A hybrid Latent Dirichlet Allocation - BERT approach for topic discovery of market places

Ibrahim Bouabdallaoui (bd.ibrahim@hotmail.com)
High School of Technology Salé, Mohammed V University in Rabat, Rabat-Salé-Kénitra
Fatima Guerouate
High School of Technology Salé, Mohammed V University in Rabat, Rabat-Salé-Kénitra
Samya Bouhaddour
High School of Technology Salé, Mohammed V University in Rabat, Rabat-Salé-Kénitra
Chaimae Saadi
High School of Technology Salé, Mohammed V University in Rabat, Rabat-Salé-Kénitra
Mohamed Sbihi
High School of Technology Salé, Mohammed V University in Rabat, Rabat-Salé-Kénitra

Research Article

Keywords: topic modeling, latent dirichlet allocation, bert, latent space representation, autoencoders

Posted Date: May 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1674353/v1

A hybrid Latent Dirichlet Allocation - BERT approach for topic discovery of market places

Ibrahim Bouabdallaoui[1*], Fatima Guerouate[1†], Samya Bouhaddour[1†], Chaimae Saadi[1†] and Mohamed Sbihi[1†]

[1]Laboratory of System Analysis, Information Processing and Industrial Management, High School of Technology Salé, Mohammed V University in Rabat, Avenue Le Prince Héritier, Salé, 11060, Rabat-Salé-Kénitra, Morocco.

*Corresponding author(s). E-mail(s): bd.ibrahim@hotmail.com;
Contributing authors: guerouate@gmail.com;
samya.bouhaddour@uit.ac.ma; chaimaesaadi900@gmail.com;
mohamed.sbihi@yahoo.fr;
†These authors contributed equally to this work.

Abstract

The tourism industry is now intimately linked to the web, as It was one of the first industries to fit on digital transformation. The tourist is more autonomous in organizing and preparing his trip, and the possibilities of leaving at competitive prices are numerous. New technologies have an impact on tourism and it is now necessary to think about communication and marketing accordingly. Market places in Morocco are reached by millions of tourists each year. Tourists share their experiences in social medias to encourage or discourage other tourists who are interested in visiting such places. The main goal of this work is to propose a hybrid architecture based on Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT) with a latent space representation technique to perform topic modeling and have as a result an idea of each tourist's review and particularly what he thinks about a specific market place in Morocco.

Keywords: topic modeling, latent dirichlet allocation, bert, latent space representation, autoencoders

1 Introduction

New information and communication technologies have brought about profound changes in today's consumer society, from where they have become a major challenge for the development of the economy of countries. Advances in communication and information technology have brought about exponential transformations in all sectors, especially tourism, hence social media has made it easier for travelers to share their needs and experiences through comments and reviews. on destinations. Finally, the massive rise of social media in the tourism sphere has profoundly transformed the terms of exchanges between travelers and brands. Tourism has moved from a basic field focused on traditional means to a new world with a new language called e-tourism and social media.[1]
TripAdvisor is the world's largest travel platform, helping 463 million travelers every month to make every trip their best trip. Travelers around the world use the TripAdvisor website and app to view more than 859 million reviews and opinions on 8.3 million accommodations, restaurants, experiences, airlines and cruises. In just a few years, artificial intelligence have completely transformed the tourism and travel industry.[2] The use of these technologies has notably made it possible to improve the customer experience both during their trip and when making their reservation. In today's world, where the amount of unstructured text is drastically increasing (comments, blog posts, etc.), it would be really useful to have tools that automatically structure information [3], so that it can be quickly accessed what interests us, filtering the noise but also detecting the appearance of new subject of interest. It is in this context that the modeling of subjects intervenes which represents the spectrum of the different approaches allowing this detection.

