2022 IEEE International Conference on
Systems, Man, and Cybernetics (SMC)

Exploring Embedding Spaces for more Coherent Topic Modeling in Electronic Health Records

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

Kalliopi Zervanou
Public Health & Primary Care
Leiden University Medical Centre
Leiden, The Netherlands
k.zervanou@liacs.leidenuniv.nl

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

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

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

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

2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-5258-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMC53654.2022.9945594

***Abstract*—The written notes in the Electronic Health Records contain a vast amount of information about patients. Implementing automated approaches for text classification tasks requires the automated methods to be well-interpretable, and topic models can be used for this goal as they can indicate what topics in a text are relevant to making a decision. We propose a new topic modeling algorithm, FLSA-E, and compare it with another state-of-the-art algorithm FLSA-W. In FLSA-E, topics are found by fuzzy clustering in a word embedding space. Since we use word embeddings as the basis for our clustering, we extend our evaluation with word-embeddings-based evaluation metrics. We find that different evaluation metrics favour different algorithms. Based on the results, there is evidence that FLSA-E has fewer outliers in its topics, a desirable property, given that within-topic words need to be semantically related.**

***Index Terms*—Topic Modeling, Natural Language Processing, Fuzzy Methods, Fuzzy Clustering, Psychiatry, Electronic Health Records, Word Embeddings, Neural Network methods**

I. Introduction

The Electronic Health Records (EHR) of a hospital are a rich source of information containing both structured fields and written notes. The written notes can be used as input to text classification algorithms. Some approaches of the classification of EHR notes include recognizing symptoms [1], [2], identifying suicidal behavior [3], [4], identification of adverse drug events [5], and violence prediction [6], [7]. Automated text classification approaches based on machine learning, utilizing clinical notes, have shown more accurate predictions than the analogue questionnaires [8]. Yet, in addition to accurate predictions, clinical providers and other decision-makers in healthcare consider the interpretability of model predictions as a priority for implementation and utilization. As machine learning applications are increasingly being integrated into various parts of the continuum of patient care, the need for prediction explanation is imperative [9]. Yet, many of these approaches are currently missing an intuitive understanding of the algorithms' inner workings. The clinical notes are represented numerically by large dense matrices with unknown semantic meaning.

One approach that carries the potential to make text classification more interpretable is using topic models. Topic modeling algorithms are a group of unsupervised natural language processing algorithms that extract latent topics in texts. Recently, FLSA-W was proposed [10], a topic modeling algorithm that outperforms other state-of-the-art algorithms in terms of the coherence (c_v)-, diversity- and interpretability score [11]. This approach represents the corpus as a term-document-matrix and then uses normal word weighting to obtain global term weights [12]. Then, singular value decomposition is used to project the global term weights into a lower-dimensional space. In this lower-dimensional space, fuzzy clustering and matrix multiplications are used to find the output matrices. The rationale behind the clustering in this space is that we assume that words are projected meaningfully so that semantically related words are located nearby each other. Instead of assuming a semantically meaningful projection, word embeddings created by algorithms such as Word2vec [13], Glove [14], and Fasttext [15] are known to project semantically related words nearby each other in an embedding space. Therefore, we hypothesize that clustering in such a space and following the same steps as with FLSA-W will likely result in better quality topics, topics with a higher coherence score. We refer to this algorithm as FLSA-E. Since topic models can make text classification of clinical notes more interpretable, the first goal of this paper is to compare FLSA-E, a new algorithm, with FLSA-W, an existing state-of-the-art algorithm. Secondly, since FLSA-E clusters in a word embedding space, we also experiment with two word embedding-based coherence measures, on top of the previously used c_v measure to evaluate how well within-topic words support each other [16], [17]. Although we hypothesize FLSA-E's topics to have a higher coherence score, we find varying results across coherence scores. FLSA-W's topics have significantly higher

c_v coherence scores, which is commonly used and is based on Normalized Pointwise Mutual Information (NPMI) [18]. Yet, with one of the word embedding-based coherence scores, FLSA-E scores best in almost all settings. In the other word embedding-based coherence score, both algorithms seem to perform equally well. This is a surprising finding indicating that various coherence scores measure different information. The structure of the paper is as follows. In Section II, we describe the various ways word embeddings are being used in topic modeling and various ways to measure topic coherence. In Section III, we describe how we obtained the data from the EHR and the experimental setup we used. In Section IV, we describe the various coherence scores resulting from our experiments and discuss their implications and limitations in Section V. Lastly, we conclude our work in Section VI and describe what should be done in future work.

