创建目标明确且可解释的主题模型，结合 LLM 生成文本增强

Anna Lieb, Maneesh Arora, Eni Mustafaraj
Wellesley College, Wellesley, USA

Keywords: large language models, GPT-4, topic modeling, text augmentation, content analysis

1

扩展摘要

在计算社会科学研究中，大规模理解自然语言文本是一个常见的挑战，特别是在研究人员可以获取大量未标注数据但缺乏标注数据的领域。无监督的机器学习技术，如主题建模和聚类，通常用于识别政治科学和社会学等领域非结构化文本数据中的潜在模式 [12, 9, 11, 5] 。这些方法克服了人类定性分析中再现性和费用高昂的常见问题。然而，主题模型的两个主要限制是其可解释性以及其在回答有针对性的、领域特定的社会科学研究问题时的实用性。在这项工作中，我们研究了使用大型语言模型 (LLM) 生成的文本增强来改善主题建模输出价值的机会。我们使用一个政治科学案例研究来在领域特定的应用中评估我们的结果。

相关工作。当将主题建模方法应用于社会科学时，研究人员必须对如何解释一组输出的关键词 [6, 12] 做出主观决定。以前尝试提高主题模型可解释性的方法使用了各种技术，包括结合从诸如 BERT 之类的预训练语言模型中获取的文档嵌入 [15, 7] ，并提供半监督的关键词种子 [3] 。随着最先进的大型语言模型（LLMs）在基于自然语言提示需要推理的任务上达到越来越高的性能 [13] ，社会科学家也在探索 LLM 驱动的计算内容分析方法，例如用于 GPT-3.5 主题生成的提示技术 [14] 。这些创新主要针对并评估在与传统主题模型相同的一般模式识别任务上设计的，不适合于探索特定领域的研究问题 [12, 1] 。它们的输出受主题生成的一般目标驱动，而没有考虑到潜在的社会科学理论。半监督的主题建模方法，如关键词种子、术语加权和先验主题词分布 [3, 4, 18] ，包含了一些特定领域的见解；然而，它们给研究人员带来了推导先验知识表示的负担。此外，研究人员的期望和偏见可能影响半监督结果的可靠性。

方法。我们提出了一种使用 LLM 生成的文本增强进行主题建模的程序，这可以为潜在文本添加有针对性的语义上下文和现实世界知识。此方法可以提高研究人员通过主题建模回答特定领域研究问题的能力。它可以结合社会科学理论，而不需要主题特定和词汇特定的先验知识，从而最大限度地减少研究人员的监督，提高可靠性。在这一改进的技术中，主题模型将 LLM 生成的描述符作为输入文本，而不是使用未经处理的原始输入文本。该方法主要适用于使用非常短的文本文档进行主题建模，例如社交媒体帖子和单句文本数据。

案例研究。我们的方法受到正在进行的政治科学研究的启发，研究美国有关批判种族理论（CRT）争议的应用。CRT 争议出现在美国的“文化战争”辩论中，始于 2020 年 9 月，当时保守派政治人物批评 K-12 学校和工作场所培训中的反种族主义课程 [17] 。为了更好地理解围绕 CRT 的党派辩论，我们对从 GDELT（全球事件、地点和基调数据）项目数据库 [10] 收集的 11,704 个与 CRT 相关的在线新闻标题进行内容分析。具体来说，我们的目标是识别新闻报道中突出的主要参与者。这可以回答政治科学问题，

例如：新闻标题是否将 CRT 框架为一个由学生、家长和教师等地方级别参与者之间自发出现的基层问题？还是新闻报道将 CRT 框定为由立法者、行政人员和新闻评论员等政治精英驱动的问题？之前的政治传播研究表明，新闻媒体对问题框架的不同可能显著影响公众对该问题的态度和理解 [2, 8, 16] 。

我们对我们的标题数据集进行了使用 LLM 生成的文本扩充的主题建模，以识别出现在标题中的角色类别。首先，我们使用 GPT-4 生成每条新闻标题中主要角色的简要描述。[1] 这是一个关键步骤，它结合了我们的研究目标（识别显著的主要角色），而不明确参考我们期望在输出主题中出现的角色或关键词。LLM 生成的文本扩充的例子如表 1 所示。基于对角色描述样本的定性审查，我们发现 GPT-4 生成的角色描述是准确的，提供了及时的真实世界信息，并提供了关于每个角色在标题背景中的角色的相关细节。然而，GPT-4 的表现可能在不同应用中有所差异，未来的工作可以更全面地评估文本扩充的质量。