2 Litterature review

Much work has recently been undertaken in mixing the Latent Dirichilet Allocation (LDA) and BERT to define topics, researchers are tending to use these approaches for sentiment analysis purposes. In first instance, a work of prospecting of monolingual and multilingual topics had been performed using these two approaches, so as to analyze topic evolution in multilingual and monolingual (Chinese and English) topic similarity settings[4]. LDA probability had been multiplied by the averaged tensor similarity of BERT embeddings to explore the evolution of the topic in scientific publications. To improve such work in a deep learning context, an unsupervised approach had been proposed for aspect term extraction from product reviews using a guided LDA model, [5] a filtered and subjected N-grams by regular expressions to multiple filtering stages.The methodology starts with a text pre-processing module and an input sequence generation and filtration module to define a strong semantic approach for aspect extraction. In second instance, Language Model Representations is being a subject of research to many researchers interested in topic discovery, a recent work has been done to analyze the challenges of using PLM

representations for topic discovery[6], and a mix of clustering approach with a latent space learning are proposed build upon PLM embeddings. Technically, the latent space, topic-word and document-topic distributions are jointly modeled so that the discovered topics can be interpreted by coherent and distinctive terms and meanwhile serve as meaningful summaries of the documents. Another work has been published to focus on knowledge distillation method LRC mixed with BERT based on contrastive learning [7] to fit the output of the intermediate layer from the angular distance aspect with a gradient perturbation-based technique applied on the training architecture to increase the robustness of the model, which the potential of knowledge distillation and then going for an experimental evaluation on 8 datasets using the General Language Understanding Evaluation (GLUE) benchmark, which gave excellent results.

3 Methodology

In this section we define the main concepts related to the proposed model. In Section 3.1 we introduce the data pre-processing method. Section 3.2 describes research methods that has been followed to build the topic modeling architecture.

3.1 Data pre-processing

3.1.1 Web Scraping

Web scraping is a technique for quickly collecting data online. The strength of this technique lies in the fact that it makes it possible to extract large quantities of data in a structured way. Given the growing importance of tourists for places and markets in Morocco, and also the emergence of experience sharing, we have developed an algorithm that automates the collection of tourist reviews on the TripAdvisor website, and more precisely in the section of shopping places in Morocco, by looping over 91 shopping places that are available, as well as a collection process is completed on the opinions of tourists on each place, taking the title of the comment, the content of the comment as well as the rating of each comment. This web scraping algorithm was developed with `BeautifulSoup` package on Python, so as to have total control of the content of the section in question (e.g. Moroccan shopping places).

3.1.2 Data management

After collecting data from TripAdvisor, a process of saving and converting data from unstructured to semi-structured and then structured, materializing the scraping code with a semi-structured architecture that can be compatible with NoSQL databases. However, there are samples which are missing and not filled in by tourists, therefore the information is encapsulated in a key-value structure and then saved in two JSON files (i.e. the first file contains comments and rates, and the second file contains titles and localization of shopping

places), in order to have a well presented and understandable architecture. The sample below shows how look like the result of web scraping process:

```
{"Name": "Jemaa el-Fnaa",
"Title": "What a hoot",
"Comment":"A chaotic cacophony greets you when you
           first enter the square and visually you are not
           sure where to look first. Yes, there are
           snake charmers but there are the monkeys, the
           musicians, the henna artists the fortune tellers,
           the food stalls, all vying for your attention.
           Don't fool yourself, they all want your money and
           morally I am not sure of the animal treatment but
           take it for what it is, take it all in and have
           fun.
"Rating": "5.0 "}
```

To have a hands-on this data, we managed to convert the JSON files to structured format so as the exploitation would be easy and figure out hidden patterns statistically.

