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

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

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 detailed and comprehensive review of selected prominent models proposed for topic modeling, including traditional models in practice and recently devised strategies from literature. An experimental evaluation of the performance of the methods under consideration using well-established datasets as well as datasets for the touristic experiences highlights their advantages and unique characteristics based on multiple evaluation parameters. Further, the study discusses open issues for the application of topic modeling strategies and future research directions and presents the conclusions.

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 [77] [25]. 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 [78]. 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 [37] [80].

Although topic modeling is being widely applied in many disciplines today, however, one of its interesting application is considered for tourism industry [60]. 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 [31] [36]. 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) [70]. 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 [27].

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

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

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.
2) Present the systemic architecture, principles and working of each of the selected novel topic modeling strategies.
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.
5) Analyze the performance of each of the topic modeling strategies and identify the potential reasons for its performance in a particular way.
6) Conclude which strategy performs 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 exclusive novel approaches. These include Latent Dirichlet Allocation (LDA) [12] , Top2Vec [5], Non-Negative Matrix Factorization (NMF) [41] and Bidirectional Encoder Representations from Transformers (BERTopic) [23]. While the second category covers the strategies which are devised from the stated novel topic modeling strategies.

These include RoBERTa [43], Contextualized Topic Model (CTM) [10], and Embedded Topic Model (ETM) [19].

Compared to the previous studies, 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.

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 Background

2.1 Definition of Terms

In this section, we provide definitions on terms and basic concepts involved in topic modeling. A typical text-based dataset is made of "documents" which are strings of variable length composed of N words, where a "word" (or "term") is considered as the fundamental unit of a sample. The set of distinct words presents in a dataset forms a "vocabulary" and a "topic" is then viewed as a probability distribution over this fixed vocabulary. Obviously, the way in which we represent words and documents has a great impact on topic modeling. We will then present the ideas that are useful to understand the approaches we are analysing in this work. We will refer to the classification of word representation's techniques proposed by S. Selva Birunda and R. Kanniga Devi [59].

Category 1: Traditional word embedding, or Count-based embedding [4]. In this class fall all those methods that use frequency of words on the whole document, co-occurrence of words, and rarity of words in documents. Traditionally, text documents are represented as a bag of words, i.e. 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. The limits of such text representation are known: the vectors are very sparse, if we add a new document with words never used before the length of the vocabulary, and so of the vectors, will increase, and the context is not considered. A first improvement is given by TF-IDF, which measures how frequent a word is in a document (TF) and how much information it provides (IDF). The well-known formula for TF-IDF is:

$$tfidf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (1)$$

where $f_{t,d}$ is the raw 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.

Category 2: Static Word Embedding. These prediction-based methods compute probabilities to the words and map them into fixed-size vectors. These embedding do not consider context, i.e. a word embedding does not change if the word appear in a different sentence. If two words often appear together, then their embeddings will be similar. This class of techniques gained in popularity after the release of Word2Vec [16] This model can utilize either of two architectures: continuous bag-of-words (CBOW), which predicts one word from the surrounding words, or Skip-gram, that, on the other end, uses one word to predict all surrounding words. Word2Vec's idea has been used to design Doc2Vec [40], 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 these models, the word representation dynamically varies based upon the surrounding words. Models that use this kind of representation, like Transformers, are SOTA for most NLP tasks.These approaches are context-dependent, i.e. they can disambiguate polysemes, thanks to the attention mechanism [67]. This means that these models can compute different embeddings for a word depending on the context. There are tons of models based on this architecture, but the most famous one is certainly BERT [18] which has been used in several applications in nlp [38] [76] and in many flavours [75]. An interesting variation of BERT for our work is SBERT [54], which, thanks to siamese and triplet network structures, can better derive sentences similarities. Since most of the proposed Transformer architectures have a limit on the number of tokens they can handle, 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.

$$tfidf = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

, where $f_{t,d}$ is the raw count of term t in the document d and D is the dataset.

2.2 Review of Recent Studies

Topic modeling has its roots in the 1980s [68], but really took off in the late 1990s thanks to methods such as LSI [17] and especially LDA [12]. Many methods based on LDA were designed over the last two decades ([7] [79] [42]). Despite their success, conventional Bayesian probabilistic topic models started to show signs of fatigue in the era of big data and deep learning. Instead, models based on the use of Deep Learning techniques are becoming more and more popular [82]. DL methods have been applied to topic modeling for document representation [81], for computing semantic representations of topics [72] and to deal with short texts [67] [48].

