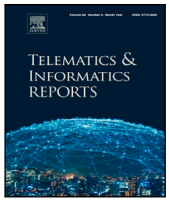
Telematics and Informatics Reports 13 (2024) 100126

Contents lists available at ScienceDirect

Telematics and Informatics Reports

journal homepage: www.elsevier.com/locate/teler



Topic modeling for conversations for mental health helplines with utterance embedding

Salim Salmi [a,*], Rob van der Mei [a], Saskia Mérelle [b], Sandjai Bhulai [c]

[a] Centrum Wiskunde & Informatica, Netherlands
[b] 113 Suicide Prevention, Netherlands
[c] Vrije Universiteit Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Topic modeling
Sentence embedding
Conversations
Mental health
Bert

ABSTRACT

Conversations with topics that are locally contextual often produces incoherent topic modeling results using standard methods. Splitting a conversation into its individual utterances makes it possible to avoid this problem. However, with increased data sparsity, different methods need to be considered. Baseline bag-of-word topic modeling methods for regular and short-text, as well as topic modeling methods using transformer-based sentence embeddings were implemented. These models were evaluated on topic coherence and word embedding similarity. Each method was trained using single utterances, segments of the conversation, and on the full conversation. The results showed that utterance-level and segment-level data combined with sentence embedding methods performs better compared to other non-sentence embedding methods or conversation-level data. Among the sentence embedding methods, clustering using HDBScan showed the best performance. We suspect that ignoring noisy utterances is the reason for better topic coherence and a relatively large improvement in topic word similarity.

1. Introduction

Topic modeling is a common method to extract latent semantic information from text in documents. It is often used to gain insight into a corpus of some kind, for example, what the current events are on social media. Latent Dirichlet Allocation (LDA) is the most popular method for topic modeling [1]. Many alternatives have been proposed that aim to outperform LDA in general cases, or for special kinds of text content such as short text.

In this article, we highlight one of the special kinds of corpora, namely conversations. Specifically, we look at conversations that are characterized by two things: 1. Information that is frequently locally contextual, and 2. utterances that do not conform to a recurring topic. Such a corpus conflicts with the assumption made by LDA and other bag-of-words models, namely that the order of words does not matter. For example, online mental health services or eHealth could use topic modeling to gain insight into their conversation data. Using more recent methods, topic modeling has been applied in counseling conversations on suicide prevention helplines [2].

Especially in long conversations, word-document co-occurrence lacks the information to describe the topics accurately. The resulting topics become too general or not coherent. In structured documents, one could split the document at some point in the hierarchy to reduce the size, for example per chapter of a book [3].

Similarly, a conversation can be viewed as a collection of utterances between participants, where an utterance is a statement or message from one of the participants. It is more likely that participants of the conversation cover a small number of topics at a time. By splitting a conversation into its utterances, one can reduce the number of topics per observation. In this way, the co-occurrence of words has more descriptive value. We hypothesize that for conversations, especially mental health conversations, where descriptive value of words is already low, splitting the conversation into utterances improves the topic model coherence. However, splitting the conversation increases the sparsity of the data. Due to the nature of statistical inference used by many topic modeling methods, this is a problem.

Several topic methods have tried to move away from the bag-of-words assumption and take sentence-level context into account [4–6]. [7] proposed the so-called Sentence Level Recurrent Topic Model to create distributed representations of sentences and topics using a Long-Term Short-Term neural network. Sentence embedding is a method to find a latent representation of a sentence in a continuous and lower-dimensional space. The most successful sentence embedding methods rely on the recent transformer model [8]. This model has

* Correspondence to: P.O. Box 94079, 1090 GB Amsterdam, Netherlands.
E-mail address: s.salmi@cwi.nl (S. Salmi).

https://doi.org/10.1016/j.teler.2024.100126
Received 1 June 2023; Received in revised form 23 January 2024; Accepted 16 February 2024
Available online 27 February 2024

shown improvement in many natural language processing tasks [9–11]. Sentence embedding can be applied to topic modeling to extract more information than simple word co-occurrence.

A downside of using sentence embedding for topic modeling is that the embedding can only be as long as the maximum input length of the model. This means that for long documents, words after the maximum length are truncated. However, conversation utterances are small enough that this maximum length is unlikely to be reached. For this reason, we believe that sentence embedding methods can be of benefit for utterance level conversation data.

A middle ground between utterances and a full conversation is to segment a conversation into groups of utterances. However, this creates an additional problem, namely: how should the conversation be segmented? A frequently used method for the purpose of text segmentation modifies the TextTiling algorithm [12]. This method employs embedding models based on BERT to embed or classify sentences to provide a lexical score for each sentence [13–15]. Lexical similarity lower than a threshold indicates where the text should be segmented.