接下来，我们使用 BERTopic 生成主题 [7] ，使用 LLM 生成的行为者描述作为输入文档。作为比较的基准，我们还使用原始的标题文本作为输入文档训练了一个 BERTopic 模型。基于我们对 BERTopic 输出的代表性关键词和文档的解释，我们为两个主题模型中的每个主题分配了标签。增强的主题建模输出及我们的定性解释显示在表 2 中，基准结果显示在表 3 中。一些相似的主题被归为一组以便于解释。

结果。我们发现，使用 GPT-4 简介（而不是未经处理的文本）进行无监督主题建模，会创建高度可解释的类别，这些类别可以在最少的人为指导下用于调查特定领域的研究问题。基线模型识别了种族主义、立法和教育等一般主题，而增强了 LLM 的模型识别了具体的参与者，如州长、立法者、教师和家长。尽管基线输出提到了一些这些参与者，但它们没有被清晰地分组。例如，最大的主题将校董事会、最高法院法官、家长和教师全部置于同一主题分组中（见表 3 ）。这种输出类型对于针对性分析出现在新闻报道中的突显人物不太有用。总的来说，这项研究表明，LLM 生成的文本可以通过语义上下文、真实世界的信息和特定领域的主题目标来增强短文档。我们提出的将 LLM 生成文本增强融入主题模型的程序，可以在保持可解释性和可重现性的同时，扩展现有无监督技术的实用性。

References

[1] Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. Three gaps in computational text analysis methods for social sciences: A research agenda. Communication Methods and Measures, 16(1):1–18, 2022.

[2] Dennis Chong and James N. Druckman. Framing theory. Annual Review of Political Science, 10(1):103–126, 2007.

[3] Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. Keyword-assisted topic models. American Journal of Political Science, 2020.

[4] Angela Fan, Finale Doshi-Velez, and Luke Miratrix. Assessing topic model relevance: Evaluation and informative priors. Statistical Analysis and Data Mining: The ASA Data Science Journal, 12(3):210–222, 2019.

[5] Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis, 18(1):1–35, 2010.

[1]我们使用以下提示从每个标题中提取演员信息：这个标题中的主要演员是什么类型的？简要描述主要演员。如果标题中没有提到演员，请说明。你不需要在回答中包含标题。

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

[7] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794, 2022.

[8] Shanto Iyengar. Is anyone responsible?: How television frames political issues. University of Chicago Press, 1994.

[9] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. Information, 10(4):150, 2019.

[10] Kalev Leetaru and Philip A Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. ISA Annual Convention, pages 1–49, 2013.

[11] Ana Macanovic. Text mining for social science–the state and the future of computational text analysis in sociology. Social Science Research, 108:102784, 2022.

[12] Laura K. Nelson. Computational grounded theory: A methodological framework. Sociological Methods & Research, 49(1):3–42, 2020.

[13] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, (...), and Barret Zoph. GPT-4 technical report, 2023.

[14] Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. TopicGPT: A prompt-based topic modeling framework. arXiv preprint arXiv:2311.01449, 2023.

[15] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! arXiv preprint arXiv:2004.14914, 2020.

[16] Andreas R.T. Schuck Sophie Lecheler, Mario Keer and Regula Hänggli. The effects of repetitive news framing on political opinions over time. Communication Monographs, 82(3):339–358, 2015.

[17] Benjamin Wallace-Wells. How a conservative activist invented the conflict over critical race theory. The New Yorker, June 2021.

[18] Justin Wood, Patrick Tan, Wei Wang, and Corey Arnold. Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pages 411–422. IEEE, 2017.

2

图表

Original headline text	GPT-4 primary actor description
KBJ and CRT	The primary actor in the headline "KBJ and CRT" refers to Ketanji Brown Jackson (KBJ), who is a Justice on the Supreme Court of the United States. She is a legal professional and a public figure, serving in the highest court in the United States.
DeSantis Bans CRT From K-12 Classrooms Despite Absence From Curriculum	The primary actor in the headline is DeSantis, referring to Ron DeSantis, the Governor of Florida. He is a political figure and a member of the Republican Party, known for his conservative policies and actions within the state of Florida.
Tennessee House Passes Bill for Black History to be Taught In Fifth Through Eighth Grade Despite State Ban on Critical Race Theory Concepts	The primary actor in this headline is the "Tennessee House." The Tennessee House refers to the lower chamber of the Tennessee General Assembly, which is the state legislature of Tennessee. This body is responsible for making and passing state laws. In this context, the Tennessee House is acting in its legislative capacity to pass a bill related to educational curriculum requirements.
Critical race theory battles are waged without enough information	The headline does not explicitly reference a specific actor. It discusses "critical race theory battles" in a general sense, implying that the primary actors could be individuals or groups engaged in debates or conflicts over critical race theory, but it does not specify who these actors are.
Critical Race Theory Seeks To Tell A More Accurate Story, Supporters Say	The primary actor in the headline is "Critical Race Theory." Critical Race Theory is not a person but an academic and legal framework that examines society and culture as they relate to categorizations of race, law, and power. It seeks to understand how racism is embedded within social structures and legal systems.

