

Automated decision classification model for tax appeals commission in Morocco using latent dirichlet allocation

Soufiane Aouichaty¹, Yassine Maleh², Abdelmajid Hajami¹, Hakim Allali¹

¹VETE Laboratory, Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco

²LaSTI Laboratory ENSA Khouribga, Sultan Moulay Slimane University, Beni Mellal, Morocco

Article Info

Article history:

Received May 6, 2023

Revised Jun 5, 2023

Accepted Jun 17, 2023

Keywords:

Latent dirichlet allocation

National tax

Natural language processing

Text classification

Topic detection

Topic modeling

ABSTRACT

This research paper focuses on extracting and classifying information from the Moroccan National Tax Appeals Commission, which is presently nonexistent in the country's legal and tax landscape. This study examines 201 decisions selected from a pool of 562, released between 1999 and 2018, pertaining to corporate tax and involving 550 disputes spanning various corporate tax classifications. The paper aims to propose latent dirichlet allocation (LDA) for topic modeling and compare it with our previous results obtained from the bidirectional encoder representations from transformers (BERT) model. The findings suggest that the rulings can be classified into two primary classifications: those that uphold or reject the tax administration's position. The proposed model shows a good performance, achieving a precision of 9.25% and an accuracy of 9.51%. This highlights the effectiveness of both LDA and BERT models for understanding and classifying topics in tax decision analysis.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Soufiane Aouichaty

VETE Laboratory, Faculty of Sciences and Techniques, Hassan First University

Settat, Morocco

Email: aouichaty.soufiane@gmail.com

1. INTRODUCTION

The national tax appeals commission (NTAC) in Morocco is a legal mechanism with authority to decide on taxpayer appeals when disputes arise from reviewing their tax returns. The NTAC functions as a safeguard for taxpayers, mainly when there may be concerns about the administration's discretionary powers when conducting tax audits. The decisions of the NTAC are important sources of information for judges, inspectors, auditors, tax specialists, accountants, chartered accountants, lawyers, researchers, and decision-makers in the tax field. Decisions on tax appeals have been criticized for needing to be more timely and resource-demanding. The database creation process includes data mining to collect valuable information from judgments to conduct in-depth research of the decision-making process and streamline future litigation, including meta-analysis.

Recent progress in natural language processing (NLP) has enabled the creation of applications that can automate parts of the tax appeals decisions (TAD) procedure. With automated systems, technical jobs may be completed with higher accuracy, shorter search times, and lower prices. However, creative and judgment-based tasks still require human involvement. In addition, cutting-edge NLP machine-learning approaches can generate novel algorithms that faithfully simulate the range of human actions needed over a TAD's life cycle [1].

Topic modeling is a modern NLP technique that enables the exploration and prediction of hidden themes within large text datasets, regardless of the languages used [2], [3]. This method proves particularly valuable when dealing with extensive documents that are impractical to read and summarize manually.

Traditional methods of identifying the threshold for an obscure theme within a corpus can be challenging. However, topic modeling offers various approaches, such as latent semantic analysis (LSA) and non-negative matrix factorization (NMF), that leverage topic space vectors or probability distributions to unveil these latent themes. By employing topic modeling, researchers gain deeper insights into the underlying structures and patterns present in textual data, contributing to enhanced understanding and knowledge discovery.

Recently, many researchers have explored text classification via topic modeling techniques. Latent dirichlet allocation (LDA) is a probabilistic generative statistical model. Therefore, given a text, the model attempts to optimize its parameters by finding the ratio of one variable to another, given its value [4]. The field of text classification has seen the widespread application of LDA [5], a topic-modeling approach. The following is a brief discussion of some relevant literature.

Asiyabi and Dacu [6] evaluated the identification of semantic information in optical and synthetic aperture radar data at varying spatial resolutions across three situations. Recently, Lee [7] analyzed newspaper articles about aging to learn more about how Korean society talks about aging. Similarly, Twinandilla *et al.* [8] and Kondath *et al.* [9] aimed to generate multi-document text summaries based on extractive topic modeling for news articles, using a unique approach combining K-means clustering and LDA. Eligüzel *et al.* [10] looked into ways to overcome feature selection and classification issues. Poushneh and Rajabi [11] used the LDA method to classify customer reviews. The authors applied the method of LDA to classify the reviews provided by customers. The outcomes of their study add to the existing body of literature by revealing the fundamental processes that customers employ to comprehend reviews and connect them with corresponding numerical ratings.