The scientific community did not focus just on designing different methods that are then applied on traditional data (text), but in the years there has been a great effort in the application of topic modeling to different fields and for many purposes [33]. Some interesting fields in which topic modeling has been used are: Marketing and Business management [49] [61] [55] [52] [28] analysis of scientific publications [3] [71], biology and medicine [42] [83], software traceability [6].

In the tourism field there are publications in which topic modeling is used to discover preferences in travel itineraries, to study customers opinions and to make recommendations. Some works in which topic modeling was used on datasets about tourism are shown in Table 1.

3 Selected Topic Modeling Approaches

[b] 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 exclusive novel strategies which are not devised or improvised from any other strategy. For this study we have considered LDA, Top2Vec, NMF and BERTopic as novel models in practice.

On the other hand, as per the category name suggests, the recently devised strategies includes the topic modeling strategies that have been devised or improvised from the novel models in practice. For this study, we have considered RoBERTa, CTM and ETM as recently devised strategies. 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) [12] is a generative probabilistic model, designed for a given corpus of text documents. The model works on the definneti 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. The generative process of LDA can be observed from the plate diagram shown in Figure . More 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:

$$P(D \mid \alpha, \beta)=\prod_{i-1}^{N} \int P\left(\vartheta_{i} \mid \alpha\right) F\left(\vartheta_{\varphi}\right) d \vartheta_{i} \tag{3}$$

By maximizing the probability, 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 a 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 model takes number of topics as input and assigns topics to documents accordingly. This may variate the for different number of topics.
3) Random assignment of topics to documents and words to topics and then iterative update. This assumes all topic assignments except the current word are correct.

3.1.2 Top2Vec

Top2Vec [5]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 locate semantically similar words, sentences, or documents within spatial proximity (Egger, 2022a). It offers pretrained embedding models. 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 (Hendry et al., 2021) [53].

Inshort, the leverages joint document and word semantic embedding to find topic vectors. Figure shows the systemic architecture based workflow of the model.

The model claims for the following assumptions:

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

3.1.3 Non-Negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) [41] 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 . Here W is represents terms mapped to topics and H represents topics mapped to documents. The values of W and H are non-negative and updates iteratively.

$$A=WH \tag{4}$$

where W and H are non-negative matrices such as W ≥ 0, and V ≥ 0. The weighted sum of the components in matrix A is:

$$A_i=\sum_{j=1}^{k} W_{ij} * H_j \tag{5}$$

The multiplicative updates for the learning part is as follows:

$$W \leftharpoonup W \frac{AH^T}{WHH^T} \tag{6}$$

Studies using TM in Tourism field				
Authors	Study Objectives	Models	Datasets	Metrics
Rossetti M. et AL [58]	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
Huy Quab Vu et Al [70]	Analysis of travel itineraries	LDA	Twitter, Foursquare	Perplexity, topic concentration
Nan Hu et Al [30]	Customers' complaints	STM	TripAdvisor	Several analysis on the topics obtained. No specific metric score
Calheiros A. at Al [13]	Sentiment Classification of Reviews	LDA	Custom dataset collected online	Several analysis on the topics obtained. No specific metri score
Takeshi Kurashima et Al [39]	Locations recommendations	Geo Topic Model	Tabelog and Flickr-sourced geotag collection	5-best accuracy
Shuhui Jiang at Al [34]	Travel recommendations	Author Topic Collaborative Filtering	Geo-tagged photos from Flickr	MAP
Yue Guo at Al [26]	Tourist satisfaction analysis	LDA	TripAdvisor	Jaccard coefficient, human analysis and Standford Topic Modelling Toolbox
Jie Bao at Al [8]	Bikesharing	LDA	Smart card data of a bike sharing system, Google Places API	Perplexity

TABLE 1: Recent studies that use topic modeling in the tourism field

$$H \leftharpoondown H\frac{W^TA}{W^TWH} \tag{7}$$

The model iterates the above two equations until it achieves convergence then it achieves final term–topic matrix W and topic–document matrix H for topics extraction. [20] [35] [73] [53]

