Topic Model Evaluation with Large Language Models

Saida Yusupova
Technische Universität Kaiserslautern, Department of Computer Science

Topic modeling is a technique for unsupervised analysis of large document corpora. It automatically learns topics, from unlabeled documents, represented as sets of words. In this work, we evaluate three topic models - one classical and two neural - on two datasets. We introduce Generative Pre-trained Transformer 3 (GPT-3) for the evaluation of topic models' interpretability and estimate its scores' relationship with automated metrics. To this end, we introduce the prompts with high accuracy or topic-terms relatedness and present the correlations between the automated and language model metrics.

1. Introduction and Related Work

Topic modeling is a set of unsupervised techniques used to analyze text in a collection of documents and identify the meaningful groups of words, topics [3]. Current evaluations of the topic model quality have fluctuated between automated and human assessments, which have shown promising results [2]. As automated coherence metrics, the topic model developers adopted the normalized pointwise mutual information (NPMI) score [10], which measures word relatedness and correlates with the interpretability of the topic [5]. While obtaining human metrics requires a reasonable number of crowdworkers in offline or online mode using survey platforms, so it is a time, energy-consuming, and costly task.
In this work, we introduce an approach of topic model evaluation using Generative Pre-trained Transformer 3 (GPT-3) [13], which is the third-generation autoregressive language model that uses deep learning to process and produce natural language text. Following the work by A. Hoyle [2], we use English articles from 20 Newsgroups and Wikipedia. For the 20 Newsgroups, we use text dataset from scikit-learn and for Wikipedia, we use Wikitext-103 [12] with the following settings: 28.5k for training, 4.2k for validation and testing. The model evaluation procedure remains the same.
We evaluate one classical model and two neural models:
Gibbs-LDA Latent Dirichlet Allocation (LDA) is a generative model optimized by Gibbs sampling [7], which represents each topic as a distribution over terms and represents each document as a mixture of topics that summarise the content [11] As a classical baseline, we are using Mallet [9] to produce topics.
Dirichlet-VAE Topic models based on Dirichlet Variational Autoencoder [4] are similar to classical LDA model, but it uses the Dirichlet distribution as a prior for the topic and word distributions.
ETM Embedded Topic Model [1] is a generative model, which uses embedding representation of terms and topics, namely it incorporates word similarity into the topic model. As a topic model, it generates the interpretable structure of the documents; as word embedding, it provides a low-dimensional representation of words, where words with similar meanings are close

	Intrusion			
Dataset →	20NG		Wiki	
Prompt ↓	μ	ρ_{spear}	μ	ρ_{spear}
1	**0.44**	<u>0.21</u>	**0.60**	0.48
2	0.41	0.15	0.59	0.39
3	0.31	0.06	0.43	<u>0.31</u>
4	0.40	0.15	0.51	0.48
5	0.38	<u>0.16</u>	0.49	**0.49**
6	0.35	0.09	0.48	0.40

	Rating			
Dataset →	20NG		Wiki	
Version ↓	μ	ρ_{spear}	μ	ρ_{spear}
1	1.87	**0.08**	2.23	**0.72**
2	1.81	0.06	2.15	0.71
3	1.89	0.03	2.25	0.71
4	1.80	0.06	2.21	0.68

Table 1: Database-wise accuracy results (μ for intrusion) and topic-terms relatedness rate (μ for rating) in a range 1-3, and spearman correlation coefficients. Values in **bold** are the highest values and <u>underlined</u> correlation coefficients have p-value<0.05.

in vector space.
As introduced by A. Hoyle [2], for each model, we generate 50 topics and calculate the topics' coherence scores (NPMI)[1].