In this article, we evaluate the application of several topic modeling methods to three conversation corpora. We evaluate each topic model for three different representations of the datasets: where each document is either the concatenated conversation, i.e. on a conversation-level, each document is a single utterance from a conversation, i.e. on a utterance-level, or each document is a segment of several utterances from a conversation, i.e. on a segment-level.

As a baseline for comparing short and conventional length topic modeling, we include GSDMM [16] and LDA, respectively. LDA sees the most frequent use when topic modeling for conversation is considered. However, we believe contextual information can be of benefit, especially when applied to utterance level data. Therefore, we also include three models using sentence embedding information, namely BERTopic [17], TopClus [18] and CombinedTM [19].

Furthermore, we compare the performance of sentence embedding trained on supervised data from a different domain, to unsupervised sentence embedding methods on the counseling corpus domain. We show that the contextual models perform better than both non contextual methods, and that BERTopic using utterance-level and segment-level data performs the best on the majority of evaluation metrics.

2. Background

2.1. Topic modeling in conversation

LDA is a generative topic model, where a document is seen as a mixture of topics and a topic is seen as a distribution over words. These topic mixtures are then sampled from a Dirichlet distribution. LDA is a basic but often powerful enough method for the goal of topic modeling. LDA has been frequently used for the purpose of basic conversation topic modeling. [20] used a variant of LDA in a crisis line, supported by expert input, to create interpretable and coherent topics. [21] proposed a model to discover topics from healthcare chat logs. Their method adapted LDA to capture noise in the form of latent personal interests of chat users.

2.2. Short texts

When topic modeling is applied to short texts like messages, the data becomes very sparse. This sparsity results in problems when models rely on document-level word co-occurrence. To tackle the sparsity problem, many different approaches have been considered. Short texts can be heuristically aggregated into longer pseudo documents as a straightforward solution [22,23]. However, sometimes this is not possible or desirable. Other methods are based on making the stronger assumption that a document covers only one topic. Biterm Topic Modeling (BTM) [24] is a popular method of topic modeling for short texts. Instead of word-document co-occurrence, BTM models the co-occurrence of word pairs (bi-terms) in the entire corpus. This method has also been extended with the use of word embeddings [25]. To deal with the problem of noise in short text, the so-called Common Semantics Topic Model implores a common topic to which it assigns noise words [26]. Rashid et al. approach the sparsity problem using a fuzzy clustering approach named Fuzzy Topic Modeling (FTM) [27]. In fuzzy clustering, words can belong to multiple clusters based on a membership function. FTM first applies dimensionality reduction through PCA and afterward uses fuzzy c-means clustering to assign words to topics. Yin et al. propose a short text clustering method called Gibbs Sampling for the Dirichlet Multinomial Mixture model (GSDMM) [16]. This generative method assumes a document corresponds to a single topic. Unlike LDA, documents are generated using the same topic.

2.3. Topic modeling using variational inference

Several methods have been proposed to incorporate neural networks for the field of topic modeling, the most successful of which are based on variational inference. ProdLDA is a method that uses variational autoencoding as a neural network approach to infer the LDA posteriors [28]. Word embeddings obtained through methods such as continuous bag-of-words by Mikolov et al. [29] are also used in topic modeling. Using word embeddings, Dieng et al. developed the Embedded Topic Model (ETM) [30]. ETM is a generative model that, like LDA, models a document as a mixture of topics. However, ETM uses distributed representations for both words and topics. The topic embeddings are inferred using variational inference. A decoder network reconstructs the words belonging to the topics. Based on this method, Combined Topic Model (CombinedTM) [19] concatenates sentence embeddings from SBERT to the bag of words representation of a document and uses the concatenation as input for the autoencoding model of ProdLDA. Using these embeddings CombinedTM shows improved performance over ETM and ProdLDA.

2.4. Topic modeling using pre-trained embedding

Top2Vec [31] is a topic modeling method uses clustering of document embeddings, where the resulting clusters form the basis for the topics. After embedding the documents, dimensionality reduction is applied to the resulting embeddings using Uniform Manifold Approximation and Projection for Dimensionality Reduction (UMAP) [32]. UMAP reduces dimensionality by approximating a higher-dimensional manifold and projecting it to a lower dimension. Afterward, HDBScan is used to cluster the resulting transformed data. HDBScan is an extension of the density-based spatial clustering of applications with noise (DBScan) algorithm [33]. Using hierarchical clustering, it can accurately detect clusters of varying densities and shapes, while avoiding noise. Furthermore, HDBScan requires little in terms of hyperparameter optimization. BERTopic [17] adapts this method, to use document embeddings obtained through Sentence-BERT (SBERT) [34], a modification of BERT using siamese networks. BERTopic combines the tokens of the documents assigned to a cluster into a single set. For each set, words with the highest TF-IDF value are used to describe the topic represented by that cluster. Sia et al. [35] used an approach similar to Top2Vec. They applied PCA dimensionality reduction along with K-means and Gaussian Mixture Models on word embeddings. Additionally they used weighted clustering and term reranking to obtain the topics. A weakly supervised approach was proposed by Meng et al. [36] where category names can be provided to increase the interpretability of the topic models. TopClus [18] uses pretrained models to learn topic representations from a latent spherical space. The spherical space benefits clustering of the embeddings while also reducing the dimensionality.

