Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling

Satya Kapoor
Amazon
Vancouver, British Columbia, Canada

Alex Gil
Amazon
Arlington, Virginia, USA

Sreyoshi Bhaduri
Amazon
Arlington, Virginia, USA

Anshul Mittal
Amazon
Arlington, Virginia, USA

Rutu Mulkar
Amazon
Seattle, Washington, USA

Abstract

Topic modeling is a widely used technique for uncovering thematic structures from large text corpora. However, most topic modeling approaches e.g. Latent Dirichlet Allocation (LDA) struggle to capture nuanced semantics and contextual understanding required to accurately model complex narratives. Recent advancements in this area include methods like BERTopic, which have demonstrated significantly improved topic coherence and thus established a new standard for benchmarking. In this paper, we present a novel approach, the Qualitative Insights Tool (QualIT) that integrates large language models (LLMs) with existing clustering-based topic modeling approaches. Our method leverages the deep contextual understanding and powerful language generation capabilities of LLMs to enrich the topic modeling process using clustering. We evaluate our approach on a large corpus of news articles and demonstrate substantial improvements in topic coherence and topic diversity compared to baseline topic modeling techniques. On the 20 ground-truth topics, our method shows 70% topic coherence (vs 65% & 57% benchmarks) and 95.5% topic diversity (vs 85% & 72% benchmarks). Our findings suggest that the integration of LLMs can unlock new opportunities for topic modeling of dynamic and complex text data, as is common in talent management research contexts.

Keywords

Topic Modeling, Large Language Models, AI in Talent Management, Qualitative Research

1 Introduction

Topic modeling is a widely-used natural language processing (NLP) technique for extracting latent thematic structures from unstructured text data, such as social media posts, news articles, or customer feedback [9][18]. By employing probabilistic models to systematically identify and categorize patterns within the text, topic modeling enables researchers to uncover themes and perspectives that may not be immediately apparent to human analysts [10]. The flexibility of topic modeling allows it to be applied across a range of theoretical and epistemological frameworks, making it a valuable tool in both quantitative and qualitative research [25].

Traditional topic modeling techniques (e.g. Latent Dirichlet Allocation) suffer from several limitations (e.g. bag-of-words limitation, specifying number of clusters) when compared to existing deep learning-based methods. They also fail to capture the contextual nuances and ambiguities inherent in natural language, as they rely heavily on predefined rules and patterns [15][24]. This can make it challenging to handle the complexities and variations present in real-world text data, and may require domain-specific knowledge or fine-tuning to achieve acceptable performance [21]. Recent advancements in LLMs have positioned methods like BERTopic as a strong alternative to traditional topic modeling methods, and have largely overcome the limitations by leveraging deep neural networks to learn rich, contextual representations from large amounts of text data [2][15]. These powerful models can capture subtle semantic and pragmatic features of language, and demonstrate strong generalization capabilities through transfer learning [24][12]. They arent without limitations though. BERTopic, for example, is a clustering based approach which suffers from limitations such as word representation overload or generation of only one topic per text.

In this paper, we present - QualIT : the Qualitative Insights Tool (pronounced "kwaa-luh-tee") to extend the capabilities of existing topic models. Our novel approach integrates pre-trained LLMs with clustering techniques to systematically address the limitations of both methods and generate more nuanced and interpretable topic representations from free-text data. Combining LLMs and clustering techniques can provide a powerful and scalable approach to automatically identify themes and patterns in large volumes of unstructured text data. LLMs offer some semantic understanding and the ability to capture contextual nuances [23], while clustering algorithms enable unsupervised grouping of similar responses into topics. Overall, the synergy between LLMs' natural language understanding and the clustering approach's ability to organize and summarize data can revolutionize topic modeling, providing a robust and insightful approach to analyzing text responses at scale.

2 Topic Modeling in Talent Management Research

Talent management researchers leverage both psychological and data/applied science to provide actionable insights to managers, leaders, and HR professionals at an organization [5][4]. Voice of Customer (VOC) is an example of a key mechanism used by Talent Management researchers to collect feedback from customers and allows researchers in talent management to close the loop. VOC via surveys offer both qualitative and quantitative response options, both sources of important information for research teams to understand how customers interact with products AND what customers expect from products [6].

