Cambridge Medium (2020), **37**, 1–8

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE TYPE

Text Embeddings to Measure Text Topics

Jianjun Yu*

Political Science Department, The University of Iowa, Iowa City, Iowa, USA
*Corresponding author. Email: Jianjyu@uiowa.edu

Abstract

Automated content analysis for measuring specific topics has become increasingly popular in social science research. This article demonstrates that text embedding, combined with cosine similarity, can accurately measure specific topics and explore the relationship between text topics and meta-information. Unlike probabilistic topic models (PTMs) such as the Structural Topic Model (STM) and Keyword-Assisted Topic Model (keyATM), the text embedding method does not require the selection of hyperparameters, keywords, or a training process. This significantly reduces computational resources and makes the method more convenient and faster to use. Additionally, the text embedding method can better capture the meaning of short texts. The findings suggest that text embedding is a superior alternative to commonly used PTMs.

Keywords: Text analysis, Text embedding, Topic model

1. Introduction

In recent years, social scientists have developed several automated content analysis methods based on PTM, such as STM and keyATM, to measure specific topics and explore the relationship between text topics and meta-information (Roberts et al. 2014; Eshima, Imai, and Sasaki 2024). These methods have been widely used for text analysis and have significantly advanced the development of social science. However, these methods require considerable effort to select key hyperparameters, such as the number of topics and keywords, complicating their usage and making it challenging to measure specific concepts of interest.

In this letter, I propose that a text embedding method offer a more convenient alternative to these topic models when scholars aim to measure specific topics and explore the relationship between text topics and meta-information. Unlike commonly used topic models, text embedding allows scholars to focus on measuring specific concepts of substantive interest without clustering texts, freeing them from selecting key hyperparameters. Moreover, text embedding methods require significantly fewer computational resources and support distributed computing, making them more suitable for big data analysis. Finally, compared with PTMs, text embedding can better capture the meaning of short text.

2. Text embedding

Text embedding in natural language processing (NLP) represents texts as vectors in a multidimensional space to capture their semantic meaning and context, facilitating efficient language data processing (Kiros et al. 2015; Conneau et al. 2017; Reimers and Gurevych 2019). By converting text into numerical vectors, machine learning models can better understand and work with language, preserving meaningful relationships and contextual nuances.

Text embeddings are generated using two main approaches: word embedding algorithms like Word2Vec and GloVe (Mikolov et al. 2013; Pennington, Socher, and Manning 2014), and fine-tuned transformer-based language models (TLMs) like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer). Appendices A and B provide more detailed discussions on how these methods generate text embeddings.

This letter focuses on two popular models: the all-mpnet-base-v2 model (Mpnet model) and the text-embedding-3-large model (GPT model).

The Mpnet model, fine-tuned on BERT and published in 2022, is compact and transparent, with publicly accessible training data, methodologies, and model structures. This openness allows scholars to scrutinize potential biases and facilitates replication. The Mpnet model is popular for its prominence on Hugging Face, a leading transformer model hub, and its top-tier performance in benchmark tests.

The GPT model, fine-tuned on OpenAI's GPT-3.5, is one of the most accurate text embedding generation models (Greene et al. 2022). However, its large size, encompassing billions of parameters, makes it impractical for use on standard personal computers, so users must generate text embeddings through OpenAI's API. Additionally, the opacity of OpenAI's GPT models poses challenges for scholars in assessing potential biases and replicating outcomes.

3. Topic score

I employ cosine similarity between the text embedding of a specific topic (topic embedding) and the text embeddings of the texts of interest to represent the relationship between the topic and the texts. A topic embedding can be the text embedding of a word, phrase, or paragraph describing the topic. I refer to this cosine similarity as the "topic score."

Cosine similarity is a metric used to measure the similarity between two vectors in a multidimensional space (Salton, Wong, and Yang 1975). It calculates the cosine of the angle between two vectors, where a similarity of 1 indicates identical vectors, and a similarity of -1 indicates diametrically opposed vectors. In NLP and machine learning, cosine similarity is used to assess the similarity between vectors representing documents, sentences, or words (Manning 2009). The formula for cosine similarity between two vectors A and B is given by:

$$CosineSimilarity(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{1}$$

Cosine similarity of text embeddings between two documents serves as a valuable metric for representing the similarity in content between documents. To illustrate this capability, consider recent tweets from President Joe Biden. Table 1 presents three tweets covering topics such as the American economy and abortion.