Table 1
Utterance and Conversation frequencies.

Corpus	Utterances	Conversations
Counseling Corpus	55,811	4,508
AnnoMI	9661	133
MultiWoz	72,022	7,679

3. Methods

This section covers the main steps for evaluating the different models. First, we highlight which models we chose to evaluate. Second, we explain which conversation corpora we used. Third, we cover the text pre-processing steps. Fourth, we show the different evaluation metrics used.

3.1. Models

We chose several topic models to evaluate on the utterance- and conversation-level data. As a baseline for short and conventional length topic modeling, we used GSDMM and LDA, respectively. For embedding-based models, we use BERTopic, TopClus and CombinedTM. Lastly, we adapted BERTopic, where instead of the BERT embeddings with averaged word2vec embeddings. We will refer to this variant as Clustering W2V. This allowed us to observe the impact of BERT embeddings had compared to a less computationally expensive embedding method. Every topic model method was tested on both utterances and the full conversation. Each topic model was optimized for 30, 60, and 90 topics. Hyperparameters for each model were optimized and the models with the best topic coherence scores were reported.

3.2. Datasets

For evaluation of the models, we used three corpora. The first corpus is a counseling corpus from a suicide prevention helpline. This corpus is difficult to model with traditional bag-of-words methods such as LDA. The second corpus we included is an English mental health dataset, of transcribed interviews using the motivational interviewing paradigm called AnnoMI [37]. The third corpus we included is the Multi-Domain Wizard-of-Oz dataset (MultiWoz). MultiWoz is a dataset of dialogues containing conversations covering topics from multiple domains. We included the MultiWoz corpus to highlight the difference between the more frequent topic modeling setting existing in literature and the mental health corpora. MultiWoz covers specific domains where local context is less important, whereas this is not the case for the other two corpora. As mentioned previously, models will be tested both on utterances and full conversations. This results in a total of six datasets. Table 1 shows the sizes of each dataset.

The counseling corpus dataset contained chat conversations from a suicide prevention helpline. In this corpus, a help seeker contacts a counselor with an issue regarding suicide. It was the counselor's job to listen to a help seeker and to explore options with this help seeker where necessary. This corpus contained conversations covering many topics. Conversations also contained interactions that were not part of recurring topics. The MultiWoz dataset is a conversation dataset spanning multiple topics and domains. Compared to the counseling corpus, MultiWoz contains shorter conversations and covers strict topics, with less interference from small talk for example. Because of this, MultiWoz was expected to have less variability in the performance.

3.3. Document representations

Topic modeling uses documents as input, where each document can potentially contain a number of topics. In this study we looked at three ways to represent single document in the context of chat conversations, and how this impacts several topic modeling methods. First we represented an entire conversation as a single document (conversation-level), by concatenating all utterances. In the context of chat data, we defined an utterance as anything a participant says until responded to by another participant. Second we represented a single utterance as a single document (utterance-level). Third, we segmented a conversation into groups of consecutive utterances and used the concatenation of a group as a single document (segment-level).

The segment level data was created using the following four steps:

- We trained a binary classifier on a RoBERTa network with a next sentence prediction (NSP) objective. We used pairs of consecutive messages and pairs of random messages from the datasets as the training data.
- Using the NSP network, consecutive messages pairs were scored.
- Message pairs that scored below a threshold were marked as the end and beginning of two sections. This threshold was set such that segments averaged 5 messages.
- Utterances belonging to the same segment were concatenated.

3.4. Pre-processing

Pre-processing of the text consisted of six steps. First, all non-alphabetic characters were removed. Second, all text was lowercased. Third, the text was lemmatized. Fourth, stop words were removed using the NLTK library of stop words. Fifth, the text was tokenized, removing any tokens that were shorter than three characters. Sixth, all but the 2,000 most frequent remaining tokens in the corpus were filtered out. For the sentence embedding, the text was only cleaned minimally, by removing special characters. On the utterance datasets, utterances with fewer than five words were removed. The chat messages and tokens were aggregated for each conversation to create the conversation dataset.

3.5. Pre-trained sentence embedding

For the sake of consistency, we used the same SBERT model for the topic modeling methods that leveraged sentence embedding. The models we used were "paraphrase-mpnet-base-v2" and "paraphrase-multilingual-mpnet-base-v2" for English and Dutch texts, respectively. The input for sentence embedding was minimally preprocessed. However, preprocessing was done to obtain tokens to describe each topic after clustering.