Cvitanic *et al.* [12] studies comparing LDA with LSA found divergent results: the former was superior in learning descriptive themes, while the latter was superior at producing a compact semantic representation of documents and words in a corpus [13]. Finally, Henderson and Eliassi-Rad [14] changed the LDA method to process graph data instead of text corpora. The contrasts between text corpora and graph data in the actual world are reflected in these prescriptions. While previous studies considered many aspects of topic modeling, LDA [15] is the most important previous work in supervised classification, which our technique addresses. Find the right amount of naturally occurring subjects in the corpus that provide evidence supporting correctness and classification.

This research aims to analyze a pool of corporate tax rulings from the Moroccan National Tax Appeals Commission using advanced techniques such as LDA and bidirectional encoder representations from transformers (BERT) model to extract and classify information. The paper makes significant contributions in several areas. Firstly, it addresses the challenging task of extracting and classifying information from the Moroccan National Tax Appeals Commission, which needs to be improved in the country's legal and tax framework. Secondly, it provides insights into customers' fundamental processes to comprehend reviews and associate them with numerical ratings. Thirdly, it demonstrates the effectiveness of advanced methodologies such as LDA and BERT model in learning and classifying modeling topics in tax decision analysis. Fourthly, it proposes a model that achieves high precision and accuracy in classifying the rulings based on their stance toward the tax administration. Overall, the research provides valuable insights into tax decision analysis and highlights the importance of using advanced techniques for understanding complex legal and tax frameworks.

The paper is organized into five sections. The next section discusses some recent studies on text classification with topic modelling. Section three is further divided into four sub-sections. The first sub-section, data collection, outlines the process used to collect the corporate tax rulings from the Moroccan National Tax Appeals Commission. The second sub-section, data pre-processing, details the steps taken to clean and prepare the data for analysis. The third sub-section, data processing, describes advanced techniques such as LDA and the BERT model to extract and classify the information from the data. The fourth sub-section, Model Evaluation, evaluates the performance of the proposed model. The fourth section, results and discussion, presents and discusses the outcomes of the analysis, including insights into the underlying mechanisms used to interpret tax rulings. Finally, the last section, conclusion, summarizes the key findings of the research and highlights its contributions to the field of tax decision analysis.

2. METHODS

2.1. Data collection

We retrieved all tax appeal decisions related to corporate tax (CT) from the Moroccan National Tax Appeals Commission's Library, covering the period from its creation to the present day. In each decision, we identified from the files archived in paper format; therefore, we excluded all tax types (income tax (IT), value added tax (VAT)...) and extracted only CT. We also excluded decisions in Arabic. For data extraction, we used a nine-step workflow. Overall, the system accepts scanned PDF choices as input and produces decision

extraction, which includes a list of important data for each data item and proposed keywords in each sentence. The nine steps are explained below.

After a thorough textual analysis and the help of regular expression–REGEX [16]. We managed to extract the paragraphs of 8 key elements that contain the arguments and judgments of each stakeholder. Out of 562 decisions between 1999 and 2018 dealing with irregularities in several types of tax, we extracted 201 decisions dealing only with corporation tax, with 550 disputes dealing with various kinds of corporation tax. Table 1 shows the dataset attributes used in our research.

Table 1. Dataset attributes

Attributes	Example
Reference	383,664
Year	2008
Tax type	Corporate tax
Litige title	Turnover
Decision_LTC	Considering that the LTC has maintained the adjustments notified by the inspector
Argu_TA	Considering that the inspector based the reconstitution of the company's turnover on the accounting records and the prices charged by the company given its invoices and that he considered the waste rates, and that the company did not establish by evidence the proof of its allegations. However, a time limit was granted to it for this purpose by the sub-commission.
Argu_taxpayer	Whereas taxpayer E contested this decision. Whereas the company informed the sub-committee that it did not proceed with this reconciliation with the inspector Whereas the inspector based the reconstruction of the company's turnover on the accounting records and the prices charged by the company given its invoices and considered the waste rate.
Decision	After hearing the two parties and deliberating, the sub-commission decided to reconstitute the LTC's decision.

