Contextualized Topic Coherence Metrics

Anonymous ACL submission

Abstract

This article proposes LLM-based topic coherence metrics inspired by standard human topic evaluations, in a family of metrics called Contextualized Topic Coherence (CTC). These metrics allow human-centered evaluation of coherence while maintaining the efficiency of automated methods. We evaluate CTC relative to five metrics and discovered that it outperforms automated topic coherence methods on seven topic models. Notably, CTC aligns with human evaluation and demonstrates excellent performance with short documents, and is not susceptible to meaningless but high-scoring topics.

1 Introduction

Topic models are a family of text-mining algorithms that identify themes in a large corpus of text data (Blei, 2012). These models (Churchill and Singh, 2022) are widely used for exploratory data analysis with the aim of organizing, understanding, and summarizing large amounts of text data (Abdelrazek et al., 2022). Numerous techniques, algorithms, and tools have been employed to develop a variety of topic models for different tasks and purposes (Srivastava and Sutton, 2017) including much recent work on neural topic models (Grootendorst, 2022). However, due to their nature as unsupervised models, comparing topic outputs, hyperparameter settings, and overall model quality has traditionally been difficult (Hoyle et al., 2022).

Topic Coherence (TC) metrics measure the interpretability of topics generated by topic models. These metrics are categorized into two classes: automated TC metrics and human-annotated TC metrics (Hoyle et al., 2021). Automated TC metrics estimate the interpretability of topic models with respect to various factors such as co-occurrence or semantic similarity of topic words. On the other hand, human-annotated TC metrics are protocols for designing surveys that rate or score the interpretability of topic models. Human judgment is often used to validate topic coherence metrics to provide an accurate assessment of the semantic coherence and meaningfulness of a given set of topics (Newman et al., 2009; Aletras and Stevenson, 2013; Mimno et al., 2011). While human-annotated TC metrics incorporate subjective human judgments and provide a more accurate and nuanced understanding of how well topic models are performing (e.g. in terms of their ability to capture the underlying themes in a text corpus), they are expensive, time-consuming, and require multiple human-subjects to avoid personal biases. On the other hand, automated metrics are more cost-effective than human-annotated methods, as they do not require the hiring and training of human annotators, which results in their ability to evaluate large amounts of data and iterate through many model comparisons.

Automated metrics are intended to align more closely with human judgment, providing a better measure of the interpretability of topic words. The risk of such approximations, however, is that they themselves become the target of optimization rather than the underlying property they were intended to measure. Several recent works suggest that this has occurred especially in the context of neural topic models. Doogan and Buntine (2021) argue that interpretability is ambiguous and conclude that current automated topic coherence metrics are unreliable for evaluating topic models in short-text data collections and may be incompatible with newer neural topic models. In a similar study, Hoyle et al. (2021) show that topics generated by neural models are often qualitatively distinct from traditional topic models while they receive higher scores from current automated topic coherence metrics. Hoyle et al. (2021) conclude that the validity of the results produced by fully automated evaluations, as currently practiced, is questionable, and they only help when human evaluations cannot be performed. Hoyle et al. (2022) in another recent work shows that neural topic models fail to improve on the traditional topic models such as Gibbs LDA (Griffiths and Steyvers, 2004; McCallum, 2002) and consider neural topic broken as they do not function well for their intended use.

To address these problems, we introduce Contextualized Topic Coherence (CTC) metrics which are a context-aware family of topic coherence metrics based on the pre-trained Large Language Models (LLM). Taking Advantage of LLMs elevates the understanding of language at a very sophisticated level incorporating its linguistic nuances, contexts, and relationships. CTC is much less susceptible to being fooled by meaningless topics that often receive high scores with traditional topic coherence metrics.

2 Automated Topic Coherence Metrics

Topic coherence (TC) metrics measure the consistency of words in a given topic to evaluate the interpretability and meaningfulness of a topic by computing the level of semantic similarity among words that are included in the topic. A high TC value indicates that the words in the topic are semantically similar and are likely to co-occur in the same circumstances.

The authors of (Newman et al., 2009, 2010b) claim that a method based on the Point-wise Mutual Information (PMI) gives the largest correlations with human ratings. They define UCI, which measures the strength of the association between pairs of words based on their co-occurrence in a sliding window of length-l words. (Mimno et al., 2011) proposes UMass, an asymmetric confirmation measure that estimates the degree of coherence between words within a given topic by calculating the $\log$ ratio frequency of their co-occurrences in the corpus of documents. UMass counts the number of times a pair of words co-occur in a given corpus and compares this number to the expected number of co-occurrences where words are randomly distributed across the whole corpus. (Aletras and Stevenson, 2013) proposes context vectors for each topic word w to generate the frequency of word co-occurrences within windows of ± 1 words surrounding all instances of w. They showed that NPMI (Bouma, 2009) has a larger correlation with human topic ratings compared to UCI and UMass. Additionally, NPMI takes into account the fact that some words are more common than others and adjusts the frequency of individual words accordingly(Lau et al., 2014). While NPMI is generally more sensitive to rare words and can handle small datasets, UMass focuses on fast computation of coherence scores over large corpora. (Stevens et al., 2012) showed that a smaller value of ϵ tends to yield better results than the default value of $\epsilon = 1$ used in the original paper since it emphasizes more the word combinations that are completely unattested. (Röder et al., 2015) proposes a unifying framework of coherence measures that can be freely combined to form a configuration space of coherence definitions, allowing their main elementary components to be combined in the context of coherence quantification. For example, they propose the C_V metric, which uses a variation of NPMI to compute topic coherence over a sliding window of size N and adds a weight γ to assign more strength to more related words. According to (Campagnolo et al., 2022), the C_V metric is more sensitive to noisy information and dirty data than C_{UMass} and C_{UCI}. (Nikolenko, 2016) and (Schnabel et al., 2015) propose the metric TC_{DWR} based on the Distributed Word Representations (DWR) (Mikolov et al., 2013b,a) which are better correlated to human judgment. Similarly, (Ramrakhiyani et al., 2017) presents a coherence measure based on grouping topic words into buckets and using Singular Value Decomposition (SVD) and integer linear programming-based optimization to create coherent word buckets from the generated embedding vectors. (Korenčić et al., 2018) proposes several topic coherence metrics based on topic documents rather than topic words. The approach essentially extracts topic documents, vectorizes them using several methods such as word embedding aggregation, and computes a coherence score based on the document vectors. (Lund et al., 2019) proposes an automated evaluation metric for local-level topic models by introducing a task designed to elicit human judgment and reflect token-level topic quality.

