTW matrix, to obtain $P(W_i, D_j)$.

3) Perform fuzzy clustering on V^T to obtain $P(T_k|W_i)^T$.

4) Calculate probability vectors:

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

$$P(W_i) = \frac{\sum_{j=1}^{N} P(W_i, D_j)}{\sum_{i=1}^{m}\sum_{j=1}^{N} P(W_i, D_j)}, \quad (7)$$

$$P(T_k) = P(T_k|W_i)P(W_i). \quad (8)$$

5) Calculate the probability of word i, given topic k

$$P(W_i, T_k) = (P(T_k|W_i) \otimes P(W_i))^T, \quad (9)$$

$$P(W_i|T_k) = \frac{P(W_i, T_k)}{\sum_{i=1}^{M} P(W_i, T_k)}. \quad (10)$$

6) Calculate the probability of topic k, given document j

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

$$P(D_j|W_i) = \frac{(P(W_i|D_j) \otimes P(D_j))^T}{P(W_i)}, \quad (12)$$

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

$$P(T_k|D_j) = \frac{(P(D_j|T_k) \otimes P(T_k))^T}{P(D_j)}. \quad (14)$$

B. FLSA-V

In this section, we describe the mathematical details of FLSA-V.

Let us define the following quantities:

M	number of unique words in the data set,
N	number of documents in the data set,
C	number of topics,
i	word index $i \in \{1, 2, 3, ..., M\}$,
j	document index $j \in \{1, 2, 3, ..., N\}$,
k	topic index $k \in \{1, 2, 3, ..., C\}$,
LTW	local-term weights (FLSA step 1).

With FLSA-V, we feed the word coordinates from the 2D mapping ($M \times 2$ into the fuzzy clustering algorithm to obtain $P(T_k|W_i)^T$. Then, the subsequent steps are as follows.

1) Calculate probability vectors.

$$P(D_j) = \frac{\sum_{i=1}^{M} LTW_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N} LTW_{ij}}, \quad (15)$$

$$P(W_i) = \frac{\sum_{j=1}^{N} LTW_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N} LTW_{ij}}, \quad (16)$$

$$P(T_k) = P(T_k|W_i)P(W_i). \quad (17)$$

2) Follow the same steps as from FLSA-W step 5 onwards.

V. Data & Experimental Setup

Our data set consists of clinical notes, written in Dutch by nurses and physicians in the psychiatry ward of the University Medical Center (UMC) Utrecht between 2012-08-01 and 2020-03-01 and is the same as in previous work [28], [29]. The 834834 notes available are de-identified for patient privacy using DEDUCE [25]. Since the goal of the topic models is to increase the understanding of the decisions made by the subsequent text classification algorithm, we maintain the same structure as in previous data sets. That is, each patient can be admitted to the psychiatry ward multiple times. In addition, an admitted patient can spend time in various sub-departments of psychiatry. The time a patient spends in each sub-department is called an admission period, and in the data set, each admission period is a data point. For each admission period, all notes collected between 28 days before and one day after the start of the admission period are concatenated and considered as a single period note. We preprocess the text by lowercasing and deaccenting all words, removing the stop words and filtering out single characters. This results in 4280 admission periods with an average length of 1481 words. Admission periods having fewer than 101 words are discarded, similar to previous work [24], [45].

Note that data points are initialized randomly with fuzzy clustering. Therefore, FLSA, FLSA-W and FLSA-V are non-deterministic. For each number of topics we train 10 different models on the entire data set. We test up to 70 topics, and choose this number since we found convergence in predictive performance in exploratory experiments. All our methods are developed in Python and are available on Github[1].

[1]https://github.com/ERijck/FLSA

1) FLSA: We used the *pyFUME* package [12] to perform c-means fuzzy clustering.

a) FLSA & FLSA-W: We used the *sparsesvd* package [2] to perform SVD on the GTW and used two factors, similar to the original work [19]. Furthermore, the reported coherence scores are based on 'normal' global term weighting, which gave the highest coherence scores amongst the four different weighting methods.

b) FLSA-V: We considered only words that appeared in at least 100 documents (2.3% of all documents) as we ran into memory issues with more words. We expect that this model will perform better if more words are used.

2) LDA: We used the Gensim package [38] to run LDA. We train with 25 epochs, use a chunk size of 100 and update after every document.