3.6. Unsupervised fine-tuning

It is worth noting that the domain of helpline conversations differs from the training data used for pre-trained networks. In addition to employing sentence embedding techniques via pre-trained networks, our study explored an alternative method for creating embeddings.

State-of-the-art sentence embedding models predominantly rely on supervised training, which involves labeled data comprising sentence pairs and their corresponding similarity scores. However, this is not always available, as is the case for the counseling corpus we are exploring in this study. To address this limitation, we implemented two unsupervised fine-tuning approaches and compared them to the outcomes using the pre-trained embeddings.

First, we used a pre-trained Dutch RoBERTa model and fine-tuned it using a triplet loss function. The triplet loss function, in this context, is defined as follows:

$$\text{Loss}(A, P, N) = \max(d(A, P) - d(A, N) + \text{margin}, 0)$$

Where d is some distance function. A, P and N are embeddings for an anchor, a positive and a negative message respectively. The positive message is of the same class as the anchor message while the negative message is of a different class. This approach encourages the network

to put at least a margin of distance between the anchor-positive pair and the anchor-negative pair.

To apply this to the counseling conversation dataset, we selected a pair of consecutive messages from a given conversation as the anchor and positive samples. Subsequently, we chose a message from a different sentence pair as the negative sample. This approach assumes that consecutive messages are more likely to be related, whereas two randomly selected messages are more likely not to be related. The training dataset consisted of all possible positive pairs. The negative message was randomly selected from the same batch. We used mean pooling to obtain the sentence embeddings, and fine-tuned the RobBERT network, a RoBERTa network trained on a large Dutch language corpus.

Second, we used the Transformer-based Denoising AutoEncoder (TSDAE) to fine-tune the same pre-trained Dutch RoBERTa model [38]. Instead of generating data through the conversation structure, TSDAE introduces noise to the data and trains an autoencoder to denoise this data. The embedding of the class token represents the sentence embedding.

We compared these two unsupervised approaches to the results of the supervised pre-trained sentence embedding model for the BERTopic model.

3.7. Evaluation metrics

We used two metrics to evaluate the models. First, we computed the topic coherence using the C_v metric [39]. This metric combines normalized pointwise mutual information and cosine similarity with a sliding window. To keep the comparison unbiased, topic coherence was calculated for the full conversation variant of each corpus. Second, we used Word2Vec embeddings for an indication of semantic relatedness. For each topic, we computed the average pairwise cosine similarity of the word embeddings using the top-5 words of the topic. We discount this within-topic relatedness by the between-topic relatedness, using the inverse of the average pairwise cosine similarity between the average Word2Vec embedding of each topic. We define the word embedding score $W = W_{\text{within}} W_{\text{between}}^{-1}$ where

$$W_{\text{within}} = \frac{1}{kl(l-1)/2} \sum_{i}^{k} \sum_{j}^{l} \sum_{m=j+1}^{l} sim(w_{ij}, w_{im}) \tag{1}$$

$$W_{\text{between}} = \frac{1}{k(k-1)/2} \sum_{i}^{k} \sum_{j=i+1}^{k} sim(\frac{\sum_{m}^{l} w_{im}}{l}, \frac{\sum_{m}^{l} w_{jm}}{l}) \tag{2}$$

We have w_{ij} as the word embedding for word j in topic i. The number of topics is denoted by k where $k = 30, 60, 90$. The number of selected words per topic is denoted by l. In our evaluation we let $l = 5$, using the top five best ranking words for each topic, according to each topic model's own metric of ranking.

4. Evaluation

4.1. Topic coherence

Table 2 shows the topic coherence for all models on the counseling conversation corpus. Here we can see that BERTopic scores higher than the other models on all evaluated topic sizes. We observe for 90 topics that the segmented conversations perform better, but the smaller topic sizes show better performance using only the utterances. Notably, the sentence embedding-based CombinedTM and TopClus improved using utterances or segments, instead of a full conversation, most likely due to the length limit of the sentence embedding model. As expected, LDA decreased in performance due to the sparsity of the smaller documents. BERTopic with average word embeddings on utterances also performed well across all topic ranges. GSDMM was not able to produce enough topics on the full conversation dataset and was left out of the results. On the full conversation corpus, LDA performed similarly to most other models.

Table 2
Topic Coherence for counseling conversation corpus.