3.1.3 Data preparation

After loading these two files, and merging them into one large data set using `Pandas` package, a cleaning process has been done: drop empty rows, undesirable data that would generate outliers in analysis (e.g. samples that contains invalid reviews), and cast rating values to float since they are recognized as string in raw data. The final shape of data has a size of 26244 rows and 4 major columns: "Name of the shop place", "Title of the review", "Review", "Rate". The name of shopping place may be duplicated in function of reviews number. Rate column contains 5 classes that we proposed to name them as follows: "terrible" for 0, "bad" for 1, "regular" for 2, "good" for 3 and "excellent" for 4. To prepare data correctly for analysis, a conversion process has been done to the comments such as lower-casing text, cleaning it from punctuation, links, emails, special characters and stop words. Stop words package is set to English language since we are working on English reviews. A parallelism mechanism has been applied to browse data and manipulate it easily without having memory issues and long processing time using multi-threading CPU pools. A pipeline of data pre-processing has been done to perform a dictionary of words based on the data prepared in the previous section, starting with applying SymSpell algorithm to finding all strings in very short time within a fixed edit distance from a large list of strings, then detecting nouns and adjectives using Part of speech tagging, and correcting typo of words, and finally applying stemming.

3.2 Research methods

This paper takes shopping places review data on TripAdvisor, and analyzes using Latent Dirichlet Allocation and BERT model (LDA-BERT) and autoencoder to define the accurate topics of each review. Thus, there are several steps of the research method.

3.2.1 Latent Dirichlet Allocation using BERT

To transform the prepared corpus into a digital representation, building a dictionary is a must: for each word, a unique index would be assigned. For this purpose, a sentence embedding process is launched: a sentence embedding is an encoding method that aims to represent the words or sentences of a text by vectors of real numbers, described in a Vector Space Model[8]. In other words, each word of the studied vocabulary V will be represented by a vector of size m. The principle of sentence embedding is to project each of these words into a vector space of a fixed size N (N being different from m). That is to say, whatever the size of the vocabulary, one must be able to project a word into its space.

The sentence embedding used in this work is based on BERT[9], which is a Transformers-like model (i.e. a model that works by performing a small constant number of steps). At each step, it applies an attention mechanism to understand the relationships between the words in the sentence, regardless of their respective positions. BERT uses Masked Language-Modeling (MLM) technique: it randomly masks words in the sentence and then tries to predict them[10]. Masking means that the model looks both ways and uses the full sentence context, left and right, to predict the masked word. Unlike old language models, it considers preceding and following words at the same time.

To perform a topic model using the proposed sentence embedding, we used the Latent Dirichlet Allocation (LDA)[11] which is a generative probabilistic model for describing collections of text documents or other types of discrete data. It belongs to a category of models called "topic models", which seek to discover thematic structures hidden in vast archives of documents. This provides effective methods for processing and organizing the documents in these archives[12]: automatic organization of documents by subject, search, understanding and analysis of the text, or even summarizing texts. The LDA is a 3-layer hierarchical Bayesian model, each document is modeled by a mixture of topics which then generates each word of the document.

And then a concatenation has been done with both LDA and BERT vector with a weight hyper-parameter to balance the relative importance of information from each source[13]. Combining sentence embeddings using BERT and topics recognition mechanism using LDA would make the model more robust as it will recognize particular patterns in tourism sphere.[14] Technically, two vectors had been initialised: the first one is the LDA vector which is based on the whole corpus, the created dictionary and a specific number of topics that the LDA should focus on, in this work, number of topics is equal to 5, to avoid

randomness in selecting related words, especially in a corpus that is hard to apply an accurate prediction task, the second vector is the sentence embedding vector that is totally collected from BERT embeddings file and then used it to encode the sentences of the corpus. These two vectors are concatenated with a weight hyper-parameter[13] of 15 applied on the LDA vector to fine tune results of the concatenation. The choice of this weight hyper-parameter is based on an experimental validation that we will introduce below.