VI. Results

Table I and Table II show the average coherence scores and standard deviation when considering 10 words and 20 words per topic, respectively. The results are based 10 runs per model. When considering 20 words per topic to calculate the coherence (Table II), we see that for almost all the number of topics, FLSA-W has the highest coherence score, except for the lowest two numbers of topics, where LDA and FLSA score better. Additionally, FLSA-V scores better than the benchmark models for the number of topics being greater than 20.

TABLE I
Coherence - 10 Words per Topic - mean & standard deviations

Num. Topics	Topic Models			
	LDA	**FLSA**	**FLSA-W**	**FLSA-V**
5	0.419 (0.038)	**0.454** (0.005)	0.435 (0.000)	0.328 (0.007)
10	0.435 (0.026)	**0.497** (0.000)	0.387 (0.009)	0.337 (0.003)
15	0.475 (0.035)	**0.493** (0.002)	0.415 (0.000)	0.315 (0.010)
20	0.456 (0.022)	**0.508** (0.010)	0.410 (0.002)	0.307 (0.001)
25	0.434 (0.024)	**0.457** (0.005)	0.408 (0.004)	0.304 (0.005)
30	0.428 (0.012)	**0.439** (0.009)	0.403 (0.008)	0.304 (0.006)
35	0.413 (0.019)	**0.426** (0.004)	0.403 (0.006)	0.305 (0.002)
40	**0.420** (0.014)	0.395 (0.004)	0.400 (0.002)	0.304 (0.003)
45	**0.400** (0.017)	0.381 (0.007)	0.385 (0.004)	0.303 (0.004)
50	0.390 (0.012)	0.360 (0.005)	**0.393** (0.005)	0.300 (0.003)
55	0.391 (0.012)	0.345 (0.006)	**0.391** (0.003)	0.295 (0.002)
60	0.376 (0.015)	0.320 (0.007)	**0.393** (0.006)	0.293 (0.002)
65	0.381 (0.007)	0.308 (0.008)	**0.400** (0.005)	0.292 (0.002)
70	0.379 (0.010)	0.298 (0.007)	**0.405** (0.003)	0.291 (0.002)

The plot of the average coherence scores (Fig. 1) shows that the FLSA-W and FLSA-V coherence scores stay roughly the same as the number of topics increases. At the same time, LDA and FLSA show decay in coherence scores for a higher number of topics. The plot with the coherence scores' standard deviation (Fig. 2) shows that all FLSA-based topics have more constant coherence scores than LDA. Especially, FLSA-V shows little variability in coherence scores across all numbers of topics.

VII. Discussion

In this work, we propose two new topic models FLSA-W and FLSA-V, to be used as interpretable topic embedding for

TABLE II
COHERENCE - 20 WORDS PER TOPIC - MEAN & STANDARD DEVIATIONS

Num. Topics	Topic Models			
	LDA	FLSA	FLSA-W	FLSA-V
5	**0.414** (0.021)	0.411 (0.006)	0.410 (0.000)	0.376 (0.004)
10	0.470 (0.035)	**0.474** (0.002)	0.462 (0.001)	0.393 (0.003)
15	0.466 (0.030)	0.473 (0.002)	**0.494** (0.000)	0.415 (0.004)
20	0.464 (0.009)	0.460 (0.003)	**0.486** (0.007)	0.429 (0.003)
25	0.435 (0.030)	0.429 (0.003)	**0.469** (0.006)	0.437 (0.002)
30	0.420 (0.019)	0.408 (0.008)	**0.470** (0.004)	0.446 (0.003)
35	0.415 (0.012)	0.392 (0.004)	**0.471** (0.007)	0.451 (0.002)
40	0.412 (0.019)	0.388 (0.007)	**0.464** (0.003)	0.456 (0.002)
45	0.389 (0.014)	0.378 (0.003)	**0.469** (0.002)	0.457 (0.002)
50	0.386 (0.013)	0.363 (0.004)	**0.470** (0.005)	0.462 (0.002)
55	0.385 (0.013)	0.349 (0.006)	**0.466** (0.003)	0.465 (0.002)
60	0.381 (0.012)	0.342 (0.006)	**0.472** (0.005)	0.466 (0.002)
65	0.380 (0.010)	0.333 (0.006)	**0.478** (0.005)	0.470 (0.002)
70	0.377 (0.013)	0.318 (0.006)	**0.482** (0.002)	0.471 (0.002)