2. Results

In our approach, we also test two tasks: intrusion and rating. As we replace human evaluation with GPT-3 assessment, Appendix A.1 describes the GPT-3 model setup. To understand the relationship between automated metrics (Figure 1 and GPT-3 scores, we estimate Spearman correlation [8] between the two sets of values for each task and dataset.
Intrusion As the first step, we test several versions of a prompt, then we test six prompts and report the accuracy (the percentage of correctly recognized intruder term by GPT-3) and correlation results, see Appendix A.2 for more details. The prompts are:

p_1 Show the least related term
p_2 Select which term is the least related to all other terms
p_3 What is the intruder term in the following terms?
p_4 Which word does not belong?
p_5 Which one of the following words does not belong?
p_6 Find the intruder term

Rating Following the same procedure we test several versions of a prompt Table 4, more details in Appendix A.3. The prompt under consideration for this task is:

p_3: Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': ['file', 'window', 'problem', 'run', 'system', 'program', 'font', 'work', 'win', 'change']
Answer: Very related
Rate how related the following terms are to each other as '3-very related', '2-somewhat

[1] gitlab.rhrk.uni-kl.de/yusupova/topics/-/tree/main/results/readable-format

related'or '1-not related': ['chip', 'clipper', 'phone', 'key', 'encryption', 'government', 'system', 'write', 'nsa', 'communication']
Answer: Somewhat related
Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': *top 10 words of a topic*
Answer:

We, database-wise merge metrics of three models, having two sets of values: the automated metrics and GPT-3 scores. Each set has 150 elements. As noted in Table 1, for intrusion task the Wikipedia corpus appears to have high accuracy, and, particularly, the p_1 in both corpora has the highest rate of correctly distinguished intruder terms. In Table 3, we show the model-wise accuracy results. While for the rating task, Wikipedia, again, has a higher rate of topic term relatedness. Correspondingly, the correlation coefficients for Wikipedia are higher compared to 20 Newsgroups, but statistically insignificant (p-values are >0.05).

3. Conclusions

As a result of this work, we can state that we have a comparably optimal prompt for the intrusion task and that the brackets-quotes pattern affects the GPT-3 response quality. The same story is seen for the rating task. For the database-wise statistics, we cannot claim some optimal prompt (version) for both tasks. However, for model-wise statistics, we can observe prompts with high correlation coefficients, which are statistically significant.

References

[1] Francisco J. R. Ruiz Adji B. Dieng & David M. Blei (2020): *Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics.*

[2] Denis Peskov Andrew Hian-Cheong Alexander Hoyle, Pravan Goel (2021): *Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. 35th Conference on Neural Information Processing Systems.*

[3] Yuening Hu Jordan Boyd-Graber & David Mimno (2017): *Applications of topic models.*

[4] Sophie Burkhardt & Stefan Kramer (2019): *Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model. In Journal of Machine Learning Research.*

[5] Karl Grieser David Newman, Jey Han Lau & Timothy Baldwin (2010): *Automatic evaluation of topic coherence. In Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.*

[6] Teven Le Scao et.al. (2022): *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.*

[7] Thomas L Griffiths & Mark Steyvers (2004): *Finding scientific topics. In Proceedings of the National Academy of Sciences. National Academy of Sciences.*

[8] Tomasz M. Kossowski Jan Hauke (2011): *Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data.*

[9] Andrew Kachites McCallum (2002): *MALLET: A machine learning for language toolkit.*

[10] Alexander Hinneburg Michael Röder, Andreas Both (2015): *Exploring the Space of Topic Coherence Measures.*

[11] Akash Srivastava & Charles Sutton (2017): *Autoencoding variational inference for topic models. In Proceedings of the International Conference on Learning Representations.*

[12] James Bradbury Stephen Merity, Caiming Xiong & Richard Socher (2017): *Pointer sentinel mixture models. In Proceedings of the International Conference on Learning Representations.*

[13] Nick Ryder Melanie Subbiah Tom B. Brown, Benjamin Mann (2020): *Language Models are Few-Shot Learners.*

[14] Radim Řehůřek (2009): *Topic Modeling for humans.*

A. Appendix

A.1. Details on GPT-3 model set up

The set up is delineated in our implementation:[2]