3 Contextualised Topic Coherence

In this article, we introduce Contextualized Topic Coherence (CTC) to refer to a new family of topic coherence metrics that benefit from the recent development of Large Language Models (LLM). This paper presents two approaches using LLMs for defining CTC metrics. The first approach uses LLMs to compute contextualized estimates of the pointwise mutual information (CPMI) between topic words. In the second approach, we use ChatGPT (OpenAI, 2022) to evaluate topic coherence similar to human-annotated metrics.

3.1 Automated CTC

CPMI. Recent work by (Hoover et al., 2021) uses conditional PMI estimates to analyze the relationship between linguistic dependencies and statistical dependencies between words. They propose Contextualized PMI (CPMI) as a new method for estimating the conditional PMI between words *in context* using a pre-trained language model. As illustrated in Figure 1, the CPMI between two words w_i and w_j in a sentence s is defined as

$$\mathrm{CPMI}(w_i, w_j \mid s) = \log \frac{p(w_i \mid s_{-w_i})}{p(w_i \mid s_{-w_{ij}})} \quad (1)$$

where p is an estimate for the probability of words in context based on a pre-trained masked language model (MLM), such as BERT. Here, s_{-w_i} represents the sentence with word w_i masked, and

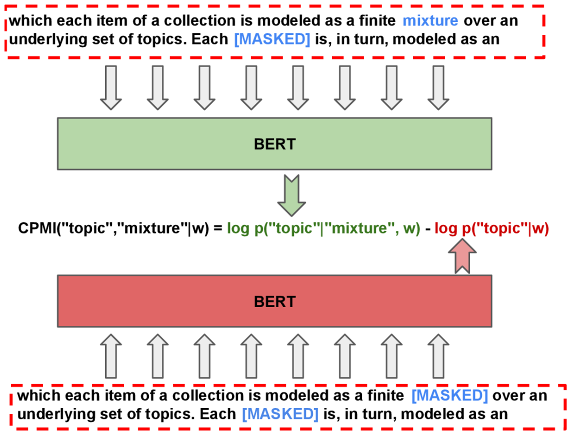


Figure 1: Calculating CPMI for two topic words in a segment of a document.

$s_{-w_{ij}}$ is the sentence with both words w_i and w_j masked.

We adopt CPMI to introduce a new automated Contextualized Topic Coherence (CTC) metric. Automated CTC estimates the statistical dependence within a topic in a corpus by computing the CPMI value for each pair of topic words along a sliding window applied to the dataset. For this, the corpus is divided into a set of window segments of length w that have k words intersecting with adjacent window segments to compute the average CPMI between each pair of words within each topic over all window segments, giving the following expression for CTC:

$$\frac{1}{n * \binom{m}{2}} \sum_{i=1}^{n} \sum_{r=2}^{m} \sum_{s=1}^{r-1} \mathrm{CPMI}(w_i^r, w_i^s \mid c^u) \quad (2)$$

where $c^u \subset$ corpus D is a window segment with length of w that has k words overlapping with its adjacent window segments, n is the number of topics and m is the number of topic words.

3.2 Semi-automated CTC

Intrusion. (Chang et al., 2009) studied the *topic words intrusion* task to assess topic coherence by identifying a coherent latent category for each topic and discovering the words that do not belong to that category. These *intruder words* are detected by human subjects to assess the quality of topic models and to measure a coherence score that assigns a low probability for intruder words to belong to a topic. We apply this idea to chatbots with a prompt (see Appendix B.1), which provides the topic words to ChatGPT (OpenAI, 2022) and asks for a category and intruder words.

Rating. While human topic ratings are expensive to produce, they serve as the gold standard for coherence evaluation (Röder et al., 2015). For example, (Syed and Spruit, 2017) uses human ratings to explore the coherence of topics generated by LDA topics across full texts and abstracts. (Newman et al., 2010a) provides human annotators with a rubric and guidelines for judging whether a topic is useful or useless. The annotators evaluate a randomly selected subset of topics for their usefulness in retrieving documents on a given topic and score each topic on a 3-point scale, where 3=highly coherent and 1=useless (less coherent). Following (Newman et al., 2010a), (Aletras and Stevenson, 2013) presented topics without intruder words to Amazon Mechanical Turk to score them on a 3-point ordinal scale. We adapt this method to chatbots with a prompt (see Appendix B.2), which provides the topic words to ChatGPT and asks to rate the usefulness of the topic words for retrieving documents on a given topic. The $\text{CTC}_{\text{Rating}}$ for a topic model is then obtained by the average sum of all ratings over all topics.

4 Experiments

In this section, we expect to observe that the baseline metrics (UCI, UMass, NPMI, C_V, DWR) rank topic models differently from CTC. We also expect CTC rankings favor interpretable topics and handle short text datasets more effectively than the baseline metrics (Doogan and Buntine, 2021; Hoyle et al., 2021). This implies that baseline metrics often yield high scores for incoherent topics, while conversely assigning low scores to well-interpretable topics. In contrast, CTC has a better model of language and can better evaluate topical similarity *as it would appear to a human reader*. Therefore, we expect to see that baseline metrics and CTC would differ at extremes of highest or lowest coherency.