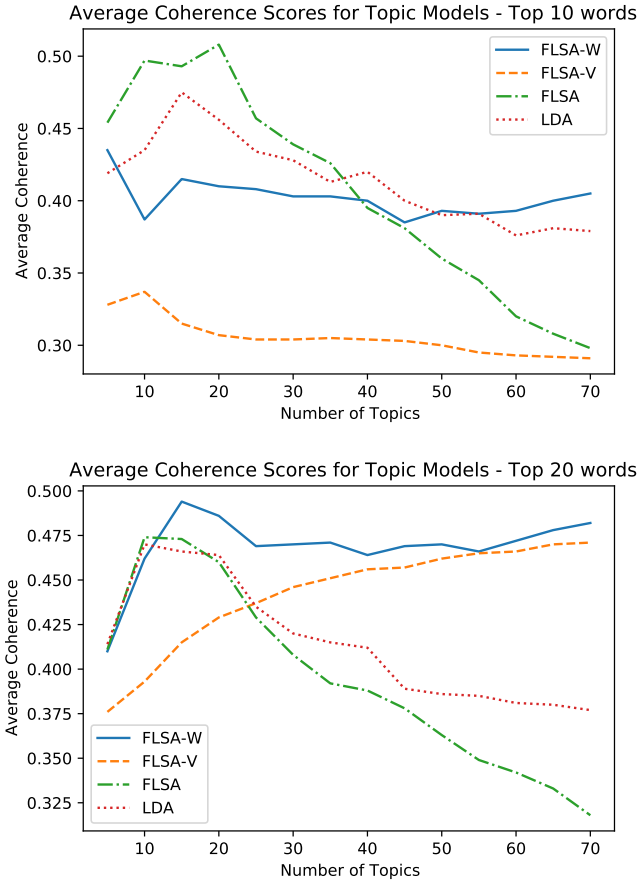

Fig. 1. Coherence scores for topic models with the top 10 and 20 words.

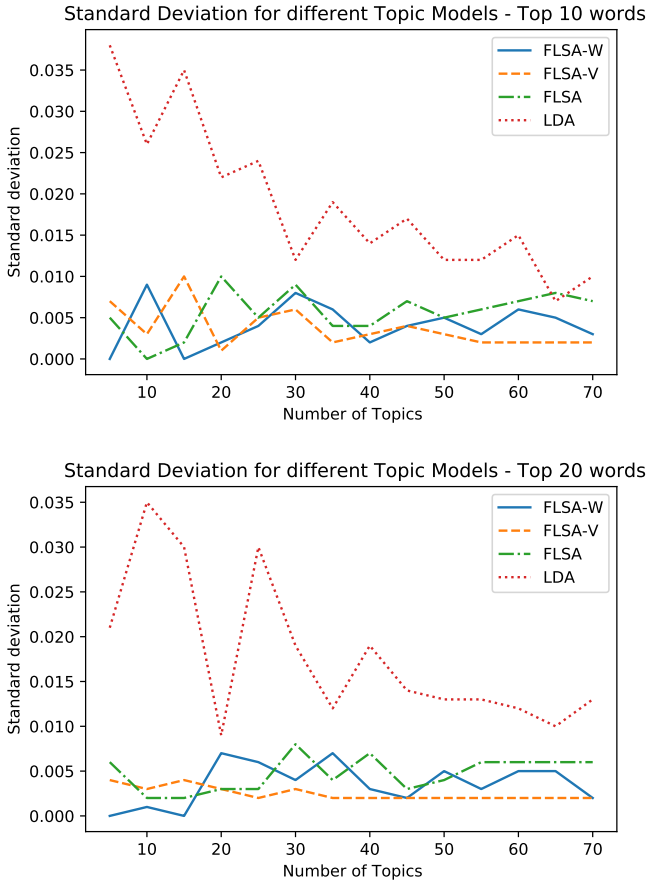

Fig. 2. Standard deviations of topic models with the top 10 and 20 words.

text classification. FLSA-W performs better than almost all models with 20-word-topics, whereas it only outperforms other models for 50+ topics with the 10-word-topics, solely based on coherence scores. In addition to focusing on coherence scores, a domain expert manually assessed several topics. For all the topics that she compared, she found the 20-word-topics better interpretable than the 10-word-topics. In addition, contrasting to the coherence score, she found LDA's topics better interpretable than FLSA-W's topics. Yet, she found FLSA-W to cover a wider variety of topic themes than LDA, which is not captured by the coherence score. Topics with high diversity are more likely to cover different themes than topics with low diversity. An indication of topic diversity is the fraction of unique words in a topic model's topics and the total number of words in these topics [8]. If the fraction is high, topics contain more unique words. Where as, a low fraction indicates that topics share many of the same words. Table III shows the average fraction for each model, based on all the models from the experiments. Our proposed models have higher topic diversity than LDA and FLSA and cover a wider variety of themes. FLSA-V has the lowest coherence score, and its coherence decays as the number of topics increase, with 10-word-topics. Yet, with the