II. Topic Modeling and Embeddings

This work compares the topics produced after fuzzy clustering on two types of word embeddings. This section discusses what topic models, word embeddings and topic coherence are.

A. Topic Models

Topic modeling is concerned with the discovery of latent semantic structure or topics within a set of documents, which can be derived from co-occurrences of words in documents [19]. Topic models are a group of unsupervised natural language processing algorithms that calculate two quantities:

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.

The top-*N* words with the highest probability per topic are typically taken to represent a topic. Topic models aim to find topics in which the top-*N* words in each topic are coherent with each other so that the topic is semantically interpretable and a common theme can be derived. Using topic embeddings for text classification, each input document is transformed into a vector of size C. Each cell indicates the extent to which the document belongs to a topic. After making predictions for each input text, interpretable classification algorithms can reveal which topics were most important for performing classifications.

B. Word Embeddings

Word embeddings are dense vectors used to represent words in a vector space. Popular algorithms for creating word embeddings from text are Word2vec [13], Glove [14] and Fasttext [15]. Word2vec was the first such algorithm to generate word vectors that explicitly encode linguistic regularities from large amounts of unstructured text [20] and it is the algorithm that we opted for in this work as a first step in our exploration of the use of word embeddings for topic modelling purposes. Word2vec is a neural network-based algorithm with two variants: the Continuous Bag-of-Words approach and the Skip-gram approach. The first approach predicts words based on their context, whereas the latter approach predicts context words based on the current word.

C. Coherence measures

Coherence is an evaluation measure to indicate how well intra-topic words, the words within a topic, support each other. Röder et al. proposed a unifying framework that represented coherence measures as a composition of parts to achieve a higher correlation with human judgements [18]. From all configurations in this space, 'c_v' coherence was found to correlate highest with human interpretation. c_v is based on NPMI (see (1)). For the calculation of the NPMI, a sliding window of size 110 is used to calculate the probabilities of word cooccurrence of the n most likely words in a topic. Then, the arithmetic mean is calculated to aggregate the scores for different topics.

$$\mathrm{NPMI(w)} = \frac{1}{n(n-1)} \sum_{q=2}^{n} \sum_{p=1}^{q-1} \frac{\log \frac{P(w_p, w_q)}{P(w_p)P(w_q)}}{-\log P(w_p, w_q)} \quad (1)$$

The NPMI formula shows that word pairs with high co-occurrence score high unless they are rare word pairs – which the denominator would normalize out. The NPMI scoring bears a remarkable resemblance to the contextual similarity produced by the inner product of word embedding vectors. Along this line of reasoning, word embedding based topic coherence was proposed [16]. We define the following properties:

D	the dimensionality of the embedding space,
E	the row-normalized word embedding matrix for a list of n words,
n	the number of (most likely) words per topic.

Then, $E \in \mathbb{R}^{n \times D}$ and $||E_i, ;|| = 1$. Let $\langle ., . \rangle$ denote the inner product. Then, we can define *pair-wise* word embedding topic coherence ($\mathrm{WEC_{pw}(E)}$) in a similar spirit as NPMI:

$$\mathrm{WEC_{pw}(E)} = \frac{1}{n(n-1)} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \langle E_{i,:}, E_{j,:} \rangle \quad (2)$$

$$= \frac{\sum \{E^T E\} - n}{2n(n-1)} \quad (3)$$

For each word pair the cosine similarity is calculated in the approach above. Since semantically related words are supposed to be located nearby each other, the average of these similarities reflects the coherence of the topic words.

Alternatively, the *centroid* word embedding topic coherence can be defined as follows:

$$\mathrm{WEC_c(E)} = \frac{1}{n} \sum \{E t^T\} \quad (4)$$

where vector $t \in \mathbb{R}^{1\times D}$ is the centroid of E, normalized to have $||t|| = 1$. With this approach, the average distance from the centroid is calculated for each dimension $\{1, 2, ..., D\}$ and the average over these averages is reported.