3.1.4 Bidirectional Encoder Representations from Transformers (BERTopic)

BERTopic [23] 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. Here BERT serves as embedder, while using sentence-transformers the BERTopic provides embedding extraction for the document corpus., with a sentence-transformers model for more than 50 languages. Similar to top2vec, the BERTopic also offers dimensionality reduction using UMAP and then clusters the documents using HDBSCAN. The architectural workflow of the approach is illustrated in Figure However, unlike Top2Vec, it applies class-based term frequency inverse document frequency (cTF-IDF), shown in eq . This efficiently evaluates the significance of terms within a cluster or class followed by the creation of term representation. Here the high score a term gets, the better it represents its topic. (Sánchez-Franco and Rey-Moreno, 2022).

$$cTF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j} \tag{8}$$

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

BERTopic, differs from other approaches as it offers continuous rather than discrete topic modeling (Alcoforado et al., 2022). The model leads to different results with repeated execution due to its stochastic nature. The model offers the following features:

1) Does not require number of topics in advance. Estimates the number of topics automatically
2) Offer several multi-lingual models to extract document embeddings. Usually in practice is 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 offers to mention outliers in the result output as Topic 0 with the label of -1.

[1] [53]

3.2 Recently Devised Strategies

3.2.1 Robustly Optimized BERT Pre-training Approach (RoBERTa)

RoBERTa is a devised strategy from BERT 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 BERT's masking strategy, 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 (Petroni et al., 2019; Vulic et al. ´ , 2020).

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 for on larger number of datasets []. 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. [22] [43] [66] [21]

3.2.2 Contextualized Topic Model (CTM)

Contextualized Topic Models (CTMs) are devised from the Neural-ProdLDA variational autoencoding approach (by Srivastava and Sutton (2017) and pre-trained embedding models []. 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 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. [10] [11]

3.2.3 Embedded Topic Model (ETM)

The embedded topic model (ETM) (Dieng et al., 2020) 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 (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). 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 dth document can be summarized as follows, where $\mathcal{LN}$(.) represents the logistic normal distribution ((Aitchison and Shen, 1980; Blei and Lafferty, 2007)):

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 intial steps of the approach, 1 and 2a, are similar to traditional LDA. The improvisation is 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 comparatively evaluated the considered novel topic models and devised topic models. The novel topic models, in this study, includes LDA, Top2Vec and NMF. While, the devised topic models include BERTopic, RoBERTa, CTM and ETM. The comparison is performed using various publicly available datasets and touristic experience focused datasets, exclusively generated for this study. The statistical summary of the datasets is mentioned in Table -. The details are mentioned in the following subsections.

4.1 Datasets

4.1.1 Benchmark Datasets

20NewsGroup (20NG) is a well-established 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 [][][].

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 [9]. It has been used in many data analysis studies such as [] [51] [46].

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 unique 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 unique 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

TABLE 2: Statistics of the datasets

Dataset	# of Docs	# of Words	Vocabulary Size	Avg. Words Per Doc
AirBnB Touristic Experiences (ATE)	737	126450	2629	68
TripAdvisor Tourist Activities (TAT)	2765	284050	4555	152
EasyTour (ET)	5724	1556416	138095	272
20 NewsGroup (20NG)	18846	3423145	29548	182
TourPedia (TP)	8000	191996	27012	24

TABLE 3: 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

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 []. 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 - shows a summary of the data preparation steps.

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 transformers based methods; 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 majorly considered English language documents. 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. 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 coherence

Topic Coherence measures the interpretablity and coherence of the topics produced by a model and its association with the considered data [56] [57]. The idea is based on distributional hypothesis of linguistics []. Unlike perplexity and predictive likelihood, which can be contrary to experts judgment [15], the topic coherence has been practiced in many studies such as [62][][]. 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. The details of these measures can be referred from [62] [56].

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

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

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:

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

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

In eq [] the context vector $\vec{v}(W')$, uses NPMI for all the word pairs. γ places more weight on larger NPMI values. 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).

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

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

4.2.2 Topic Diversity

Topic diversity is a significantly impactful evaluation parameter to assess the topics produced by a topic model[] [63]. 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 [][][]. 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 [].

$$TopicDiversity = \frac{n(U)}{K * n(T)} \tag{13}$$

Here 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.