2.2. Data pre-processing

To classify texts based on their topics, unstructured data must be pre-processed to preserve relevant keywords, as shown in Figure 1 [17]:

- Markup tag filter: Eliminates column-specific tags in markdown. Filter input strings completely. Title and section screening will be done for all incoming documents.
- Stanford tagger: Tags document terms using part-of-speech (POS) tags. French, English, German, Spanish, and Arabic texts apply. Stanford NLP models underlie taggers.
- Stanford lemmatizer: Lemmatizes terms in the input documents with the stanford core NLP library.
- Punctuation eraser: Removes all punctuation characters of terms in the input documents.
- Number filter: Removes from the input documents all phrases that only number.
- N Char filter: Removes words from input documents with fewer than N characters.
- Stop word filter: Removes any terms included in the second input table and/or the stop word list from the input documents.
- Case converter: Converts all terms in the input documents to lower or upper case.

2.3. Data processing

To demonstrate the relevance of the LDA and BERT models for pre-training the language model, we trained using LDA BERT on a dataset of 550 tax disputes. Both models were trained on various dispute points to capture the nuances of the domain. Our specific focus was on disputes related to corporate taxation, allowing us to delve deeper into the intricacies of this particular area.

By narrowing the scope to corporate taxation disputes, we aimed to enhance the specificity and accuracy of the models in addressing the challenges and complexities unique to this field. This focused approach enabled us to analyze and classify the language used in these disputes more effectively, leading to a more robust and reliable automated decision classification model. Through our research and training process, we were able to establish the applicability and effectiveness of the LDA and BERT models in the context of tax dispute resolution. By leveraging the power of these models, we can gain valuable insights into the language patterns, arguments, and legal concepts surrounding corporate taxation disputes, ultimately contributing to more efficient and accurate decision-making within the National Tax Appeals Commission in Morocco.

2.3.1. The LDA

The LDA model, developed by Blei *et al.* [18]. A three-layer Bayesian probabilistic model of words, subjects, and documents is a document topic generation model used in unsupervised machine learning algorithms. It learns the underlying top to make the subject more meaningful and meaningful probabilities of words that appear in a series of texts to infer the distribution of words that define subjects. Bag-of-words convert each page to a vector of word frequencies. This simplifies text modeling. The bag-of-words approach

ignores word order, reduces complexity, and improves model building. Each document and subject are probability distributions of themes and words, respectively. Table 2 displays our study's hyperparameters used for the LDA model. These hyperparameters are crucial for controlling the behavior and performance of the LDA model during the training process.

Table 2. Hyperparameters LDA model

Attribute	Description	Values
Alpha	Smoothing parameter for document-topic distributions.	0.1
Beta	Smoothing parameter for each topic.	0.01
numIterations	The number of iterations of Gibbs sampling.	1,000
numThreads	The number of threads for parallel training.	1
numTopWords	After model estimation, the number of most probable words to print for each topic.	20
numTopics	Specifies the number of topics	10

2.3.2. Bidirectional encoder representations from transformers (BERT)

In 2018, Google's Jacob Devlin and his team announced a new language representation paradigm called BERT [19]. Since its inception, it has been the standard against which all other natural language processing studies are measured [20]. Unlike earlier language representation algorithms, BERT can predict words from both left and right contexts. Unsupervised models like BERT were also created, which can be trained with the enormous web-based plain text corpus for most languages. BERT excels at text classification and other natural language processing tasks due to this collection of abilities.

a) Hyperparameters for BERT-base

The 110M-parameter, 12-layer, 768-hidden, 12-self-attention-head Bert multi-cased L-12 H-768 A-12/2 base pre-trained model served as the basis for our experimentations. The Adam optimizer, considered one of the most reliable and popular options in deep learning, was used [21]. The model convergence was aided by Adam's optimizer with the warmup steps, which were low learning-rate updates. After many failed tries due to memory constraints, we finally settled on a hyper-parameter configuration that worked. The initialization ratio for warming up was 0.1, and the basic learning rate was $3e-5$. We used an empirically determined maximum of 7 epochs and a batch size of 32, and we preserved the best model from the validation set for further use. Figure 1 details our proposed model based on LDA.