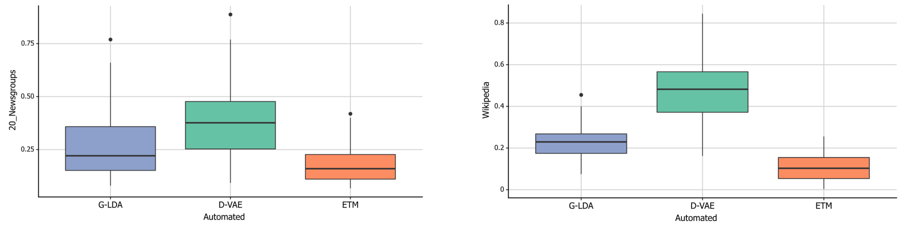
- We provide the API-key to GPT-3 in file 'gpt3-api-key.txt'.
- We use 'text-davinci-003' engine.
- We set the temperature to 0 for deterministic responses.
- The maximum number of response tokens we set is 10 for the intrusion task, and 60 for the rating task.
- We set 'frequency_penalty' to 0 so that we don't penalize new tokens based on their existing frequency in the text.
- Also, we set 'presence_penalty' to 0 and don't penalize new tokens based on their appearance in the text.

A.2. Details on prompt variations for intrusion task

The prompt versions, versions with respect to the p_1, are:

v_1: Show the least related term: ['construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains']
v_2: Show the least related term: 'construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains'
v_3: Show the least related term: [construction, locomotives, cantata, coaster, railway, trains]
v_4: Show the least related term: construction, locomotives, cantata, coaster, railway, trains
v_5: Show the least related term
Terms: ['construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains']
Answer:
v_6: Show the least related term
Terms: 'construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains'
Answer:
v_7: Show the least related term
Terms: [construction, locomotives, cantata, coaster, railway, trains]

[2] gitlab.rhrk.uni-kl.de/yusupova/topics



(a) 20 Newsgroups NPMI-values (b) Wikipedia NPMI-values

Figure 1: Automated evaluations (NPMI) suggest a clear winner between the three models. NPMI declares D-VAE as a winner for topics derived from both datasets, with G-LDA in second place.

Version → Prompt ↓	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
1	0.68	0.70	0.66	0.68	0.82	**0.86**	0.82	0.84
2	0.70	0.66	0.64	0.64	0.74	**0.76**	0.72	0.74
3	0.40	0.46	0.42	0.42	0.52	**0.60**	0.56	0.52

Table 2: Accuracy results for three prompts of D-VAE model for topics derived from Wikipedia in intrusion task.

Answer:
v_8: Show the least related term
Terms: construction, locomotives, cantata, coaster, railway, trains
Answer:

Versions from v_1 to v_4 represent different setups of square brackets and single quotes, while v_5 to v_8 represent a more structured way. We randomly choose two prompts and one prompt, p_2, similar to the question the crowdworkers were asked to determine intruder term[3]. Additionally, as the D-VAE fared better on the intrusion task in the original experiment [2], we perform our initial evaluations on this model for Wikipedia dataset. We test all eight versions and the Table 2 shows the accuracy results. We see that v_6 has the highest accuracy in all three versions, so in further tests of the prompts, we use v_6. To note, in the initial stage of our experiment, the following prompt was also under consideration, as the transformer was given reasonable responses only if the prompt consisted of the responses that we were expecting:

Show the least related term
Terms: image, object, pixel, face, scene, privacy
Answer: privacy
Show the least related term
Terms: graph, edge, message, cell, vertex, propagation

[3]The topic-intruder files along with the GPT-3 responses can be found at gitlab.rhrk.uni-kl.de/yusupova/topics/-/tree/main/transformer-tests

	20 Newsgroups											
Model →	G-LDA				D-VAE				ETM			
Prompt ↓	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
1	**0.52**	0.25	<u>0.36</u>	0.27	**0.38**	0.24	**0.27**	0.27	**0.42**	0.24	0.25	0.23
2	**0.52**	0.25	<u>0.35</u>	0.25	0.34	0.22	-0.13	-0.09	0.38	0.24	<u>**0.45**</u>	0.37
3	0.40	0.24	<u>**0.44**</u>	0.33	0.14	0.12	-0.03	-0.002	0.38	0.24	0.15	0.11
4	0.50	0.25	<u>0.39</u>	0.27	0.28	0.20	-0.01	-0.02	**0.42**	0.24	<u>0.42</u>	0.40
5	0.46	0.25	<u>0.40</u>	0.30	0.30	0.21	0.03	0.04	0.38	0.24	<u>0.32</u>	0.29
6	0.42	0.24	<u>0.34</u>	0.35	0.22	0.17	0.12	0.13	0.40	0.24	0.16	0.11