Models		Number of topics		
		30	60	90
Utterances	LDA	0.45	0.45	0.45
	GSDMM	0.50	0.51	0.50
	CombinedTM	0.52	0.52	0.52
	BERTopic	**0.60**	**0.61**	0.59
	Clustering W2V	0.54	0.56	0.55
	TopClus	0.53	0.52	0.53
Segmented	LDA	0.47	0.47	0.47
	GSDMM	0.53	0.54	0.53
	CombinedTM	0.56	0.46	0.46
	BERTopic	0.53	0.59	**0.62**
	Clustering W2V	0.54	0.53	0.57
	TopClus	0.55	0.50	0.53
Full conversation	LDA	0.52	0.52	0.52
	GSDMM	–	–	–
	CombinedTM	0.46	0.46	0.46
	BERTopic	0.51	0.52	0.51
	Clustering W2V	0.50	0.52	0.52
	TopClus	0.45	0.45	0.46

Table 3
Topic Coherence for AnnoMI.

Models		Number of topics		
		30	60	90
Utterances	LDA	0.59	0.59	**0.60**
	GSDMM	0.57	0.60	0.59
	CombinedTM	0.45	0.46	0.44
	BERTopic	0.59	0.59	0.57
	Clustering W2V	0.58	0.57	0.57
	TopClus	0.50	0.51	0.50
Segmented	LDA	0.53	0.54	0.52
	GSDMM	0.57	0.60	0.59
	CombinedTM	0.42	0.43	0.44
	BERTopic	**0.59**	**0.61**	0.59
	Clustering W2V	0.58	0.59	0.57
	TopClus	0.47	0.46	0.47
Full conversation	LDA	0.48	0.48	0.48
	GSDMM	0.48	0.48	0.48
	CombinedTM	0.41	0.38	0.35
	BERTopic	0.50	0.49	0.50
	Clustering W2V	0.52	0.51	0.49
	TopClus	0.40	0.43	0.43

Table 3 shows the coherence scores for the AnnoMI corpus. Models trained on utterances and segmented also do better than on the full conversation for this dataset. However, LDA and GSDMM perform much better on the AnnoMI compared to the counseling conversations. BERTopic shows the best performance for 30 and 60 topics and LDA show the best performance for 90 topics.

Table 4 contains the coherence scores for the MultiWoz corpus. On this corpus, BERTopic also outperformed the other algorithms that were tested. However, this time the full conversation performed better than the utterances. Notably, all the clustering methods on both utterances and conversation performed better than the generative methods.

4.2. Average pairwise word embedding similarity

The word embedding similarity scores can be found in Table 5. For this metric, the differences are more pronounced compared to the topic coherence. On the counseling conversation corpus, we again find BERTopic to outperform other models using utterance data. Similarly, BERTopic with word2vec embeddings also show good performance. It also performs better than both sentence embedding methods on the full corpus. This is most likely due to the length limitation of sentence embedding. Low scores for LDA demonstrate the difficulty this method has with this type of noisy data. Even on the full conversation corpus,

Table 4
Topic Coherence for counseling MultiWoz corpus.

Models		Number of topics		
		30	60	90
Utterances	LDA	0.42	0.43	0.43
	GSDMM	0.56	0.57	0.57
	CombinedTM	0.63	0.60	0.61
	BERTopic	0.72	0.72	0.67
	Clustering W2V	0.72	0.70	0.65
	TopClus	0.65	0.65	0.63
Segmented	LDA	0.49	0.49	0.49
	GSDMM	0.47	0.47	0.48
	CombinedTM	0.60	0.60	0.58
	BERTopic	0.68	0.75	0.77
	Clustering W2V	0.71	0.69	0.66
	TopClus	0.67	0.69	0.65
Full conversation	LDA	0.57	0.60	0.59
	GSDMM	0.54	0.54	0.53
	CombinedTM	0.62	0.62	0.61
	BERTopic	**0.80**	**0.81**	**0.80**
	Clustering W2V	0.75	0.72	0.70
	TopClus	0.70	0.71	0.69

Table 5
Word embedding similarity scores for counseling conversation corpus.

Models		Number of topics		
		30	60	90
Utterances	LDA	0.47	0.37	0.37
	GSDMM	0.45	0.46	0.50
	CombinedTM	0.78	0.73	0.78
	BERTopic	**1.70**	**1.87**	**1.77**
	Clustering W2V	1.28	1.44	1.37
	TopClus	0.67	0.66	0.66
Segmented	LDA	0.33	0.34	0.34
	GSDMM	0.45	0.45	0.46
	CombinedTM	0.70	0.66	0.65
	BERTopic	0.77	0.79	0.78
	Clustering W2V	0.80	0.81	0.80
	TopClus	0.65	0.59	0.63
Full conversation	LDA	0.52	0.39	0.38
	GSDMM	–	–	–
	CombinedTM	0.65	0.66	0.65
	BERTopic	0.52	0.58	0.59
	Clustering W2V	0.89	0.80	0.81
	TopClus	0.67	0.66	0.66

LDA does not perform well. Furthermore, most models perform better when using the utterance datasets.

Word embedding scores for the AnnoMI dataset can be found in Table 6. LDA on full conversation showed overall the best scores. This is in contrast to the poor performance for topic coherence LDA obtained. For the other methods, utterance and segmented approaches performed better.