4.2.3 Inverted RBO

Another interesting parameter used to evaluate the quality of topics is Inverted RBO (IRBO). It is a topic diversity estimation parameter that illustrates to what extent topics differ from each other [10] [65][]. It ranges from 0 to 1, where 0 means fully identical and 1 means fully diverse topics. It uses Ranked-Based Overlap measure [74] and compute the how disjoint are topics based on word-ranking for top K words. The parameter has been used by many value studies to estimate the quality of topics such as [14] [48] and [64].

4.3 Experiment and Results

In this subsection, we have illustrated the results obtained through the conducted experiemental 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. 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. Note that we pre-defined the number of topics for LDA, NMF , CTM and ETM using elbow method as suggest by [][], while Top2Vec, BERTopic and RoBERTa are modeled to decided best suitable number of topic by themselves [][].

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. Table - shows the results obtained in this regard for both topic diversity and Inverted RBO (IRBO). Moreover, Figure - illustrates a comparison of the models with respect to average topic diversity and Figure - shows 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 -, which illustrates reduced variation of topic diversity among models and BERTopic outperforms others. Similarly it is interesting to observe from Figure - that BERTopic and RoBERTa shows much less IRBO when applied for small-sized dataset ATE with shorter document lengths. Note that although Top2Vec provides higher topic diversity on average but the number of clusters (topics) it has produced is also less as compared to others for almost each dataset (Table -). This might also indicate a high intra-cluster similarity which is expected to be less for a good topic model [].

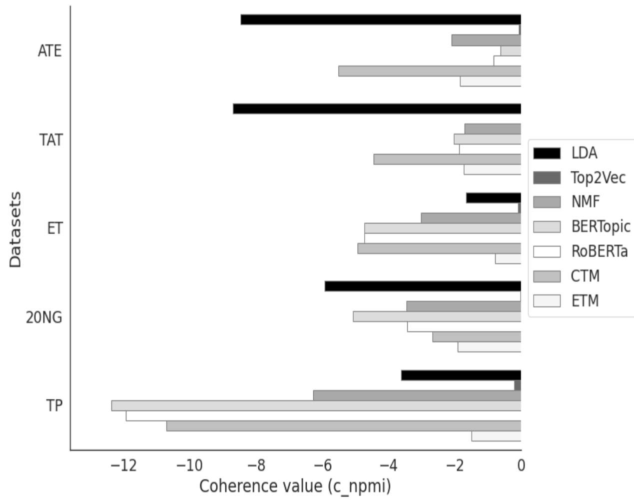
Topic Coherence: Further, we analyzed the coherence parameters (C_{uci}, C_v, C_{umass} and C_{npmi}) to determine the semantic coherence of the topics generated by each model for the datasets under consideration. Figure - depicts a comparative analysis of all the models for the given datasets for each 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 [2].

Notice from Figure - that although NMF shows better C_{uci} for comparatively smaller sized datasets; ATE and TAT, but with size growth of datasets, 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 -, while NMF shows better coherence on average and for 3 out of 5 datasets, it is worth noting that for largest sized dataset; 20NG, Top2Vec exhibits better C_v than others. This may imply to the variation in results due to dataset size, where NMF seems suitable for small to medium sized datasets. An interesting observation can be made from Figure - for C_{umass}, where Top2Vec outperforms others on average and also individually for each dataset including Italian language dataset "EasyTour". Note that Top2Vec shows better C_{umass} in each case irrespective to size and type of data implying better stability of the coherence parameter. A similar interesting result can be visualized from Figure -, where ETM illustrates better C_{npmi} in case of every dataset and on average as a whole. Considering the c_v as the closest coherence measure to human judgement [][][], 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, as supported by [][]. This behavior can be observed from Table -, where results are mentioned in detail.