20-topic-model, its coherence score grows as the number of topics increase. Since FLSA-V was trained with words that appeared in at least 100 documents, we expect better topics if the model is trained with more words. The FLSA topics that were found to be interpretable based on the manual assessment had minimal variation and a low topic diversity. This finding supports our hypothesis that words should be clustered rather than documents. Additionally, the domain expert found FLSA-W to be better interpretable than FLSA. Furthermore, we observed the following surprising patterns in coherence scores.

1) Unlike LDA and FLSA, neither FLSA-W's nor FLSA-V's coherence score decay as the number of topics increases. We cannot explain why this is the case. Yet, this could be a valuable feature for text classification since topic embeddings typically contain 50+ topics.
2) The direction of FLSA-V's coherence score changed drastically after the number of words per topic changed. Also, FLSA-W's coherence scores are higher for almost all the number of topics, with 20-word-topics. These changes indicate that the number of words per topic may impact a topic's quality, depending on the used topic model.
3) LDA has a much higher variation in coherence score than the other models for almost all settings. This indicates that our proposed models and FLSA are more stable topic models than LDA, since there is less variability in the models that they produce. The above findings are mainly based on the coherence score. However, not all our findings support that coherence should be used as a measure of interpretability. A finding favouring coherence for interpretability is that FLSA-V's topics were found to be better interpretable with 20-word-topics than with 10-word-topics, similar to the coherence scores. Yet, in contrast to the coherence score, the domain expert found LDA topics to be more interpretable than FLSA-W.

Note that, with text classification of nurse notes the topic model depends on the context, and the expressions of people from different geographical locations change. This is an open question that we do not address with our proposed models.

TABLE III
AVERAGE FRACTION OF UNIQUE WORDS IN TOPICS

Words per Topic	Topic Models			
	LDA	FLSA	FLSA-W	FLSA-V
10	0.503	0.607	0.998	0.958
20	0.457	0.635	0.997	0.894

VIII. CONCLUSION

This work proposes two new topic models derived from FLSA as an intermediary step towards more explainable text classification. We found that the number of words per topic can strongly influence a topic's quality. The domain expert found topics with 20 words per topic most interpretable. FLSA-W has a higher coherence score than other models for almost all number of topics with 20-word-topics. For both the 10-word-topics and 20-word-topics, FLSA-W has the highest coherence score for the number of topics greater than 45. This is a valuable characteristic for text classification since topic embeddings typically contain more than 50 topics. However, we have found an inconsistent alignment between coherence scores and interpretability, as indicated by the domain expert. Future work should therefore aim at formulating topic evaluation metrics more adequately. A recent paper defines topic interpretability as the product of topic coherence and topic diversity [8]. With this definition of interpretability, FLSA-W and FLSA-V would significantly outperform LDA and FLSA. In a future study we will analyze the interpretability further and compare FLSA-W and FLSA-V to other topic models. Also, our findings are based on a domain-specific data set. Therefore, it should also be tested on other data sets. In addition, exploratory work should be conducted to assess how topic models are affected by changes in the number of words per topic. Finally, we will study the performance of our predictive violent risk assessment in actual clinical practice in the future.

ACKNOWLEDGMENT

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

REFERENCES

[1] Alonso, J. M., Castiello, C., & Mencar, C. (2015). Interpretability of fuzzy systems: Current research trends and prospects. Springer handbook of computational intelligence, 219-237.
[2] Berry, Michael W. (1992). Large scale sparse singular value computations. International Journal of Supercomputer Applications, 6, 13–49.
[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
[4] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).
[5] Blei, D., & Lafferty, J. (2006). Correlated topic models. Advances in neural information processing systems, 18, 147.
[6] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
[7] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (4171–4186)
[8] Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453.
[9] Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 0210-0215). IEEE.
[10] Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Scientific reports, 8(1), 1-10.
[11] Flach, P. (2012). Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press.
[12] Fuchs, C., Spolaor, S., Nobile, M. S., & Kaymak, U. (2020). pyFUME: a Python package for fuzzy model estimation. In 2020 IEEE international conference on fuzzy systems (FUZZ-IEEE) (pp. 1-8). IEEE.
[13] Guillaume, S. (2001). Designing fuzzy inference systems from data: An interpretability-oriented review. IEEE Transactions on fuzzy systems, 9(3), 426-443.
[14] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

[15] Heerschop, B., van Iterson, P., Hogenboom, A., Frasincar, F., & Kaymak, U. (2011). Accounting for negation in sentiment analysis. In 11th Dutch-Belgian Information Retrieval Workshop (DIR 2011) (pp. 38-39).

