A Review on Topic Modeling Techniques and Experimental Evaluation in Analysis of Touristic Experience

Maryam Kamal, Gianfranco Romani, Aris Anagnostopoulos and Ioannis Chatzigiannakis

Abstract—Topic Modeling is a well-adopted text mining strategy that discovers potential topics for documents that best characterize them. It identifies the semantic structures of the documents and based on document clusters it detects suitable words or phrases that may topicalize the documents. Topic Modeling has a distinct significance in a broad range of information systems such as social media, e-commerce and tourism sectors. This study presents a comprehensive review of selected prominent topic models specifically in the context of touristic experiences, including novel models in practice and recently devised strategies from literature. An experimental evaluation of the performance of considered models, using five well-established datasets (four related to the touristic experiences and one generic), highlights their advantages and unique characteristics based on multiple evaluation parameters. Further, the study discusses quantitative and qualitative findings along with conclusive deductions, open issues for the tourism context related application of topic models and future research directions.

Index Terms—Topic Modeling, Text Mining, Comparative analysis, Experimental evaluation, Touristic Experiences.

1 Introduction

The escalated adoption of web-applications, such as recommender system, social networks and QA systems, have accelerated the diversity and volumes of digital data exponentially in the recent years [1] [2] [3]. It has become evidently significant and challenging to accomplish intelligent tasks such as clustering, classification, sentiment analysis and delivery of online advertisements based on user interests [4]. Topic Modeling is a well-adopted data mining strategy that discovers potential topics for documents that best characterize them. It identifies the hidden semantics in the unstructured documents and based on document clusters it classifies and detects suitable words or phrases that are potential latent topics for the documents [5] [6].

Over the past few decades, the scientific community has endeavoured to come up with topic models that best fit the purpose for various domains of study serving diverse purposes [7]. Some interesting applications of topic modeling can be found in various fields including marketing and business management [8] [9] [10], analysis of scientific publications [11] [12], biology and medicine [13] [14] and software traceability [15]. Although topic modeling is being widely applied in many disciplines today, however, one of its interesting application is considered for tourism industry [16]. In recent years, the trend of personalized travel recommendations and automated content analysis of online posted travel offerings and reviews requires identification of topics for tourists' satisfaction and travel businesses [17] [18] [19] [20] [21]. This has made topic modeling one of the most in-demand techniques in the domain of tourism, where topics and labels are required to associate diverse preferences of tourists to related offerings by the travel business, considering the travellers' reviews and user-generated content (UGC) [22]. Such a context requires extraction of diverse and coherent tourists' interest topics, that are evaluated in this study through topic coherence and topic diversity parameters. Even though numerous valuable knowledge models have been designed to accomplish such machine learning tasks, however, the insufficiency of automation in ontology engineering leaves a gap for the field of tourism in this regard [23].

In this study, we aim to present a thorough and meticulous review on various promising topic modeling strategies along with their experimental evaluation in the context of touristic experience. The objectives of this study are as follows:

1) Discuss the preliminaries and important concepts related to topic modeling, in consideration of touristic experiences.
2) Present the systemic architecture, principles and working of each of the selected novel topic models suitable for touristic experiences.
3) Present the structure, working and optimized attributes of various topic modeling strategies devised from novel topic modeling strategies.
4) Experimentally explore the performance of topic modeling strategies based on multiple evaluation parameters for multiple benchmark and devised tourism datasets having different characteristics.
5) Analyze the performance of each of the topic model and identify the potential reasons for its performance in a particular way for a given condition.
6) Provide insights on about the strategy that performs

• *The authors are with the Department of Computer, Control and Management Engineering (DIAG), Sapienza University of Rome, 00185 Rome RM, Italy*
E-mail: kamal@diag.uniroma1.it, romani.1814407@studenti.uniroma1.it, aris@diag.uniroma1.it and ichatz@diag.uniroma1.it.

Manuscript received April 19, 2005; revised August 26, 2015.

better in the given context and discuss open issues in the application of topic models in the given context.

Since topic modeling has improvised with time, so the keen interest of this study is to review and explore only those strategies which have shown promising results in past and from which new promising strategies have been devised. We have categorized the selected topic modeling strategies into two categories. The first category covers the well-known novel approaches. These include Latent Dirichlet Allocation (LDA) [24], Top2Vec [25] and Non-Negative Matrix Factorization (NMF) [26]. The second category covers the strategies which are devised from the stated novel topic modeling strategies. These include Bidirectional Encoder Representations from Transformers (BERTopic) [27], RoBERTa [28], Contextualized Topic Model (CTM) [29], and Embedded Topic Model (ETM) [30].

Compared to the previous studies [31] [7] [32], our survey discusses relatively newer devised strategies along with in-practice novel strategies, goes deeper into the algorithms and provide fine-grained understanding of each associated concept. The study also demonstrates a detailed experimental exploration and evaluation of the strategies under consideration along with highlighting the potential reasons behind the particular performance trend of each of the strategies. Note that our study involve models evaluation based on significantly valuable four types of topic coherence parameters and two types of topic diversity parameters. While topic coherence reports for the interpretability of topics, the topic diversity measures how distinctive and varied are the topics produced. Both of these are crucial categories of evaluation for the context of our study, as it deals with diverse preferences of tourists related to touristic experiences that needs to be conveniently interpretable.

The rest of the paper is organized as follows. In Section 2, we have discussed some preliminaries and important concepts related to topic modeling along with a brief overview. In Section 3 we have presented a detailed survey on the selected novel topic modeling strategies followed by the survey on selected devised strategies. Section 3 mentions the experimental exploration of the strategies along with introduction of the datasets and evaluation parameters. Section 4 presents the results of the experimental evaluation followed by the discussion and analysis in Section 5. In Section 6, we present a conclusion. Section 7 presents open issues in the application of topic models in tourism context along with future research directions.

2 Definition of Terms and Notations

In this section, we provide definitions on terms, notations and basic concepts involved in topic modeling. Note that a text-based dataset is composed of a set of ***"documents" (D)*** which are strings of variable length composed of N words. Here a ***"word" (W) or "term" (T)*** is considered as the fundamental unit of a sample data. The set of distinct words present in a dataset forms the ***"vocabulary" (V)*** and a ***"topic"*** is then viewed as a probability distribution over this fixed vocabulary, it represents a label for a cluster of documents from a given dataset. Obviously, the way in which we represent words and documents has a great impact on topic modeling. Topic models traditionally work on vector representation of words and documents as input, known as ***"Word Embedding"*** and ***"Document Embedding"*** respectively. Here embeddings usually real-value vectors, representing words or documents in vector space such that similar words or documents appear closer to each other in spatial proximity. In this subsection, we refer to the classification of word embedding and representation techniques to establish a background knowledge related to topic models involved in this study. The classification presented is majorly based on the study by S. Selva Birunda and R. Kanniga Devi [33].

TABLE 1
Notation definition

Symbols	Meanings
D	Set of datasets
d	Single document
V	Vocabulary
W	Single word
T	Single term
BoW	Bag of words representation
TF-IDF	Term frequency - inverse document frequency
θ	Topic-document distribution
φ	Term-topic distribution

Category 1: *Traditional word embedding, or Count-based embedding [34].* This class comprises of methods that use frequency of words, co-occurrence of words and rarity of words for documents representation. A traditional representation of documents from this class is a ***"bag of words" (BoW)***. In BoW, each document is described by a vector of dimension equal to the vocabulary size, where each dimension represents the number of times a certain word appears in a document. However, such a text representation have limitations; the vectors tend to be very sparse, addition of new document having unknown vocabulary may cause technical difficulties or elevation in vector lengths, and the context is not considered. Another frequently adopted representation method from this class is ***"Term Frequency–Inverse Document Frequency" (TF-IDF)*** where *TF* measures how frequently a word appears in a document and *IDF* how much importance weight it carries. Note that *IDF* is introduced to suppress the weight of terms that occur very frequently in many documents, this also helps to magnify the weight of terms that occur rarely and are important. TF-IDF can be estimated using Eq. (1) as follows:

$$tfidf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \times \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{1}$$

where $f_{t,d}$ is the count of term t in the document d and D is the dataset. The i-th document is then represented as $d_i = [tfidf_{0,i}, ..., tfidf_{N,i}]$, where N is the number of words in the vocabulary V.

Category 2: *Static Word Embedding*. This category of representation involve prediction-based methods that compute probabilities of occurrence of the words and map those into fixed-size vectors. The embeddings produced by this category do not consider context, that is, a word embedding does not change if the word appear in sentences with different semantics. If two words often appear together, then embeddings generated for these are similar. This class of methods gained in popularity after the release of ***Word2Vec*** [35]. Word2Vec represents words into numeric vectors and also learn words association from the corpus. It may utilize either of its two architectures; ***(a) Continuous Bag-of-Words (CBOW)*** and ***(b) Skip-gram***. While CBOW predicts one target word from the surrounding context words, Skip-gram, on the other hand, uses one target word to predict surrounding context words. Word2Vec method has been used to design Doc2Vec [36], an algorithm that can create a numeric representation of a document, regardless of its length.

Category 3: *Contextualized Word Embedding*. Since context is considered in this class of methods, the word representation dynamically varies based upon the surrounding words. Methods that use this class of representation, such as Transformers based embeddings, are considered as state-of-the-art for most NLP tasks. These approaches are context-dependent, that is, these can disambiguate polysemes, thanks to the attention mechanism [37]. This means that methods from this category can compute different embeddings for a word depending on the context. There are plethora of representation methods based on contextual architecture. One of the well-known from this category is BERT [38] which has been used in several applications in NLP [39] [40] and with multiple variations [41]. An interesting variation of BERT used in topic models from our study is SBERT [42], that uses siamese and triplet network structures. Since most of the proposed transformer based architectures have a limit on the number of tokens to be processed at a time, document embeddings can be computed by dividing the text in chunks, finding the average of all the word embeddings in every chunk, and then averaging the chunks embeddings.