4.1 Experimental setup

Datasets. The experiments incorporate two datasets including the 20Newsgroups dataset (Lang, 1995) and a collection of 17K tweets by Elon Musk published between 2017 and 2022 by (Raza, 2023).

Topic Models. The experiments involve six different topic models including Gibbs LDA (Griffiths and Steyvers, 2004), Embedded Topic Model (ETM) (Dieng et al., 2020), Adversarial-neural Topic Models (ATM) (Wang et al., 2019), Top2Vec (Angelov, 2020), and Contextualized Topic Model (CTM) (Bianchi et al., 2021), and BERTopic (Grootendorst, 2022).

Topic Coherence Metrics. The topics generated by the topic models are evaluated using the proposed Contextualized Topic Coherence (CTC) metrics, which are then compared to the well-established automated topic coherence metrics C_V, UCI, UMass, NPMI, and DWR. For CTC_{CPMI}, we segmented the 20Newsgroup and Elon Musk's Tweets datasets into chunks of 15 and 20 words, respectively, without intersections. We then extracted the CPMI for all word pairs in each segment using the pre-trained language models *bert-base-uncased* and *Tesla K80 15 GB GPU* from Google Colab (Bisong and Bisong, 2019). This pre-computing step took about 7 hours but allowed us to compute CTC_{CPMI} for any topic model in the order of a few seconds. For evaluating $\text{CTC}_{\text{Intrusion}}$ and $\text{CTC}_{\text{Rating}}$, we made a request for each topic to *ChatGPT* with *GPT 3.5 Turbo*, which cost less than a dollar for all the experiments.

4.2 Results

Tables 1 and 2 represent the results of the evaluation of the topic models obtained from the 20Newsgroup and Elon Musk's Tweets datasets, respectively, using CTC and the baseline metrics. To allow us to compare the models in terms of topic coherence metrics, the highest value for each metric is shown in bold. the highest values for each metric within each topic model are noted in *italic* font. This helps us determine the optimal

Table 1: Scores of Topic Coherence Metrics on 20Newsgroup dataset.

Topic Models		Baseline Metrics					CTC Metrics		
	#T	UCI	UMass	NPMI	C_V	DWR	Rating	Intrusion	CPMI
Gibbs LDA (2003)	20	*0.260*	*-2.338*	*0.043*	*0.512*	*0.211*	*1.3*	0.225	*9.92*
	50	-0.121	-2.771	0.023	0.479	0.191	1.16	0.220	5.99
	100	-0.690	-3.030	0.002	0.450	0.149	1.14	*0.267*	3.25
ETM (2020)	20	*0.478*	-2.08	*0.067*	*0.563*	0.292	0.7	***0.452***	19.16
	50	0.380	***-1.903***	0.054	0.532	*0.330*	1.22	0.348	20.35
	100	0.351	-1.962	0.049	0.522	0.312	*1.23*	0.41	***22.58***
ATM (2019)	20	-1.431	-3.014	-0.059	0.338	*0.151*	0.92	0.305	0.03
	50	-0.940	-2.902	-0.046	0.342	0.077	*1.15*	0.275	0.18
	100	*-0.735*	*-2.741*	*-0.032*	*0.362*	0.053	1.12	*0.340*	*1.72*
CTM (2021)	20	-1.707	-4.082	0.005	*0.601*	*0.268*	1.25	0.385	*5.93*
	50	*-0.724*	*-3.008*	*0.046*	0.590	0.236	*1.56*	0.380	7.02
	100	-0.926	-3.118	0.027	0.561	0.210	1.31	*0.392*	6.16
Top2Vec (2020)	85	***0.910***	*-2.449*	***0.192***	***0.785***	***0.473***	***1.670***	*0.399*	*3.77*
BERTopic (2022)	145	*-1.023*	*-5.033*	*0.098*	*0.681*	*0.309*	*1.517*	*0.359*	*2.91*

Table 2: Scores of Topic Coherence Metrics on Elon Musk's Tweets dataset

Topic Models		Baseline Metrics					CTC Metrics		
	#T	UCI	UMass	NPMI	C_V	DWR	Rating	Intrusion	CPMI
Gibbs LDA (2003)	10	*-0.441*	*-3.790*	*0.016*	*0.498*	*0.838*	*1.6*	0.29	*2.19*
	20	-1.834	-5.415	-0.049	0.395	0.798	1.5	0.225	1.04
	30	-3.068	-6.390	-0.099	0.336	0.783	1.466	*0.33*	0.86
ETM (2020)	10	***0.205***	-3.209	***0.051***	*0.560*	0.952	1.1	*0.24*	***5.41***
	20	0.155	***-3.079***	0.028	0.538	0.974	*1.433*	0.233	4.48
	30	0.025	-3.215	0.022	0.515	***0.978***	1.05	0.195	4.30
ATM (2019)	10	-9.021	-12.859	-0.324	*0.364*	0.730	*1.2*	0.211	-0.004
	20	-7.967	-11.770	-0.283	0.343	0.694	1.1	0.177	*0*
	30	*-7.278*	*-11.301*	*-0.258*	0.350	*0.753*	0.933	*0.214*	-0.03
CTM (2021)	10	*-2.614*	*-7.049*	*-0.030*	***0.580***	*0.888*	***2.0***	***0.439***	1
	20	-3.720	-8.336	-0.070	0.534	0.880	1.45	0.185	*3.04*
	30	-3.589	-8.063	-0.064	0.573	0.873	1.766	0.276	2.56
Top2Vec (2020)	164	*-6.272*	*-10.536*	*-0.152*	*0.401*	*0.847*	*1.481*	*0.274*	*2.08*
BERTopic (2022)	217	*-4.131*	*-11.883*	*-0.020*	*0.432*	*0.541*	*1.539*	*0.276*	*1.52*

number of topics for all models except Top2Vec and BERTopic, which don't need this parameter as an input.

