The Journal of Open Source Software

Turftopic: Topic Modelling with Contextual Representations from Sentence Transformers

Márton Kardos [1], **Kenneth C. Enevoldsen** [1], **Jan Kostkan** [1], **Ross Deans Kristensen-McLachlan** [1,3], **and Roberta Rocca** [2]

1 Center for Humanities Computing, Aarhus University, Denmark **2** Interacting Minds Center, Aarhus University, Denmark **3** Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Denmark

DOI: 10.21105/joss.08183

Software

- Review
- Repository
- Archive

Editor: Abhishek Tiwari

Reviewers:

- @PetrKorab
- @Jemoka
- @marccanby

Submitted: 18 March 2025
Published: 01 July 2025

Summary

Topic models are machine learning techniques that are able to discover themes in a set of documents. Turftopic is a topic modelling library including a number of recent developments in topic modelling that go beyond bag-of-words models and can understand text in context, utilizing representations from transformers. Turftopic focuses on ease of use, providing a unified interface for a number of different modern topic models, and boasting both model-specific and model-agnostic interpretation and visualization utilities. While the user is afforded great flexibility in model choice and customization, the library comes with reasonable defaults, so as not to needlessly overwhelm first-time users. In addition, Turftopic allows the user to: a) model topics as they change over time, b) learn topics on-line from a stream of texts, c) find hierarchical structure in topics, d) learning topics in multilingual texts and corpora. Users can utilize the power of large language models (LLMs) to give human-readable names to topics. Turftopic also comes with built-in utilities for generating topic descriptions based on key-phrases or lemmas rather than individual words.

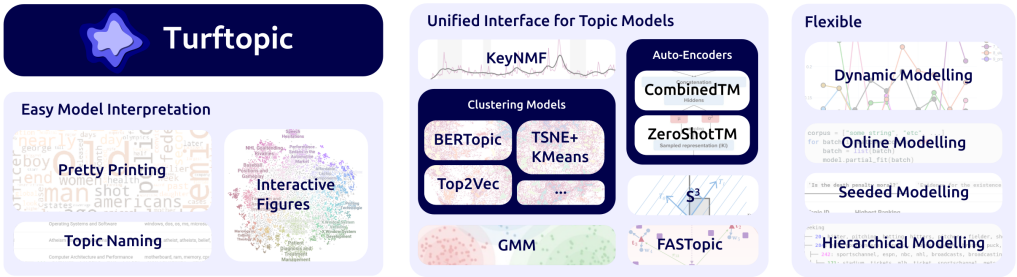

Figure 1: An Overview of Turftopic's Functionality

Statement of Need

While a number of software packages have been developed for contextual topic modelling in recent years, including BERTopic (Grootendorst, 2022), Top2Vec (Angelov, 2020), CTM (Bianchi, Terragni, & Hovy, 2021), these packages include implementations of one or two topic models, and most of the utilities they provide are model-specific. This has resulted in the unfortunate situation that practitioners need to switch between different libraries and adapt to their particularities in both interface and functionality. Some attempts have been made at creating unified packages for modern topic models, including STREAM (Thielmann

et al., 2024) and TopMost (Wu, Pan, et al., 2024). These packages, however, have a focus on neural models and topic model evaluation, have abstract and highly specialized interfaces, and do not include some popular topic models. Additionally, while model interpretation is a fundamental aspect of topic modelling, the interpretation utilities provided in these libraries are fairly limited, especially in comparison with model-specific packages, like BERTopic.

Turftopic unifies state-of-the-art contextual topic models under a superset of the `scikit-learn` (Pedregosa et al., 2011) API, which many users may be familiar with, and can be readily included in `scikit-learn` workflows and pipelines. We focused on making Turftopic first and foremost an easy-to-use library that does not necessitate expert knowledge or excessive amounts of code to get started with, but gives great flexibility to power users. Furthermore, we included an extensive suite of pretty-printing and model-specific visualization utilities that aid users in interpreting their results. In addition, we provide native compatibility with `topicwizard` (Kardos et al., 2025), a model-agnostic topic model visualization library. The library also includes three topic models that, to our knowledge, only have implementations in Turftopic: KeyNMF (Kristensen-McLachlan et al., 2024), Semantic Signal Separation (S^3) (Kardos et al., 2024), and GMM, a Gaussian Mixture model of document representations with a soft-c-tf-idf term weighting scheme.