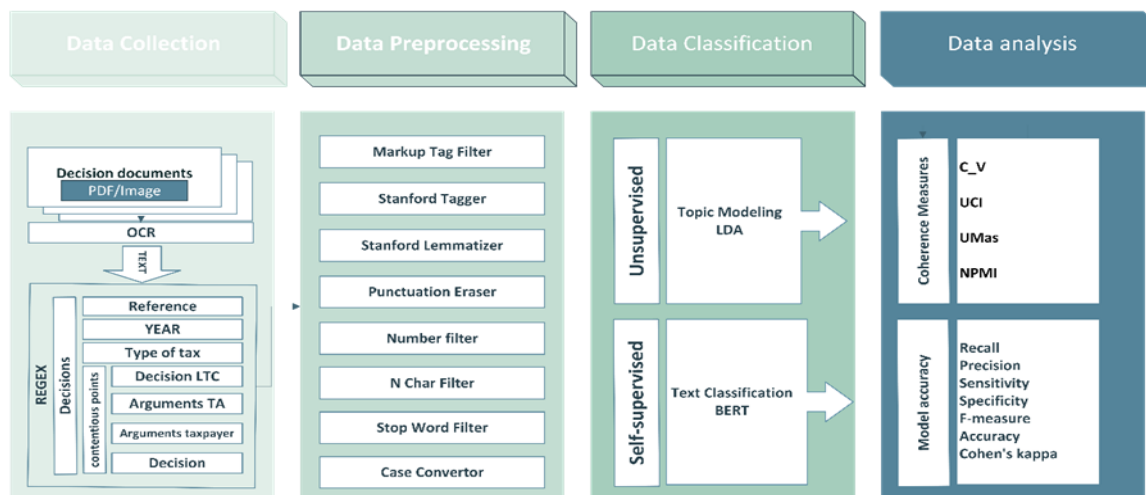


Figure 1. The proposed classification model

2.4. Model evaluation

2.4.1. Coherence measures

Topic coherence is a metric that evaluates topics' semantic coherence and interpretability by examining the degree of semantic similarity among the top words within each topic [22]. This measure is crucial in distinguishing between statistically derived topics and meaningful interpretations. In recent studies, four coherence measures for LDA have demonstrated strong correlations with human judgments when applied

to the same dataset [23], [24]. These coherence measures provide a quantitative assessment of the quality and coherence of topics generated by the LDA model, facilitating the evaluation and comparison of different topic models in a reliable and interpretable manner. Here is a short rundown of the various coherence measures and how they are arrived at:

- a) C_v method uses a sliding window, normalized pointwise mutual information (NPMI), and cosine similarity to indirectly establish a one-set segmentation of the essential phrases as shown in (1) [25].

$$coherence(V) = \sum_{(v_i, v_j) \in V} score(v_i, v_j, \epsilon) \tag{1}$$

V is a set of words describing the topic and ϵ indicates a smoothing factor that guarantees that score returns real numbers. (We will explore the effect of choosing ϵ ; the original authors used $\epsilon = 1$.)

- b) C_{uci} measure uses a sliding window, where PMI is the information communicated between every potential pair of top keywords as shown in (2).

$$score(v_i, v_j, \epsilon) = \log \frac{p(v_i, v_j) + \epsilon}{p(v_i)p(v_j)} \tag{2}$$

- c) C_{umass} uses document co-occurrence counts, one-preceding segmentation, and logarithmic conditional probability to affirm as shown in (3) [26].

$$score(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)} \tag{3}$$

where D(x, y) counts the number of documents with x and y, and D(x) counts x-containing documents. Instead of an external corpus, the UMass metric computes these counts over the same corpus used to train the topic models. More inherent. It checks if the models learned corpus data [26].

- d) C_{npmi} is a refined variant of C_{uci} coherence that makes use of NPMI as shown in (4).

$$score(v_i, v_j, \epsilon) = \frac{\log \frac{p(v_i, v_j) + \epsilon}{p(v_i)p(v_j)}}{-\log (p(v_i, v_j) + \epsilon)} \tag{4}$$

2.4.2. Model accuracy

Text classification evaluation measure uses precision, recall, and F-measure. Only evaluating a classifier on recall and accuracy is illogical. Analysis categories determine success. Documents are classified by a set of criteria. True positive (TP), false positive (FP), true negative (TN), and false negative (FN) are the relevant measures (FP). The assessment metrics obtained from these inputs are shown below to evaluate a prediction model. Conventional information retrieval measures including recall, precision, F-measure, accuracy, and Cohen's Kappa are presented in the (5)-(9) [27].