Table 1. Tweets from Biden

Tweets from Biden	
Tweet 1	I know that some prices are still too high for too many. I am doing everything in my power to lower costs from energy bills and medicine to addressing hidden junk fees companies use to rip you off.I won't stop fighting for American workers and American families.
Tweet 2	Our economy created 2.7 million new jobs in 2023 while the unemployment rate was consistently below 4%. That's more jobs than during any year of the prior Administration.This morning's report confirms that it was a great year for American workers
Tweet 3	Today's Supreme Court order allows Idaho's abortion ban to go back into effect, denying women emergency abortion care required by law. These bans threaten women's health, force them to travel, and make it harder for doctors to provide care. This should never happen in America.

Table 2 displays the cosine similarity between these tweets[1]. The first two tweets, focused on

[1] All text embeddings are calculated by the Mpnet model

Biden's efforts to address economic issues, exhibit higher cosine similarity with each other compared to their similarity with the tweet addressing abortion.

Table 2. Tweets Cosine Similarity Between Tweets

Tweets	Tweet 1	Tweet 2	Tweet 3
Tweet 1	1	0.458	0.158
Tweet 2	0.458	1	0.128
Tweet 3	0.158	0.128	1

Table 3. Tweets Cosine Similarity with Topics

	Economy	Job market	Abortion
Tweet 1	0.267	0.186	0.084
Tweet 2	0.382	0.405	0.090
Tweet 3	0.091	0.103	0.503

Table 3 delineates the cosine similarity between the tweets and three topics: 'economy,' 'job market,' and 'abortion.' The calculated cosine similarities illustrate the proximity of each tweet to a particular topic. It shows that the first two tweets are more closely aligned with the 'economy' topic and less so with 'abortion.' Additionally, it reveals that the second tweet is thematically nearer to 'job market' than 'economy,' reflecting its specific focus on job market issues. These results show that the topic score can properly reflect the relationship between certain topics and text content.

The topic score can then be used as a feature of texts, which scholars can use as a variable in their research to explore the relationship between text topics and their meta-information.

Measuring topics with topic scores offers several advantages compared to PTM. This method allows scholars to measure specific, research-critical concepts. Commonly used methods, such as STM, may not necessarily capture specific concepts of substantive interest. They might create multiple topics with similar content or merge distinct themes into a single topic, leading to miscategorization and obscuring topic interpretation. Topic scores, on the contrary, allow scholars to measure any topics they are interested in.

Second, topic scores are easier to calculate, freeing researchers from the intensive work of selecting hyperparameters and text pre-processing. While some PTMs, such as keyATM, allow users to assign topics before clustering, they often require selecting keywords, a process with unclear best practices. Moreover, PTMs need a predetermined number of topics, and clustering results are sensitive to this number. Selecting the number of topics involves fitting models with many different values, which is time-consuming, and even then, only provides a range requiring further validation. Additionally, PTMs need extensive text pre-processing, including noise reduction, tokenization, and removal of stop words. Calculating text embeddings and topic scores eliminates the need for these steps.

Third, text embedding is better at measuring the meaning of short text, such as tweets. PTMs often falter with short texts due to their reliance on word occurrences to estimate topic distributions. Short texts, with their limited word occurrences, pose a challenge in accurately determining their topic distributions (Yan et al. 2013), prompting researchers to adopt specific text preprocessing strategies or exclude short texts altogether (Barberá et al. 2019; Ying, Montgomery, and Stewart 2022). The generation of text embedding, on the contrary, will consider word meaning, semantic relationships, and contextual information learned during pre-training and fine-tuning, ensuring consistent quality across text lengths.