General observations. Before analyzing the results in Tables 1 and 2 in detail, we examine the relationship between the CTC metrics and the baseline metrics by performing Pearson's correlation coefficient analysis (Sedgwick, 2012) on the results from Tables 1 and 2 similar to (Doogan and Buntine, 2021). As shown in Figure 2a (see Appendix C), for 20Newsgroup, the baseline metrics UCI and UMass are highly correlated with CPMI but not with $\text{CTC}_{\text{Rating}}$ and $\text{CTC}_{\text{Intrusion}}$, which are more correlated with the baseline measures NPMI and C_V and DWR (which are also highly correlated). On the other hand, for the short text EM Tweets dataset, Figure 2b (see Appendix C) shows that CPMI has a high correlation with all baseline methods, while $\text{CTC}_{\text{Intrusion}}$ and $\text{CTC}_{\text{Rating}}$ are completely independent of CPMI and the baseline measures.

Concerning our expectation that baseline metrics rank topic models differently from CTC metrics, Table 1 reports that the baseline metrics (except for UMass) point to Top2Vec while CTC metrics (except for $\text{CTC}_{\text{Rating}}$) point to ETM for achieving the highest scores. Similarly, Table 2 reports that the baseline metrics (except for C_V) point to ETM while CTC metrics (except for CTC_{CPMI}) point to CTM for achieving the highest scores. These contradictions between CTC and baseline metrics are aligned with our expectations and we will explore them with a meta-analysis of topics generated by these topic models and the scores they have received from CTC and baseline

Table 3: Top-2 and bottom-2 topics of ETM$^{(100)}$ and Top2Vec on 20Newsgroup

Topic Model	Ranked By	Topics	C_V	CPMI
ETM$^{(100)}$ (2020)	Highest C_V	god, christian, people, believe, jesus	0.740	0.017
		drive, card, scsi, disk, mb,	0.739	0.037
	Lowest C_V	book, number, problem, read, call	0.369	0.018
		line, use, power, bit, high	0.458	0.018
	Highest CPMI	year, time, day, one, ago, week	0.559	0.709
		game, year, team, player, play	0.706	0.242
	Lowest CPMI	new, number, also, well, call, order, used	0.340	-0.007
		people, right, drug, state, world, country	0.529	-0.002
Top2Vec (2020)	Highest C_V	dsl, geb, cadre, shameful, jxp	0.995	0.009
		tor, nyi, det, chi, bos	0.989	0.012
	Lowest C_V	hacker, computer, privacy, uci, ethic	0.255	-0.0001
		battery, acid, charged, storage, floor	0.344	0.006
	Highest CPMI	mailing, list, mail, address, send	0.792	0.154
		icon, window, manager, file, application	0.770	0.076
	Lowest CPMI	lc, lciii, fpu, slot, nubus, iisi	0.853	-0.004
		ci, ic, incoming, gif, edu	0.644	-0.002

metrics.

Meta-analysis. To check the performance of different coherence metrics, we will compare the intepretability of their high and low-scoring topics. Note that CTC metrics observe contextual patterns between topic words, and therefore, we expect them to provide more consistent coherence scores according to the interpretability of the generated topics for all topic models.

To verify the consistency of some representative scores in Table 1, we examine the topics for 20 Newsgroup generated by Top2Vec, which have high and low scores for baseline metrics, and ETM, which have high and low scores for CTC metrics. Table 3 compares the top-2 and bottom-2 topics ranked by C_V and CTC_{CPMI}. The motivation behind choosing these metrics is from our correlation analysis in Figure 2a(see Appendix C), which in CTC_{CPMI} and C_V has the least correlation among CTC and baseline metrics. First, we notice that the top-2 topics returned by C_V for Top2Vec are not readily interpretable but are statistically meaningful: *dsl, geb, cadre, shameful, jxp* are fragments of an email signature that occurs 82 times, while *tor, nyi, det, chi, bos* are abbreviations for hockey teams. This is not surprising, since Top2Vec produces what we call "trash topics", which is a common problem for clustering-based topic models that cannot handle so-called "trash clusters" (Giannotti et al., 2002). While CTC_{CPMI} returns a more coherent ranking for Top2Vec (the top 2 topics appear coherent, while the bottom topics are incoherent for human evaluation). This supports our assumption that traditional topic coherence metrics such as C_V fail to evaluate neural topic models and, in this case, even give the highest scores to trash topics. This happens because they only consider the syntactic co-occurrence of words in a window of text and cannot observe the underlying relationship between topic words. CTC_{CPMI}, on the other hand, can detect these trash topics and score them more accurately because it is contextual and accompanied by LLMs that have rich information about linguistic dependencies between topic words. CTC_{CPMI} then also could be a good measure to filter out these topics. The second observation in Table 3 is that all eight topics returned for ETM are coherent. This is because ETM, which is a semantically-enabled probabilistic topic model, produces decent topics that are overall highly ranked by CTC_{CPMI}, as shown in Figure 4b (see Appendix C).

In the same way, we check the consistency of some representative scores in Table 2 by checking the interpretability of topics for Elon Musk's tweets generated by ETM, which has high baseline scores, and by CTM, which has high CTC scores. As shown in Table 4, we compare the top 2 and bottom 2 topics ranked by NPMI and CTC_{Rating}. As shown in Figure 2b (see Appendix C), these metrics are among those with the lowest correlation between CTC and baseline metrics. A notable finding for CTM topics is that top-

Table 4: Top-2 and bottom-2 topics of ETM$^{(30)}$ and CTM$^{(30)}$ on Elon Musk's Tweets