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = Sensitivity = \frac{TP}{TP+FN} \tag{6}$$

$$F-Measure = \frac{2*(Precision)*Recall}{Precision+Recall} \tag{7}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$Cohen's\ Kappa = \frac{2*(TP*TN-FN*FP)}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)} \tag{9}$$

3. RESULTS AND DISCUSSION

Examining how each person assigns a probability mass (or "weight") to their favorite words might illuminate who they are. Two representations of the relative importance of the subjects' initial words will be used to accomplish this. Each line in Figure 2 represents one of our 10 categories. Take note of the precipitous decline in topic weights as we proceed down the list of most significant terms in descending order. Figure 3 shows the overall weight allocated to each subject's top 10 words, revealing that the top 2 words in each topic are given substantially more weight than the remaining words.

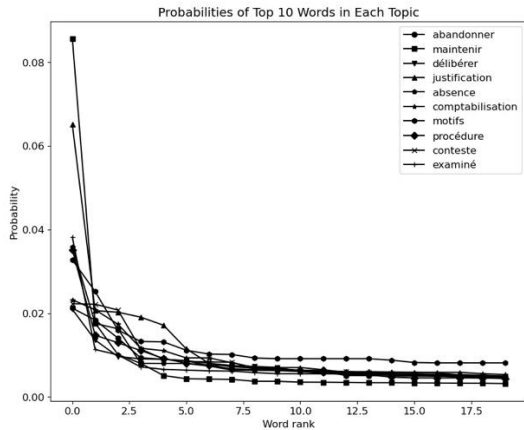


Figure 2. Probabilities of top 20 words in each topic

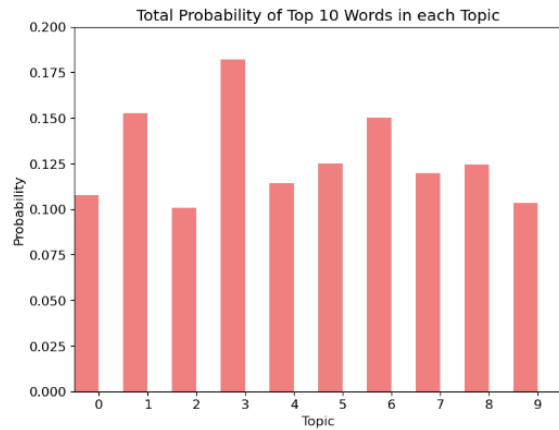


Figure 3. Total probability of top 10 words in each topic

Here we can see that the top 10 terms in our topic model only make up a modest percentage (between 15% and 18%) of the entire probability mass of their topic. So, although we may utilize the most frequently occurring terms to determine overarching themes for each issue, this part aims to understand how modifying these parameters impacts the topic model's features. Let's begin with the alphabet. Plotting the topic weight distribution for the same document under models fit with different alpha values should reveal the effect of modifying its value, as alpha is responsible for smoothing document preferences over subjects. We load topic models trained with varying values of alpha. Then we present the (sorted) topic weights for the TAD decision on a sample document using the models trained with high (alpha=10), original (alpha=0.1), and low (alpha=0.001) values. Figure 4 shows how the alpha parameter smoothes out the topic distribution for this article, where low alpha values result in a disproportionate amount of weight being given to a single subject, and high values of alpha result in a much more level distribution of weight amongst topics.