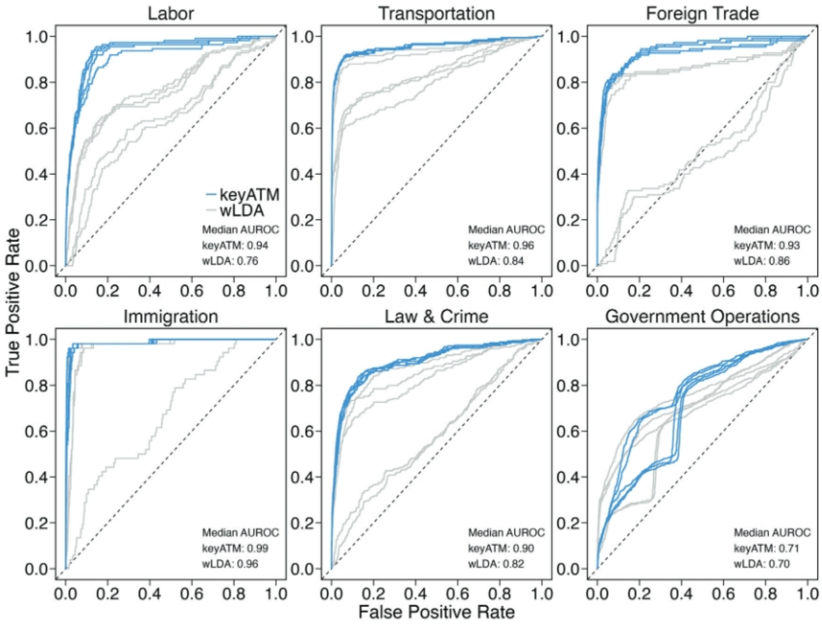
Finally, unlike other methods, calculating topic scores does not require a training process, which demands high computational power and memory. For large text data analysis, users can generate topic scores in chunks, avoiding the need to load all data into memory simultaneously, and allowing distributed computing. This significantly reduces the memory needed and facilitates the analysis process, allowing users without access to high-performance clusters to conduct research. Generating text embeddings and calculating topic scores can be done on a mid-level personal CPU or GPU. In the first replication study, I will report the time needed for generating embeddings on an Intel Core i5-13600 CPU, a mid-range CPU, and a V100 NVIDIA GPU, an older and readily available GPU published in 2018. to demonstrate the modest resource usage for topic score calculation.

4. Empirical study

In the following section, I apply this topic score method to three datasets: American Congressional bills, open-ended survey responses from , and . The American Congressional bills dataset was used by Eshima and his coauthors to show that their recently published keyATM achieves state-of-the-art performance. The other two datasets represent typical data formats on which social scientists conduct automated content analysis.

4.1 Replication 1: Congressional bills

In early 2024, Eshima et al (2024) proposed a new topic model, keyATM, which uses a small number of keywords to guide the topic generation process, enabling the model to generate topics with specific research interests, they cluster over 4000 American Congressional bills to demonstrate its state-of-the-art performance. These bill was assigned a primary policy topic from among 21 topics by human coders. They used the area under the receiver operating characteristic (AUROC) of topic probabilities from the document-topic distribution as the evaluation metric. To compare with keyATM, I calculated topic scores for all 21 topics. The topic embeddings were calculated using short descriptions of the 21 topics provided by the Congressional Bills Project, which Eshima and his coauthors used to generate keywords for keyATM. No text pre-processing was performed. Both text embedding models truncated input texts longer than their input length limit. Although best practices involve splitting long texts into chunks and averaging their embeddings, truncation provided sufficiently comparable results in this research.



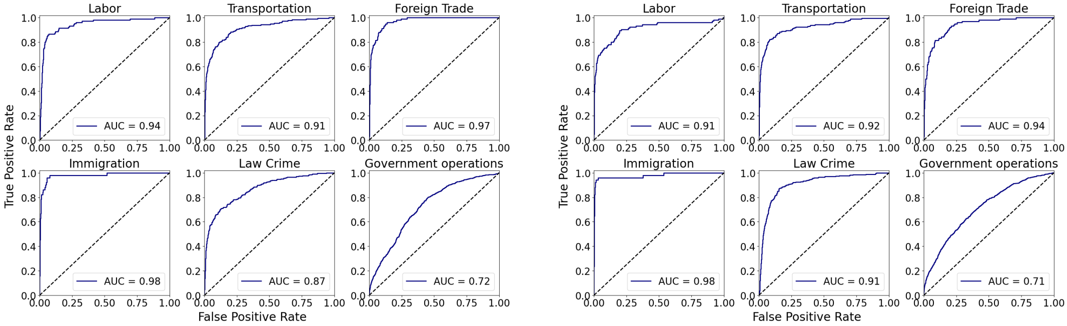
Note: Each line represents the ROC curve from one of the five Markov chains with different starting values for keyATM (blue lines) and wLDA (gray lines). The median AUROC indicates the median value of AUROC among five chains for each model.