	Wikipedia											
Model →	G-LDA				D-VAE				ETM			
Prompt ↓	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
1	0.68	0.22	0.23	0.26	**0.86**	0.12	0.17	0.16	0.26	0.19	-0.05	-0.03
2	**0.72**	0.20	0.07	0.14	0.76	0.18	0.19	0.19	**0.28**	0.20	**0.07**	0.09
3	0.48	0.25	0.04	0.05	0.60	0.24	0.20	0.20	0.20	0.16	-0.09	-0.09
4	0.64	0.23	**0.24**	0.27	0.76	0.18	0.19	0.19	0.14	0.12	-0.22	-0.21
5	0.60	0.24	**0.24**	0.28	0.74	0.19	**0.24**	0.24	0.12	0.11	-0.20	-0.19
6	0.56	0.25	0.11	0.07	0.70	0.21	0.23	0.23	0.18	0.15	-0.16	-0.15

Table 3: GPT-3 metrics (intrusion task) for 20 Newsgroups and Wikipedia. <u>Underlined</u> values are spearman correlation coefficients between GPT-3 scores and automated metrics, which have p-value < 0.05. **Bold** values have the highest accuracy and correlation coefficient. p_s and p_p are spearman and pearson correlation coefficients, respectively.

> Answer: cell
> Show the least related term
> Terms: construction, locomotives, cantata, coaster, railway, trains
> Answer:

With the new model, 'text-davinci-003', there was no need for further use of this prompt as the responses received using this engine were reasonable and in a one-word format for intrusion tasks.

In Table 3, we show the results of six prompts using the version with high accuracy. For the intrusion task, we can state that p_1 and p_2, which ask for the least related term, have high accuracies, and the prompt formulation matters.

A.3. Details on prompt variations for rating task

Same as in the intrusion task, we test the effect of brackets and single quotes on GPT-3 responses. The prompts are:

> p_1: Rate how related the following terms are to each other as 'very related', 'somewhat related'or 'not related': *top 10 words of a topic*

	20 Newsgroups											
Model →	G-LDA				D-VAE				ETM			
p_3 ↓	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
v_1	2.14	0.80	0.36	0.18	1.44	0.65	**0.27**	0.25	2.02	0.74	**0.14**	0.22
v_2	2.04	0.84	0.38	0.23	1.38	0.60	**0.27**	0.26	2.00	0.76	0.11	0.17
v_3	2.12	0.83	0.35	0.18	1.44	0.69	0.19	0.14	2.12	0.87	0.09	0.15
v_4	1.96	0.80	**0.40**	0.25	1.40	0.64	0.22	0.16	2.04	0.92	0.10	0.18

	Wikipedia											
Model →	G-LDA				D-VAE				ETM			
p_3 ↓	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
v_1	2.88	0.19	0.25	0.27	2.80	0.32	**0.50**	0.50	nan	nan	nan	nan
v_2	2.78	0.29	**0.35**	0.36	2.68	0.50	0.59	0.58	nan	nan	nan	nan
v_3	2.96	0.08	0.24	0.30	2.80	0.32	0.49	0.49	nan	nan	nan	nan
v_4	2.92	0.15	0.32	0.34	2.70	0.45	0.46	0.43	nan	nan	nan	nan

Table 4: GPT-3 metrics (rating task) for 20 Newsgroups and Wikipedia datasets. **Bold** values have the highest correlation coefficient. Underlined values are spearman correlation coefficients between gpt-3 scores and automated metrics, which have p-value < 0.05. p_s and p_p are spearman and pearson correlation coefficients, respectively

p_2: Rate how related the following terms are to each other in a range from 1 to 3: *top 10 words of a topic*