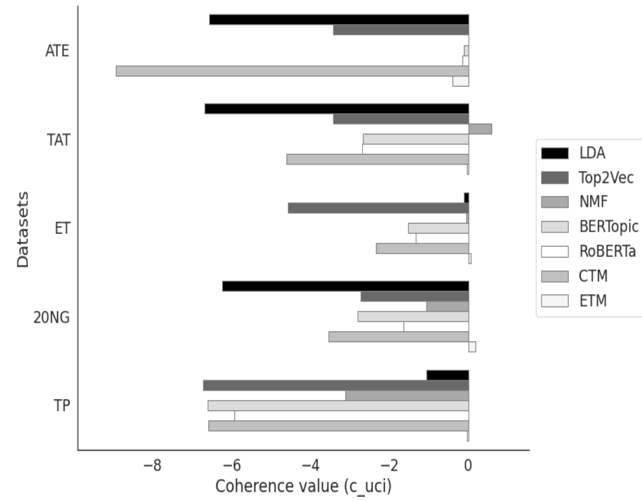
4.4 Validation of Analysis

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

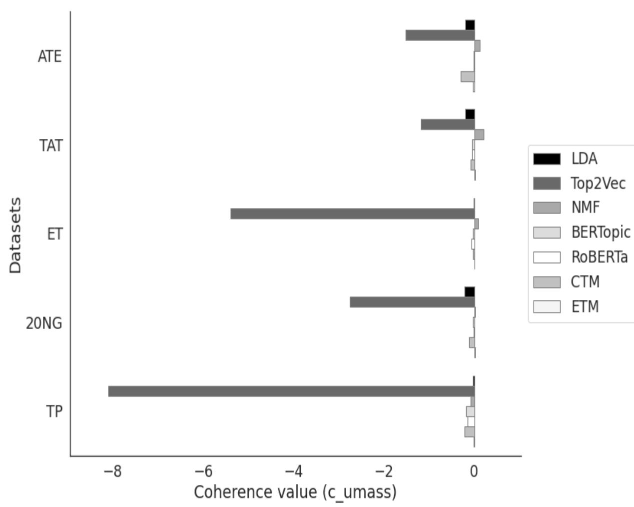
The shortcomings of the study reveals that Top2Vec generates topics with better diversity for majority of the datasets under consideration. This has been found for both parameters; topic diversity and IRBO. Such behavior for Top2Vec generating better topic diversity has been found by [69] [2] and [29]. 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