3.2.2 Auto-encoder

Autoencoders are somewhat special neural networks that have exactly the same number of neurons on their input layer and their output layer. The goal for an autoencoder is to have an output closest to the input.[15] It contains two elements: Encoder and Decoder.
The encoder is made up of a set of layers of neurons, which process the data in order to build new so-called "encoded" representations. In turn, the layers of neurons in the decoder receive these representations and process them in an attempt to reconstruct the original data. The differences between the reconstructed data and the initial data make it possible to measure the error made by the autoencoder. The training consists in modifying the parameters of the auto-encoder in order to reduce the reconstruction error measured on the different samples of the dataset.
The proposed model is to make a latent space representation using autoencoders, whereby the first stack of layers contains 32 layers and it's considered for encoding purposes, the "bottleneck" part is for the latent space representation where the topic modeling features are recognized and well matched between each other, and then the last stack of layers are for decoding, while respecting the shape of the output.

4 Results and discussion

In this work, a topic modeling algorithm has been implemented using a mix between Latent Dirichlet Allocation and BERT embeddings, results are based on the auto-encoding approach to match accurate patterns according each review.[16]
Using the methodology described in the previous section, we aim to compare the estimation and fit of various models that might do the same work as the Latent Dirichlet Allocation and BERT combination. To evaluate the proposed model, a parameter Γ has been initialized to give a relative importance of LDA according the BERT embedding to balance weights between these two patterns.[13] We looped over a range of Γ values to define the optimal point, the plot below shows the results of learning and predictions of the auto-encoder model based on a combination of gammas that we proposed :

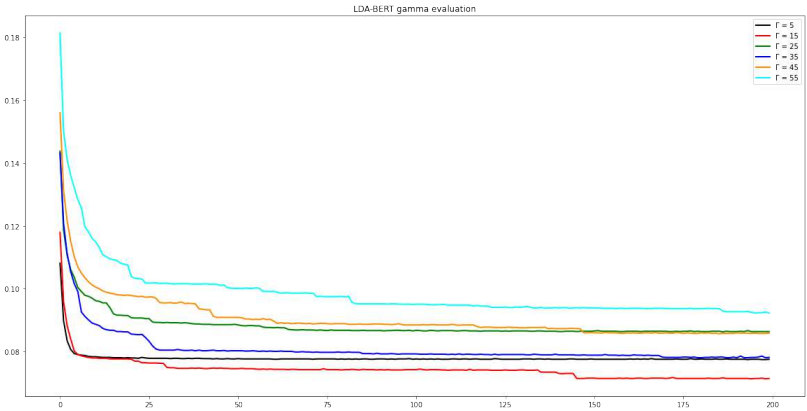

Fig. 1 Validation loss curve for each Gamma

As can be seen in Fig.1 the curves whose Γ values are greater than or equal to 25, have loss values that start respectively with 0.18 and decrease by a step of 2 for each Gamma value. For $\Gamma = 55$ which is the largest value, we see that the curve starts from the value 0.18 and drains completely towards the 25 epochs and it continues to decrease slowly to reach about 0.09 in the 200th epoch. For $\Gamma = 45$, the loss curve starts to decrease from 0.158 and then it drains in 15th epoch and then, it continues to decrease very fast to the 45th epoch steady pace around the 200th epoch. For $\Gamma = 35$, the work is done in the same way as in $\Gamma = 45$ except that at the beginning the loss curve starts on the value 0.14 and takes the same values as its previous one in the last epochs. For $\Gamma = 25$, the loss curve starts from the same starting point as that of $\Gamma = 35$, but this time it decreases exponentially in the first 27 epochs, and takes a constant pace until the 200th epoch. For $\Gamma = 15$, the curve starts at 0.12, and drains until the 23rd epoch and registers a small drop around the 27th epoch, and another around the 140th. And finally, for $\Gamma = 5$, it starts from the value 0.11, which is the small value and quickly drains towards the 5th epoch and after it takes a constant pace until the end of the learning, touching the same values than that of $\Gamma = 25$.
As a result, we can deduce that the $\Gamma = 15$ coefficient is the most balanced coefficient so that we can bring to a balance between BERT embeddings and LDA results. If we vary the values around 15, it will give practically the same results, but the most optimal result is the one marked in the figure above. Which tells us the following accuracy table which gives the results according to each Γ value:

