Incremental Mixture of Normalizing Flows for Dynamic Topic Modelling

Federico Ravenda[1,*], Andrea Raballo[2], Antonietta Mira[3] and Fabio Crestani[1]

[1]Faculty of Informatics, Università della Svizzera italiana, Lugano, Switzerland
[2]Faculty of Biomedical Sciences, Università della Svizzera italiana, Lugano, Switzerland
[3]Faculty of Economics, Università della Svizzera italiana, Lugano, Switzerland
[3]Department of Science and High Technology, Insubria University, Italy

Abstract

With the increasing availability of large-scale textual data in various domains, understanding the evolution of topics over time has become crucial for extracting meaningful insights and capturing the dynamic nature of information flow. In this work, we propose a model based on Mixtures of Normalizing Flows, able to extract topics from a collection of time-varying documents in a dynamic way. Our model takes as input embeddings generated by a pre-trained transformer-based language model for each timestamp, and learns the parameters of a complex high dimensional mixture density distribution, grouping documents into clusters, each representing a different topic. The parameters of the mixture distribution are updated at each timestamp and topic representations are generated using the TF-IDF procedure. A preliminary analysis shows that the topics generated using this procedure are competitive with those generated by state-of-the-art dynamic topic modelling.

Keywords

Topic Modelling, BERT, Temporal Clustering, Probabilistic Deep Learning

1. Introduction

Topic models are useful tools for uncovering hidden thematic patterns and structures within a collection of documents. Among the many models proposed, *Latent Dirichlet Allocation (LDA)* [1], stands out as a prevalent and widely-used topic model. This probabilistic model represents topics as distributions over words, while, simultaneously, characterizing each document as a mixture of topics. By employing *LDA*, we can effectively capture the underlying structure and relationships between words and topics, enabling a comprehensive understanding of the content and composition of documents within a corpus.

However, topics often change over time and it would be unrealistic to assume that documents from very different time instants would be generated by the exact same distribution. Thus, in this paper we aim to design an architecture that would enable the modeling of topics that rise and fall in popularity in time. Alternatively, topics that emerge and disappear, enabling the exploration of the relationship between topics and external temporal factors. In fact, unlike traditional static

IIR2023: 13th Italian Information Retrieval Workshop, June 8th - 9th, 2023, Pisa, Italy
*Corresponding author.
federico.ravenda@usi.ch (F. Ravenda); andrea.raballo@usi.ch (A. Raballo); antonietta.mira@usi.ch (A. Mira); fabio.crestani@usi.ch (F. Crestani)

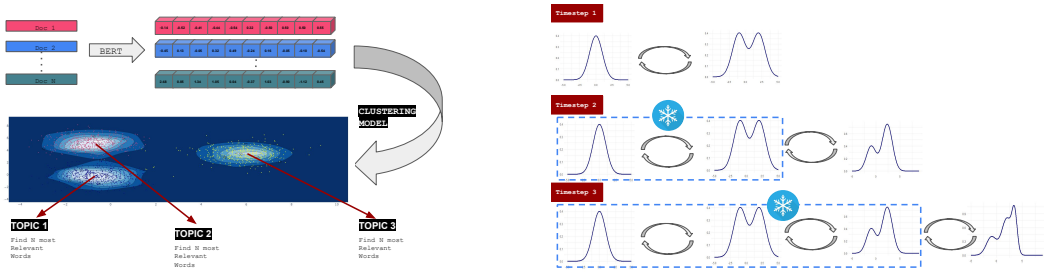

Figure 1: *Panel (left)* shows the working pipeline of a model like BERTopic, without considering the temporal aspect. Documents are transformed into embeddings, and clustering is performed. Each cluster represents a topic from which the N most representative words will be extracted.
Panel (right) shows how the Mixture of Normalizing Flows is updated at each considered time instant. A new layer is added, and new parameters must be learnt. On the other hand previous layers are frozen.[1]. The intuition is that if there are small topic changes, as expected from one time instant to the next one, the model benefits from the learning of the previous time distribution and has enough freedom to update the parameters with the new collection of documents and to find new insights in topics.

topic models, *Dynamic Topic Models (DTMs)* are designed to capture the temporal dynamics and changing patterns of topics within a document collection [2].