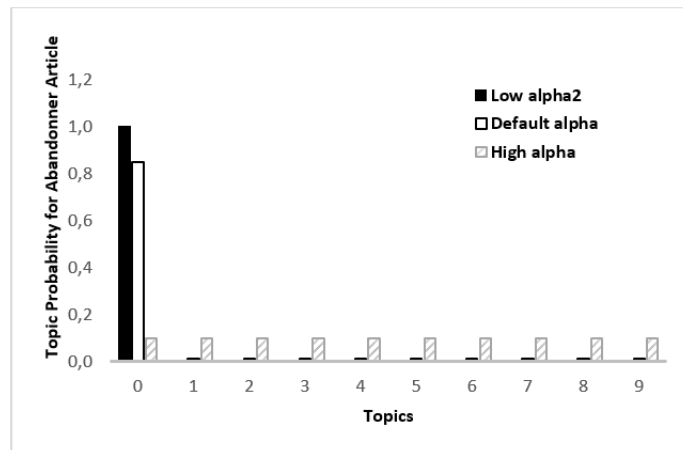


Figure 4. Topic probability for abandonner article

We anticipate that, similar to how charting topic weights for a text allowed us to observe the influence of alpha, plotting word weights for each subject will allow us to show the effect of varying eta. However, we just don't have the requisite number of words at our disposal. Instead, we will display the combined importance of each topic's most important 10 words and least important 500 words using the original eta model's weights of 0.1 and the low eta model's weight of 0.001.

Figures 5 and 6 show how the low eta model favors the most often occurring words within a subject while giving less importance to the least frequently occurring terms (or more intuitively, topics are composed of few words). High eta models, however, give less importance to the most often used words and greater importance to the least frequently used terms. Therefore, topics with a higher eta have a more even importance distribution over the whole lexicon.

The topic coherence metric is useful for evaluating the human interpretability of various topic models. The coherence score quantifies the degree to which the reader may understand a set of subjects, and the topic coherences capture this sweet spot. Figure 7 shows that the coherence score also increases when the number of topics increases. This suggests that selecting the model that produced the greatest coherence score before it plateaued or dropped significantly would be wise. In this example, there are two subjects. Coherence scores are measures used to evaluate the quality of topics generated by topic models. Figure 7(a) Coherence score C_V quantifies topic coherence by calculating the pointwise mutual information (PMI) between word pairs within a topic. Figure 7(b) Coherence score U_{MASS} also uses PMI but estimates word probabilities differently by considering unigram frequencies. Figure 7(c) Coherence score C_{UCI} estimates word probabilities based on relative frequencies and typically yields lower scores. Figure 7(d) Coherence score C_{NPMI} is a normalized version of PMI that addresses its limitations.

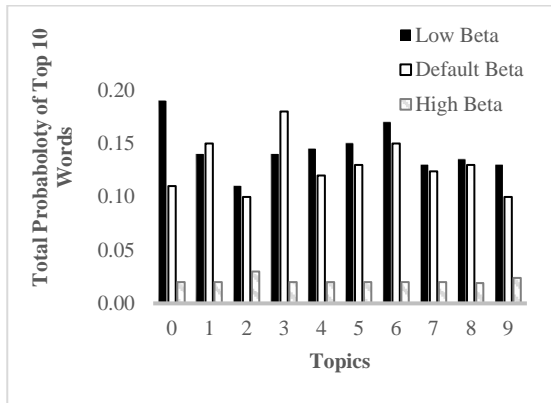


Figure 5. Total probability of top 10 words

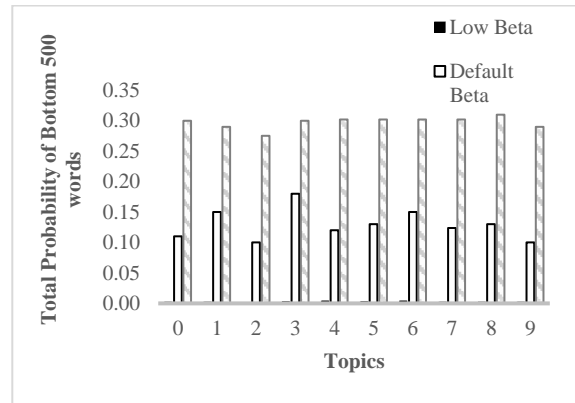
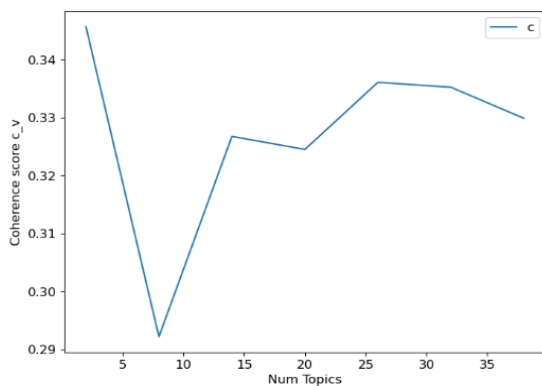
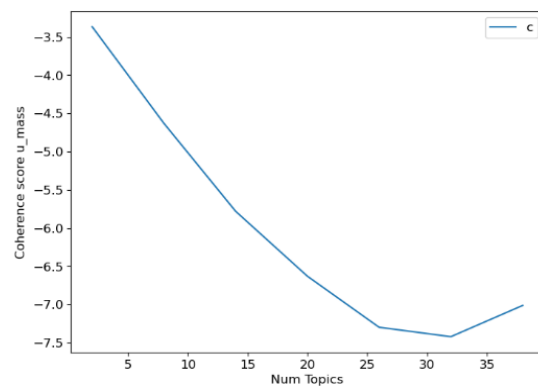