Table 1 Validation accuracy for each Gamma

Γ=5	Γ=15	Γ=25	Γ=35	Γ=45	Γ=55
0.8062	0.9981	0.9977	0.9966	0.9961	0.9928

In the table above, the autoencoder model $\Gamma = 15$ gave the highest accuracy, with a big difference in comparison to $\Gamma = 5$, which gave loss results that are much better, and thus, for the other Gamma values, the results are much closer than the result of the best score, in spite of the loss curves which were poor in comparison with $\Gamma = 15$ and $\Gamma = 5$. These results explain that in spite of the results of accuracy reach a score perfect in terms of learning and testing, the loss curves are the only ones to demonstrate the ability of the model to correctly predict the sequences of words and in particular the different topics defined for a sentence, using the technique of Latent Space Representation which is mainly based on autoencoders.
To do this, we must go through the visual presentation of the samples so that we can be sure of the results defined by the model and see the consistency of the predicted topics, as well as their semantics with the context of the sentence.[17] Let's take the example of the sample presented in section 3.1.2:

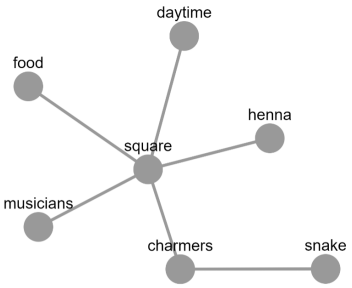

Fig. 2 Graph of predicted topics

The graph above is the prediction of the sample shown below using the implemented LDA-BERT auto-encoder and the Dash Cytoscape for graph visualisation. Results of predictions are shown as tokens, that are stored in a list of 7 elements, these elements are the topics defined by the model. To put these results in graph visualization, we used Dash Cytoscape that has as a logic to define nodes, which are the predicted topics, and define edges, that we defined them as categories: these categories are a set of words that are compared to the predicted words, and then a word similarity [18] is calculated using `SpaCy` package, the highest similarity value is the appropriate word to define the semantic of the word. These categories are defined as an id of each word.
The graph below shows that all the words related to the sample are attached to the square, which means that the LDA-BERT autoenconder with $\Gamma = 15$ has proven a high comprehension of the touristic aspect of the dataset, the general idea of the sentence is exactly the same as the predicted topic with the appearance of time element which is not mentioned in the sentence literally, but that the model understood it as a chronological element for the description of the square in question.

5 Conclusion

In this paper, a Latent Dirichlet Allocation and BERT embedding had been mixed to build a Latent Space Representation model based on Auto Encoders to predict topics of a sentence that talks about a touristic shopping place. We have seen through the previous sections, the results of the model while variating the Gamma coefficient that is able to calibrate the embedding weights and LDA results, and we chose the best score based on the accuracy and loss results.
Furthermore, this work would introduce a new challenge related to clustering[19], whereby we will present a bunch of clustering models based on meta-heuristic approaches and clustering algorithms to evaluate accurately the most frequent topics in a giving shopping place.

Acknowledgments. We would like to address our acknowledgements to the Moroccan National Center for Scientific and Technical Research (CNRST) and the Moroccan Institute for Scientific and Technical Information (IMIST) for giving us the access to use their High-Performance Computing (Marwan).

Declarations

- Funding: Research reported in this article was supported by Al-Khawarizmi program of the Moroccan National Center for Scientific and Technical Research (CNRST) and the Moroccan Digital Development Agency (ADD) to promote artificial intelligence projects in Moroccan industries.
- Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.
- Ethical Approval: This study was approved by Moroccan National Center for Scientific and Technical Research (CNRST).
- Research involving Human Participants and/or Animals: Research involving Human Participants and/or Animals: There is no experimental procedures involving humans and animals that were conducted in this research.
- Informed consent: Informed consent is not applicable
- Author contributions: Ibrahim Bouabdallaoui implemented the solution using Tensorflow framework and wrote the main manuscript text, Samya Bouhaddour made data collection on TripAdvisor website, Fatima Guerouate, Chaimae Saadi and Mohamed Sbihi did the work of supervision as they are professors and experts in Artificial Intelligence providing us guidelines and strategies to follow up in the research work.