Table 7 contains the word embedding scores for MultiWoz. The highest scores are seen in the models using utterance data, shared between the clustering models.

4.3. Pre-trained and unsupervised sentence embedding

Table 8 shows the topic coherence using BERTopic with pre-trained embeddings and fine-tuned embeddings using the triplet loss and TSDAE unsupervised methods. Both fine-tuning methods show a marginal improvement over only pre-trained embeddings. Since the difference is small, the additional fine-tuning of the dataset could be omitted.

4.4. Topic words

Table 9 the top 5 topic words for 15 topics using the best performing BERTopic model. Clear topics can be discerned where this

Table 6
Word embedding similarity scores for AnnoMI.

Models		Number of topics		
		30	60	90
Utterances	LDA	0.76	0.77	0.77
	GSDMM	0.80	0.79	0.79
	CombinedTM	0.58	0.60	0.55
	BERTopic	0.76	0.75	0.72
	Clustering W2V	0.75	0.76	0.70
	TopClus	0.59	0.61	0.58
Segmented	LDA	0.79	0.76	0.77
	GSDMM	0.75	0.76	0.78
	CombinedTM	0.59	0.58	0.55
	BERTopic	0.70	0.70	0.69
	Clustering W2V	0.71	0.69	0.70
	TopClus	0.59	0.57	0.60
Full conversation	LDA	**0.83**	**0.82**	**0.82**
	GSDMM	–	–	–
	CombinedTM	0.57	0.56	0.55
	BERTopic	0.75	0.78	0.76
	Clustering W2V	0.69	0.71	0.68
	TopClus	0.62	0.64	0.61

Table 7
Word embedding similarity scores for MultiWoz corpus.

Models		Number of topics		
		30	60	90
Utterances	LDA	0.40	0.26	0.26
	GSDMM	0.23	0.23	0.25
	CombinedTM	0.53	0.51	0.50
	BERTopic	0.52	**0.62**	0.60
	Clustering W2V	**0.58**	0.52	**0.66**
	TopClus	0.54	0.54	0.51
Segmented	LDA	0.42	0.38	0.31
	GSDMM	0.24	0.20	0.27
	CombinedTM	0.36	0.37	0.46
	BERTopic	0.50	0.52	0.59
	Clustering W2V	0.53	0.48	0.55
	TopClus	0.53	0.54	0.54
Full conversation	LDA	0.28	0.33	0.34
	GSDMM	0.19	0.19	0.17
	CombinedTM	0.32	0.34	0.31
	BERTopic	0.44	0.45	0.44
	Clustering W2V	0.36	0.43	0.45
	TopClus	0.29	0.34	0.31

Table 8
Topic Coherence for different sentence embedding methods using BERTopic on the counseling conversation corpus.

Models	Number of topics		
	30	60	90
Pre-trained SBERT	0.60	0.61	0.59
Tripletloss	0.61	0.61	0.60
TSDAE	0.60	0.61	0.61

was not possible using classical LDA methods on either utterance- or conversation-level data.

4.5. Discussion

This study explores topic modeling in mental health conversations, focusing on the challenge posed by the lack of words with descriptive power. We hypothesized that utilizing the full text of the conversation might be more challenging due to this limitation. Additionally, mental health conversation corpora may exhibit topics occurring very frequently across many documents, as well as topics occurring very infrequently.

To address these challenges, we proposed that dividing documents into smaller sections could enhance the coherency of topic modeling.

Table 9
Top 5 topic words for BERTopic on the counseling corpus.

Topic	Word 1	Word 2	Word 3	Word 4	Word
1	parents	mother	my	and	you
2	to sleep	tired	sleep	bed	me
3	friends	who	friend	you	with
4	music	listening	hearing	to	the
5	at home	lonely	room	alone	house
6	mindfulness	search	a	practice	breathing exercise
7	medication	pills	me	drinking	have
8	wound	blood	bleeding	knife	care
9	writing	reading	book	me	what
10	url	link	site	website	urge
11	safe	keep	safety	yourself	you
12	watch	series	film	netflix	youtube
13	eating	dinner	eating disorder	me	cooking
14	sports	gaming	exercise	games	fifa
15	thoughts	my	suicide	suicide thoughts	that

Our analysis, based on the results, revealed that BERTopic outperformed other methods for most of the different corpora and topic sizes. Notably, the highest performance was consistently observed in the utterance and segmented datasets.

Furthermore, we see that methods using HDBScan performed well. The property of HDBScan to deal with noisy segments, potentially contributed to increased flexibility in modeling conversations. The results not only supported our initial hypotheses but also indicated that hierarchical clustering yielded the best performance. This outcome aligns with the observation that topics in mental health conversations have great variance in their occurrence.