3 Recent Studies and Selected Models

In the recent years, many studies and researches have found topic models significantly helpful to cater active tourism related concerns. For instance, topic modeling is used to discover preferences in travel itineraries, to study customers opinions and to make recommendations. Since our study involves application of topic models in context of touristic experiences, we have summarized some recent relevant studies for topic modeling in tourism, in Table 2.

Although topic modeling finds initial roots in the 1980s [50], it gained prominence in 1990s due to appreciable performance recorded by topic models such as LSA [51], NMF [26] and, in particular way, LDA [24]. Over the past two decades, models such as LDA have been used to devise various other promising models such as [52], [53] and [13]. However, despite their success, conventional Bayesian probabilistic topic models started to show signs of fatigue and could not meet the expectations of big data handling in the era of big data and deep learning [54] [31]. Instead, models based on deep learning are attaining more popularity and appreciation. Deep learning based models are now applied for topic modeling, document representation [55], computing semantic representations of topics [56] and to deal with short texts [37] [57].

With the aim to comprehensively review and compare topic modeling approaches in the context of touristic experiences, we initially categorized the approaches into two categories, namely, "Novel Models in Practice" and "Recently Devised Strategies". The novel models in practice includes exclusively designed strategies which are not evolved or improvised from any other strategy. For this study we have considered LDA, Top2Vec and NMF as novel models in practice.

On the other hand, as per the category name suggests, the recently devised strategies includes the topic models that have been evolved or improvised from the novel models in practice. For this study, we have considered BERTopic, RoBERTa, CTM and ETM for this category. We have reviewed each of the above approaches in the following subsections as per their category.

3.1 Novel Models in practice

3.1.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [24] is a generative probabilistic model, designed for a given corpus of text documents. The model works on the De Finetti theorem and considers that K latent topics exists in the given N documents corpus, where a multinomial distribution represents each topic over the M words in the vocabulary extracted from the document corpus. It assumes a document consists of sampling variant proportional mixture of these topics and the topics samples various words representing those topics. Precisely, the algorithm in a nutshell is illustrated as follows:

1) For the ith document d in the document corpus D, (where $i = 1, 2, ..., N$), choose $\theta_i \sim$ Dirichlet(α).
2) For each word $w_{i,m}$ in the document d:
   a) Draw topic $z_{i,m} \sim$ multinomial(θ_i)
   b) Estimate topic distribution $\varphi_{z_{i,m}} \sim$ Dirichlet(β)
   c) Estimate word $w_{i,m} \sim$ multinomial ($\varphi_{z_{i,m}}$)

Here α and β are Dirichlet hyper-parameters. These are used to estimate probability of document corpus D as follows:

$$P(D \mid \alpha, \beta) = \prod_{i-1}^{N} \int P(\theta_i \mid \alpha) F(\theta, \varphi) \, d\theta_i \tag{2}$$

By maximizing the probability in Eq. (2), the model learns topic-document distribution θ and term-topic distribution φ, thus generating suitable topics for documents. The model considers following assumptions for its processing:

1) Each document is an unordered collected of words, namely bag-of-words (BOWs). This indicates that that model does not consider grammatical and contextual structure of the sentences.
2) Number of topics are pre-decided. This indicates that the model takes number of topics as input and assigns topics to documents accordingly. This may variate the for different number of topics.
3) The assignments of topics to documents and words to topics is random and the updates are iterative.

TABLE 2
Recent studies that use topic modeling in the tourism field

Studies using TM in Tourism field				
Study	Objectives	Model(s) Used	Datasets	Evaluation Metrics
Takeshi Kurashima et Al [43] (2013)	Locations recommendations	Geo Topic Model	Tabelog and Flickr-sourced geotag collection	5-best accuracy
Shuhui Jiang at Al [44] (2015)	Travel recommendations	Author Topic Collaborative Filtering	Geo-tagged photos from Flickr	MAP
Rossetti M. et AL [45] (2016)	Rating prediction and recommendation, suggest ratings for reviews and interpretation of users and items	LDA, Topic-Sentiment Criteria	TripAdvisor, Yelp	RMSE, two-sample Kolmogorov-Smirnov test
Calheiros A. at Al [46] (2017)	Sentiment Classification of Reviews	LDA	Custom dataset collected on-line	Several analysis on the topics obtained. No specific metri score
Yue Guo at Al [47] (2017)	Tourist satisfaction analysis	LDA	TripAdvisor	Jaccard coefficient, human analysis and Standford Topic Modelling Toolbox
Jie Bao at Al [48] (2017)	Bikesharing	LDA	Smart card data of a bike sharing system, Google Places API	Perplexity
Huy Quab Vu et Al [22] (2019)	Analysis of travel itineraries	LDA	Twitter, Foursquare	Perplexity, topic concentration
Nan Hu et Al [49] (2019)	Customers' complaints	STM	TripAdvisor	Several analysis on the topics obtained. No specific metric score

This assumes all topic assignments except the current word are correct.

3.1.2 Top2Vec

Top2Vec [25] is a relatively new topic model that uses word embeddings to discover latent semantic structure from the corpus of text documents. The model offers text data vectorization to identify semantically similar documents, words or sentences within joint embedding spatial proximity [58]. As word vectors that appear semantically nearest to the document vectors best describe the documents' topic, the number of documents clusters represents the number of topics, where each topic is represented by multiple closest words [32]. In short, it leverages joint document and word semantic embedding to find topic vectors.

The model claims for the following assumptions:

1) It considers joint document and word vectors, keeping the track of semantics rather than bag-of-words (BOW).
2) It automatically suggests the number of topics.
3) It does not require data pre-processing such as stopwords removal, lemmatization and stemming.

3.1.3 Non-Negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) [26] is a unsupervised learning model based on linear algebra that transforms the high-dimensional data into a reduced semantic space with non-negative hidden matrix structures. It works on the TF-IDF transformed data and decomposes the term-document matrix A, form of the original document matrix, into the product of two matrices W and H as denoted in Eq. (3):

$$A = WH \tag{3}$$

where W and H are non-negative matrices such as $W \geq 0$, and $H \geq 0$. Here W represents terms mapped to topics and H represents topics mapped to documents.

Eq. (4) shows the weighted sum of the components in matrix A is:

$$A_i = \sum_{j=1}^{k} W_{ij} * H_j \tag{4}$$

The values of W and H are updated iteratively as follows:

$$W \leftharpoonup W \frac{AH^T}{WHH^T} \tag{5}$$

$$H \leftharpoonup H \frac{W^T A}{W^T W H} \tag{6}$$

The model iterates the above Eq. (5) and (6) until it achieves convergence then it achieves final term–topic matrix W and topic–document matrix H for topics extraction. [59] [60] [61].

The model works on the following assumptions:

1) Considers original documents as matrix that is a inner product of two matrices, say W and H. Here W represents Documents-Topics matrix, while H represents Topics-Terms matrix.
2) Considers non-negative matrices values.
3) It requires pre-defining of number of topics as input
4) It requires data pre-processing such as stopwords removal, lemmatization, special characters removal and stemming.

3.2 Recently Devised Strategies

3.2.1 Bidirectional Encoder Representations from Transformers (BERTopic)

BERTopic [27] is a recent promising embedding based topic modeling approach that uses BERT embeddings and transformer embeddings. It is similar to Top2Vec regarding its algorithmic structure. BERTopic provides embedding extraction for the document corpus with a sentence-transformers model for more than 50 languages. Similarly to Top2Vec, BERTopic also offers dimensionality reduction using UMAP and then clusters the documents using HDBSCAN. However, unlike Top2Vec, it applies a variation of TF-IDF, Eq. (1), called class-based term frequency inverse document frequency (cTF-IDF), shown in Eq. (7). This variation efficiently evaluates the significance of terms within a cluster or class followed by the creation of term representation [62]. Here the higher score a term gets, the better it represents its topic [63].

$$cTF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_{j}^{n} t_j} \tag{7}$$

Where, t is the frequency of each word for each class i, w is the total number of words, and m is the total number of documents being divided by the total frequency of word t across all classes n.

BERTopic offers continuous instead of discrete topic modeling [64], that makes it different from other approaches. The model leads to different results with repeated execution due to its stochastic nature. The model offers the following features:

1) It does not require number of topics in advance. Estimates the number of topics automatically
2) It offers several multi-lingual models to extract document embeddings. Usually in practice it uses sentence-transformers package [13] with two default models; Distilbert for English and XLM-R for any other language. The XLM-R models support 50+ languages.
3) The approach mentions outliers in the result output as Topic 0 with the label of -1.

3.2.2 Robustly Optimized BERT Pre-training Approach (RoBERTa)

RoBERTa [28], is a devised strategy from BERT embedding model. It is, infact, a robustly optimized variant of BERT model. It is transformers based model that takes into consideration the context of a given word for its each occurrence. RoBERTa uses a dynamic version of BERT's masking strategy [65] [38], where the model learns to predict hidden sections and topics for the text documents and modifies key hyper-parameters of BERT. The model, like BERT, encodes substantial information about lexical semantics [66].

In comparison to BERT, RoBERTa is equipped with dynamic mask generation, full-sentences without Next Sentence Prediction (NSP) objective, larger batches and a larger byte-level byte pair encoding (BPE). It has been trained for longer and on larger number of datasets [67]. Although the original study of RoBERTa found it outperforming BERT and XLNet, however, it is interesting to observe how it performs in the context of touristic experiences, which is the scope of this study.

