DTECT: Dynamic Topic Explorer & Context Tracker

Suman Adhya, Debarshi Kumar Sanyal

Indian Association for the Cultivation of Science, Kolkata, India
adhyasuman30@gmail.com, debarshi.sanyal@iacs.res.in

Abstract

To address the challenge of interpreting evolving themes in temporal text, we present **DTECT** (**D**ynamic **T**opic **E**xplorer & **C**ontext **T**racker), an interactive, end-to-end system for uncovering thematic dynamics. The system integrates a complete pipeline that supports data preprocessing, multiple model architectures, and dedicated metrics to analyze temporal topic quality. To enhance interpretability, DTECT features LLM-driven automatic topic labeling, trend analysis, interactive visualizations with document summarization, and a natural language chat interface. This cohesive platform empowers users to intuitively explore how topics change over time.

Code — https://github.com/AdhyaSuman/DTECT

Interactive Demo — https://huggingface.co/spaces/AdhyaSuman/

Demo Video — https://youtu.be/B8nNfxFoJAU

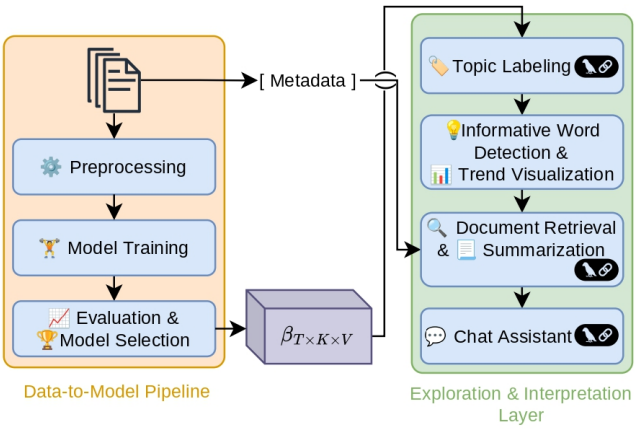

Figure 1: System architecture and workflow of DTECT.

Introduction

Understanding how topics evolve is crucial for interpreting large, temporally structured text corpora such as news archives, academic literature, and social media. The Dynamic Topic Model (DTM) (Blei and Lafferty 2006) extends Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) to model topic trajectories over time.

Subsequent advancements include the Dynamic Embedded Topic Model (DETM) (Dieng, Ruiz, and Blei 2019), the Chain-Free Dynamic Topic Model (CFDTM) (Wu et al. 2024), and evaluation metrics specific to dynamic topic models by Karakkaparambil James et al. (2024).

Several toolkits such as GENSIM (Řehůřek and Sojka 2010), OCTIS (Terragni et al. 2021), and TOPMOST (Wu, Pan, and Luu 2024) support topic modeling workflows.

Despite these advances, a unified end-to-end system for dynamic topic modeling is still missing. Semantic interpretability remains a key challenge (Doogan and Buntine 2021), as topics are typically unlabeled distributions over words. This makes it tedious to inspect and track topic shifts across time.

To address these limitations, we present **DTECT** (**D**ynamic **T**opic **E**xplorer and **C**ontext **T**racker), a unified, interactive system for dynamic topic modeling and interpretation. DTECT provides an end-to-end pipeline for preprocessing, modeling, and evaluation. It supports LLM-enhanced topic labeling to interpret evolving topics, a scoring-based method to surface temporally informative words and visualize their trajectories, and context-specific document retrieval via keyword–timestamp queries with document highlighting. In addition, DTECT enables LLM-based summarization of retrieved documents to explain temporal trends and offers a conversational interface for grounded, follow-up exploration based on retrieved evidence.

System Overview

As illustrated in Figure 1, DTECT comprises two main components: (1) *Data-to-Model Pipeline* and (2) *Exploration & Interpretation Layer*.

Data-to-Model Pipeline. Handles (a) *Preprocessing* of timestamped corpora; (b) *Training* with three dynamic topic models; and (c) *Evaluation & Selection* via temporal scores and runtime.

Exploration & Interpretation Layer. Provides (a) *Topic Labeling* with a chosen LLM using keyword trajectories; (b) *Trend Exploration* by scoring and visualizing temporally salient words; (c) *Retrieval & Summarization* of timestamped documents into concise highlights; and (d) a *Chat*

Assistant for natural language queries grounded in summaries.

Data-to-Model Pipeline

DTECT provides a flexible pipeline for processing timestamped corpora and fitting dynamic topic models. This section covers datasets, preprocessing, supported models, and evaluation.

Datasets and Preprocessing. We demonstrate DTECT on three diverse corpora—ACL Anthology (78,622 docs, 16 bins), UN Debates (7,507 docs, 46 bins), and TCPD-IPD Finance (20,361 docs, 21 bins). All are processed with a modular OCTIS-based script that produces tokenized texts, vocabulary, BoW matrices, and timestamp indices. Crucially, this standardized pipeline generalizes beyond these examples, *allowing any timestamped text dataset to be incorporated without manual intervention*.

Supported Models. DTECT supports three models: DTM (Blei and Lafferty 2006) with temporal dependencies, DETM (Dieng, Ruiz, and Blei 2019) using RNN-based evolution, and CFDTM (Wu et al. 2024) with contrastive learning. All outputs a temporal topic-word tensor $\beta \in \mathbb{R}^{T \times K \times V}$, the basis of DTECT's exploration layer.