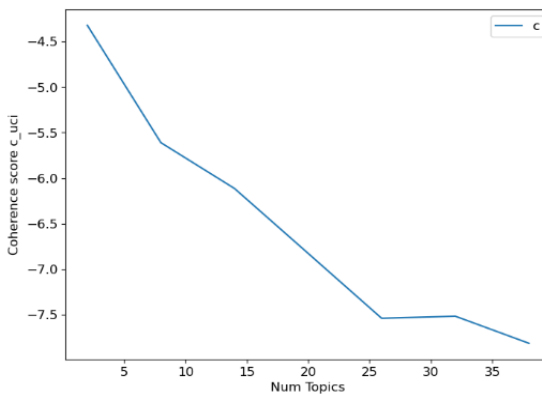
Figure 6. Total probability of bottom 500 words



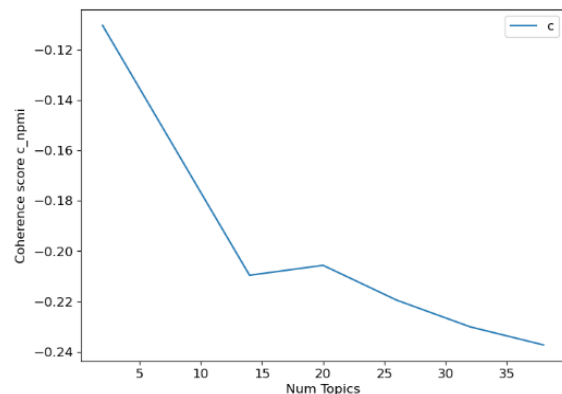
(a)



(b)



(c)



(d)

Figure 7. Coherence measures for topic models, including; (a) coherence score C_V , (b) coherence score U_{MASS} , (c) coherence score C_{UCI} , and (d) coherence score C_{NPMI}

3.1. Comparison with BERT

Figure 8 shows the accuracy of LDA model in comparison with BERT. BERT achieves 0.991 in terms of precision and 0.986 in accuracy. LDA model achieves 0.925 in terms of precision, and 0.951 for accuracy. Figure 9 shows the execution time of the two models: BERT (55.5146 minutes for BERT Learner and 0.959316 minutes for BERT Predictor) and 0.78 minutes for the LDA Model. As this is hardware-dependent, other users may have completely different run times. Since the measures are very close to 1, both models are efficient for text classification with processing advantages in time for LDA.



Figure 8. Accuracy of BERT and LDA

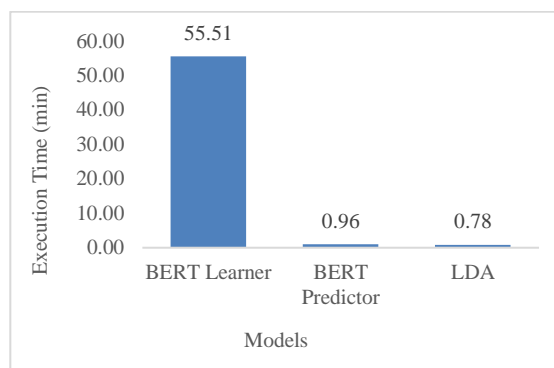


Figure 9. Models training time

The LDA approach used in this work has some significant limitations. The distribution and interpretation of topics are the main challenges. Topic modeling is sensitive to input data and analysis; adding additional documents and using text mining methods such as tokenization and stemming can produce completely different topics. As a result, the topics are generally a mixture of heterogeneous elements, and it is difficult to attribute truth to interpretation and validation in our corpus. However, depending on the source data and the collection period, different topics and themes can be deduced from our results. Thus, our results can provide an accurate picture of the decisions of NTAC.