3.2.3 Contextualized Topic Model (CTM)

Contextualized Topic Models (CTMs) are devised from the Neural-ProdLDA variational autoencoding approach and pre-trained embedding models [29]. The two major categories of CTM include Combined Topic Model (CombinedTM) and Zero-Shot Topic Model (ZeroShortTM). CombinedTM uses contextual embeddings, SBERT, with the bag of words (BOW) to produce coherent topics. The framework trains a neural inference network that maps the BoW document representation into a continuous latent representation. Then, a decoder network reconstructs the BoW by generating its words from the latent document representation. A hidden layer represents documents with the same dimensions as the vocabulary size and the BOW representation.

On the other hand, ZeroShotTM [68] is a variation of CTM that works for missing words in data and also offers multilingual topic modeling (if trained with multi-lingual embeddings). It is a neural variational topic model that combines deep learning based topic models with embeddings techniques such as SBERT. Once the model is trained by reconstructing BOW from neural network, it can generate the representations of the documents and predict their topic distributions even for the unknown words in test data. Although CTMs are a promising addition, however, these have some constraints including the maximum of size of BOW (not to be more than 2000 elements), multi-lingual model not be trained on English data and pre-processing required to generate BOW.

3.2.4 Embedded Topic Model (ETM)

The embedded topic model (ETM) [69] is a generative topic model devised from LDA. It combines LDA with variational auto-encoder (VAE). The basic idea is to optimize and use LDA with word embeddings (word2vec). It produces word embedding similar to the CBOW word embeddings. However, ETM uses assigned topic vector instead of context vector. ETM offers two version, native ETM which learns its own topics and words embeddings and ETM SG that uses the pre-trained word embeddings.

ETM functions in a simple manner. It uses categorical distribution to model each word. The parameter for each modeled word is the inner product between a word embedding and its assigned topic embedding. The fitting of model uses amortized variational inference algorithm. The generative process ETM for d-th document can be summarized as follows, where $\mathcal{LN}$(.) represents the logistic normal distribution:

1) Draw topic proportions $\theta_d \sim \mathcal{LN}$(0,I).
2) For each word n in the document:
    a) Generate topic assignment $z_{dn} \sim$ Cat(θ_d).
    b) Generate $w_{dn} \sim$ softmax($\rho^T \alpha_{z_{dn}}$)

Note that the initial steps of the approach, 1 and 2a, are similar to traditional LDA. The difference can be found in step 2b, where the model uses vocabulary embedding ρ and assigned topic embedding $\alpha_{z_{dn}}$ to get the words from the topic z_{dn}.

4 Comparative Evaluation

In this section, we have reported the comparative evaluation of the considered novel topic models and devised topic models. The novel topic models, in this study, includes LDA, Top2Vec and NMF. On the other hand, the devised topic models include BERTopic, RoBERTa, CTM and ETM. The comparison is performed using 1 generic dataset and 4 touristic experience focused datasets, out of which 3 are exclusively generated for this study. The statistical summary of the datasets is mentioned in Table 3. The details of the experimental evaluations are mentioned in the following subsections.

TABLE 3
Statistics of the datasets

Dataset Labels	# of Docs	# of Words	Vocabulary Size	Avg. Words Per Doc
ATE	737	126,450	2,629	68
TAT	2,765	284,050	4,555	152
ET	5,724	1,556,416	138,095	272
TP	8,000	191,996	27,012	24
20NG	18,846	3,423,145	29,548	182

TABLE 4
Pre-processing done on each dataset

Models	Data Pre-processing			
	Stopwords Removal	Lemmat- -ization	Removal of Punctuations, Special Charc. Hastags, Emojis URLs, Numbers	Part of Speech
LDA	Yes	Yes	Yes	Nouns
Top2Vec	No	No	No	All
NMF	Yes	Yes	Yes	Nouns
BERTopic	No	No	No	All
RoBERTa	No	No	No	All
CTM	No	No	No	All
ETM	Yes	Yes	Yes	Nouns

4.1 Datasets

4.1.1 Benchmark Datasets

20NewsGroup (20NG) is a well-established generic benchmark dataset having more than 18000 newsgroup articles based on 20 different topics. The dataset is primarily in English language and is versatile to serve a split for training and testing data. It has been widely used to evaluate topic models in many studies such as [70] and [71].

TourPedia (TP) is a publicly available dataset related tourism places and reviews about those places. The places include accommodations, restaurants, points of interest, and attractions. The dataset contains more than 490,000 places and 577,000 reviews. It is based on 8 cities; Amsterdam, Barcelona, Berlin, Dubai, London, Paris, Rome and Tuscany. TourPedia was contributes by the project OpeNER, funded by the 7th Framework Program of the European Commision [72]. It has been used in many data analysis studies such as [73] and [74].

4.1.2 Touristic Experience Datasets

We have established three datasets, exclusively, for this study. These datasets are extracted from various web-based tourism platforms and contains data related to touristic experiences and products offered online. Since online tourism services are a growing market, where diverse-topics based online services are published on tourism platforms, it is interesting to analyze how these intelligent topic modeling strategies perform in context of online touristic experiences and products.

TripAdvisor Tourist Activities (TAT): We have devised a dataset from TripAdvisor which consists of data about all the tourist activities offered online for the region of Rome, Italy. The activities are extracted from the *"Things to do"* section of the website. The dataset contains 2765 entries where each entry contains text data related to 7 attributes, including an activity's title, description, popular mentions, price, duration, ratings and itinerary.

AirBnB Touristic Experiences (ATE): We have established a dataset from AirBnB which consists of data related to touristic experiences mentioned on the AirBnB website. The data is mined from the *"Experiences"* module of the web-portal for the region of Rome, Italy. This dataset is based on 737 records where each record is about a touristic experience published on AirBnB. Each record holds textual data related to 8 attributes; title, description, price, ratings, number of pictures, location, number of reviews, video availability.

EasyTour (ET): To analyze the multi-lingual aspect of the topic models, we have devised a unique dataset based on Italian Language. It has 5724 entries, each having 30 attributes such as id, document type, title, description, locations, duration, images, distance, publishing date and more. The dataset consists of data related to tourist services and POIs, for the Italian touristic experiences. The dataset is obtained from the beta testing phase of the app KuriU for the EasyTour project, which is in the development phase.

4.1.3 Data Pre-processing and Preparation

Data preprocessing is an important phase for many topic models [75]. Some topic models work on the principle of *"Garbage in garbage out"*, so it is significantly crucial to learn what a model feeds on. Suitably preprocessed data will get best out of a topic model while inappropriately preprocessed data may fail the performance of even a highly well-performing topic model. Hence in this subsection, we mention the categories of data pre-processing applied to the datasets for each model as per its requirements. Table 4 shows a summary of the data preparation steps for each technique.

Note the context of our study requires nouns as topics rather than adjectives or verbs. For instance, a topic such as "Museum" or "Cuisine" is a more insightful topic for touristic experience interests rather than a topic such as "Beautiful" or "Walking". Hence data is processed in such as way for the models which require pre-processing. Moreover, since some methods included in the study; Top2Vec, BERTopic, RoBERTa and CTM are recommended to be used without data preprocessing, hence no pre-processing is applied to datasets for these models.

For the purpose of experimentation, we considered English language documents for 4 out of 5 datasets. Hence from the devised datasets, AirBnB Touristic Experiences (ATE), we considered 611 documents that are in English language, and from TripAdvisor Tourist Activities (TAT) we considered 1860 documents that are in English language. To analyze the behavior of models on multi-lingual aspect, all 5724 documents from Italian Language dataset, EasyTour (ET), are considered. While all documents are considered from the benchmark datasets, that is, 18,846 documents from 20 Newsgroup (20NG) and 8,000 documents from Tourpedia (TP). We have considered the text description of all the documents for the purpose of analyzing the topic models.

4.2 Evaluation Parameters

4.2.1 Topic Diversity

Topic diversity (TD): It is a significantly impactful evaluation parameter to assess the topics produced by a topic model [76]. It measures the distinctiveness of the document clusters produced by the models. Topic diversity has been used in multiple studies to support the evaluation such as [77] and [78]. It simply estimates the percentage of constituent unique words in given K top words for all topics. The value of topic diversity usually ranges between 0 and 1, where a value close to 1 means higher topic diversity while a value closer to 0 means a lower topic diversity. A model is appreciated if it produces higher topic diversity for a given dataset.

$$TD = \frac{n(U)}{K * n(T)} \tag{8}$$

Here, in Eq. (8), n(U) represents the cardinality of the set of unique words U. K represents the top K words for all topics. T represents the set of topics generated by the model where n(T) is the cardinality of the set T.

Inverted RBO (IRBO): Another interesting parameter used to evaluate the diversity of topics is Inverted RBO (IRBO). It is a recently introduced metric [29] that has already been used in several works to estimate the quality of topics such as [79] [57] and [80]. It illustrates to what extent topics differ from each other [81]. It ranges from 0 to 1, where 0 means fully identical and 1 means fully diverse topics. It uses Ranked-Based Overlap measure [82] and compute the how disjoint are topics based on word-ranking for top K words. We decided to use this metric because, differently from the standard topic diversity measure, it penalizes topics with common words at different rankings less than topics sharing the same words at the highest ranks [83].

4.2.2 Topic coherence

Topic Coherence measures the interpretablity and coherence of the topics produced by a model and its association with the considered data [84] [85]. The idea is based on distributional hypothesis of linguistics. Unlike perplexity and predictive likelihood, which can be contrary to experts judgment [86], the versions of topic coherence we are using are considered as the best approximation for human ratings [84] and have been practiced in many studies such as [87], [88] and [89]. Note that a higher value of topic coherence represents better results of a topic model in terms of producing coherent topics. We have used the following variants of the topic coherence, for the purpose of evaluation, for given N top words of a topic, $P(w_i, w_j)$ refers to the probability of occurrence of words w_i and w_j together, while $P(w_i)$ and $P(w_j)$ is the probability of occurrence of these words individually. The details of these measures can be referred from [90].