Topic Model	Ranked By	Topics	NPMI	Rating	Intrusion
CTM$^{(30)}$ (2021)	Highest NPMI	erdayastronaut, engine, booster, starship, amp	0.122	3	0.1
		year, week, next, month, wholemarsblog	0.057	2	0.1
	Lowest NPMI	transport, backup, ensure, installed, transaction	-0.480	2	0.1
		achieving, transition, late, transport, precision	-0.459	1	0.1
	Highest Rating	tesla, rt, model, car, supercharger	-0.152	3	0.5
		spacex, dragon, launch, falcon, nasa	-0.283	3	0.4
	Lowest Rating	ppathole, soon, justpaulinelol, yes, sure	-0.330	1	0.5
		achieving, transition, late, transport, precision	-0.459	1	0.1
ETM$^{(30)}$ (2020)	Highest NPMI	amp, time, people, like, would, many	0.001	2	0.7
		engine, booster, starship, heavy, raptor	-0.023	2	0.1
	Lowest NPMI	amp, rt, tesla, im, yes	-0.283	1	0.1
		amp, tesla, year, twitter, work	-0.228	1	0.1
	Highest Rating	amp, twitter, like, tesla, dont	-0.186	2	0.8
		amp, time, people, like, would	0.001	2	0.7
	Lowest Rating	amp, tesla, year, twitter, work	-0.228	1	0.1
		amp, tesla, one, like, time	-0.204	1	0.1

Table 5: Topic Coherence Scores of Gibbs LDA, DVAE, ETM on NYT News

Topic Models		Baseline Metrics				Human Evaluation		CTC Metrics		
	#T	UCI	UMass	C_V	NPMI	Intrusion	Rating	Intrusion	Rating	CPMI
Gibbs LDA	50	1.42	-7.6	0.69	0.15	0.71	**2.66**	**2.12**	0.62	**4.18**
DVAE	50	**2.43**	-15	**0.84**	**0.25**	**0.74**	2.48	2.05	**0.67**	0.61
ETM	50	1.01	**-7.4**	0.60	0.11	0.64	2.38	2.06	0.64	3.72

Table 6: Top-5 topics among the topics generated by Gibbs LDA, DVAE and ETM on NYT News

Top-5 Sorted by	Model	Topic	Scores		
			C_V	Human	CTC
C_V	DVAE	inc, 9mo, earns, otc, qtr, rev	0.98	1.2	0.9
	DVAE	inc, 6mo, earns, otc, rev, qtr	0.98	1.2	1.3
	DVAE	inc, otc, qtr, earns, rev, 6mo	0.97	1.3	0.8
	DVAE	arafat, hamas, gaza, palestinians, west_bank	0.97	2.1	1.5
	DVAE	condolences, mourns, mourn, board_of_directors, heartfelt, deepest	0.97	0.6	1.3
Human Score	Gibbs LDA	film, theater, movie, play, director, films	0.73	3	2.7
	DVAE	skirts, dresses, chanel, couture, fashion	0.91	3	1.3
	DVAE	tenants, tenant, zoning, rents, landlords, developers	0.86	3	1.2
	DVAE	paintings, sculptures, galleries, picasso, sculpture, drawings,	0.91	2.9	2.1
	DVAE	television, network, news, cable, nbc, year, cbs	0.68	2.8	1.9
CTC	Gibbs LDA	film, theater, movie, play, director, films	0.73	3	2.7
	ETM	court, judge, law, case, federal, lawyer, trial	0.80	2.8	2.6
	Gibbs LDA	court, law, judge, case, state, federal, legal,	0.72	2.6	2.2
	Gibbs LDA	music, dance, opera, program, work, orchestra, performance	0.73	1.1	2.1
	ETM	film, movie, story, films, directed, movies, star, character	0.79	2.7	2.1

ics ranked highest by the $\text{CTC}_{\text{Rating}}$ metric tend to be more interpretable compared to those ranked highest by NPMI, and similarly, topics ranked lowest by the $\text{CTC}_{\text{Rating}}$ metric tend to be less interpretable compared to those ranked lowest by NPMI. The above observation also holds true for ETM, as the $\text{CTC}_{\text{Rating}}$ metric is not affected by the scarcity of short text records. This is because $\text{CTC}_{\text{Rating}}$ is complemented by a chatbot that mitigates the impact of limited data availability. It is also interesting to note that the topics generated by CTM are overall more interpretable and coherent than those generated by ETM. This demonstrates the validity of $\text{CTC}_{\text{Rating}}$ and $\text{CTC}_{\text{Intrusion}}$ over baseline metrics, as we observed in Table 2. It also reveals the superiority of CTM over ETM, as shown in Figure 4d (see Appendix C), in short text datasets as a result of a contextualized element in its architecture.

5 Human Evaluation

The goal of automated topic coherence metrics is to accurately approximate human reactions to topics without the need for expensive, time-consuming studies that require multiple annotators to avoid bias. In this section we compare the proposed metric with published human evaluation metrics based on data provided by Hoyle et al. (2021). This data includes three topic models (Gibbs LDA (McCallum, 2002), DVAE (Srivastava and Sutton, 2017), and ETM (Dieng et al., 2020)) models with 50 topics generated on the (New York Times) dataset, along with human evaluation (intrusion and ranking). We evaluate the generated topics with CTC_{CPMI}, $\text{CTC}_{\text{intrusion}}$ and $\text{CTC}_{\text{ranking}}$, which are comparable to human intrusion and human ranking.

As shown in Table 5, human evaluators tend to see little quantifiable difference between Gibbs LDA and DVAE, while traditional metrics show pronounced differences. In contrast, we find that CTC metrics more closely match human preferences (or lack thereof). It is possible that this result is simply due to a miscalibration of relative scores. To show that humans and CTC rank topics similarly, we also report Spearman's Rank Correlation (Myers and Sirois, 2004) to assess the strength and direction of the monotonic relationship between the ranking of topics in each metric. As shown in Figure 3, the CTC metrics have an overall higher correlation with human ratings than the baselines.