Functionality

Turftopic includes a wide array of contextual topic models from the literature, these include: FASTopic (Wu, Nguyen, et al., 2024), Clustering models, such as BERTopic (Grootendorst, 2022) and Top2Vec (Angelov, 2020), auto-encoding topic models, like CombinedTM (Bianchi, Terragni, & Hovy, 2021) and ZeroShotTM (Bianchi, Terragni, Hovy, Nozza, et al., 2021), KeyNMF (Kristensen-McLachlan et al., 2024), S^3 (Kardos et al., 2024) and GMM. At the time of writing, these models are representative of the state of the art in contextual topic modelling and intend to expand on them in the future.

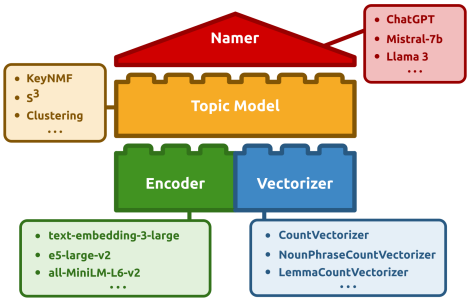

Figure 2: Components of a Topic Modelling Pipeline in Turftopic

Each model in Turftopic has an *encoder* component, which is used for producing continuous document-representations (Reimers & Gurevych, 2019), and a *vectorizer* component, which extracts term counts in each documents, thereby dictating which terms will be considered in topics. The user has full control over what components should be used at different stages of the topic modelling process, thereby having fine-grained influence on the nature and quality of topics.

The library comes loaded with numerous utilities to help users interpret their results, including *pretty printing* utilities for exploring topics, *interactive visualizations* partially powered by the `topicwizard` (Kardos et al., 2025) Python package, and *automated topic naming* with LLMs.

To accommodate a variety of use cases, Turftopic can be used for *dynamic* topic modelling,

where topics are expected to change over time. Turftopic is also capable of extracting topics at multiple levels of granularity, thereby uncovering *hierarchical* topic structures. Some models can also be fitted in an *online* fashion, where documents are accounted for as they come in batches. Turftopic also includes *seeded* topic modelling, where a seed phrase can be used to retrieve topics relevant to the specific research question.

Use cases

Topic models can be and have been utilized for numerous purposes in both academia and industry. They are a key tool in digital/computational humanities, mainly as an instrument of quantitative text analysis or *distant reading* (Nielbo et al., 2024), as topic models can pick up on macro-patterns in corpora, at times missed by close readers (JOCKERS, 2013), and might be able to provide a more impartial account of a corpus's content. Topic models can also aid discourse analysis by facilitating exploratory data analysis, and quantitative modelling of information dynamics (Jacobs & and, 2019). Industry analysts might make use of topic models for analyzing customer feedback (Nguyen & Ho, 2023) or social media data related to a company's products (Huang et al., 2025).

Since topic models learn topically informative representations of text, they can also be utilized for down-stream applications, such as content filtering, recommendation (Bergamaschi & Po, 2015), unsupervised classification (Thielmann et al., 2023), information retrieval (Yi & Allan, 2009) and pre-training data curation (Peng et al., 2025).

The Turftopic framework has already been utilized by Kristensen-McLachlan et al. (2024) for analyzing information dynamics in Chinese diaspora media, and is currently being used in multiple ongoing research projects, including one concerning the media coverage of the HPV vaccine in Denmark, and another studying Danish golden-age literature. We provide examples of correct usage and case studies as part of our documentation.

Target Audience

Turftopic's utility has already been demonstrated for computational scholars in digital humanities, and political science, and we expect that it will be of utility to a diverse audience of researchers in social sciences, medicine, linguistics and legal studies. It can furthermore prove valuable to business analysts working with text-based data to generate qualitative insights.