Figure 1. Comparison of the ROC Curves between keyATM and wLDA for Six Selected Topics

For the topic score of Mpnet model, generating all topic scores took around 20 minutes on an Intel Core i5-13600 CPU and 1 minute on a V100 GPU. I used a V100 GPU on Google Colab, which cost approximately 0.009 USD. For topic score of GPT model, I request text embedding of all bills through OpenAI's API, which cost around 2 USD. I also replicated the Eshima et al's keyATM code on the i5 CPU. It took around 16 hours to fit all five chains of the keyATM model to cluster the 4421 bills[2].

Figure 1 copies the ROC curves from the original keyATM paper, comparing keyATM and weighted LDA for six selected topics. Figure 2 shows the ROC curves for topic scores of the same six topics: Figure 2a for the Mpnet model and Figure 2b for the GPT model. ROC curves for all 21 topics are in Appendix C.

Comparing Figures 1 and 2, keyATM does not outperform the topic scores even generated by the 2022 available Mpnet model. The average AUROC for all 21 topics is 90.00 for keyATM, 90.00 for Mpnet topic scores, and 90.21 for GPT topic scores. These results show that the topic score method provides accuracy comparable to the state-of-the-art keyATM.



(a) Generated from the Mpnet Model

(b) Generated from the GPT Model

Figure 2. The ROC Curves for Topic Scores of the Six Topics

Moreover, keyATM relies on selecting keywords for each topic and uses time-consuming and memory-intensive Markov chain Monte Carlo (MCMC) algorithms. In contrast, the topic score method requires less memory and is much faster, in partiuclar for large data. The generation of topic embeddings is also very flexible and easier to apply than selecting keywords. Thus, compared to keyATM, the topic score method is more useful when a researcher is interested in measuring specific topics.

4.2 Replication 2: Open-Ended Survey Responses

In their seminal work on the gender gap in climate change attitudes, Bush and Clayton argued that people in wealthy countries perceive greater costs to climate change mitigation than those in lower-income countries, with men in wealthy countries, such as the USA, being more sensitive to these material costs than women. To support their argument, they conducted a survey across ten countries with an open-ended question about how efforts to stop climate change would harm respondents. Using STM, they found that people in lower-income countries were more likely to be clustered into the topic: 'no harm.' They also showed that male respondents in the USA were more likely to discuss the rising costs of acting to stop climate change.

To replicate Bush and Clayton's work, I created two topic scores: one with the topic embedding 'climate change does not harm me' (no harm) and another with 'climate change raises material costs' (rising cost). No hyperparameter selection or text preprocessing was done. I used these topic scores as dependent variables to analyze the influence of countries' income and gender.

Table 4 reports the results. The first two columns show regression results for topic scores generated from the Mpnet model, while the last two columns show results for the GPT model. The

[2] Eshima and his coauthors ran five independent Markov chains with different random starting values for each topic due to PTM sensitivity to initial values

regression shows that for topic scores generated from both models, income has a negative effect on the 'no harm' topic score, and being female has a negative effect on the 'rising cost' topic score. These findings are consistent with the original work and the effects found from text embedding method is also more significant.

This replication shows the ability of the text embedding method to explore the relationship between text meaning and meta information. It also shows the advantage of text embedding method to analysis open-ended response from cross countries survey. Most PTMs cannot handle multiple languages without translation, whereas many text embedding models can analyze multilingual texts and convert multiple languages into the same vector space, eliminating the need for translation.