However, insights from qualitative text largely remains a missed opportunity due to being time, labor, and resource intensive to analyze manually [3]. Typically, qualitative research projects take three months, end to end for a team of researchers, including sampling participants, collecting data, and analyzing documents. An

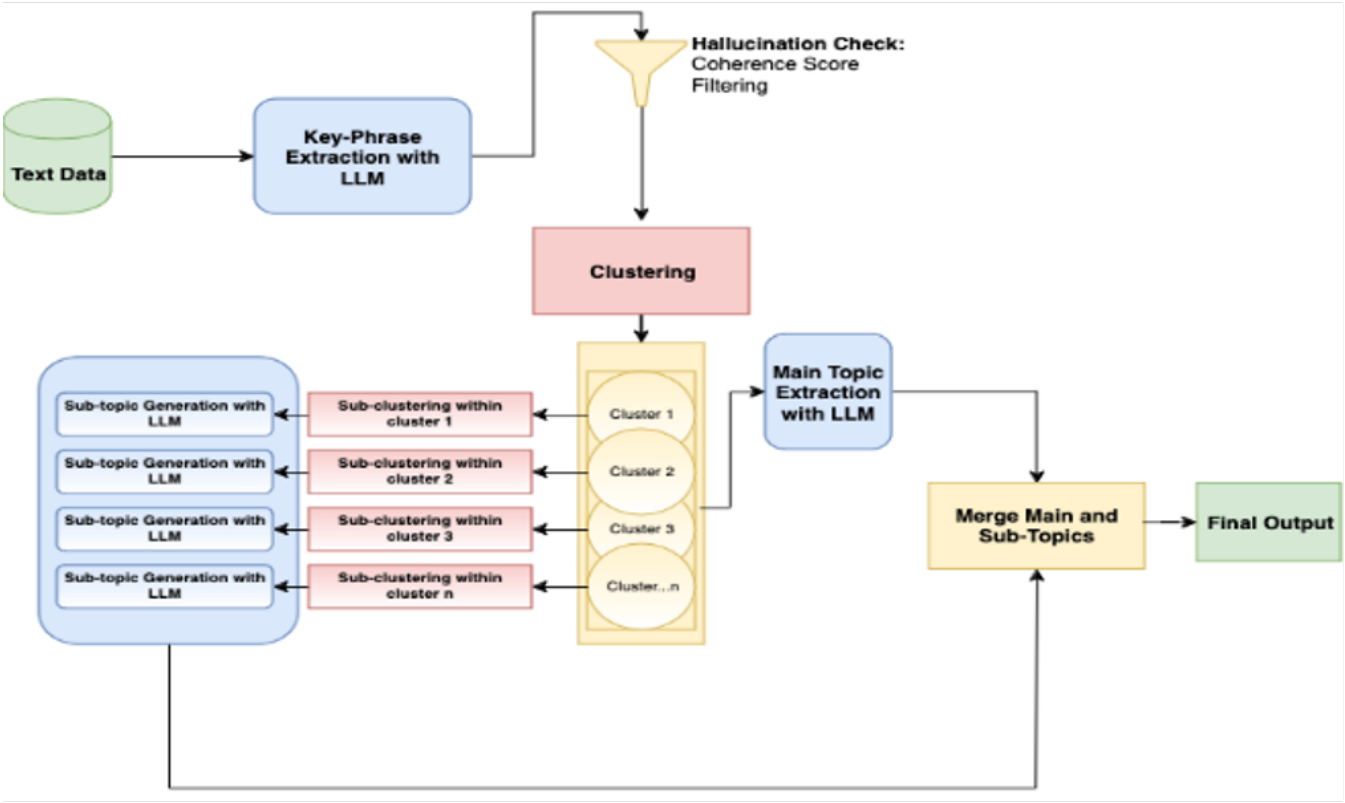

Figure 1: QualIT : Qualitative Insights Tool

automated topic modeling tool, such as QualIT, compliments the analysis step which is estimated to take one month of a researcher's time per project. In comparison, a pre-trained LLM model may take 1:30 minutes to process 2500 documents. Of course, human in the loop deep-dives and quality checks would still be necessitated, however, such automated quantative approaches can provide research direction and support for qualitative researchers.

Further, the lack of familiarization/expertise in qualitative research methods limits easy analysis or sharing of insights. For program or product owners, topic modeling tools can democratize access to insights by automating and augmenting analysis of qualitative documents and thematically distilling the information. For researchers, these tools do not aim to replace manual deep-dives, but rather serves as a novice qualitative research assistant [5] [7] and reduces the amount of time it takes to manually analyze open text documents and surfaces effective trends in the data.