As the focus on pre-training data mixing techniques is on the rise, we expect that Turftopic will help facilitate foundational language model research. The library's design, wide array of models, and flexibility are also aimed at enabling usage in more extended production pipelines for retrieval, filtering or content recommendation, and we thus expect the package to be a most valuable tool for the industry NLP practitioner.

Turftopic is also an appropriate choice for educational purposes, providing instructors with a single, user-friendly framework for students to explore and compare alternative topic modelling approaches.

References

Angelov, D. (2020). *Top2Vec: Distributed representations of topics*. https://doi.org/10.48550/arXiv.2008.09470

Bergamaschi, S., & Po, L. (2015). Comparing LDA and LSA topic models for content-based movie recommendation systems. In V. Monfort & K.-H. Krempels (Eds.), *Web information systems and technologies* (pp. 247–263). Springer International Publishing. https://doi.org/10.1007/978-3-319-27030-2_16

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 759–766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.96

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 1676–1683). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.143

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv Preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Huang, Y., Li, M., Tsung, F., & Chang, X. (2025). Mining social media data via supervised topic model: Can social media posts inform customer satisfaction? *Decision Sciences*. https://doi.org/10.1111/deci.12660

Jacobs, T., & and, R. T. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485. https://doi.org/10.1080/13645579.2019.1576317

JOCKERS, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press. ISBN: 9780252037528

Kardos, M., Enevoldsen, K. C., & Nielbo, K. L. (2025). *Topicwizard – a modern, model-agnostic framework for topic model visualization and interpretation*. https://doi.org/10.48550/arXiv.2505.13034

Kardos, M., Kostkan, J., Vermillet, A.-Q., Nielbo, K., Enevoldsen, K., & Rocca, R. (2024). *S^3 – semantic signal separation*. https://doi.org/10.48550/arXiv.2406.09556

Kristensen-McLachlan, R. D., Hicke, R. M. M., Kardos, M., & Thunø, M. (2024). Context is key(NMF):: Modelling topical information dynamics in chinese diaspora media. In W. Haverals, M. Koolen, & L. Thompson (Eds.), *Proceedings of the computational humanities research conference 2024* (Vol. 3834, pp. 829–847). CEUR-WS. https://doi.org/10.48550/arXiv.2410.12791

Nguyen, V.-H., & Ho, T. (2023). Analysing online customer experience in hotel sector using dynamic topic modelling and net promoter score. *Journal of Hospitality and Tourism Technology*, *14*(2), 258–277. https://doi.org/10.1108/jhtt-04-2021-0116

Nielbo, K. L., Karsdorp, F., Wevers, M., Lassche, A., Baglini, R. B., Kestemont, M., & Tahmasebi, N. (2024). Quantitative text analysis. *Nature Reviews Methods Primers*, *4*(1). https://doi.org/10.1038/s43586-024-00302-w

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, J., Zhuang, X., Jiantao, Q., Ma, R., Yu, J., Bai, T., & He, C. (2025). *Unsupervised topic models are data mixers for pre-training language models*. https://doi.org/10.48550/arXiv.2502.16802

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992).

Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Thielmann, A., Reuter, A., Weisser, C., Kant, G., Kumar, M., & Säfken, B. (2024). STREAM: Simplified topic retrieval, exploration, and analysis module. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 435–444). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.41

Thielmann, A., Weisser, C., Krenz, A., & and, B. S. (2023). Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling. *Journal of Applied Statistics*, *50*(3), 574–591. https://doi.org/10.1080/02664763.2021.1919063

Wu, X., Nguyen, T. T., Zhang, D. C., Wang, W. Y., & Luu, A. T. (2024). FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2405.17978

Wu, X., Pan, F., & Luu, A. T. (2024). Towards the TopMost: A topic modeling system toolkit. In Y. Cao, Y. Feng, & D. Xiong (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: System demonstrations)* (pp. 31–41). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-demos.4

Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy (Eds.), *Advances in information retrieval* (pp. 29–41). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00958-7_6