1) C_{uci} uses sliding window and the pointwise mutual information (PMI) of all word pairs for top words as shown in Eq. (9).

$$C_{uci} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \log \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)} \tag{9}$$

2) C_v uses sliding window, top words' one-set segmentation with an indirect confirmation measure, using cosine similarity with normalized pointwise mutual information (NPMI) using the following set of equations:

$$\vec{v}(W') = \left\{ \sum_{w_i \epsilon W'} \left(\frac{log \frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log(P(w_i, w_j)+\epsilon)} \right)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{10}$$

$$\Phi_{s_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i . w_i}{\|\vec{u}\|_2 . \|\vec{w}\|_2} \tag{11}$$

In Eq. (10) the context vector $\vec{v}(W')$, uses NPMI for all the word pairs. γ places more weight on larger NPMI values. In Eq. (11) Φ is the confirmation measure that measures the vector cosine similarity of all the context vectors

3) C_{umass} uses count of document co-occurrences, one-preceding segmentation and confirmation measure (logarithmic conditional probability), following the computation from Eq. (12).

$$C_{umass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, wj) + \epsilon}{P(wj)} \tag{12}$$

4) C_{npmi} is an improvisation of the C_{uci} coherence that uses normalized pointwise mutual information (NPMI).

4.3 Experiment and Results

In this subsection, we have illustrated the results obtained through the conducted experimental exploration. The implementations are conducted using Python version 3.9.7 on Jupyter Notebook and re-implemented on Google Colab for cross validation. The coherence evaluation parameters are estimated using Gensim toolkit. while topic diversity measures are estimated using Octis toolkit. Each model is experimented with ten iterative runs and the results mentioned in this section are average recorded for each experiment.

The workstation is equipped with Intel(R) Core(TM) i5-10210U CPU@1.60GHz, 2.11 GHz, 20GB RAM. For the experimentation we have used the default text embedding models for each strategy, which are; *Doc2Vec* for Top2Vec, *roberta-base-nli-stsb-mean-tokens* for RoBERTa and *all-MiniLM-L6-v2* for BERTopic (English datasets) while *paraphrase-multilingual-MiniLM-L12-v2* for Italian language dataset. Note that we pre-defined the number of topics for LDA, NMF , CTM and ETM using elbow method as used in [91] and [92], while Top2Vec, BERTopic and RoBERTa are modeled to decided best suitable number of topic by themselves.

4.3.1 Topic Diversity:

An interesting quality determinant explored in this study is topic diversity. A model is well-appreciated if it estimates higher topic diversity with a suitable number of topics. Figure 1 shows the results obtained in this regard, where Figure 1a illustrates a comparison of the models with respect to average topic diversity (TD) and Figure 1b shows average Inverted RBO (IRBO) achieved for each dataset. Here Top2Vec shows higher topic diversity on average, for both cases, considering all datasets. An interesting finding is for TP dataset from Figure 1a, which illustrates a reduced variation of topic diversity among models and BERTopic as best method. Similarly it is interesting to observe from Figure 1b that BERTopic and RoBERTa shows much less IRBO when applied to small-sized dataset with shorter document lengths like ATE. Note that although Top2Vec provides higher topic diversity on average, the number of clusters (topics) it has produced is also considerably less for almost each dataset (Figure 1a). This might also indicate a high diversity within a topic cluster which is expected to be less for a good topic model.

4.3.2 Topic Coherence:

Further, we analyzed the coherence parameters (C_{uci}, C_v, C_{umass} and C_{npmi}) to determine the semantic coherence of the topics generated by each model for the datasets under consideration. Figure 2 depicts a comparative analysis of all the models for the given datasets for each coherence parameter. Note that the higher the coherence score, the better coherent are the topics, except for C_{umass}, where a lower value represents better coherence, according to Gensim implementaton [93].

Notice from Figure 2a that NMF shows better C_{uci} for comparatively smaller sized datasets as ATE and TAT, but when the size of the datasets grows, ETM starts depicting better results. On average ETM concludes to delivers maximum coherence as compared to the others, in terms of C_{uci}. Considering the C_v coherence measure from Figure 2b, while NMF shows better coherence on average for 3 out of 5 datasets, its performance are worse when it is used on largest dataset, 20NG, where, instead, Top2Vec exhibits better C_v than others. This may imply a sensibility of NMF to the datasets sizes, where this model seems suitable for small to medium sized datasets when considering C_v. An interesting observation can be made from Figure 2c for C_{umass}, where LDA outperforms others on average, while Top2Vec shows better performance for the Italian language dataset ET. Note that although LDA shows better C_{umass} on average, BERTopic outperforms all in case of the medium-sized English dataset TP. This implies adoption of LDA for small and large sized English dataset when considering C_{umass} coherence. Top2Vec might be applied if dealing with multi-lingual medium sized dataset while BERTopic is suggested for medium-sized English dataset when C_{umass} is concerned. Another interesting result can be visualized from Figure 2d, where NMF illustrates better C_{npmi} in almost every dataset (except for TP) and on average as a whole. Notice that for TP, ETM outperforms all in terms of C_{npmi}. Also for 20NG, ETM and NMF delivers same readings. Hence we can state that NMF performs better for small to medium sized datasets, while ETM performs better for medium to large sized datasets if C_{npmi} is concerned. Considering the c_v as the closest coherence measure to human judgement [87] [90], we can state that NMF produces more human interpretable topics as compared to others. However, the diverse shortcoming points to insightful implicit findings of the study that the coherence of topic models are significantly influenced by the type and size of the datasets along with number of topics the model uses. This behavior can be observed from Tables 5-9, where results are mentioned in detail.

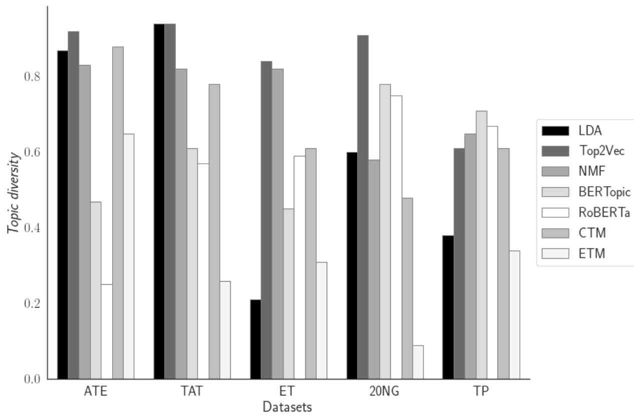
4.4 Validation of Analysis

In this subsection, we aim to validate the findings of the study by relating to behaviors of models from previous studies or providing rationale for an unexpected behavior.

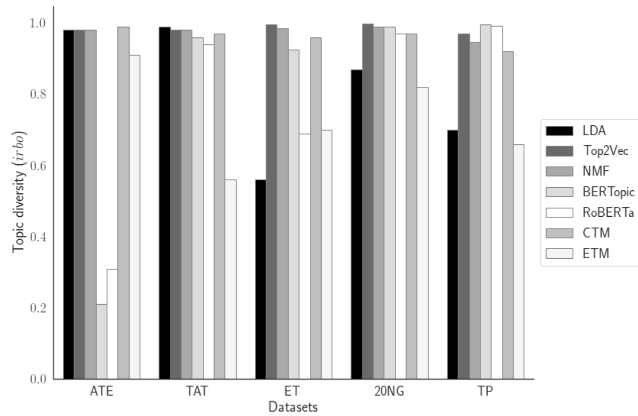
The shortcomings of the study reveals that Top2Vec generates topics with better diversity for majority of the datasets under consideration. This has been found for both parameters, topic diversity and IRBO. Such behavior for Top2Vec generating better topic diversity has been found by [94] [95] and [96]. At this point it is important to justify the use of Doc2Vec embedding for Top2Vec instead of other variants. Note that we conducted a sub-analysis among the other embedding variants for Top2Vec and found Doc2Vec performing better than the others on average for our datasets. We compared Doc2Vec, universal-sentence-encoder-multilingual and distiluse-base-multilingual-cased for two variants of documents; chunked and not chunked, to analyze the impact of length of documents also. Figure 3 show a partial visualization of results for C_v and C_{npmi} obtained for ET dataset. Since ET is a unique Italian language dataset, multilingual settings has been used for it.

For the C_{uci} parameter, that measures point-wise mutual information, we observed ETM depicts better results for majority of the datasets for our study. The appreciable results by ETM for C_{uci} can also be found in studies like [97] and [98]. Note that as mentioned earlier, ETM is a devised strategy from LDA with Word2Vec improvement. As LDA is already a well-established strategy delivering considerable C_{uci} coherence [99], an improved version of it is expected to perform even better.