III. Data & Experimental Setup

A. Datasets

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 [6], [7], [10], [21]. The 834834 notes available are de-identified for patient privacy using DEDUCE [22]. 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. 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 admission period are concatenated and considered 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 [8], [23].

B. Methodology

1) Experimental Setup: This work compares the topics arising from clustering in two different spaces. For the first approach, we use the FLSA-W algorithm [10] (see Figure 1). This approach represents the corpus as a term-document-matrix ($M \times N$) and then uses normal word weighting [12] to obtain global term weights ($M \times N$). Then, singular value decomposition is used to project the global term weights into a lower-dimensional space ($M \times S$). In this lower-dimensional space, fuzzy clustering (fuzzy c-means clustering [24]) and matrix multiplications are used to find the output matrices. The rationale behind the clustering in this space is that we assume that words are projected meaningfully so that semantically related words are located nearby each other. Since word embeddings generated by algorithms such as Word2vec, FastText and Glove are known to project semantically related words nearby each other, we use this as the space to cluster in and find topics for our second approach. In the second approach, which we refer to as FLSA-E (see Figure 2), we train a Word2Vec word embedding from the corpus. Then, we use this word embedding as the input to the fuzzy clustering algorithm and follow similar steps as in FLSA-W to find the output matrices. To compare both methods, we train topic models with the following number of singular values (FLSA-W) and embedding sizes (FLSA-E): 2,3,...,20. For each of these settings, we train models with the following number of topics: 5, 10, 15, ..., 50.

2) Evaluation: The output of topic modeling algorithms consists of a set of topics, each consisting of a set of words. The quality of the topics can be measured both at the intra- and inter-topic levels. Since we are interested in the information captured within topics, we focus on the intra-topic level only and use the three coherence scores as described in Section II-C. We report the average score based on five runs for each of the tested settings. All scores range between zero and one, where one means perfect coherence and zero means no coherence.

IV. Results

Tables I, II and III show the different coherence scores for FLSA-W and FLSA-E. For each setting (coherence method, number of topics and embedding size/number of singular values), we compare FLSA-W with FLSA-E and show the best performance in boldface. Note that some settings are in boldface for both algorithms. In that case, the numbers rounded to three decimals are the same. From these tables, we can see that choosing the best-performing model highly depends on the used type of coherence method. With c_v coherence, FLSA-W outperforms FLSA-E in all settings, with WEC_{pw} coherence, both methods perform equally well approximately, and with WEC_c, FLSA-E outperforms FLSA-W in almost all settings. Therefore, it seems that the different coherence score seem to capture different kinds of information. Furthermore, with WEC_{pw} we cannot find a clear pattern explaining when which algorithm performs better. Lastly, the differences in performance between FLSA-W and FLSA-E are relatively small with WEC_c, whereas a significant difference occurs with c_v coherence.

V. Discussion

This work compares the topics produced from clustering in word embedding space (FLSA-E) with topics from singular value decomposition space (FLSA-W). We hypothesized that topics from the first would produce higher topics as word embeddings are known to locate semantically related words nearby each other. From the results, it can be seen that different metrics favour different algorithms. Therefore, we cannot yet draw conclusions based on our experiments. Ultimately, the coherence measure that correlates highest with human interpretation should be favoured over other coherence scores. If indeed c_v correlates highest with human interpretation, then this means that FLSA-W is the best performing algorithm. This would be a surprising finding indicating that the semantic information is better captured in singular value decomposition space than in word embedding space. However, such a correlation might vary over the analyzed data, and notes from electronic health records are likely to be differently structured than language in different domains. Therefore, comparing correlations of coherence scores with human interpretation should be done with topics from electronic health records.