Table 1: 使用 GPT-4 描述进行样本文本增强。GPT-4 被提示生成每个新闻标题中主要行为者的简要描述。这些行为者的描述提供了与我们的研究问题特别相关的额外的现实世界背景和语义信息。

Primary actor interpretation	# of docs in topic	KeyBERT representation
CRT ideology (itself)	2128	examines society, political, activists, groups, debate, concept, opposing, topic
School administration	664	school boards, school districts, school district, public schools, superintendent, overseeing public
Teachers	271	teachers, educators, employed educators, new teachers, educational roles
	167	educators, teachers, teachers professors, educators working, responsible educating
	121	teachers union, teachers role, federation teachers, unions organizations
Governors	201	governor, mississippi governor, governor political, lieutenant governor, state executive
	197	florida governor, governor desantis, ron desantis
	144	glenn youngkin, youngkin politician, governor glenn, virginia governor
Legislators	383	proposed legislation, legislature, legislative action, state legislature, legislators
	137	senator, states senator, senators elected, senate senate
Parents	538	group parents, parents, black father, individuals children, father individual
News media	486	fox news, news channel, cable news, msnbc, tucker carlson, news coverage, political commentator, television host
Republicans	442	republican party, gop, political party, republicans
Joe Biden	278	joe biden, biden administration, president joe, political figure, president
Florida	220	florida education, florida state, florida school
Military	173	military officer, military advisor, general milley, joint chiefs, chairman, secretary defense
Attorneys general	116	montana attorney, attorney general, attorney, missouri attorney, indiana attorney, overseeing state
Southern Baptists	111	baptist convention, southern baptists, baptist organization, convention sbc
No assignment	2795	Outlier documents
	2132	Rule-based exclusion from model (contains “does not reference” or “does not explicitly reference”)

Table 2: BERTopic 的结果和解释使用了 GPT-4 生成的文章描述。模型考虑了一元词组和二元词组，最小主题大小为 100 个文档。BERTopic 生成了 19 个可解释的主题，并识别出了特定的主要参与者。展示了 10 个 KeyBERT 词语中的一个代表性子集。

Topic interpretation	# of docs in topic	KeyBERT representation
Misc. educational and political actors	2723	school board, school district, public schools, brown jackson, schools, parents, curriculum, teachers
Racial conflict & controversy	1394	racial, racism, racist, controversy, white people, debate, critics, diversity, fox news, discussion
State-specific coverage	221	textbooks florida, florida classrooms, florida schools, schools desantis, florida desantis, desantis proposes
(FL, TX, SD, VA)	192	texas legislature, texas schools, texas lawmakers, banning texas, approval texas, texas bill, texas public
	149	florida bans, bans classrooms, florida education, florida news, florida classrooms, taught florida
	117	dakota bans, dakota legislature, dakota noem, dakota lawmakers, dakota gov, dakota colleges, gov kristi,
	100	glenn youngkin, youngkin campaign, gov youngkin, youngkin bans, youngkin virginia
Values in the classroom	658	america classrooms, racism, ethnic studies, fight schools, american schools, civics education
Passing policies	351	bills banning, bills ban, bill senate, mississippi senate, transparency bill, anti bill, ban teaching
	172	bans schools, schools banning, bans teaching, black students, teachers protest
Parents	381	black parents, parents protest, parents fight, parents rally, parents push, parents concerned, black mother
Republicans & Democrats	192	gop reps, gop attacks, biden proposal, issue gop, gop, gop lawmakers, back biden, biden, reps push
	110	biden education, schools biden, biden pushing, biden fight, biden administration, biden democrats
Teachers	266	pledge teach, teachers pledging, sign pledge, new teachers, teach controversial
Military	208	professor defends, academy professor, defends teaching, military academies, air force, cadets, force officer
Protest & elections	134	board races, school board, condemning racism, board candidates, meeting debate, board meeting, dismay students, questions school
	97	letter criticism, racist, letters opposing, issue letter
Southern Baptists	117	southern baptists, baptist convention, baptist leaders
No assignment	4122	Outlier documents

Table 3: BERTopic 使用原始文章标题的结果和解释。该模型考虑了单字和双字组，最小主题大小为 90 个文档。BERTopic 产生了 18 个主题，解释性不如表 2 中的 GPT-4 增强结果。展示了代表性的一组 10 个 KeyBERT 词语。