Considering the mean value of C_v coherence parameter for all topic models, NMF shows significantly better results. Such a behavior of NMF has been supported by multiple studies such as [59] and [100]. NMF outperform others soley for 3 out of 5 dataset, ATE, TAT and ET, while for TP it preceded with a marginal variation in readings. The interesting

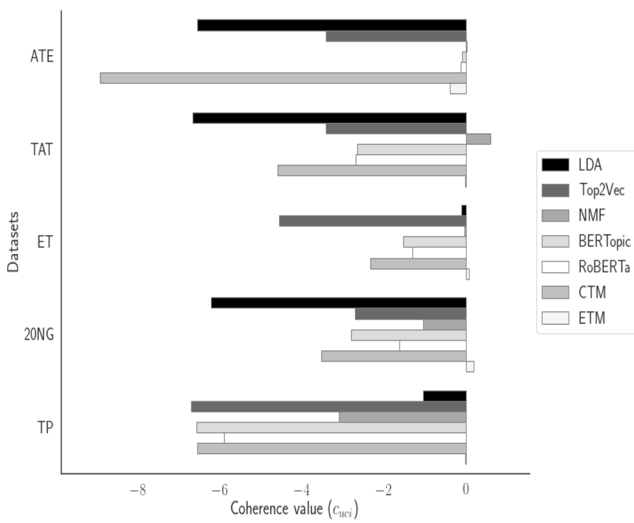


(a) *Comparison of the methods using topic diversity*

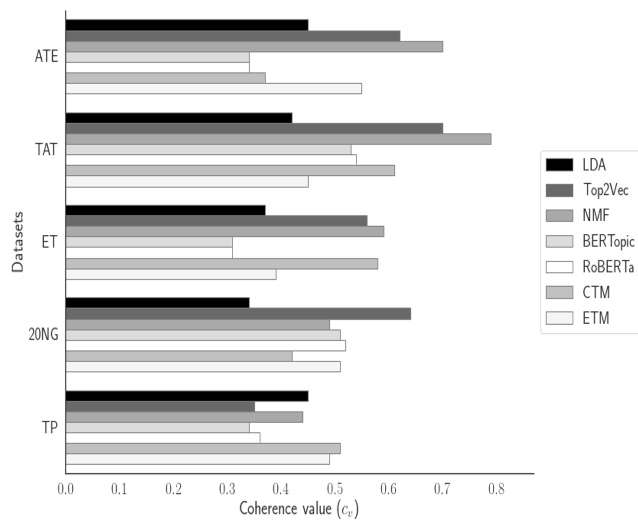


(b) *Comparison of the methods using IRBO score*

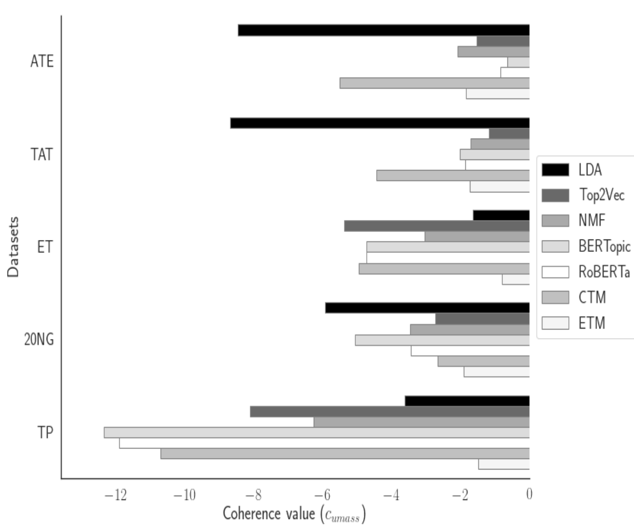
Fig. 1. Topic modeling evaluation based on Diversity metrics



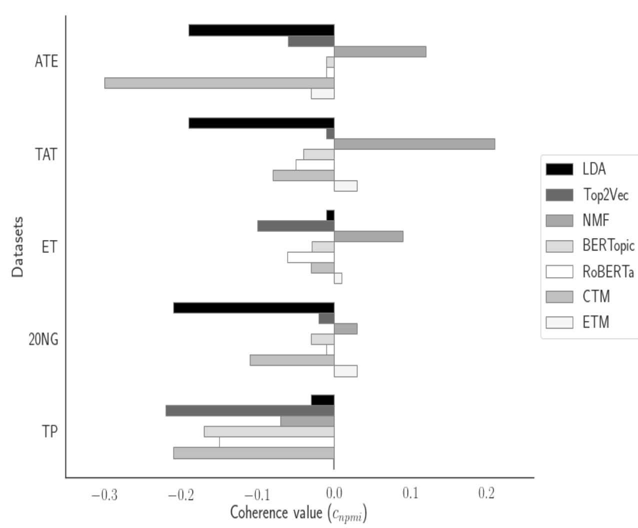
(a) *Comparison of the methods using C_{uci} score*



(b) *Comparison of the methods using C_v score*

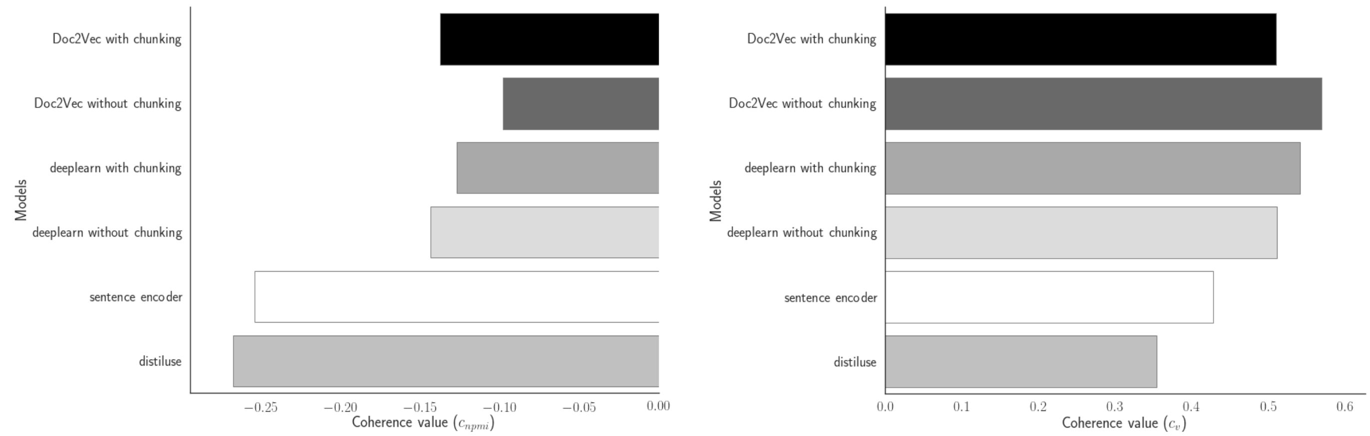


(c) *Comparison of the methods using C_{umass} score*



(d) *Comparison of the methods using C_{npmi} score*

Fig. 2. Topic modeling evaluation based on Coherence metrics



(a) *Average C_{npmi} score*

(b) *Average C_v score*

Fig. 3. Comparisons of variations for Top2Vec

observation can be made for 20NG dataset, where NMF was outperformed by others with a considerable variation. As the size of 20NG dataset makes it different from others, we can relate that NMF may not be suitable for larger datasets, as also supported by [101] [102].

Further, we observed that C_{umass} is rather a different parameter where a lower value signifies better coherence [93]. Note that LDA outperforms others on average in this regard, as similarly found by [103]. The probable reason for this could be that LDA considers document as a mixture of topics and words, occurring together with considerable probability, while C_{umass} involves counting of co-document appearance [104] which are more likely to be supported by topics and words occurring together in a document, that is mixture of topics and words, hence the topics produced by LDA are likely to have better C_{umass} scores.

Finally the study finds that NMF delivers better C_{npmi} score for majority of the datasets compared to all other models. Since C_{npmi} uses normalized version of PMI score (known as NPMI) and C_v is also estimated based on NPMI score, along with cosine similarity. It is most-likely for a technique performing better for C_v to also perform better on C_{npmi}, which is observable in case of NMF for majority of the datasets.

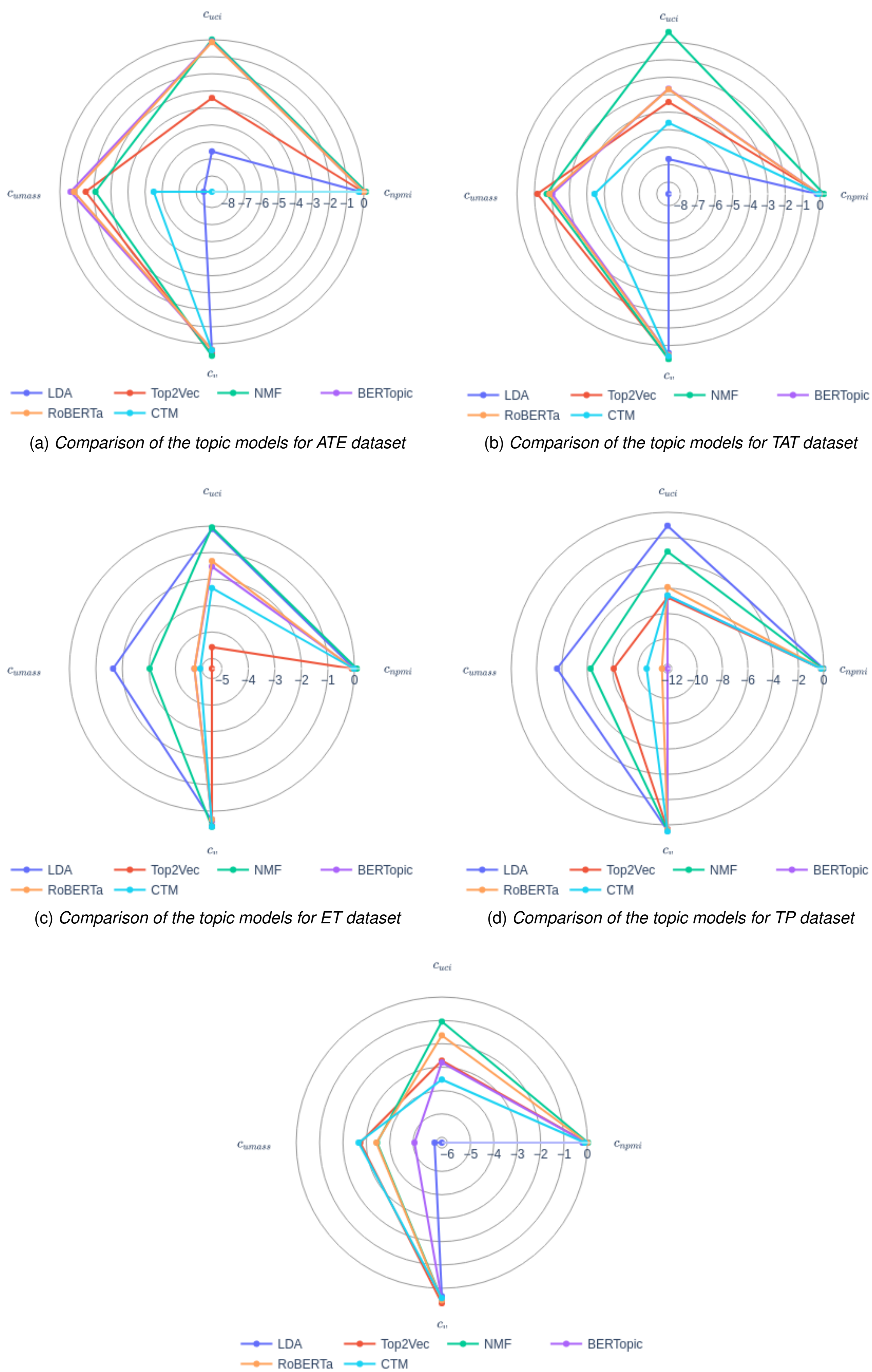
5 Discussion

This section presents a thorough discussion about the achieved findings. Notice that in addition to the quantitative results, this section also includes some implicit qualitative findings.