The exception was the coherence scores for the MultiWoz dataset. We believe that this might be due to the length and general topic amount being lower for MultiWoz than they are for the counseling corpus. On the word embedding similarity metric, all models except LDA improved in performance when using utterances instead of the full conversation. The downside of this method is that relationships between utterances belonging to the same conversation are not considered. Therefore, a topic can only be as specific as can be expressed in a single message. However, we also found that ignoring noisy messages leads to better topic coherence and word similarity.

Between the utterances datasets and the segmented datasets, we observed that both obtained the highest topic coherence on multiple occasions. However, for the word embedding similarity, the utterance dataset outperformed the segmented dataset.

Our study suggests several opportunities for future research. An important limitation of this study is the limited generalizability, and including multiple helplines would help in this respect. While we did include two mental health datasets, the AnnoMI dataset is not as extensive as the suicide counseling dataset.

To address the sparsity constraints of short text, while considering noisy utterances. A possible solution is to use hierarchically constructed features from both utterance and conversation level data. Generative neural models such as ProdLDA and CombinedTM function through variational autoencoders [19,28]. Hierarchical autoencoders have also shown good performance on several tasks [40]. This could extend the variational autoencoder for topic modeling to take local context into account. A variant of this has been proposed by [41] using hierarchical LSTM models.

5. Conclusion

Topic modeling can be difficult on datasets like conversation data, where local context, emotion and subtext is important. However, by reducing granularity, a local context can be incorporated using sentence embedding. We found that clustering of sentence embedding with noise using BERTopic results in more coherent topic models for conversation data when compared to other topic modeling methods. For the domain of topic modeling in conversation, we saw that BERTopic over utterances outperforms other models on conversation in both counseling corpus as well as the easier to model. Furthermore, based on the inspection of the topic models, we found the topics to be the most interpretable. This method is particularly useful when the overall topics within a corpus are of interest. Finally, we believe this can especially be of use for mental health services to gain insight into their conversations.

CRediT authorship contribution statement

Salim Salmi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Rob van der Mei:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Saskia Mérelle:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Sandjai Bhulai:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

Research was funded by 113 zelfmoordpreventie (113 suicide prevention) which itself is funded by the Ministry of Health, Welfare & Sports of the Netherlands. All authors report no financial relationships with commercial interests.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (null) (2003) 993–1022, Number of pages: 30 Publisher: JMLR.org tex.issue_date: 3/1/2003.

[2] S. Salmi, S. Mérelle, R. Gilissen, R. van der Mei, S. Bhulai, Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID-19 pandemic: in-depth analysis using encoder representations from transformers, BMC Public Health 22 (1) (2022) 530–539, http://dx.doi.org/10.1186/s12889-022-12926-2.

[3] D.M. Blei, M.I. Jordan, T.L. Griffiths, J.B. Tenenbaum, Hierarchical topic models and the nested Chinese restaurant process, in: Proceedings of the 16th International Conference on Neural Information Processing Systems, NIPS '03, MIT Press, Cambridge, MA, USA, 2003, pp. 17–24, event-place: Whistler, British Columbia, Canada.

[4] A. Gruber, Y. Weiss, M. Rosen-Zvi, Hidden topic markov models, in: M. Meila, X. Shen (Eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, in: Proceedings of machine learning research, Vol. 2, PMLR, San Juan, Puerto Rico, 2007, pp. 163–170, tex.pdf: http://proceedings.mlr.press/v2/gruber07a/gruber07a.pdf.

[5] H. Wang, D. Zhang, C. Zhai, Structural topic model for latent topical structure analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, Association for Computational Linguistics, USA, 2011, pp. 1526–1535, Number of pages: 10 Place: Portland, Oregon.

[6] L. Du, W. Buntine, H. Jin, C. Chen, Sequential latent Dirichlet allocation, Knowl. Inf. Syst. 31 (3) (2012) 475–503, http://dx.doi.org/10.1007/s10115-011-0425-1.

[7] F. Tian, B. Gao, D. He, T.-Y. Liu, Sentence level recurrent topic model: Letting topics speak for themselves, 2016, arXiv:1604.02038, [cs].

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv:1706.03762, [cs].

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, arXiv:1810.04805, [cs].

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, arXiv:1907.11692, [cs].

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q.V. Le, XLNet: Generalized autoregressive pretraining for language understanding, 2020, arXiv:1906.08237, [cs].

[12] M.A. Hearst, TextTiling: segmenting text into multi-paragraph subtopic passages, Comput. Linguist. 23 (1) (1997) 33–64.

[13] H. Gao, R. Wang, T.-E. Lin, Y. Wu, M. Yang, F. Huang, Y. Li, Unsupervised dialogue topic segmentation with topic-aware utterance representation, 2023.

[14] L. Xing, G. Carenini, Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring, 2021.

[15] A. Solbiati, K. Heffernan, G. Damaskinos, S. Poddar, S. Modi, J. Cali, Unsupervised topic segmentation of meetings with BERT embeddings, 2021.