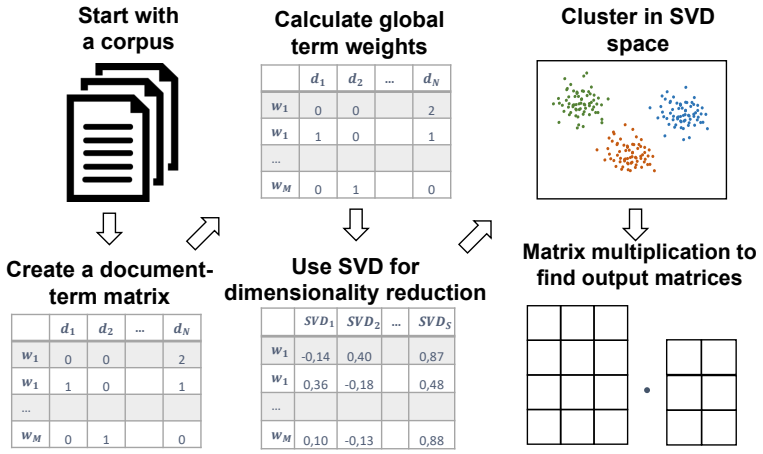

Fig. 1. Visual representation of FLSA-W's steps

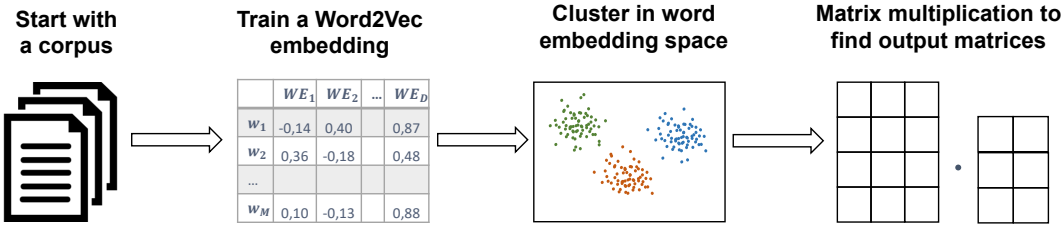

Fig. 2. Visual representation of FLSA-E's steps

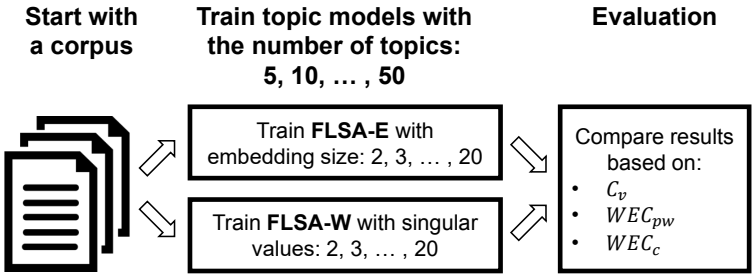

Fig. 3. Visual representation of Experimental setup

Furthermore, we assume WEC_c to be more sensitive to outliers than WEC_{pw}. An outlier would move the centroid away from other coordinates in the word embedding space and therefore the aggregated distances are larger. Since, FLSA-E's WES_c scores are almost all higher than FLSA-W, we expect FLSA-E to have fewer outliers than FLSA-W.

A. Limitations

A limitation of this work is that it used the Word2vec algorithm only as a space to cluster in and find the coherence scores, while other algorithms might be more suitable. Furthermore, we have only used one specific dataset. Therefore, our findings cannot be generalized. Also, since a topic modeling algorithm's output consists of a set of topics, each a set of words; the quality of produced topics should both indicate how well the words within topics support each other, intra-topic quality, and how unique the words in each topic are, inter-topic quality. In this study we focused on intra-topic quality only and will be extended in future work. Lastly, the dimensionality of the singular value decomposition- / word embedding space ranges between two and twenty. Word embeddings are typically trained with much higher dimensionality. However, clustering in such a high-dimensional space might not make sense as the number of dimensions would be close to, or much higher than, the number of topics. Also, the training time would be longer for a higher dimensionality.

VI. Conclusion

There are many applications of text classification based on electronic health records in the clinical domain. For these tasks, classification interpretability is imperative. Using topic modeling algorithms as topic embeddings for text classification might make a model more explainable. The experiments and comparisons between FLSA-W and the newly proposed

TABLE I

C_v COHERENCE FOR VARIOUS EMBEDDING SIZES AND NUMBER OF TOPICS. THE COLUMNS SHOW THE NUMBER OF SINGULAR VALUES (FLSA-W)/ THE EMBEDDING DIMENSIONALITY (FLSA-E) AND THE ROWS SHOW THE NUMBER OF TOPICS.