We also examine the consistency of the scores obtained by different coherence metrics and compare the coherence of high and low scoring topics from different topic models and CTC metrics. As shown in Table 6, Table 7, C_V is not able to score topics correctly. For example, the topic *inc, 9mo, earns, otc, qtr, rev* gets the highest score, even though it has little clear interpretability and has been rated relatively low by human evaluators. On the other hand, CTC metrics score topics relative to their contextual relationship and are very close to human scores. For example, the topic *film, theater, movie, play, director, movies* receives the highest score by both CTC and human scoring.

6 Conclusion

This paper introduces a new family of topic coherence metrics called Contextualized Topic Coherence Metrics (CTC) that benefits from the recent development of Large Language Models (LLM). CTC includes two approaches that are motivated to offer flexibility and accuracy in evaluating neural topic models under different circumstances. Our results show automated CTC outperforms the baseline metrics on large-scale datasets while semi-automated CTC outperforms the baseline metrics on short-text datasets. After a comprehensive comparison between recent neural topic models and dominant classical topic models, the results indicate that some neural topic models, which optimize traditional topic coherence metrics, often receive high scores for topics that are overly sensitive to idiosyncrasies such as repeated text, and lack face validity. We show with our experiments that CTC is not susceptible to being deceived by these meaningless topics by leveraging the abilty of LLMs to better model human expectations for language and evaluate topics within and outside their contextual framework.

Limitations

CTC metrics come with several limitations, such as latency, accuracy, and the potential for biased results. For instance, CPMI can be a time-consuming process, as it involves running all sentences through LLMs and calculating word co-occurrences for every pair of words across all topics. Additionally, the results for Rating and Intrusion may vary with each query to LLMs. Therefore, it is necessary to configure the LLM's temperature and iterate through multiple queries to obtain normalized values. Furthermore, it's important to be aware that LLMs can exhibit bias, and their utilization in this application could potentially perpetuate such biases.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2022. Topic modeling algorithms and applications: A survey. *Information Systems*, page 102131.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*, pages 13–22.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Ekaba Bisong and Ekaba Bisong. 2019. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

João Marcos Campagnolo, Denio Duarte, and Guillherme Dal Bianco. 2022. Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management*, 13(4).

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Fosca Giannotti, Cristian Gozzi, and Giuseppe Manco. 2002. Clustering transactional data. In *Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19–23, 2002 Proceedings 6*, pages 175–187. Springer.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jacob Louis Hoover, Wenyu Du, Alessandro Sordoni, and Timothy J. O'Donnell. 2021. Linguistic dependencies and statistical dependence. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in Neural Information Processing Systems*, 34:2018–2033.

Alexander Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? *arXiv preprint arXiv:2210.16162*.

Damir Korenčić, Strahil Ristov, and Jan Šnajder. 2018. Document-based topic coherence measures for news media text. *Expert systems with Applications*, 114:357–373.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality. *arXiv preprint arXiv:1905.13126*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for languagetoolkit. *http://mallet. cs. umass. edu*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12.

David Newman, Sarvnaz Karimi, and Lawrence Cavedon. 2009. External evaluation of topic models. In *Proceedings of the 14th Australasian Document Computing Symposium*, pages 1–8. University of Sydney.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010a. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010b. Evaluating topic models for digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, page 215–224, New York, NY, USA. Association for Computing Machinery.

Sergey I. Nikolenko. 2016. Topic quality metrics based on distributed word representations. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1029–1032, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2022. Chatgpt: Engaging and dynamic conversations. `https://openai.com/blog/chatgpt`.

Nitin Ramrakhiyani, Sachin Pawar, Swapnil Hingmire, and Girish Palshikar. 2017. Measuring topic coherence through optimal word buckets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 437–442.

Yasir Raza. 2023. Elon musk tweets dataset (17k): Dataset of elon musk tweets till now (17k). `https://www.kaggle.com/datasets/yasirabdaali/elon-musk-tweets-dataset-17k`.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.

Philip Sedgwick. 2012. Pearson's correlation coefficient. *Bmj*, 345.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 952–961.

Shaheen Syed and Marco Spruit. 2017. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, pages 165–174. IEEE.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098.

A Automated Coherence Metrics

Topic Models were initially evaluated with held-out perplexity as an automated metric (Blei et al., 2003). Perplexity quantifies how well a statistical model predicts a sample of unseen data and is computed by taking the inverse probability of the test set, normalized by the number of words in the dataset. According to (Chang et al., 2009), perplexity has been found to be inconsistent with human interpretability. As a result, the field shifted towards adopting automated topics coherence metrics that rely on word co-occurrence-based methods like Point-wise Mutual Information (PMI) (Cover, 1999).

A.1 Definition

As defined as follows, Topic coherence over PMI (TC_{UCI}) is defined as the average of the $\log_2$ ratio of co-occurrence frequency of word w_i^r and w_i^s within a given topic i.

$$\text{TC}_{\text{UCI}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\binom{m}{2}}\sum_{r=2}^{m}\sum_{s=1}^{r-1}\text{PMI}(w_i^r, w_i^s) \quad (3)$$

with

$$\text{PMI}(w^i, w^j) = \log_2 \frac{P(w^i, w^j) + \epsilon}{P(w^i)P(w^j)} \quad (4)$$

where n is the number of topics with m topic words and PMI represents the pointwise mutual information between each pair of words (w_i^r and w_i^s) in the topic i. PMI is computed by taking the logarithm of the ratio of the joint probability of two words $P(w_i^r, w_i^s)$ appearing together to the individual probabilities of the words $P(w_i^r)$, $P(w_i^s)$ occurring separately. Note that $\epsilon = 1$ is added to avoid the logarithm of zero.

On the other hand, UMass (Mimno et al., 2011) computes the co-document frequency of word w_i^r and w_i^s divided by the document frequency of word w_i^s.

$$\text{UMass}(w_i^r, w_i^s) = \log \frac{D(w_i^r, w_i^s) + \epsilon}{D(w_i^s)} \quad (5)$$

where n and m are the numbers of topics and topic words respectively. The smoothing parameter ϵ was initially introduced to be equal to one and avoid the logarithm of zero.

