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** Interactive Minds Center, Aarhus University, Denmark **3** Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Denmark

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Turftopic is a topic modelling library including a number of recent topic models that go beyond bag-of-words 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. The user is afforded great flexibility in model choice and customization, but the library comes with reasonable defaults, not to overwhelm first-time users with a plethora of choices. In addition, Turftopic allows you to model topics, as they change over time, learning themes from streams of texts, finding hierarchical topics, and multilingual usage. 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.

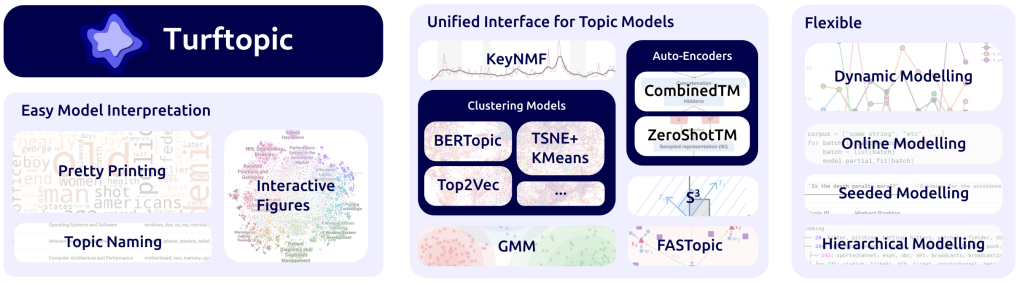

Figure 1: An Overview of Turftopic's Functionality

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 an incredibly important 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 users are likely already 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, included an extensive suite of pretty-printing and visualization utilities that aid users in interpreting their results. The library also includes three topic models, which to our knowledge only have implementations in Turftopic, these are: KeyNMF (Kristensen-McLachlan et al., 2024), S^3 (Kardos et al., 2024), and GMM.

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), Semantic Signal Separation (Kardos et al., 2024) and GMM. We believe these models to be representative of the state of the art in contextual topic modelling and intend to expand on them in the future.

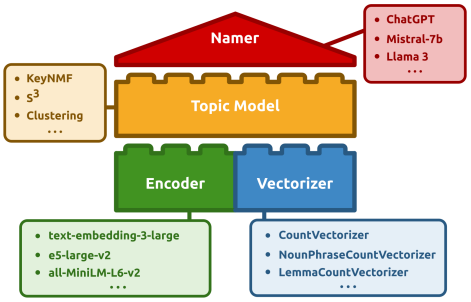

Figure 2: Components of a Topic Modelling Pipeline in Turftopic

Each model in Turftopic has an *encoder* component, which is used for producing continuous document-representations, 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 a lot of utilities to help users interpret their results, including *pretty printing* utilities for exploring topics, *interactive visualizations* partially powered by the `topicwizard` (Kardos, 2023) Python package, and *automated topic naming* with LLMs.

To accommodate a variety of use-cases, Turftopic can be used for dynamic topic modelling, where we expect topics to change over time, can be used for uncovering hierarchical structure in topics. Some models can also be fitted in an *online* fashion, where documents are accounted for as they come in by 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 utilized in a number of research settings, including exploratory data analysis, discourse analysis of diverse domains, such as newspapers, social media or policy documents. Turftopic 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 analyzing discourse on the HPV vaccine in Denmark, and studying Danish golden-age literature.

Target Audience

We expect that Turftopic will prove useful to a diverse user base including computational researchers in digital humanities and social sciences, and industry NLP professionals. Due to ease of use, Turftopic is also an appropriate choice for educational purposes.

Angelov, D. (2020). *Top2Vec: Distributed representations of topics*. https://arxiv.org/abs/2008.09470

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*. https://arxiv.org/abs/2203.05794

Kardos, M. (2023). *topicwizard: Pretty and opinionated topic model visualization in Python* (Version 0.5.0). https://github.com/x-tabdeveloping/topic-wizard

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

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.

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

Wu, X., Nguyen, T., Zhang, D. C., Wang, W. Y., & Luu, A. T. (2024). *FASTopic: A fast, adaptive, stable, and transferable topic modeling paradigm*. https://arxiv.org/abs/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