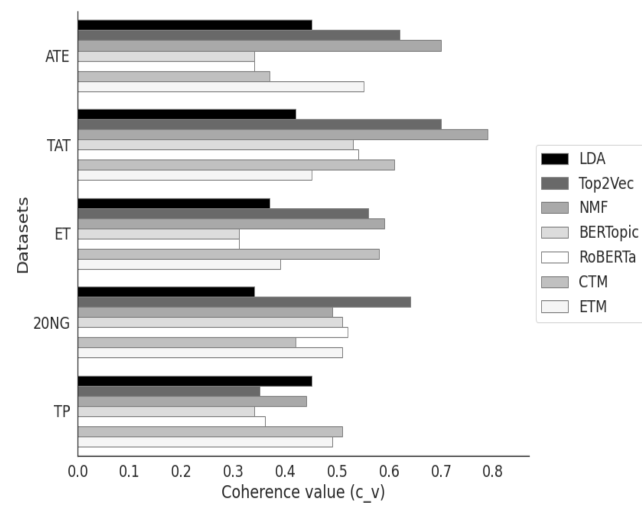
(a) *Results for c_npmi*



(b) *Results for c_uci*

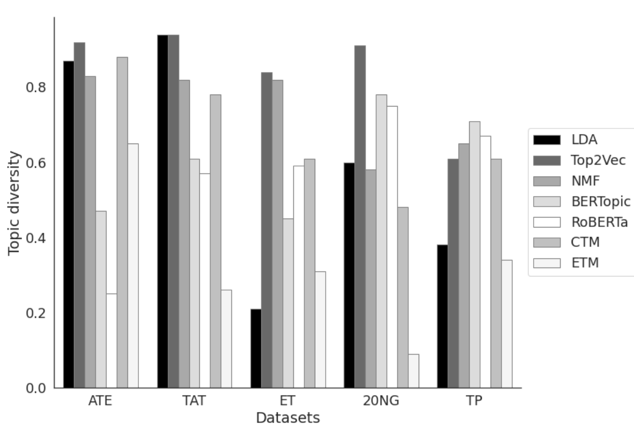


(c) *Results for c_umass*

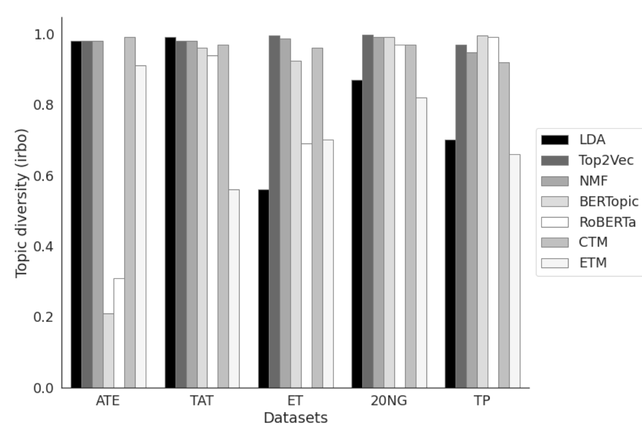


(d) *Results for c_v*

Fig. 1: Results for the coherence metrics



(a) *Results for topic diversity*



(b) *Results for IRBO*

Fig. 2: Results for the diversity metrics

TABLE 4: Results for ATE

Models	Coherence (C_uci)	Coherence (C_V)	Coherence (u_mass)	Coherence (c_nmpi)	Topic Diversity	IRBO	Number of topics
LDA	-6.56	0.45	-8.45	-0.19	0.87	0.98	14
Top2Vec	-3.42	0.62	-0.06	-1.52	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 5: Results for TAT

Models	Coherence (C_uci)	Coherence (C_V)	Coherence (c_npmi)	Coherence (u_mass)	Topic Diversity	IRBO	Number of topics
LDA	-6.68	0.42	-8.68	-0.19	0.94	0.99	16
Top2Vec	-3.42	0.70	-0.01	-1.17	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 6: Results for ET

Datasets	Coherence (C_uci)	Coherence (C_V)	Coherence (u_mass)	Coherence (c_nmpi)	Topic Diversity	IRBO	Number of Topics
LDA	-0.11	0.37	-1.65	-0.01	0.21	0.56	22
Top2Vec	-4.57	0.56	-0.10	-5.38	0.84	0.995	50.1
NMF	-0.05	0.59	-3.03	0.09	0.82	0.986	22
BERTopic	-1.53	0.31	-4.72	-0.029	0.45	0.925	74.9
RoBERTa	-1.32	0.31	-4.72	-0.061	0.59	0.69	14.1
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 7: Results for 20NG

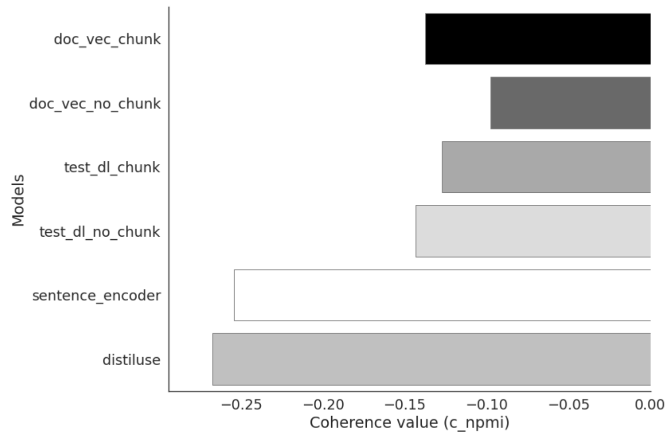
Datasets	Coherence (C_uci)	Coherence (C_V)	Coherence (u_mass)	Coherence (c_nmpi)	Topic Diversity	IRBO	Number of Topics
LDA	-6.23	0.34	-5.92	-0.21	0.60	0.87	111
Top2Vec	-2.72	0.64	-0.02	-2.74	0.91	0.998	83
NMF	-1.05	0.49	-3.46	0.03	0.58	0.99	111
BERTopic	-2.80	0.51	-5.06	-0.03	0.78	0.99	216
RoBERTa	-1.64	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