Different methodologies have emerged to streamline the creation of topics by leveraging word and document embeddings clustering [3, 4]. In particular BERTopic [5] is a *state-of-the-art* widely used topic modeling algorithm that utilizes *BERT* [6] embeddings and clustering techniques to identify and extract meaningful topics from text data. It combines the power of transformer-based language models and efficient clustering to provide a comprehensive approach to topic modeling.

In this work, a BERTopic-inspired approach is discussed and the pipeline of work is summarised in Fig. 1 and in depth explained in Section 2. Unlike *BERT* in which clustering of documents is carried out independently of their temporal nature, in the approach we propose the idea is to study how the mixture distributions evolves over time (i.e. how components of the mixture, which represent topics, change progressively) using time-varying collection of documents.

2. Methodology

This section describes briefly the method we propose for the extraction of dynamic topics from data.

2.1. Mixture of Normalizing Flows

Normalizing Flows (NFs) [7] have emerged as a prominent class of generative models, offering flexible and tractable approaches for modeling complex data distributions. In this work, we present a novel approach based on Mixture of Normalizing Flows [8, 9]. By combining the

benefits of mixture models and normalizing flows, based on preliminary analysis, the proposed approach enables effective modeling of multivariate data with complex dependencies.

Formally, a Mixture of Normalizing Flows represents a probability distribution as a mixture of transformed distributions, where each component in the mixture is associated with a particular latent variable. The latent variable determines the choice of the component in the mixture, and the transformed distribution captures the data distribution after applying a series of *bijective* transformations.

Consider a random variable $X \in \mathbb{R}^D$ representing the observed data, and a latent variable Z that follows a categorical distribution with K categories. Mixture of Normalising Flow is defined as:

$$P(X) = \sum_{k=1}^{K} P(Z = k)P(X|Z = k)$$

where $P(Z = k)$ represents the prior probability of latent variable Z taking the value k, and $P(X|Z = k)$ represents the likelihood of observed data X conditioned on the value of $Z = k$.

Each $P(X|Z = k)$ is modeled as a transformed distribution, typically using a normalizing flow. A normalizing flow is a series of invertible transformations applied to a base distribution, such that the resulting transformed distribution captures the complex data structure. For our implementation we chose a *Masked Autoregressive Flow (MAF)* [10] since it is well-suited for modeling data with complex conditional distributions. Specifically we work with the following mixture model:

$$p_X(x; parameters) = \pi_1 MAF_1(x; parameters_1) + \ldots + \pi_k MAF_k(x; parameters_k)$$

where the k components are fixed *a priori*. In practice, Mixture of Normalizing Flows are trained by maximizing the *log-likelihood* of the observed data using techniques such as maximum likelihood estimation or variational inference.

2.2. From Static to Dynamic Topic Modelling

The proposed model architecture consists of a mixture of normalizing flows, which is updated incrementally as new time steps as it is processed. At each time step, the documents in the collection are fed into the existing model to learn the distribution. The initial model is trained using the first time step, and subsequent time steps update the model incrementally. The scheme used to train our model and update the parameters of the distribution over time is described in Algorithms 1 and 2 below.

Algorithm 1 Training Process

1: *Initialization:* Initialize the model with a single layer representing the first time step. Each layer consists of a mixture of normalizing flows capturing specific topics.
2: *Training:* Train the initial layer using the documents from the first time step. Update the parameters of the mixture of normalizing flows using gradient-based optimization to maximize the likelihood of the observed data.
3: *Clustering:* Compute the likelihoods of the documents for each component in the mixture. Apply clustering techniques to identify the dominant topics for each time step based on the highest likelihood.
4: *Extracting Representative Words:* All documents belonging to the same cluster are merged into a single large document. Once the large documents have been created, a TF-IDF [11] matrix is calculated to represent the frequency of terms within the documents belonging to the different clusters.

Algorithm 2 Updating Scheme

1: **for** timestamp i in [timestamp 2, . . . , timestamp T] **do**
2: *Freezing and Adding Layers:* Freeze the weights of the initial layer to preserve the learned knowledge. Add a new layer to the model to capture the topics in the next time step.
3: *Training and Update:* Train the new layer using the documents from the next time step. Optimize the parameters of the mixture of normalizing flows in the new layer using the same training process as described earlier.

In summary, the proposed methodology involves training a mixture of normalizing flows for each time step, identifying dominant topics through clustering, extracting representative words, and incrementally updating the model by freezing previous layers and adding new ones. This iterative process allows to capture topics evolution and to analyse temporal dynamics in the document collections.