3 Related Work

One of the most widely-used topic modeling approaches is Latent Dirichlet Allocation (LDA) [10]. LDA is a generative probabilistic model that operates on the principle that each document in a corpus is composed of a mixture of latent topics, with each topic being represented by a unique probability distribution over the vocabulary [9]. The model learns these topic-word distributions by leveraging the co-occurrence patterns of words within the documents, allowing it to uncover the underlying thematic structure of the corpus.

A key step in the LDA modeling process is determining the appropriate number of topics to be extracted from the data. This number is not something LDA can automatically infer, but rather must be provided by the researcher as an input parameter [17]. The choice of the number of topics can have a significant impact on the interpretability and performance of the LDA model, as too few topics may fail to capture the nuances of the data, while too many topics can lead to overfitting and poor generalization [1]. Unfortunately, there is no universally optimal approach for selecting the number of topics, and the appropriate choice often depends on the specific research objectives and characteristics of the text corpus [14]. As a result, practitioners must carefully explore and validate different topic count configurations to arrive at model parameters that best suits their needs.

The limitations of LDA have motivated researchers to explore the use of more advanced natural language processing methods, particularly those based on LLMs. LLMs, such as BERT [15], Generative Pre-trained Transformers (GPT) [24], and T5 [13], have demonstrated remarkable performance on a wide range of NLP tasks by leveraging deep neural networks to learn rich, contextual representations from massive amounts of text data. Unlike traditional topic modeling approaches that rely on hand-crafted features or simple statistical patterns, LLMs can capture complex semantic and syntactic relationships within language, allowing them to better handle the nuances and ambiguities inherent in real-world text [22]. Moreover, these models can be fine-tuned on domain-specific

data or downstream tasks, enabling them to adapt to the particular characteristics of a given text corpus [27].

Recent studies have explored the integration of LLMs into topic modeling frameworks, demonstrating significant improvements in performance and interpretability compared to traditional methods. For example, some authors [8] have proposed a topic modeling approach that uses BERT embeddings to initialize the topic-word distributions, leading to more coherent and semantically meaningful topics. Other researchers have investigated the use of LLMs for various aspects of the topic modeling pipeline, such as document representation [28], and topic labeling [19]. These advancements have shown that integration of LLMs can make a more powerful topic modeling tool for uncovering insights from large text corpora.

4 Methodology

We propose a new method LLM Enhanced Topic Modeling, which consists of multiple steps to generate the main topics, which are then used towards determining sub-topics from documents. The three key steps in this approach are Key Phrase Extraction, Hallucination Check, and Clustering.

4.1 Key-Phrase Extraction

In this step, we prompt the LLM to extract key-phrases representing the individual document. The LLM analyzes the content, discerning patterns and topics that frequently occur within the text. Guided by the prompt, the LLM pinpoints the key-phrases related to the defined role. The LLM prompt can extract multiple key phrases from the document, depending on its content. These key-phrases are essential for understanding the more nuanced aspects of the main subject matter. Adding this step provides an advantage over alternative methods. Alternative methods (e.g.BERTopic) assume that each document only contain a single topic, when in reality a document may contain more than a single topic.

4.2 Hallucination Check

To ensure the reliability of extracted key-phrases, a coherence score is calculated for each phrase. This score assesses how well the key-phrase aligns with the actual text, serving as a metric for consistency and relevance of the subsequent topics. Key-phrases with the lowest coherence scores may be flagged for 'hallucination', indicating potential errors in topic extraction, and are removed accordingly. For our approach, phrases with coherence scores below 10% were excluded. The coherence score was calculated using cosine similarity (Equation 1), with the texts first converted to embeddings using the Amazon Titan model.

$$C_i = \frac{1}{n} \sum_{j=1}^{n} \frac{\left(V_{\text{input},ij} \cdot V_{\text{keyphrases},ij}\right)}{\left|V_{\text{input},ij}\right| \cdot \left|V_{\text{keyphrases},ij}\right|} \tag{1}$$

Where:

- C_i is the coherence score for the i-th document.
- n is the number of dimensions in the embedding space.
- $V_{\text{input},ij}$ is the j-th dimension of the normalized embedding vector for the input text of the i-th document.
- $V_{\text{keyphrases},ij}$ is the j-th dimension of the normalized embedding vector for the theme text of the i-th document.
- $\left(V_{\text{input},ij} \cdot V_{\text{keyphrases},ij}\right)$ denotes the dot product of the two vectors.
- $\|v\|$ denotes the Euclidean norm (or length) of vector v.