Evaluation & Model Selection. We adopt three metrics from Karakkaparambil James et al. (2024): **Temporal Topic Coherence (TTC)** for semantic coherence; **Temporal Topic Smoothness (TTS)** for gradual evolution across timestamps; **Temporal Topic Quality (TTQ)**, a composite balancing both. These metrics guide training and model selection.

Exploration & Interpretation Layer

DTECT offers a Streamlit interface connecting topic model *outputs* with tools for labeling, trend detection, retrieval, summarization, and conversational analysis. It supports LangChain-compatible models such as OpenAI, Gemini, and Anthropic.

Topic Labeling. Since topic models yield unlabeled word distributions, DTECT forms *temporal keyword trajectories* from top-N words per timestamp and uses an LLM to generate concise labels, cached via SHA-256 to prevent redundant calls.

Temporally Informative Word Detection & Trend Visualization. Stable top words often miss evolving trends. DTECT highlights salient terms by scoring each word's probability trajectory within a topic using three criteria: (i) *burstiness*, the extent to which a word spikes at any time compared to its average; (ii) *specificity*, its prominence in one topic relative to others; and (iii) *uniqueness*, how rarely it appears among top words across topics. The combined score surfaces emerging words, whose temporal trajectories can then be visualized.

Document Retrieval and Summarization. Users can select any word-timestamp pair to retrieve relevant documents. Candidates are retrieved with FAISS and diversified with MMR. The final set is summarized into bullet points by an LLM.

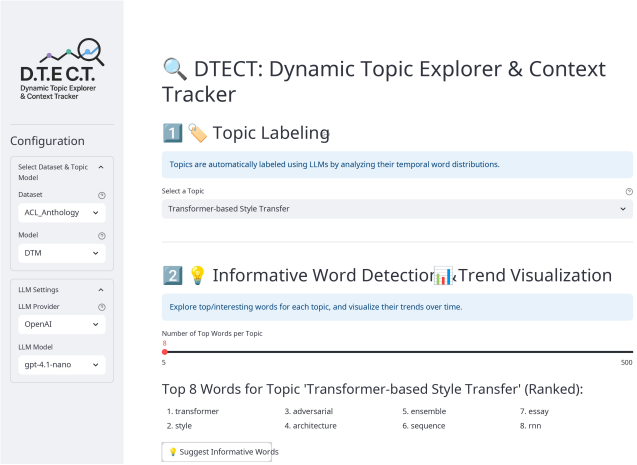

Figure 2: DTECT interface. After selecting the dataset, model, and LLM, users can explore topics by choosing from LLM-generated labels.

Conversational Exploration. DTECT offers a chat interface for follow-up queries, built on LangChain's `ConversationChain`. Responses are grounded in retrieved documents via a system prompt that enforces reliance on the provided content and transparency when information is missing.

User Interface

DTECT provides an interactive web interface, accessible locally or via https://huggingface.co/spaces/AdhyaSuman/DTECT, shown in Figure 2. The workflow begins with dataset preprocessing, followed by model training and evaluation. Once complete, users can interactively explore results by selecting the dataset, model, and LLM settings. DTECT automatically generates and caches inverted indices and topic labels to ensure efficient reuse.

Topic exploration proceeds by selecting a topic and inspecting its temporally salient words, visualized through interactive charts. Clicking a word at a given timestamp retrieves matching documents, which can be summarized or further explored via a document-grounded chat. All components run in real time for smooth, end-to-end exploration.

Conclusion

Navigating evolving themes in large text collections is challenging. We introduce DTECT, an interactive system that combines dynamic topic models for scalable analytics with LLMs for nuanced understanding. It offers a unified interface, allowing users to effortlessly explore automatically labeled topic trends and delve into the underlying documents. Features like one-click temporal summarization and a conversational assistant streamline analysis, transforming data into an intuitive, exploratory experience that reveals the temporal narratives within any large text corpus.

References

Blei, D. M.; and Lafferty, J. D. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 113–120. New York, NY, USA: Association for Computing Machinery. ISBN 1595933832.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan.): 993–1022.

Dieng, A. B.; Ruiz, F. J.; and Blei, D. M. 2019. The Dynamic Embedded Topic Model. *arXiv preprint arXiv:1907.05545*.

Doogan, C.; and Buntine, W. 2021. Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3824–3848. Online: Association for Computational Linguistics.

Karakkaparambil James, C.; Nagda, M.; Haji Ghassemi, N.; Kloft, M.; and Fellenz, S. 2024. Evaluating Dynamic Topic Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 160–176. Bangkok, Thailand: Association for Computational Linguistics.

Řehůřek, R.; and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Terragni, S.; Fersini, E.; Galuzzi, B. G.; Tropeano, P.; and Candelieri, A. 2021. OCTIS: Comparing and Optimizing Topic models is Simple! In Gkatzia, D.; and Seddah, D., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–270. Online: Association for Computational Linguistics.

Wu, X.; Dong, X.; Pan, L.; Nguyen, T.; and Luu, A. T. 2024. Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 3088–3105. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Wu, X.; Pan, F.; and Luu, A. T. 2024. Towards the TopMost: A Topic Modeling System Toolkit. In Cao, Y.; Feng, Y.; and Xiong, D., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 31–41. Bangkok, Thailand: Association for Computational Linguistics.