3. Analysis and Results

In order to assess the goodness of the model we propose, we want to compare the results obtained with those in BERTopic's *Table 3* paper with respect to the *Tweet's Trump* dataset (a collection of *44'253* tweets, excluding retweets, from 2009 to 2021) in which *BERTopic* and *LDA Sequence* (inspired by [12]) models are considered. *Topic Coherence* and *Diversity* metrics are used as measures of model's goodness.

- Topic Coherence [13] measures the semantic coherence or interpretability of topics. It quantifies how closely related the words within a topic are to each other. Higher values indicate stronger semantic relatedness among the words in a topic, reflecting a more coherent topic.
- Topic Diversity measures the distinctiveness or uniqueness of topics in a topic model. As discussed in [14], it measures the percentage of unique words for all topics.

The results we obtained are shown in Table 1. They are slightly worse than those obtained by BERTopic, but overperform LDA Sequence in both metrics. One of the reasons why metrics scores are lower than those of BERTopic could be that no fine-tuning operation is performed in the topic representations generated by our approach (in BERTopic different models are implemented e.g., GPT, KeyBertInspired, and Zero Shot Classifier).

	Topic Coherence	Topic Diversity
I-MoNF	0.072	0.818
LDA Sequence	0.009	0.715
BERTopic	0.079	0.863

Table 1
Topic Coherence and Topic Diversity scores were calculated on dynamic topic modelling tasks. The first row higlights performances of our implemented model (*Incremental-M*ixture *o*f *N*ormalizing *F*lows) (I-MoNF) while the following 2 rows represent the performances reported in paper [5] on the same dataset.

4. Discussion and Future Works

Our proposed model presents a temporal clustering approach based on mixtures of normalizing flows, whose parameters dynamically evolve over time. One significant advantage of this clustering model is its probabilistic nature, allowing us to infer the characteristics of topics not only from the generated qualitative representations but also from the parameters of the learned distribution over time. For instance, a change in the estimated mean could signify topic evolution, while a small variance might represent a cohesive cluster, that is a coherent topic. Moreover, normalizing flows are particularly suitable for high-dimensional data, without any intermediate dimensionality reduction step.

Even if preliminary results from our analysis demonstrate that the implemented model performs to the level of the state-of-the-art dynamic topic models, future developments are to be considered to be able to scale the model to larger collections of data and to improve the topic representation. More specifically:

- It is necessary to evaluate the model on datasets traditionally used to assess the quality of dynamic topic modeling models, to properly compare the model performance with recognised standards.
- In order to obtain more consistent topic representations, a sliding window approach would be useful.
- Considering the rise and fall of a new topic, think of adjusting the number of mixture components based on the parameters provided by the learned distribution at each iteration.
- Consider exploring alternative deep probabilistic models instead of normalizing flows, i.e., mixture of density networks, which are faster learning approaches.

All of these will be part of future work.

References

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[2] S. Bahrainian, I. Mele, F. Crestani, Modeling discrete dynamic topics, in: Proceedings of the ACM Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, April 3-7, 2017, 2017, pp. 858–865.

[3] S. Sia, A. Dalmia, S. J. Mielke, Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!, arXiv preprint arXiv:2004.14914 (2020).

[4] D. Angelov, Top2vec: Distributed representations of topics, arXiv preprint arXiv:2008.09470 (2020).

[5] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] D. Rezende, S. Mohamed, Variational inference with normalizing flows, in: International conference on machine learning, PMLR, 2015, pp. 1530–1538.

[8] G. G. Pires, M. A. Figueiredo, Variational mixture of normalizing flows, arXiv preprint arXiv:2009.00585 (2020).

[9] S. Ciobanu, Mixtures of normalizing flows, in: Proceedings of ISCA 34th International Conference on, volume 79, 2021, pp. 82–90.

[10] G. Papamakarios, T. Pavlakou, I. Murray, Masked autoregressive flow for density estimation, Advances in neural information processing systems 30 (2017).

[11] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, volume 242, Citeseer, 2003, pp. 29–48.

[12] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 113–120.

[13] G. Bouma, Normalized (pointwise) mutual information in collocation extraction, Proceedings of GSCL 30 (2009) 31–40.

[14] A. B. Dieng, F. J. Ruiz, D. M. Blei, Topic modeling in embedding spaces, Transactions of the Association for Computational Linguistics 8 (2020) 439–453.