4.3 Clustering

We use a partitional clustering algorithm (K-Means) to group the key-phrases identified in the previous step. The aim of this step is to group documents into multiple clusters, each representing a collection of documents with similar semantic content. In our approach, we implement two layers of clustering to find main topics and sub-topics. Unlike existing methods that directly use documents to create clusters, this approach leverages key phrases as they represent a more condensed form of the documents.

4.3.1 Clustering for Main Topics Create the primary cluster and pass the key-phrases of each cluster to another LLM prompt to distill a main theme from the grouped documents. This synthesis involves extracting a comprehensive topic that encapsulates the essence of each cluster, providing a clear and consolidated view of the overarching subjects within the documents.

4.3.2 Clustering for Sub Topics Implement a secondary level of clustering within each primary cluster to uncover more specific sub-topics. This will involve reapplying the clustering algorithm to each main cluster, separating the documents into sub-clusters based on finer nuances and more detailed thematic differences. For each sub-cluster, we prompt the LLM again to analyze condensed documents within sub-clusters. The model extracts sub-topics, for each sub-cluster.

There are several advantages of using this novel approach for LLM extracted key-phrases as features for clustering, as opposed to direct grouping of documents. Primarily, it reduces noise and the influence of irrelevant data, allowing the algorithm to operate on the distilled thematic essence of the documents. Another key benefit is that this approach can avoid the need for manual data exploration and cleaning steps, such as stopwords or punctuation removal, as the LLM is able to extract only the most meaningful content from the documents. This results in a set of clusters that are thematically concentrated, facilitating a more nuanced analysis and understanding of the document collection. Our approach ensures that documents with shared underlying topics are clustered together, even if they are not textually identical. One major drawback of K-Means is that it requires the number of clusters as a parameter to perform clustering. To address this drawback, we utilized the length of data and calculate a Silhouette score (Equation 2) to automatically determine the number of clusters. Silhouette score is a metric used to calculate the ideal number of clusters [26].

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{\max a(i),b(i)}, & \text{if } |C_I| > 1 \\ 0, & \text{if } |C_I| = 1 \end{cases} \tag{2}$$

Where:

- s(i): Silhouette score
- a: The mean distance between a sample and all other points in the same cluster.
- b: The mean distance between a sample and all other points in the nearest cluster that the sample is not a part of.

comp.graphics	talk.politics.guns	rec.sport.baseball	sci.crypt
comp.os.ms-windows.misc	talk.politics.mideast	rec.sport.hockey	sci.electronics
comp.sys.ibm.pc.hardware	misc.forsale	rec.autos	sci.med
comp.sys.mac.hardware	talk.politics.misc	alt.atheism	sci.space
comp.windows.x	talk.religion.misc	soc.religion.christian	rec.motorcycles

Table 1: Ground-truth topics from 20 NewsGroup

	List of Words
Output One	['sale', 'discount', 'price', 'pricing', 'prices', 'purchase', 'dealer', 'sales', 'offer', 'shipping']
Output Two	['space', 'launch', 'spacecraft', 'lunar', 'nasa', 'satellite', 'orbit', 'rocket', 'moon', 'satellites']
Output Three	['encryption', 'security', 'cryptography', 'key', 'privacy', 'crypto', 'decryption', 'data', 'secure', 'vulnerabilities']

Table 2: Example output from QualIT

5 Experimental Setup

AWS Sagemaker and AWS Bedrock were used to preprocess data, run experiments and validate results. For the benchmark models, Gensim's LDA implementation and BERTopic were used.

5.1 Dataset

The 20 NewsGroups dataset was used to run the experiments. This dataset was chosen as a benchmark because it is widely used for these type of experiments with a recent example in the BERTopic paper. The 20 NewsGroups dataset3 contains 20,000 news articles across 20 categories. The dataset was pre-processed minimally: lowered tokens, special characters, stop words and tokens smaller than length 3 were removed, as well as tokens were lemmatized. This pre-processed dataset was used in all methods in this experiment for fair comparisons.