[16] Hogenboom, A., van Iterson, P., Heerschop, B., Frasincar, F., & Kaymak, U. (2011). Analyzing sentiment while accounting for negation scope and strength. In 23rd Benelux Conference on Artificial Intelligence (BNAIC 2011) (pp. 327-328). BNAIC.

[17] Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., & Kaymak, U. (2015). Exploiting emoticons in polarity classification of text. J. Web Eng., 14(1&2), 22-40.

[18] Jurafsky, D., Martin, J. H. (2009). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Pearson Prentice Hall. ISBN: 9780131873216 0131873210

[19] Karami, A., Gangopadhyay, A., Zhou, B., & Kharrazi, H. (2018). Fuzzy approach topic discovery in health and medical corpora. International Journal of Fuzzy Systems, 20(4), 1334-1345.

[20] Kaya, M. Y., & Karsligil, M. E. (2010). Stock price prediction using financial news articles. In 2010 2nd IEEE International Conference on Information and Financial Engineering (pp. 478-482). IEEE.

[21] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

[22] Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584).

[23] Mahmud, M., Kaiser, M. S., McGinnity, T. M., & Hussain, A. (2021). Deep learning in mining biological data. Cognitive Computation, 13(1), 1-33.

[24] Menger, V., Scheepers, F., & Spruit, M. (2018). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. Applied Sciences, 8(6), 981.

[25] Menger, V., Scheepers, F., van Wijk, L. M., & Spruit, M. (2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. Telematics and Informatics, 35(4), 727-736.

[26] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. Unpublished

[27] Molnar, C. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

[28] Mosteiro, P., Rijcken, E., Zervanou, K., Kaymak, U., Scheepers, F., & Spruit, M. (2020). Making sense of violence risk predictions using clinical notes. In International Conference on Health Information Science (pp. 3-14). Springer, Cham.

[29] Mosteiro, P., Rijcken, E., Zervanou, K., Kaymak, U., Scheepers, F., & Spruit, M. R. (2021). Machine learning for violence risk assessment using Dutch clinical notes. Journal of Artificial Intelligence for Medical Sciences.

[30] Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 100-108).

[31] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2), 103-134.

[32] Ohno-Machado, L., & Séroussi, B. (2019). Identifying suicidal adolescents from mental health records using natural language processing. In MEDINFO 2019: Health and Wellbeing e-Networks for All: Proceedings of the 17th World Congress on Medical and Health Informatics (Vol. 264, p. 413). IOS Press.

[33] Olkin, I., & Rubin, H. (1964). Multivariate beta distributions and independence properties of the Wishart distribution. The Annals of Mathematical Statistics, 261-269.

[34] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[35] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365. Unpublished

[36] Pradhan, Ligaj, Chengcui Zhang, and Steven Bethard. "Towards extracting coherent user concerns and their hierarchical organization from user reviews." 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI). IEEE, 2016.

[37] Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In Twenty-fourth international joint conference on artificial intelligence.

[38] Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2)

[39] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

[40] Rumshisky, A., et. al. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational psychiatry, 6(10), e921-e921.

[41] Soni, A., Van Eck, N. J., & Kaymak, U. (2007). Prediction of stock price movements based on concept map information. In 2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (pp. 205-211). IEEE.

[42] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. In China National Conference on Chinese Computational Linguistics (pp. 194-206). Springer, Cham.

[43] Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In 2017 IEEE International conference on data science and advanced analytics (DSAA) (pp. 165-174). IEEE.

[44] Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. scientometrics, 84(2), 523-538.

[45] Van Le, D., Montgomery, J., Kirkby, K. C., & Scanlan, J. (2018). Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. Journal of biomedical informatics, 86, 49-58.

[46] Van Loan, C. F. (1976). Generalizing the singular value decomposition. SIAM Journal on numerical Analysis, 13(1), 76-83.

[47] Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.

[48] Wang, X., & McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 424-433).

[49] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic modeling of short texts: A pseudo-document view. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2105-2114).