FLSA_W	2	4	6	8	10	12	14	16	18	20
5	**0.410**	**0.557**	**0.402**	**0.509**	**0.417**	**0.460**	**0.513**	**0.464**	**0.468**	**0.468**
10	**0.462**	**0.424**	**0.428**	**0.479**	**0.412**	**0.447**	**0.508**	**0.456**	**0.506**	**0.495**
15	**0.492**	**0.447**	**0.454**	**0.515**	**0.433**	**0.442**	**0.508**	**0.452**	**0.489**	**0.490**
20	**0.487**	**0.489**	**0.460**	**0.503**	**0.430**	**0.450**	**0.502**	**0.459**	**0.491**	**0.491**
25	**0.471**	**0.491**	**0.496**	**0.509**	**0.450**	**0.432**	**0.513**	**0.471**	**0.489**	**0.489**
30	**0.470**	**0.492**	**0.487**	**0.531**	**0.466**	**0.453**	**0.515**	**0.471**	**0.499**	**0.488**
35	**0.471**	**0.485**	**0.481**	**0.533**	**0.468**	**0.460**	**0.511**	**0.470**	**0.500**	**0.495**
40	**0.465**	**0.501**	**0.504**	**0.530**	**0.488**	**0.462**	**0.509**	**0.464**	**0.498**	**0.497**
45	**0.469**	**0.498**	**0.501**	**0.523**	**0.503**	**0.469**	**0.511**	**0.463**	**0.497**	**0.497**
50	**0.470**	**0.498**	**0.489**	**0.538**	**0.489**	**0.490**	**0.506**	**0.462**	**0.495**	**0.497**
FLSA_E	2	4	6	8	10	12	14	16	18	20
5	0.228	0.234	0.242	0.242	0.248	0.245	0.259	0.255	0.254	0.253
10	0.223	0.225	0.236	0.236	0.241	0.239	0.239	0.236	0.235	0.241
15	0.206	0.22	0.25	0.253	0.247	0.245	0.249	0.248	0.248	0.251
20	0.216	0.239	0.242	0.239	0.242	0.232	0.235	0.23	0.237	0.235
25	0.219	0.243	0.276	0.276	0.281	0.286	0.28	0.276	0.286	0.282
30	0.218	0.231	0.237	0.24	0.246	0.251	0.246	0.244	0.247	0.259
35	0.22	0.242	0.264	0.27	0.278	0.277	0.272	0.282	0.281	0.284
40	0.219	0.236	0.278	0.299	0.28	0.278	0.282	0.27	0.277	0.281
45	0.219	0.224	0.237	0.24	0.237	0.24	0.229	0.243	0.24	0.248
50	0.221	0.227	0.247	0.246	0.253	0.247	0.247	0.246	0.249	0.248

TABLE II

PAIRWISE NEURAL COHERENCE SCORES (WEC_{pw}) FOR VARIOUS EMBEDDING SIZES AND NUMBER OF TOPICS. THE COLUMNS SHOW THE NUMBER OF SINGULAR VALUES (FLSA-W)/ THE EMBEDDING DIMENSIONALITY (FLSA-E) AND THE ROWS SHOW THE NUMBER OF TOPICS.

FLSA_W	2	4	6	8	10	12	14	16	18	20
5	0.462	0.483	0.481	0.496	0.495	0.492	0.495	0.494	0.493	0.493
10	0.466	0.481	**0.478**	**0.488**	**0.494**	**0.493**	**0.495**	**0.493**	**0.492**	**0.491**
15	0.469	0.482	0.474	**0.497**	**0.496**	0.493	0.495	0.493	0.491	0.491
20	**0.476**	**0.480**	0.466	**0.483**	**0.488**	**0.494**	**0.494**	0.493	0.491	0.491
25	**0.477**	**0.475**	0.466	0.479	0.489	0.495	0.495	0.493	0.492	0.491
30	**0.474**	0.473	0.464	**0.492**	**0.491**	**0.493**	**0.495**	**0.493**	**0.492**	**0.491**
35	**0.477**	**0.469**	0.463	0.484	0.493	0.492	0.495	0.493	0.492	0.491
40	**0.476**	0.475	0.468	0.484	0.489	0.493	0.495	0.493	0.492	0.492
45	**0.476**	**0.477**	**0.466**	0.475	**0.488**	**0.492**	**0.495**	**0.494**	0.492	**0.492**
50	**0.478**	0.479	0.465	0.484	0.479	0.491	**0.494**	**0.493**	0.492	**0.492**
FLSA_E	2	4	6	8	10	12	14	16	18	20
5	**0.465**	**0.495**	**0.494**	**0.499**	**0.498**	**0.499**	**0.499**	**0.499**	**0.500**	**0.501**
10	**0.482**	**0.483**	0.473	0.472	0.473	0.475	0.475	0.474	0.473	0.474
15	**0.471**	**0.486**	**0.495**	0.494	**0.496**	**0.495**	**0.497**	**0.494**	**0.497**	**0.497**
20	0.464	0.469	**0.479**	0.471	0.469	0.468	0.470	0.467	0.468	0.468
25	0.465	0.4706	**0.490**	**0.496**	**0.498**	**0.497**	**0.499**	**0.500**	**0.500**	**0.499**
30	0.473	**0.484**	**0.490**	0.475	0.473	0.469	0.469	0.469	0.471	0.469
35	0.467	**0.469**	**0.499**	**0.503**	**0.500**	**0.500**	**0.501**	**0.500**	**0.501**	**0.495**
40	0.464	**0.478**	**0.499**	**0.497**	**0.496**	**0.496**	**0.498**	**0.497**	**0.498**	**0.497**
45	0.467	0.474	**0.466**	0.479	0.472	0.478	0.476	0.483	**0.493**	**0.492**
50	0.469	**0.487**	**0.495**	**0.493**	**0.491**	**0.493**	0.493	0.492	**0.496**	**0.492**