As the results obtained are diverse, our study does not indicate one model to be better than all others, rather it suggests suitability of models as per the size and type of dataset. Notice that LDA performs visibly better on average in case of the TAT dataset (Figure 4b), considering one coherence and both diversity parameters, C_{umass}, TD and IRBO, recall that TAT is a medium sized English dataset. While NMF performs better on average considering the coherence parameters, C_{uci}, C_v and C_{npmi} for TAT as well as for ATE, both being small to medium sized English datasets. Due to the visible difference obtained in coherence readings and a marginal difference in the diversity readings, we suggest that NMF outperforms LDA, which in turn performs better than all others for small to medium English datasets. Moreover, the qualitative implications find NMF to be faster, more consistent and producing more human-interpretable topics for ATE (Figure 4a) and TAT (Figure 4b) datasets as compared to others.

Furthermore, Figure 4c shows that Top2Vec outperforms others on average for ET dataset, that is a medium-sized Italian language dataset, in terms of C_{umass}, TD and IRBO. On the other hand, NMF performs appreciable for C_v and C_{npmi}. Hence, Top2Vec is suitable for multi-lingual medium sized dataset, followed by NMF which may be adopted if only coherence is under consideration. A similar behavior of Top2Vec can be observed in Figure 4e, which exhibits results for the 20NG dataset, that is large-size English dataset. Here Top2Vec illustrated better results for C_v, TD and IRBO, followed by NMF that outperforms others for C_{npmi} and IRBO parameters. Notice that although BERTopic is devised from Top2Vec architecture, still it marginally underperforms compared to Top2Vec in terms of stated parameters for 20NG dataset. While BERTopic exhibits considerable results for only TP dataset (Figure 4d), that is a medium sized English dataset for C_{umass}, TD and IRBO parameters. Notice that there is a marginal difference between BERTopic and RoBERTa for these parameters. The qualitative analysis found BERTopic to be much stochastic in nature for small to medium sized datasets, producing insufficient number of topics over multiple runs often illustrating the inclusion of stopwords in the topic words for short-lengthened documents. Hence we suggest Top2Vec suitable for large English datasets while RoBERTa may be suitable for medium sized English dataset instead of BERTopic due to its better stability, consistency and efficiency.

Another interesting implicit finding of this study indicates an observable relation between number of topics and topic diversity (TD). We noticed increased TD when the number of topics generated by a model are comparatively lesser. This may be because of the fact that lesser number of



(a) *Comparison of the topic models for ATE dataset*

(b) *Comparison of the topic models for TAT dataset*

(c) *Comparison of the topic models for ET dataset*

(d) *Comparison of the topic models for TP dataset*

(e) *Comparison of the topic models for 20NG dataset*

Fig. 4. Evaluation of the topic models based on the results on each dataset

TABLE 5
Comparisons of the results on ATE dataset.

Models	Coherence (C_{uci})	Coherence (C_v)	Coherence (C_umass)	Coherence (C_{npmi})	Topic Diversity (TD)	IRBO	Number of topics *
LDA	-6.56	0.45	**-8.45**	-0.19	0.87	0.98	14
Top2Vec	-3.42	0.62	-1.52	-0.06	**0.92**	0.98	6
NMF	**0.01**	**0.70**	-2.09	**0.12**	0.83	0.98	14
BERTopic	-0.10	0.34	-0.63	-0.01	0.47	0.21	3
RoBERTa	-0.14	0.34	-0.83	-0.01	0.25	0.31	10
CTM	-8.93	0.37	-5.51	-0.30	0.88	**0.99**	14
ETM	-0.40	0.55	-1.85	-0.03	0.65	0.91	14

TABLE 6
Comparisons of the results on TAT dataset.

Models	Coherence (C_{uci})	Coherence (C_v)	Coherence (C_{umass})	Coherence (C_{npmi})	Topic Diversity (TD)	IRBO	Number of topics *
LDA	-6.68	0.42	**-8.68**	-0.19	**0.94**	**0.99**	16
Top2Vec	-3.42	0.70	-1.17	-0.01	**0.94**	0.98	6
NMF	**0.59**	**0.79**	-1.70	**0.21**	0.82	0.98	16
BERTopic	-2.66	0.53	-2.02	-0.04	0.61	0.96	45
RoBERTa	-2.69	0.54	-1.86	-0.05	0.57	0.94	44
CTM	-4.61	0.61	-4.44	-0.08	0.78	0.97	16
ETM	-0.03	0.45	-1.72	0.03	0.26	0.56	16

TABLE 7
Comparisons of the results on ET dataset.

Datasets	Coherence (C_{uci})	Coherence (C_v)	Coherence (C_{umass})	Coherence (C_{npmi})	Topic Diversity (TD)	IRBO	Number of Topics *
LDA	-0.11	0.37	-1.65	-0.01	0.21	0.56	22
Top2Vec	-4.57	0.56	**-5.38**	-0.10	**0.84**	**0.99**	50
NMF	-0.05	**0.59**	-3.03	**0.09**	0.82	0.98	22
BERTopic	-1.53	0.31	-4.72	-0.03	0.45	0.93	75
RoBERTa	-1.32	0.31	-4.72	-0.06	0.59	0.69	14
CTM	-2.34	0.58	-4.94	-0.03	0.61	0.96	22
ETM	**0.06**	0.39	-0.79	0.01	0.31	0.70	22

TABLE 8
Comparisons of the results on TP dataset.

Datasets	Coherence (C_{uci})	Coherence (C_v)	Coherence (C_{umass})	Coherence (C_{npmi})	Topic Diversity	IRBO	Number of Topics *
LDA	-1.06	0.45	-3.62	-0.03	0.38	0.70	14
Top2Vec	-6.72	0.35	-8.10	-0.22	0.61	0.97	41
NMF	-3.10	0.44	-6.27	-0.07	0.65	0.95	14
BERTopic	-6.59	0.34	**-12.35**	-0.17	**0.71**	**0.99**	142
RoBERTa	-5.91	0.36	-11.91	-0.15	0.67	**0.99**	106
CTM	-6.57	**0.51**	-10.70	-0.21	0.61	0.92	14
ETM	**-0.03**	0.49	-1.49	**-0.01**	0.34	0.66	14

TABLE 9
Comparisons of the results on 20NG dataset.

Datasets	Coherence (C_{uci})	Coherence (C_v)	Coherence (C_{umass})	Coherence (C_{npmi})	Topic Diversity (TD)	IRBO	Number of Topics *
LDA	-6.23	0.34	**-5.92**	-0.21	0.60	0.87	111
Top2Vec	-2.72	**0.64**	-2.74	-0.02	**0.91**	**0.99**	83
NMF	-1.05	0.49	-3.46	**0.03**	0.58	**0.99**	111
BERTopic	-2.80	0.51	-5.06	-0.03	0.78	**0.99**	216
RoBERTa	-1.64	0.52	-3.43	-0.01	0.75	0.97	90
CTM	-3.53	0.42	-2.67	-0.11	0.48	0.97	111
ETM	**0.19**	0.51	-1.91	**0.03**	0.09	0.82	111

* The all the values mentioned in the table are rounded off average from 10 runs.

topics indicates lesser number of clusters, where if a dataset is clustered with comparatively smaller number, the chances of obtaining better inter-cluster distance are increased compared to when clusters are overlapping. Hence the more disjoint the topic clusters are, the increased is the TD. Further, the study finds transformer based models, Top2Vec, BERTopic and RoBERTa, stochastic in nature. This is because of the utilization of UMAP that produces variations in results for repetition of the same experiment [105]. However, out of these three models, Top2Vec shows comparatively lesser variation in results followed by RoBERTa, where BERTopic requires several iteration to produce stable results, in the case of our datasets which comprises of usually short-lengthened documents and are small to medium size. An inferred rationale for such BERTopic's behavior can be that a lesser number of documents in dataset might have resulted in not much distinguishable clusters formation, and since it uses cluster level TF-IDF ($cTF-IDF$), it may have resulted in same words in multiple topics (in case of overlapping clusters) or much lesser number of clusters (topics) which have degraded the performance overall. BERTopic often lacks the accurate identification of all the topics present in our medium-sized datasets, as also mentioned in [106].

6 Conclusion

Our study delineates a comprehensive review of promising novel and devised topic modeling strategies. These include LDA, NMF and Top2Vec as novel strategies while BERTopic, RoBERTa, CTM and ETM as devised strategies. Further, our study presents an in-detail experimental evaluation based comparative analysis of these models in touristic experiences context. The analysis is conducted based on topic coherence and topic diversity in terms of multiple significant parameters. We considered four topic coherence parameters: C_{uci}, C_v, C_{umass} and C_{npmi} along with two diversity parameters: Topic Diversity (TD) and Inverted RBO (IRBO). The experimental evaluations are conducted over five variant and contextually diverse datasets where four are related to touristic experiences, out of which three are exclusively designed for the purpose of this study. The study contributes significant conclusive quantitative results and reveals many valuable implicit deductions.