Similarly, (Aletras and Stevenson, 2013) proposes context vectors for each topic word w to generate the frequency of word co-occurrences within windows of ± 1 words surrounding all instances of w.

$$\text{NPMI}(w_i^r, w_i^s) = \frac{\log_2 \frac{P(w_i^r, w_i^s) + \epsilon}{P(w_i^r)P(w_i^s)}}{-\log_2(P(w_i^r, w_i^s) + \epsilon)} \quad (6)$$

(Röder et al., 2015) proposes C_V, which is a variation of NPMI.

$$\text{C}_{\text{V}}(w_i^r, w_i^s) = \text{NPMI}^{\gamma}(w_i^r, w_i^s) \quad (7)$$

One way to estimate TC_{DWR} is to compute the average pairwise cosine similarity between word vectors in a topic as follows.

$$\text{DWR}(w_i^r, w_i^s) = \frac{w_i^r \cdot w_i^s}{\|w_i^r\| \cdot \|w_i^s\|} \quad (8)$$

B LLM Prompts

In this section, we present LLM prompts used in our experiments. The descriptions of the prompts for the ratings and intrusion task are as follows.

B.1 Intrusion

System prompt: *I have a topic that is described by the following keywords: [topic-words]. Provide a one-word topic based on this list of words and identify all intruder words in the list with respect to the topic you provided. Results be in the following format: topic: <one-word>, intruders: <words in a list>*

The number of intrusion words ($|I_i|$) returned by chatbot for each topic i, is used to define $\text{CTC}_{\text{Intrusion}}$ as follows:

$$\text{CTC}_{\text{Intrusion}} = \sum_{i=1}^{n} \frac{1 - \frac{|I_i|}{m}}{n} \tag{9}$$

where n is the number of topics and m is the number of topic words.

B.2 Rating

System prompt: *I have a topic that is described by the following keywords: [topic-words]. Evaluate the interpretability of the topic words on a 3-point scale where 3 = "meaningful and highly coherent" and 0 = "useless" as topic words are usable to search and retrieve documents about a single particular subject. Results be in the following format: score: <score>*

B.3 Normalized CPMI

To improve comparability, we also propose a normalized version of CPMI that extend its generalizability and allows to mitigate potential biases that may arise due to specific dataset characteristics or idiosyncrasies. Additionally, it facilitates threshold determination and provides a consistent scale that allows researchers to set thresholds based on desired coherence levels, ensuring the metric is effectively utilized in practical applications.

B.3.1 Definition

Given a set of n topics $\text{TM} \mapsto \{t_1, t_2, \ldots, t_n\}$ with m words $t_i \mapsto \{w_1^i, w_2^i, \ldots, w_m^i\}$ as an output of topic model TM on the corpus of e documents $D = \{d_1, d_2, \ldots, d_e\}$, the CTC based on Normalized CPMI (NCPMI) called $\text{CTC}_{\text{NCPMI}}$ is defined as follows.

$$\frac{1}{e * n * m} \sum_{d=1}^{e} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{NCPMI}(w_j^i, t^i \mid c^d) \tag{10}$$

while $\text{NCPMI}(w_j^i, t^i \mid c^d)$ is:

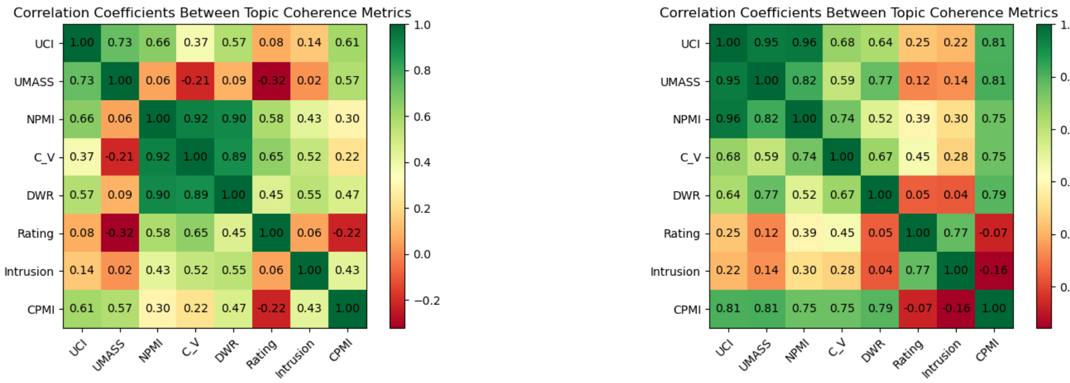
$$\frac{log \frac{P(w_j^i | c^d_{-w_j^i})}{P(w_j^i | c^d_{-t^i})}}{-log(P(w_j^i \mid c^d_{-w_j^i}) \times P(t^i \mid c^d_{-t^i}))} \tag{11}$$

where P is an estimate for the probability of words given context based on language model LM. The $c^d_{-w_i}$ is the document d with word w_i masked, and $c^d_{-t_j}$ is the document d with words of topic t^i masked.