model FLSA-E in this work are aimed to find better interpretable topic models by exploring word embedding spaces as the basis for clustering. Given the varying outcomes per coherence measure, we cannot yet conclude which algorithm is preferred. Yet, there seems to be evidence that the topics produced in FLSA-E have fewer outliers, which is a desirable property.

Future work will experiment with more algorithms than Word2vec only (such as Glove and Fasttext) to train the word embeddings. Also, correlations between produced topics and human interpretation will indicate which algorithm is preferred by humans. Furthermore, further analyses will be done on the context in which the coherence score is most appropriate. Additionally, inter-topic quality measures will be used in the analysis in future studies. Lastly, topics derived from higher-dimensional embedding spaces might lead to better topics.

REFERENCES

[1] E. Scheurwegs, M. Sushil, S. Tulkens, W. Daelemans, and K. Luyckx, "Counting trees in random forests: predicting symptom severity in psychiatric intake reports," *Journal of biomedical informatics*, vol. 75, pp. S112–S119, 2017.

[2] D. Chandran, D. A. Robbins, C.-K. Chang, H. Shetty, J. Sanyal, J. Downs, M. Fok, M. Ball, R. Jackson, R. Stewart *et al.*, "Use of natural language processing to identify obsessive compulsive symptoms in

TABLE III
CENTROID-BASED NEURAL COHERENCE SCORES (WEC_c) FOR VARIOUS EMBEDDING SIZES AND NUMBER OF TOPICS. THE COLUMNS SHOW THE NUMBER OF SINGULAR VALUES (FLSA-W)/ THE EMBEDDING DIMENSIONALITY (FLSA-E) AND THE ROWS SHOW THE NUMBER OF TOPICS.