[16] J. Yin, J. Wang, A dirichlet multinomial mixture model-based approach for short text clustering, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, New York, USA, 2014, pp. 233–242, http://dx.doi.org/10.1145/2623330.2623715.

[17] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022, http://dx.doi.org/10.48550/arXiv.2203.05794, arXiv preprint arXiv:2203.05794.

[18] Y. Meng, Y. Zhang, J. Huang, Y. Zhang, J. Han, Topic discovery via latent space clustering of pretrained language model representations, in: The Web Conference, 2022.

[19] F. Bianchi, S. Terragni, D. Hovy, Pre-training is a hot topic: Contextualized document embeddings improve topic coherence, 2021, arXiv:2004.03974, [cs].

[20] K. Dinakar, J. Chen, H. Lieberman, R. Picard, R. Filbin, Mixed-initiative real-time topic modeling & visualization for crisis counseling, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, ACM, Atlanta, Georgia, USA, 2015, pp. 417–426, http://dx.doi.org/10.1145/2678025.2701395.

[21] T. Wang, Z. Huang, C. Gan, On mining latent topics from healthcare chat logs, J. Biomed. Inform. 61 (2016) 247–259, http://dx.doi.org/10.1016/j.jbi.2016.04.008.

[22] J. Weng, E.-P. Lim, J. Jiang, Q. He, TwitterRank: finding topic-sensitive influential twitterers, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10, ACM Press, New York, New York, USA, 2010, p. 261, http://dx.doi.org/10.1145/1718487.1718520.

[23] L. Hong, B.D. Davison, Empirical study of topic modeling in Twitter, in: Proceedings of the First Workshop on Social Media Analytics - SOMA '10, ACM Press, Washington D.C., District of Columbia, USA, 2010, pp. 80–88, http://dx.doi.org/10.1145/1964858.1964870.

[24] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the 22nd International Conference on World Wide Web - WWW '13, ACM Press, Rio de Janeiro, Brazil, 2013, pp. 1445–1456, http://dx.doi.org/10.1145/2488388.2488514.

[25] X. Li, A. Zhang, C. Li, L. Guo, W. Wang, J. Ouyang, Relational biterm topic model: short-text topic modeling using word embeddings, Comput. J. 62 (3) (2019) 359–372, http://dx.doi.org/10.1093/comjnl/bxy037.

[26] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi, J. Ouyang, Filtering out the noise in short text topic modeling, Inform. Sci. 456 (2018) 83–96, http://dx.doi.org/10.1016/j.ins.2018.04.071.

[27] J. Rashid, S.M.A. Shah, A. Irtaza, Fuzzy topic modeling approach for text mining over short text, Inf. Process. Manage. 56 (6) (2019) 102060, http://dx.doi.org/10.1016/j.ipm.2019.102060.

[28] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, 2017, arXiv:1703.01488, [stat].

[29] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS '13, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 3111–3119, Number of pages: 9 Place: Lake Tahoe, Nevada.

[30] A.B. Dieng, F.J.R. Ruiz, D.M. Blei, Topic modeling in embedding spaces, Trans. Assoc. Comput. Linguist. 8 (2020) 439–453, http://dx.doi.org/10.1162/tacl_a_00325.

[31] D. Angelov, Top2Vec: Distributed representations of topics, 2020, arXiv:2008.09470, [cs, stat].

[32] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, 2020, arXiv:1802.03426, [cs, stat].

[33] R.J.G.B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: J. Pei, V.S. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 2013, pp. 160–172.

[34] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, 2019, arXiv:1908.10084, [cs].

[35] S. Sia, A. Dalmia, S.J. Mielke, Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too!, 2020, CoRR arXiv:2004.14914.

[36] Y. Meng, J. Huang, G. Wang, Z. Wang, C. Zhang, Y. Zhang, J. Han, Discriminative topic mining via category-name guided text embedding, in: Proceedings of the Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2121–2132, http://dx.doi.org/10.1145/3366423.3380278.

[37] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato Recupero, D. Riboni, Anno-MI: A dataset of expert-annotated counselling dialogues, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2022, pp. 6177–6181, http://dx.doi.org/10.1109/ICASSP43922.2022.9746035.

[38] K. Wang, N. Reimers, I. Gurevych, TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 671–688, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.59.

[39] 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, ACM, Shanghai, China, 2015, pp. 399–408, http://dx.doi.org/10.1145/2684822.2685324.

[40] A. Vahdat, J. Kautz, NVAE: A deep hierarchical variational autoencoder, 2021, arXiv:2007.03898, [cs, stat].

[41] J. Li, M.-T. Luong, D. Jurafsky, A hierarchical neural autoencoder for paragraphs and documents, 2015, arXiv:1506.01057, [cs].