The diverse quantitative findings of the study implicitly reveal that there is no conclusive winner among the considered models and the performance and suitability of the models are correlated to the size and type of data. For this reason, we have concluded the suitability of the models as per the mentioned attributes of the datasets. From Table 5, we observed that for ATE, NMF performs better as compared to others for 3 out of 6 parameters, C_{uci}, C_v and C_{npmi}, followed by LDA, Top2Vec and CTM which performed better for 1 parameter each, C_{umass}, TD and IRBO respectively. Similarly Table 6 illustrates results for TAT where NMF performs better on 3 out of 6 parameters, C_{uci}, C_v and C_{npmi}. While LDA also shows better performance for 3 out 6 parameters, C_{umass}, TD and IRBO. Here LDA outperforms others majorly for diversity while NMF outperforms others majorly for coherence. Top2Vec produces equal TD as LDA for TAT and it is also delivers better TD for ATE. Hence we conclude that use of NMF is preferred for small to medium sized datasets where document length is moderately shorter on average for better coherence, while Top2Vec or LDA delivers better diversity in such cases.

Further, from Table 7, we conclude that on average Top2Vec outperforms others for medium sized dataset having multi-lingual documents. Since Top2Vec outperforms others for 3 out of 6 parameters, C_{umass}, TD and IRBO, followed by NMF that outperformed others for C_v and C_{npmi}, we suggest the suitability of Top2Vec for such cases if moderate coherence is preferred along with high diversity. Conversely, NMF is preferred if good coherence is required irrespective of high diversity. Moreover, from Table 9 we conclude that Top2Vec performs better on average for large sized English datasets as it delivers better results for 3 out of 6 parameters C_v, TD and IRBO. Although Table 8 reveals that BERTopic outperforms others quantitatively for majority parameters (C_{umass}, TD and IRBO) for medium-sized datasets, however RoBERTa exhibits considerably better qualitative aspects than BERTopic for such datasets with marginal difference in readings in terms of C_{umass}, TD and IRBO. Hence, we suggest the use of Top2Vec for large sized English datasets and RoBERTa for medium-sized English datasets. In both cases ETM may also be used if only the coherence parameter is of concern, since it delivers better coherence for both cases in terms of C_{uci} and C_{npmi}.

7 Open Issues and Future Research Directions

The diverse domain of touristic experiences comprises of heterogeneous related issues in regard of topic modeling as reported by several studies including [107] [18] and [108]. Firstly, the context of tourism lacks a comparative standard for datasets compared to other fields [109]. Secondly, new topic modeling approaches based on deep learning needs large and often labelled data, which are often not available for this field. Furthermore, the text or documents describing touristic experiences, tourism products or tourist reviews are often particularly short in length, which can be challenging for topic models. Although there exist promising attempts to cater this concern [110] [111], still the issue persists.

Moreover, as neural network based topic models are often stochastic black boxes, their use may lead to a loss of interpretability of the results or unexpected behavior for different iterative runs, as we experienced in our study. Another important issue in the context of touristic experience is unavailability of versatile and diverse public or benchmark datasets, which can be used to establish a judgement for topic models. This current limitation can be one of the possible future directions of this study, where other methods can help alleviate this problem, such as the usage of knowledge graphs [112] [113] and transfer learning [114] for the approaches based on deep learning and the usage of side information [115] or multimodal data [116].

Another interesting future direction in this particular context of study is the consideration of connection between data consulted by the tourists and the period in which such content is consulted. Here the continuation of our work can consider the dynamic aspect of the data to detect which topics are important in a determined period of time and forecast the topics potentially important for similar future events.

Acknowledgments

The authors are grateful to the EasyTour Project established by Amarena SRL.

References

[1] Feng Yi, Bo Jiang, and Jianjun Wu. Topic modeling for short texts via word embedding and document correlation. *IEEE Access*, 8:30692–30705, 2020.

[2] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568, 2022.

[3] Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas, and Grammati Pantziou. Mobile recommender systems in tourism. *Journal of Network and Computer Applications*, 39:319–333, 2014.

[4] Chen Yong, Zhang Hui, Liu Rui, Ye Zhiwen, and Lin Jianying. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019.

[5] Damir Korenčić, Strahil Ristov, Jelena Repar, and Jan Šnajder. A topic coverage approach to evaluation of topic models. *IEEE Access*, 9:123280–123312, 2021.

[6] Peng Zhang, Suge Wang, Deyu Li, Xiaoli Li, and Zhikang Xu. Combine topic modeling with semantic embedding: Embedding enhanced topic model. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2322–2335, 2020.

[7] H. Jelodar, Y. Wang, and C. et al. Yuan. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimed Tools Appl 78*, page 15169–15211, 2019.

[8] Mekhail Mustak, Joni Salminen, Loïc Plé, and Jochen Wirtz. Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, 124:389–404, 2021.

[9] M Reisenbichler and T. Reutterer. Topic modeling in marketing: recent advances and research opportunities. *J Bus Econ*, page 327–356, 2019.

[10] Nicolas Pröllochs and Stefan Feuerriegel. Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management*, 57(1):103070, 2020. Big data and business analytics: A research agenda for realizing business value.

[11] Andreas Älgå, Oskar Eriksson, and Martin Nordberg. Analysis of scientific publications during the early phase of the covid-19 pandemic: Topic modeling study. *J Med Internet Res*, 22(11):e21559, Nov 2020.

[12] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 448–456, New York, NY, USA, 2011. Association for Computing Machinery.

[13] L Liu, L Tang, W Dong, S Yao, and W Zhou. An overview of topic modeling and its current applications in bioinformatics. *Springerplus*, 2016.

[14] B. Zheng, D.C. McLean, and X. Lu. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, 2006.

[15] Hazeline U. Asuncion, Arthur U. Asuncion, and Richard N. Taylor. Software traceability with topic modeling. In *2010 ACM/IEEE 32nd International Conference on Software Engineering*, volume 1, pages 95–104, 2010.

[16] Wafa Shafqat and Yung-Cheol Byun. A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis. *Sustainability*, 12(1), 2020.

[17] Chao Huang, Qing Wang, Donghui Yang, and Feifei Xu. Topic mining of tourist attractions based on a seasonal context aware lda model. *Intelligent Data Analysis*, 22:383–405, 03 2018.

[18] P. Andrei Kirilenko, Svetlana O. Stepchenkova, and Dai Xiangyi. Automated topic modeling of tourist reviews: Does the anna karenina principle apply? *Tourism Management*, 83:104241, 2021.

[19] Lorena Bourg, Thomas Chatzidimitris, Ioannis Chatzigiannakis, Damianos Gavalas, Kalliopi Giannakopoulou, Vlasios Kasapakis, Charalampos Konstantopoulos, Damianos Kypriadis, Grammati Pantziou, and Christos Zaroliagis. Enhancing shopping experiences in smart retailing. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–19, 2021.

[20] Ioannis Chatzigiannakis, Georgios Mylonas, and Andrea Vitaletti. Urban pervasive applications: Challenges, scenarios and case studies. *Computer Science Review*, 5(1):103–118, 2011.

[21] Thomas Chatzidimitris, Damianos Gavalas, Vlasios Kasapakis, Charalampos Konstantopoulos, Damianos Kypriadis, Grammati Pantziou, and Christos Zaroliagis. A location history-aware recommender system for smart retail environments. *Personal and Ubiquitous Computing*, 24(5):683 – 694, 2020. Cited by: 11.

[22] Huy Quan Vu, Gang Li, and Rob Law. Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management*, 75:435–446, 2019.

[23] Valentinus R. Hananto, Uwe Serdült, and Victor Kryssanov. A tourism knowledge model through topic modeling from online reviews. In *2021 7th International Conference on Computing and Data Engineering*, ICCDE 2021, page 87–93, New York, NY, USA, 2021. Association for Computing Machinery.

[24] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. volume 3, pages 601–608, 01 2001.

[25] Dimo Angelov. Top2vec: Distributed representations of topics. 2020.

[26] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[27] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[29] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *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)*, pages 759–766, Online, 2021. Association for Computational Linguistics.

[30] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[31] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. 02 2021.

[32] Egger R and Yu J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front Sociol*, 2022.

[33] S. Selva Birunda and R. Kanniga Devi. A review on word embedding techniques for text classification. In Jennifer S. Raj, Abdullah M. Iliyasu, Robert Bestak, and Zubair A. Baig, editors, *Innovative Data Communication Technologies and Application*, pages 267–281, Singapore, 2021. Springer Singapore.

[34] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *ArXiv*, abs/1901.09069, 2019.

[35] KENNETH WARD CHURCH. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[36] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org, 2014.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[39] M. V. Koroteev. Bert: A review of applications in natural language processing and understanding, 2021.

[40] Patrick Xia, Shijie Wu, and Benjamin Van Durme. Which *bert? a survey organizing contextualized encoders, 2020.

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics.

[42] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[43] Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya, and Ko Fujimura. Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 375–384, New York, NY, USA, 2013. Association for Computing Machinery.

[44] Shuhui Jiang, Xueming Qian, Jialie Shen, and Tao Mei. Travel recommendation via author topic model based collaborative filtering. In Xiangjian He, Suhuai Luo, Dacheng Tao, Changsheng Xu, Jie Yang, and Muhammad Abul Hasan, editors, *MultiMedia Modeling*, pages 392–402, Cham, 2015. Springer International Publishing.

[45] M. Rossetti, Stella, and M. F. & Zanker. Analyzing user reviews in tourism with topic models. *Inf Technol Tourism*, page 5–21, 2016.

[46] Ana Catarina Calheiros, Sérgio Moro, and Paulo Rita. Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing and Management*, 26(7):675–693, 2017.