References

[1] Artemenko, O., Pasichnyk, V., Kunanets, N., Shunevych, K.: Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems. In: COLINS, pp. 259–271 (2020)

[2] Zsarnoczky, M.: How does artificial intelligence affect the tourism industry? VADYBA **31**(2), 85–90 (2017)

[3] Wang, T., Liu, Y.: Jsea: A program comprehension tool adopting lda-based topic modeling. International Journal of Advanced Computer Science and Applications **2**(3) (2017)

[4] Xie, Q., Zhang, X., Ding, Y., Song, M.: Monolingual and multilingual topic analysis using lda and bert embeddings. Journal of Informetrics **14**(3), 101055 (2020)

[5] Venugopalan, M., Gupta, D.: An enhanced guided lda model augmented with bert based semantic strength for aspect term extraction in sentiment analysis. Knowledge-Based Systems **246**, 108668 (2022)

[6] Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Han, J.: Topic discovery via latent space clustering of pretrained language model representations. In: Proceedings of the ACM Web Conference 2022, pp. 3143–3152 (2022)

[7] Fu, H., Zhou, S., Yang, Q., Tang, J., Liu, G., Liu, K., Li, X.: Lrc-bert: latent-representation contrastive knowledge distillation for natural language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12830–12838 (2021)

[8] Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S.J., Goodman, N.D.: Evaluating compositionality in sentence embeddings. arXiv preprint arXiv:1802.04302 (2018)

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

[10] Wu, J.-L., Chung, W.-Y.: Sentiment-based masked language modeling for improving sentence-level valence–arousal prediction. Applied Intelligence, 1–17 (2022)

[11] Chauhan, U., Shah, A.: Topic modeling using latent dirichlet allocation: A survey. ACM Computing Surveys (CSUR) **54**(7), 1–35 (2021)

[12] Rinke, E.M., Dobbrick, T., Löb, C., Zirn, C., Wessler, H.: Expert-informed topic models for document set discovery. Communication Methods and Measures **16**(1), 39–58 (2022)

[13] Nambiar, R.S., Gupta, D.: Dedicated farm-haystack question answering system for pregnant women and neonates using corona virus literature. In: 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 222–227 (2022). IEEE

[14] Yang, N., Jo, J., Jeon, M., Kim, W., Kang, J.: Semantic and explainable research-related recommendation system based on semi-supervised methodology using bert and lda models. Expert Systems with Applications **190**, 116209 (2022)

[15] Yu, M., Quan, T., Peng, Q., Yu, X., Liu, L.: A model-based collaborate filtering algorithm based on stacked autoencoder. Neural Computing and Applications **34**(4), 2503–2511 (2022)

[16] Palani, S., Rajagopal, P., Pancholi, S.: T-bert–model for sentiment analysis of micro-blogs integrating topic model and bert. arXiv preprint arXiv:2106.01097 (2021)

[17] Müller, V., Sieg, C., Linsen, L.: Uncertainty-aware topic modeling visualization. In: 2021 IEEE 6th Workshop on Visualization for the Digital Humanities (VIS4DH), pp. 12–18 (2021). IEEE

[18] Prakoso, D.W., Abdi, A., Amrit, C.: Short text similarity measurement methods: a review. Soft Computing **25**(6), 4699–4723 (2021)

[19] Thompson, L., Mimno, D.: Topic modeling with contextualized word representation clusters. arXiv preprint arXiv:2010.12626 (2020)