Table 4. Comparison of Methods for Exploring the Effects of Meta Information on TED Talk Topics

	Dependent variable:					
Topic generator	The Mpnet model		The GPT model		STM	
lnGDPPC	-0.017***		-0.019***		-0.013***	
	(0.002)		(0.002)		(0.002)	
Female		-0.039***		-0.031***		-0.017**
		(0.012)		(0.008)		(0.008)
Observations	11,849	975	11,849	975	11,849	975
R^2	0.007	0.036	0.010	0.052		
Adjusted R^2	0.006	0.031	0.010	0.047		

Table note
a *p<0.1; **p<0.05; ***p<0.01
b

4.3 Replication 3: Better Ability of Capture Short text meaning

I compare the performance of the text embedding method and the probabilistic topic model in analyzing short texts by using the two methods to cluster short text. The idea is that if text embeddings better capture the text meanings, it should provide a more coherence clustering result than PTMs.

The dataset used is the Twitter Financial News dataset from Hugging Face, one of the few publicly available Twitter datasets with topic labels [3]. Its training set contains 17,000 tweets about financial news labeled into 20 topics. For the comparative analysis of clustering methods, purity is employed as the evaluation metric. Purity is specifically designed for assessing the performance of clustering methods (Meilă 2007). It measures the extent to which clusters contain a dominant class. Purity hinges on the premise that while clustering methods may partition data into a different number of groups than those determined by human labeling, data points sharing the same human label should be more likely to be grouped together than those with different labels.

Table 5. Purity score

Number of topic	Mpnet	GPT	STM
20	0.513	0.563	0.439
40	0.560	0.605	0.456
60	0.605	0.634	0.493
80	0.615	0.632	0.472
100	0.620	0.660	0.487

[3] For more details about the data: https://huggingface.co/datasets/zeroshot/twitter-financial-news-topic

Table 5 presents the comparison of clustering performance using Mpnet and GPT generated text embeddings against STM. The performance of different models is compared across 5 different topic numbers. Similar to other evaluation metrics, purity tends to increase with the number of topics but plateaus after surpassing a certain threshold, which is often considered the optimal topic number for a clustering method. The results consistently indicate superior performance of clustering on text embeddings compared to STM. The findings suggest that clustering utilizing text embeddings outperforms STM in generating clusters characterized by significantly greater semantic coherence.

5. Validation

Scholars might be concerned about the accuracy of topic scores. Two common validation approaches can address this. The first is to select texts with high and low topic scores and manually examine whether the scores are valid. The second approach involves generating a word frequency matrix weighted by topic scores. Words that frequently appear in texts with high topic scores are selected, and the word intrusion method is used to evaluate whether human coders agree that these high-frequency words reflect the topic's meaning.

6. Conclusion

This letter demonstrates that when measuring specific topics and exploring the relationship between text topics and meta-information, the text embedding method, which generates topic scores, provides a more convenient alternative to commonly used methods such as keyATM and STM. Text embedding is not limited to measuring specific topics. For example, combined with clustering methods like k-means, text embedding can also be used to cluster texts and explore content. While demonstrating the advantages of text embedding in other applications is beyond the scope of this letter, it encourages social scientists to further explore its use in their research.

Acknowledgement

Funding Statement No Fund

Competing Interests No Competing Interests

Notes

References

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review* 113 (4): 883–901.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364.*

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2024. Keyword-assisted topic models. *American Journal of Political Science* 68 (2): 730–750.

Greene, Ryan, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model. Accessed June 7, 2023. https://openai.com/blog/new-and-improved-embedding-model.

Kiros, Ryan, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems* 28.

Manning, Christopher D. 2009. *An introduction to information retrieval.* Cambridge university press.

Meilă, Marina. 2007. Comparing clusterings—an information based distance.*Journal of multivariate analysis* 98 (5): 873–895.

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

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. Glove: global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp),* 1532–1543.

Reimers, Nils, and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084.*

Roberts, Margaret E, 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): 1064–1082.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11): 613–620.

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on world wide web,* 1445–1456.

Ying, Luwei, Jacob M Montgomery, and Brandon M Stewart. 2022. Topics, concepts, and measurement: a crowdsourced procedure for validating topics as measures. *Political Analysis* 30 (4): 570–589.