FLSA_W	2	4	6	8	10	12	14	16	18	20
5	0.852	0.837	0.849	0.842	0.847	0.828	0.841	0.827	0.82	0.821
10	0.899	**0.910**	**0.885**	**0.852**	0.846	0.841	0.846	0.819	0.814	0.803
15	0.857	0.910	0.894	0.870	0.856	0.841	0.844	0.818	0.801	0.801
20	0.865	0.879	0.883	0.898	0.850	0.847	0.843	0.816	0.802	0.802
25	0.891	0.854	0.881	0.893	0.846	0.862	0.848	0.819	0.805	0.802
30	0.884	0.855	0.872	0.884	0.843	0.865	0.848	0.817	0.807	0.806
35	0.883	0.846	0.870	0.865	0.860	0.855	0.846	0.818	0.806	0.807
40	0.886	0.87	0.882	0.862	0.854	0.866	0.848	0.823	0.806	0.806
45	0.874	0.873	0.863	0.861	0.830	0.863	0.850	0.825	0.807	0.807
50	0.890	0.877	0.867	0.860	0.832	0.858	0.849	0.821	0.806	0.807
FLSA_E	2	4	6	8	10	12	14	16	18	20
5	**0.899**	**0.897**	**0.881**	**0.897**	**0.902**	**0.907**	**0.907**	**0.902**	**0.914**	**0.930**
10	**0.944**	0.883	0.858	0.850	**0.874**	**0.872**	**0.866**	**0.862**	**0.862**	**0.870**
15	**0.926**	**0.911**	**0.943**	**0.901**	**0.956**	**0.925**	**0.946**	**0.959**	**0.953**	**0.943**
20	**0.912**	**0.888**	**0.913**	**0.908**	**0.900**	**0.903**	**0.903**	**0.892**	**0.902**	**0.891**
25	**0.909**	**0.908**	**0.944**	**0.918**	**0.921**	**0.921**	**0.952**	**0.920**	**0.936**	**0.927**
30	**0.915**	**0.918**	**0.908**	**0.915**	**0.900**	**0.920**	**0.909**	**0.903**	**0.924**	**0.916**
35	**0.917**	**0.892**	**0.898**	**0.912**	**0.906**	**0.907**	**0.907**	**0.916**	**0.924**	**0.892**
40	**0.918**	**0.900**	**0.928**	**0.910**	**0.881**	**0.877**	**0.895**	**0.871**	**0.885**	**0.892**
45	**0.924**	**0.894**	**0.883**	**0.887**	**0.869**	**0.893**	**0.881**	**0.894**	**0.914**	**0.913**
50	**0.913**	**0.895**	**0.897**	**0.870**	**0.852**	**0.864**	**0.865**	**0.849**	**0.867**	**0.857**

patients with schizophrenia, schizoaffective disorder or bipolar disorder," *Scientific reports*, vol. 9, no. 1, pp. 1–7, 2019.

[3] A. C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran, "Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[4] N. J. Carson, B. Mullin, M. J. Sanchez, F. Lu, K. Yang, M. Menezes, and B. L. Cook, "Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records," *PloS one*, vol. 14, no. 2, p. e0211116, 2019.

[5] E. Iqbal, R. Mallah, R. G. Jackson, M. Ball, Z. M. Ibrahim, M. Broadbent, O. Dzahini, R. Stewart, C. Johnston, and R. J. Dobson, "Identification of adverse drug events from free text electronic patient records and information in a large mental health case register," *PloS one*, vol. 10, no. 8, p. e0134208, 2015.

[6] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers, and M. Spruit, "Making sense of violence risk predictions using clinical notes," in *International Conference on Health Information Science*. Springer, 2020, pp. 3–14.

[7] ——, "Machine learning for violence risk assessment using Dutch clinical notes," *Journal of Artificial Intelligence for Medical Sciences*, vol. 2, no. 1-2, pp. 44–54, 2021.

[8] V. Menger, F. Scheepers, and M. Spruit, "Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text," *Applied Sciences*, vol. 8, no. 6, p. 981, 2018.

[9] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018, pp. 559–560.

[10] E. Rijcken, F. Scheepers, P. Mosteiro, K. Zervanou, M. Spruit, and 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, pp. 1–8.

[11] E. Rijcken, U. Kaymak, F. Scheepers, P. Mosteiro, K. Zervanou, and M. Spruit, "FLSA-W as an Interpretable Topic Modeling Algorithm," *Submitted*, 2022.

[12] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy approach topic discovery in health and medical corpora," *International Journal of Fuzzy Systems*, vol. 20, no. 4, pp. 1334–1345, 2018.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431. [Online]. Available: https://aclanthology.org/E17-2068

[16] R. Ding, R. Nallapati, and B. Xiang, "Coherence-aware neural topic modeling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 830–836. [Online]. Available: https://aclanthology.org/D18-1096

[17] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[18] M. Röder, A. Both, and 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.

[19] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[21] E. Rijcken, U. Kaymak, F. Scheepers, P. Mosteiro, K. Zervanou, and M. Spruit, "Topic modeling for interpretable text classification from EHRs," *Frontiers in Big Data*, p. 846930, 2022.

[22] V. Menger, F. Scheepers, L. M. van Wijk, and M. Spruit, "DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text," *Telematics and Informatics*, vol. 35, no. 4, pp. 727–736, 2018.

[23] D. Van Le, J. Montgomery, K. C. Kirkby, and J. Scanlan, "Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting," *Journal of Biomedical Informatics*, vol. 86, pp. 49–58, 2018.

[24] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media, 2013.