[47] Yue Guo, Stuart J. Barnes, and Qiong Jia. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59(C):467–483, 2017.

[48] J. Bao, C. Xu, and P. et al. Liu. Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. *Netw Spat Econ*, 17:1231–1253, 2017.

[49] Nan Hu, Ting Zhang, Baojun Gao, and Indranil Bose. What do hotel customers complain about? text analysis using structural topic model. *Tourism Management*, 72:417–426, 2019.

[50] Ike Vayansky and Sathish A.P Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[51] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[52] Daniel Backenroth, Zihuai He, Krzysztof Kiryluk, Valentina Boeva, Lynn Petukhova, Ekta Khurana, Angela Christiano, Joseph D. Buxbaum, and Iuliana Ionita-Laza. Fun-lda: A latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: Methods and applications. *The American Journal of Human Genetics*, 102(5):920–942, 2018.

[53] Dongjin Yu, Dengwei Xu, Dongjing Wang, and Zhiyong Ni. Hierarchical topic modeling of twitter data for online analytical processing. *IEEE Access*, 7:12373–12385, 2019.

[54] Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International conference on machine learning*, pages 190–198. PMLR, 2014.

[55] Wenyue Zhang, Yang Li, and Suge Wang. Learning document representation via topic-enhanced lstm model. *Knowledge-Based Systems*, 174:194–204, 2019.

[56] Rui Wang, Deyu Zhou, and Yulan He. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098, 2019.

[57] Riki Murakami and Basabi Chakraborty. Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts. *Sensors*, 22(3), 2022.

[58] Roman Egger. *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*. Springer Nature, 2022.

[59] Shini George and Srividhya Vasudevan. Comparison of lda and nmf topic modeling techniques for restaurant reviews. 03 2021.

[60] Pooja Kherwa and Poonam Bansal. Topic modeling: A comprehensive review. *ICST Transactions on Scalable Information Systems*, 7:159623, 07 2018.

[61] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

[62] Abeer Abuzayed and Hend Al-Khalifa. Bert for arabic topic modeling: An experimental study on bertopic technique. *Procedia Computer Science*, 189:191–194, 2021. AI in Computational Linguistics.

[63] Manuel J Sánchez-Franco and Manuel Rey-Moreno. Do travelers' reviews depend on the destination? an analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing*, 39(2):441–459, 2022.

[64] Alexandre Alcoforado, Thomas Palmeira Ferraz, Rodrigo Gerber, Enzo Bustos, André Seidel Oliveira, Bruno Miguel Veloso, Fabio Levy Siqueira, and Anna Helena Reali Costa. Zeroberto: Leveraging zero-shot text classification by topic modeling. In *International Conference on Computational Processing of the Portuguese Language*, pages 125–136. Springer, 2022.

[65] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[66] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[67] Anna Glazkova. Identifying topics of scientific articles with bert-based approaches and topic modeling. In Manish Gupta and Ganesh Ramakrishnan, editors, *Trends and Applications in Knowledge Discovery and Data Mining*, pages 98–105, Cham, 2021. Springer International Publishing.

[68] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online, 2021. Association for Computational Linguistics.

[69] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[70] Rob Churchill and Lisa Singh. Topic-noise models: modeling topic and noise distributions in social media post collections. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 71–80. IEEE, 2021.

[71] Rebecca Taylor and Johan A du Preez. Simlda: A tool for topic model evaluation. In *Proceedings of the Future Technologies Conference*, pages 534–554. Springer, 2023.

[72] Alessio Bechini, Davide Gazzè, Andrea Marchetti, and Maurizio Tesconi. Towards a general architecture for social media data capture from a multi-domain perspective. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 1093–1100, 2016.

[73] Jeel Patel and Siddhaling Urolagin. Sentiment analysis and prediction of point of interest-based visitors' review. In Srikanta Patnaik, Xin-She Yang, and Ishwar K. Sethi, editors, *Advances in Machine Learning and Computational Intelligence*, pages 393–401, Singapore, 2021. Springer Singapore.

[74] Ram Krishn Mishra, Siddhaling Urolagin, and J. Angel Arul Jothi. Sentiment analysis for poi recommender systems. In *2020 Seventh International Conference on Information Technology Trends (ITT)*, pages 174–179, 2020.

[75] Rob Churchill and Lisa Singh. textprep: A text preprocessing toolkit for topic modeling on social media data. In *DATA*, pages 60–70, 2021.

[76] Jian Tang, Cheng Li, Ming Zhang, and Qiaozhu Mei. Less is more: Learning prominent and diverse topics for data summarization, 2016.

[77] Hosein Azarbonyad, Mostafa Dehghani, Tom Kenter, Maarten Marx, Jaap Kamps, and Maarten De Rijke. Hitr: Hierarchical topic model re-estimation for measuring topical diversity of documents. *IEEE Transactions on Knowledge and Data Engineering*, 31(11):2124–2137, 2018.

[78] Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, 77(5):4375–4388, 2021.

[79] Ginevra Carbone and Gabriele Sarti. Etc-nlg: End-to-end topic-conditioned natural language generation. *IJCoL*, 2020.

[80] Silvia Terragni and Elisabetta Fersini. OCTIS 2.0: Optimizing and comparing topic models in italian is even simpler! In Elisabetta Fersini, Marco Passarotti, and Viviana Patti, editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[81] Silvia Terragni, Elisabetta Fersini, and Enza Messina. Word embedding-based topic similarity measures. In Elisabeth Métais, Farid Meziane, Helmut Horacek, and Epaminondas Kapetanios, editors, *Natural Language Processing and Information Systems*, pages 33–45, Cham, 2021. Springer International Publishing.

[82] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), nov 2010.

[83] Silvia Terragni, Ismail Harrando, Pasquale Lisena, Raphael Troncy, and Elisabetta Fersini. One configuration to rule them all? towards hyperparameter transfer in topic models using multi-objective bayesian optimization. *arXiv preprint arXiv:2202.07631*, 2022.

[84] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery.

[85] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. Evaluating topic coherence measures, 2014.

[86] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. volume 32, pages 288–296, 01 2009.

[87] Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174, 2017.

[88] Derek O'callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

[89] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941, 2014.

[90] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[91] Ranjit Vijayan. Teaching and learning during the covid-19 pandemic: A topic modeling study. *Education Sciences*, 11(7):347, 2021.

[92] Andrei P Kirilenko, Svetlana O Stepchenkova, and Xiangyi Dai. Automated topic modeling of tourist reviews: does the anna karenina principle apply? *Tourism Management*, 83:104241, 2021.

[93] Turki Alenezi and Stephen Hirtle. Normalized attraction travel personality representation for improving travel recommender systems. *IEEE Access*, 2022.

[94] Daniela Vianna and Edleno Silva de Moura. Organizing portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3388–3392, New York, NY, USA, 2022. Association for Computing Machinery.

[95] Turki Alenezi and Stephen Hirtle. Normalized attraction travel personality representation for improving travel recommender systems. *IEEE Access*, 10:56493–56503, 2022.

[96] Darell Hendry, Fariz Darari, Raditya Nurfadillah, Gaurav Khanna, Meng Sun, Paul Constantine Condylis, and Natanael Taufik. Topic modeling for customer service chats. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–6, 2021.

[97] Viet Huynh, He Zhao, and Dinh Phung. Otlda: A geometry-aware optimal transport approach for topic modeling. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18573–18582. Curran Associates, Inc., 2020.

[98] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3143–3152, New York, NY, USA, 2022. Association for Computing Machinery.

[99] Shaymaa H. Mohammed and Salam Al-augby. Lsa & lda topic modeling classification: comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science*, 19:353–362, 2020.

[100] Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

[101] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1087–1099, 2012.

[102] E Mejía-Roa, D Tabas-Madrid, J Setoain, C García, F Tirado, and A Pascual-Montano. Nmf-mgpu: non-negative matrix factorization on multi-gpu systems. *BMC Bioinformatics*, 2015.

[103] Poonam Tijare and P Jhansi Rani. Exploring popular topic models. In *Journal of Physics: Conference Series*, volume 1706, page 012171. IOP Publishing, 2020.

[104] Zikai ZHOU and Kei WAKABAYASHI. Topic modeling using jointly fine-tuned bert for phrases and sentences.

[105] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[106] Anton Frederik Thielmann, Christoph Weisser, Thomas Kneib, and Benjamin Saefken. Coherence-based document clustering. 2021.

[107] Qin Li, Shaobo Li, Sen Zhang, Jie Hu, and Jianjun Hu. A review of text corpus-based tourism big data mining. *Applied Sciences*, 9(16), 2019.

[108] Big data in tourism: General issues and challenges. *J Tourism Hospit*, 10:1–6, 2021.

[109] Cyril Cappi, Camille Chapdelaine, Laurent Gardes, Eric Jenn, Baptiste Lefevre, Sylvaine Picard, and Thomas Soumarmon. Dataset definition standard (dds), 2021.

[110] Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online, nov 2020. Association for Computational Linguistics.

[111] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. Topic memory networks for short text classification, 2018.

[112] Dingcheng Li, Siamak Zamani, Jingyuan Zhang, and Ping Li. Integration of knowledge graph embedding into topic modeling with hierarchical Dirichlet process. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 940–950, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[113] Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, and Qinfei Chen. Incorporating knowledge graph embeddings into topic modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[114] Jeon-Hyung Kang, Jun Ma, and Yan Liu. Transfer topic modeling with ease and scalability, 2013.

[115] Changwei Hu, Piyush Rai, and Lawrence Carin. Non-negative matrix factorization for discrete data with hierarchical side-information. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1124–1132, Cadiz, Spain, 09–11 May 2016. PMLR.

[116] Hao Zhang, Gunhee Kim, and Eric P. Xing. Dynamic topic modeling for monitoring market competition from online text and image data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1425–1434, New York, NY, USA, 2015. Association for Computing Machinery.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.