For the above two prompts, the GPT-3 does not give reasonable responses. For topic terms

['rower', 'hammersmith_bridge', 'rowed', 'mile_post', 'rowing', 'cambridge', 'boat_race', 'chiswick_steps', 'oxford', 'university_of_oxford'],

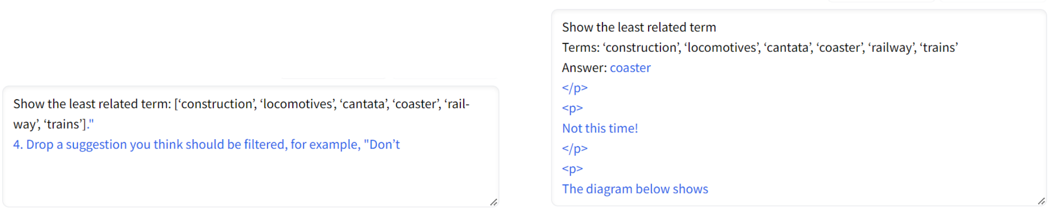
the following are the examples of such responses (only two terms were considered by GPT-3 or terms were considered pairwise):

Rower and Rowing: Very Related OR *rower: 3hammersmith_*

rower and rowing: very relatedhammersmith bridge and chiswick steps: very relatedmile post and cambridge: not relatedboat race and university of oxford: somewhat related

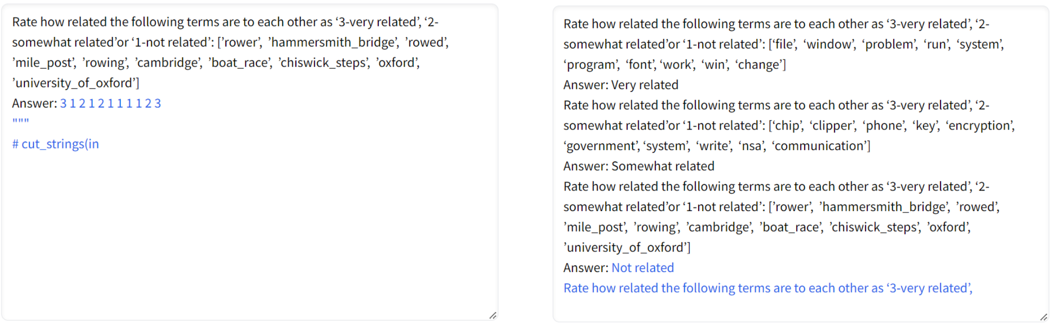
For the following prompt p_3, where we show the GPT-3 the format of the response we are expecting, we also manipulate brackets-quotes, and Table 4shows the results:

p_3: Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': ['file', 'window', 'problem', 'run', 'system', 'program', 'font', 'work', 'win', 'change']
Answer: Very related
Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': ['chip', 'clipper', 'phone', 'key', 'encryption', 'government', 'system', 'write', 'nsa', 'communication']



(a) (b)

Figure 2: Intrusion task



(a) (b)

Figure 3: Rating task

Answer: Somewhat related
Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': *top 10 words of a topic*
Answer:

Note, for the ETM model for Wikipedia corpus, we get 'nan' values. This is due to the fact that GPT-3 gave 'Not Related' response for all 50 topics, name the elements of the second set of values are all the same (in our case 1s) and *scipy.stats* raises *ConstantInputWarning* (The correlation coefficient is not defined in this case).

A.4. Multilingual Language Model BLOOM

Our main Language Model was GPT-3, however, we also tested the existing prompts on the multilingual language model BLOOM [6]. Figure 2 and 3 are examples of some responses for intrusion and rating tasks. As you can see, in the case of structured prompt Figure 2(b) the BLOOM model understood the task and returned the possible intruder term, while for unstructured case Figure 2(a), the task failed.
For the rating task, we can observe a similar pattern, the BLOOM model did a good job if we first showed it the format of response we were expecting.