C Correlation Study

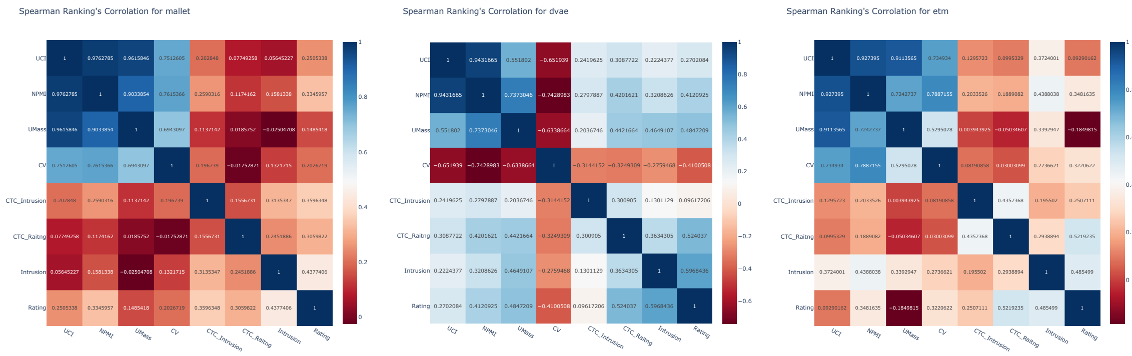
Pearson correlation is a statistical measure used to assess the degree of linear association between sets of data. As shown Figure 2, we applied this method to the results of topic coherence metrics on the topic models to evaluate how closely related or similar the quality of topics generated by these models is. A high positive Pearson correlation coefficient indicates that the topic models produce similar results in terms of topic coherence, suggesting that they are consistent and reliable. Conversely, a low or negative correlation suggests inconsistency or divergence in the quality of topics generated by the different models.

On the other hand, Spearman's rank correlation coefficient is a statistical measure used to assess the strength and direction of the monotonic relationship between sets of data. As show in Figure 3, we applied this method to evaluation topic coherence metrics for human evaluation to determine if there is a consistent ranking of these models in terms of their performance across different metrics. A high positive Spearman's rank correlation coefficient suggests that the rankings of the three models across the evaluation metrics are similar, indicating consistency in their performance. Conversely, a low or negative correlation suggests variability in the rankings, indicating that different metrics may lead to different model preferences.



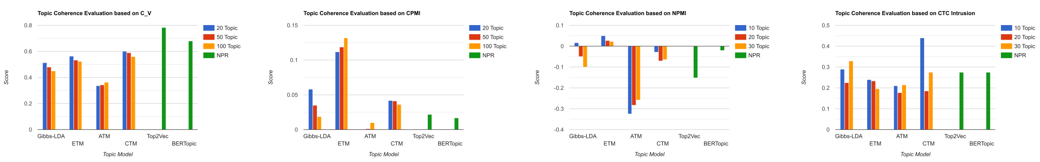
(a) 20Newsgroup (b) Elon Musk Tweets

Figure 2: Pearson's correlation coefficient on CTC and baseline



(a) Gibbs LDA (b) DVAE (c) ETM

Figure 3: Spearman's rank correlation coefficients between evaluation metrics for three topic models



(a) 20Newsgroup | C_V (b) 20Newsgroup | CPMI (c) Twitter | NPMI (d) Twitter | Intrusion

Figure 4: Comparison Between Topic Models based on Topic Coherence Evaluation

Table 7: Bottom-5 topics among the topics generated by Gibbs LDA, DVAE and ETM on NYT News

Botton-5 Sorted by	Model	Topic	Scores		
			C_V	Human	CTC
C_V	DVAE	spade, derby, belmont, colt, spades, dummy, preakness	0.23	1.5	0.4
	ETM	like, making, important, based, strong, including, recent	0.35	2	0.3
	ETM	time, half, center, open, away, place, high	0.37	1.6	0.2
	ETM	today, group, including, called, led, known, began, built, early,	0.37	2	0.3
	Gibbs LDA	people, editor, time, world, good, years, public, long,	0.37	0.1	1.1
Human Score	Gibbs LDA	people, editor, time, world, good, years, public,	0.37	0.1	1.1
	ETM	week, article, page, march, tuesday, june, july	0.57	0.4	1.3
	Gibbs LDA	street, tickets, sunday, avenue, information, free	0.75	0.4	0.3
	ETM	new_york, yesterday, director, manhattan, brooklyn, received	0.49	0.4	1
	Gibbs LDA	bedroom, room, bath, taxes, year, market, listed, kitchen, broker	0.72	0.4	1.3
CTC	Gibbs LDA	city, mayor, state, new_york, new_york_city, officials	0.61	2.5	0.1
	ETM	power, number, control, according, increase, large	0.44	0.9	0.2
	Gibbs LDA	officials, board, report, union, members, agency, yesterday	0.51	0.8	0.3
	ETM	time, half, center, open, away, place, high, day, run	0.37	1.2	0.3
	ETM	net, share, inc, earns, company, reports, loss, lead	0.73	1.8	0.3

D More Results

Figure 4 compares overall rating of the mentioned topic models in Section 4 over 20Newsgroup and the twitter dataset based on C_V, CPMI, NPMI, and Intrusion. The details of this figure are explained in Section 4.2.

Table 7 presents bottom-5 topics among the topics generated by Gibbs LDA, DVAE, and ETM on the NYT News dataset for better comparison between scores generated by CTC metrics against baseline and human evaluation.

E Python Package

CTC is implemented as a service for researchers and engineers who aim to evaluate and fine-tune their topic models[1]. The source code of this python package is provided in *./ctc* and a notebook named *example.ipynb* is prepared to explain how to use this python package as follows.

E.0.1 Automated CTC

```
from ctc.main import Auto_CTC
#initiating the metric
eval=Auto_CTC(segments_length
    =15, min_segment_length=5,
    segment_step=10,device="mps")

# segmenting the documents
docs=documents
eval.segmenting_documents(docs)

# creating cpmi tree including
    all co-occurence values
    between all pairs of words
eval.create_cpmi_tree()
#eval.load_cpmi_tree()

# topics=[["game","play"],["man
    ","devil"]] for instance
eval.ctc_cpmi(topics)
```

E.0.2 Semi-automated CTC

```
from ctc.main import
    Semi_auto_CTC

openai_key="YOUR OPENAI KEY"

y=Semi_auto_CTC(openai_key,
    topics)

y.ctc_intrusion()

y.ctc_rating()
```

[1] https://anonymous.4open.science/r/CTC-39DB