5.2 Model

LLM Enhanced Topic Modeling will be compared with LDA and BERTopic. Both LDA and BERTopic ran with default parameters from their respective model providers. Our LLM Enhanced Topic Modeling utilized Claude-2.1, with a temperature setting of 0, top_k of 50, and top_p of 0. These parameters were chosen with the aim of achieving the highest coherence scores, reflecting its superior semantic understanding and contextual analysis at the individual document level.

5.3 Evaluation

$$i_n(x,y) = \frac{\left(\ln \frac{p(x,y)}{p(x)p(y)}\right)}{-\ln p(x,y)} \tag{3}$$

We use topic coherence and topic diversity to measure performance of this experiment. We chose these metrics as they allow for comparisons between our benchmark models. Topic coherence (TC) (Equation 3) is a measure used to evaluate how meaningful a topic is based on the degree of semantic similarity among the top most frequent words within the topic. It ranges from [-1, 1] and is estimated by normalized pointwise mutual information [11]. A score of 1 in topic coherence indicates a perfect association. The importance of this metric is its semblance to human judgement with reasonable performance [20]. Topic diversity (TD) is the percentage of distinct words for all topics produced. [16] This metric ranges from [0, 1], with 1 signaling diversity in topics. Each model was evaluated by using TC & TD at intervals of 10 topics, within the search space from 10 to 50 topics, for a total of 5 evaluations. The results were averaged across all runs.

Table 1 is an example of the ground-truth topics. Table 2 is an example of output topic words by our proposed method. Each evaluated method outputs a similar list of words. The words in Table 2 are used to describe and understand what each output is meant to be about. We evaluate these outputs on how well they map to the 20 ground-truth categories in Table 1. This step is performed by manual human classification of each method output to the 20 ground-truth mapping. This step is important because it signals how well each method is able to correctly cluster to the ground truth and how easy it is for humans to classify and agree on the output's classification. For quantifying how well the methods do, we calculate the percentage of agreement between evaluators on categorized topics.

6 Results

The results from this experiment can be found in Table 3 and human evaluation of outputs can be found in Table 4. Table 3 shows that on average, LLM Enhanced Topic Modeling outperforms both LDA and BERTopic in the 20 NewsGroups dataset in terms of TC and TD. The proposed method exhibits greater performance for both Topic Coherence and Topic Diversity for topics in the 10-30 range. The 10-30 topic range is ideal for this dataset because 20 topics is the ground truth number for this dataset. All three methods achieved their highest level of coherence and diversity when the number of topics was set to 20. This finding is consistent with the structure of the dataset. However, when the number of topics increases to 40 and 50, BERTopic minimally outperforms LLM Enhanced Topic Modeling in terms of Topic Coherence, but not for Topic Diversity.

Further, human evaluators were tasked with categorizing topic words (outputs) from each method into one of the 20 ground-truth categories in Table 1. We found that human evaluators were able to

	No. of Topics	Topic Coherence	Topic Diversity
LDA	10	47.0%	69.0%
	20	57.0%	72.0%
	30	52.0%	67.3%
	40	52.0%	79.1%
	50	49.0%	77.0%
	Avg	51.4%	72.7%
BERTopic	10	56.0%	82.0%
	20	65.0%	85.0%
	30	62.0%	88.3%
	40	**62.0**%	88.8%
	50	**60.2**%	87.2%
	Avg	61.0%	86.3%
QualIT	10	**66.0**%	**95.0**%
	20	**70.0**%	**95.5**%
	30	**65.0**%	**93.0**%
	40	61.0%	**93.0**%
	50	60.0%	**92.0**%
	Avg	**64.4**%	**93.7**%

Table 3: Each model was evaluated by using Topic Coherence & Topic Diversity at intervals of 10 topics, within the search space from 10 to 50 topics, for a total of 5 evaluations. The results were averaged across all runs.

	Percentage of Agreement for Categorized Topics		
	At least 2 evaluators agreed	At least 3 evaluators agreed	All 4 evaluators agreed
LDA	50%	25%	20%
BERTopic	45%	25%	20%
QualIT	80%	50%	35%

Table 4: Agreement of topic classification by human evaluators to ground truth. A higher value indicates greater classification overlap between independent evaluators.

agree in topic classification more often with the outputs of QuallT vs the benchmarks, as shown in Table 4. For example, the column "At least 3 evaluators" shows us how many lists of topic words (out of 20) were identically mapped to the same topics by at least 3 evaluators independently. We see that the percentage of agreement for categorized topics is consistent across the number of evaluators and across methods. This means that based on our human evaluation, the output from our method is less ambiguous to humans and easier to classify into topics vs the benchmarks.