TABLE 8: Results for TP

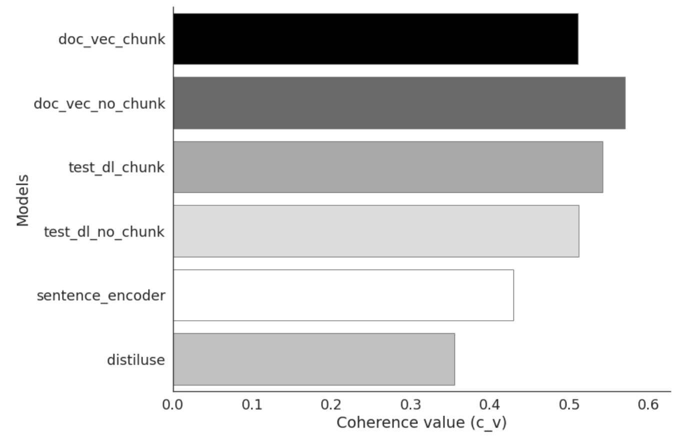
Datasets	Coherence (C_uci)	Coherence (C_V)	Coherence (u_mass)	Coherence (c_nmpi)	Topic Diversity	IRBO	Number of Topics
LDA	-1.06	0.45	-3.62	-0.03	0.38	0.70	14
Top2Vec	-6.72	0.35	-0.22	-8.10	0.61	0.97	40.8
NMF	-3.10	0.44	-6.27	-0.07	0.65	0.947	14
BERTopic	-6.59	0.34	-12.35	-0.17	0.71	0.996	142.4
RoBERTa	-5.91	0.36	-11.91	-0.15	0.67	0.991	105.5
CTM	-6.57	0.51	-10.70	-0.21	0.61	0.92	14
ETM	-0.03	0.49	-1.49	-0.001	0.34	0.66	14

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 - 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 depicting better results for all the datasets for our study, while producing highly appreciable C_{uci} results for [32] and [45]. Note that as mentioned earlier, ETM is a devised strategy from LDA with word-to-vector improvement. As LDA is already a well-established



(a) *C_npmi score in several tests of Top2Vec*



(b) *C_v score in several tests of Top2Vec*

Fig. 3: Comparisons between several parameters for Top2Vec

strategy delivering considerable C_{uci} coherence [47], 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 [20] and [50]. 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 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 in [24] [44].

Further, we observed that C_{umass} is rather a different parameter where a lower value signifies better coherence in case of Gensim usage []. Note that Top2Vec outperforms others, the probable reason for this could be that it generates topic vectors from joint document-word embedding spaces, occurring together with considerable probability [], while C_{umass} involves counting of co-document appearance [84] which are more likely to be supported by joint document-word embedding spaces, hence the topics produced by Top2Vec are likely to have better C_{umass} scores.

Finally the study finds that ETM delivers better C_{npmi} score for each dataset compared to all other models. Since C_{npmi} uses normalized version of C_{uci} which is in turn based on PMI score. It is most-likely for a technique performing better for C_{uci} to also perform better on C_{npmi}, which is observable in case of ETM for all the datasets.

5 Discussion and Analysis

6 Conclusion

7 Open Issues and Future Research Directions

mds
August 26, 2015

7.1 Subsection Heading Here

Subsection text here.

7.1.1 Subsubsection Heading Here

Subsubsection text here.

8 Conclusion

The conclusion goes here.

Appendix A Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgments

The authors would like to thank...

References

[1] 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.

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

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] Piyush Ghasiya and Koji Okamura. Investigating covid-19 news across four nations: A topic modeling and sentiment analysis approach. *IEEE Access*, 9:36645–36656, 2021. Publisher Copyright: © 2013 IEEE.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Timothy R. Hannigan, Richard F. J. Haans, Keyvan Vakili, Hovig Tchalian, Vern L. Glaser, Milo Shaoqing Wang, Sarah Kaplan, and P. Devereaux Jennings. Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, 13(2):586–632, 2019.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

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

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

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

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

[56] 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.

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

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

[59] 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.

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

[61] Vanitha Swaminathan, H. Andrew Schwartz, Rowan Menezes, and Shawndra Hill. The language of brands in social media: Using topic modeling on social media conversations to drive brand strategy. *Journal of Interactive Marketing*, 57(2):255–277, 2022.

[62] 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.

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

[64] 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.

[65] 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.

[66] Laure Thompson and David Mimno. Topic modeling with contextualized word representation clusters, 2020.

[67] 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.

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

[69] 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.

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

[71] 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.

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

[73] 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.

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

[75] 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.

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

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

[78] 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.

[79] 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.

[80] 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.

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

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

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

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

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.