7 Limitations and Future Work

The current method presents certain limitations that could be addressed in future research. Firstly, the runtime of the model is a significant constraint for large dataset; efforts should be made to reduce it to be more in line with BERTopic, which completes in approximately 30 minutes as opposed to current 2-3 hours for our LLM Enhanced Topic Modeling. This enhancement in efficiency would greatly benefit scalability and usability in practical applications. Additionally, while our method currently utilizes K-Means clustering, there is an opportunity to explore HDBSCAN. Preliminary comparisons suggest that BERTopic's use of HDBSCAN demonstrates more effective results than K-Means. Adding a similar approach may enhance the model's ability to discern and group more complex and nuanced data patterns, thus offering a promising avenue for advancing the robustness and accuracy of the topic modelling performed by our method.

8 Closing Thoughts

Qualitative Inights Tool (QualIT) presented in this paper represents a significant advancement in the extraction and analysis of qualitative insights from unstructured text data. By leveraging state-of-the-art pre-trained large language models and a novel topic modeling framework, QualIT is able to surface both high-level topic insights as well as more granular subtopics from qualitative feedback data.

Crucially, our experiments demonstrate that QualIT's approach produces more coherent and diverse topics compared to benchmark topic modeling techniques.

As organizations increasingly rely on qualitative data to drive strategic decision making especially for talent management, tools like QualIT that can efficiently and effectively extract meaningful insights will become increasingly invaluable. We believe the QualIT framework represents an important step forward in empowering researchers, product teams, and decision-makers to uncover the rich insights hidden within their qualitative datasets. Going forward, further enhancements to the language modeling capabilities (such as languages beyond English, especially low resources ones) and topic clustering algorithms underpinning QualIT hold promise to unlock even more powerful qualitative analysis capabilities.

References

[1] Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14*. Springer, 391–402.

[2] Vaswani Ashish. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017), I.

[3] Sreyoshi Bhaduri. 2018. NLP in Engineering Education-Demonstrating the use of Natural Language Processing Techniques for Use in Engineering Education Classrooms and Research. (2018).

[4] Sreyoshi Bhaduri, Marina Dias, Amulya Mysore, Robert Pulvermacher, Amelia Rivera-Burnett, Shahriar Sadighi, and Wanqun Zhao. 2023. Using science to support and develop employees in the tech workforce-an opportunity for multi-disciplinary pursuits in engineering education. (2023).

[5] Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Anshul Mittal, and Rutu Mulkar. 2024. Reconciling Methodological Paradigms: Employing Large Language Models as Novice Qualitative Research Assistants in Talent Management Research. *arXiv preprint arXiv:2408.11043* (2024).

[6] Sreyoshi Bhaduri, Kenneth Ohnemus, Jess Blackburn, Anshul Mittal, Yan Dong, Savannah Laferriere, Robert Pulvermacher, Marina Dias, Alex Gil, Shahriar Sadighi, et al. 2024. (Multi-disciplinary) Teamwork makes the (real) dream work: Pragmatic recommendations from industry for engineering classrooms. (2024).

[7] Sreyoshi Bhaduri, Michelle Soledad, Tamoghna Roy, Homero Murzi, and Tamara Knott. 2021. A Semester Like No Other: Use of Natural Language Processing for Novice-Led Analysis on End-of-Semester Responses on Students' Experience of Changing Learning Environments Due to COVID-19. In *2021 ASEE Virtual Annual Conference Content Access*.

[8] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974* (2020).

[9] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

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

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

[12] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[13] Raffel Colin. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21 (2020), 140–1.

[14] Romain Deveaud, Eric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17, 1 (2014), 61–84.

[15] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

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

[17] T Griffiths and M Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America. *vol* 101 (2004), p9.

[18] Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21, 3 (2013), 267–297.

[19] 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 (2021), 2018–2033.

[20] 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*. 530–539.

[21] Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124* (2019).

[22] Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[23] Ali Maatouk, Nicola Piovesan, Fadhel Ayed, Antonio De Domenico, and Merouane Debbah. 2024. Large language models for telecom: Forthcoming impact on the industry. *IEEE Communications Magazine* (2024).

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[25] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science* 58, 4 (2014), 1064–1082.

[26] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE, 747–748.

[27] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8968–8975.

[28] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. 2021. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10623–10633.