4. CONCLUSION

In conclusion, this pilot study successfully explores the application of LDA topic modeling and compares it with the BERT model for text classification. The study specifically focuses on analyzing rulings from the Moroccan National Administrative Council (NTAC) to uncover hidden themes. By utilizing the beta and alpha posterior probabilities of LDA, we are able to identify key phrases associated with each topic and determine the relevance of documents to specific topics. The research findings highlight that taxpayers often face the choice of either accepting the tax administration's stance or pursuing litigation, indicating the potential of using LDA topic modeling to classify judgments by decision type.

It is worth noting that BERT exhibits greater effectiveness in tasks requiring a deep understanding of language, while LDA proves more effective for tasks centered around topic modeling and corpus structure comprehension. Each technique has its own set of strengths and weaknesses, and the selection between them will depend on the specific requirements of the project at hand. For future research, we plan to further evaluate and refine our LDA and BERT models by applying them to diverse datasets and languages. By conducting these experiments, we aim to demonstrate the efficiency and robustness of our models across various contexts and language domains. Additionally, exploring the integration of these models with other advanced techniques and approaches could potentially enhance their performance and broaden their applications in text classification and topic modeling tasks.




REFERENCES

- [1] N. A. I. Omoregbe, I. O. Ndaman, S. Misra, O. O. Abayomi-Alli, and R. Damaševičius, "Text messaging-based medical diagnosis using natural language processing and fuzzy logic," *Journal of Healthcare Engineering*, vol. 2020, pp. 1–14, Sep. 2020, doi: 10.1155/2020/8839524.
- [2] U. Chauhan and A. Shah, "Topic modeling using latent dirichlet allocation: a survey," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–35, Sep. 2022, doi: 10.1145/3462478.
- [3] A. J. Rawat, S. Ghildiyal, and A. K. Dixit, "Topic modelling of legal documents using NLP and bidirectional encoder representations from transformers," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1749–1755, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1749-1755.





- [4] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, and M. Cabanillas-Carbonell, "Sentiment analysis of tweets using unsupervised learning techniques and the k-means algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 571–578, 2022, doi: 10.14569/IJACSA.2022.0130669.
- [5] O. Iparraguirre-Villanueva et al., "Search and classify topics in a corpus of text using the latent dirichlet allocation model," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 30, no. 1, p. 246, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp246-256.
- [6] R. M. Asiyabi and M. Datcu, "Earth observation semantic data mining: latent dirichlet allocation-based approach," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2607–2620, 2022, doi: 10.1109/JSTARS.2022.3159277.
- [7] S. C. Lee, "Topic modeling of Korean newspaper articles on aging via latent dirichlet allocation," *Asian Journal for Public Opinion Research*, vol. 10, no. 1, pp. 4–22, 2022.
- [8] S. Twinandilla, S. Adhy, B. Surarso, and R. Kusumaningrum, "Multi-document summarization using k-means and latent dirichlet allocation (LDA) - significance sentences," *Procedia Computer Science*, vol. 135, pp. 663–670, 2018, doi: 10.1016/j.procs.2018.08.220.
- [9] M. Kondath, D. P. Suseelan, and S. M. Idicula, "Extractive summarization of Malayalam documents using latent dirichlet allocation: an experience," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 393–406, Jan. 2022, doi: 10.1515/jisys-2022-0027.
- [10] N. Eligüzel, C. Çetinkaya, and T. Dereli, "A novel approach for text categorization by applying hybrid genetic bat algorithm through feature extraction and feature selection methods," *Expert Systems with Applications*, vol. 202, p. 117433, Sep. 2022, doi: 10.1016/j.eswa.2022.117433.
- [11] A. Poushneh and R. Rajabi, "Can reviews predict reviewers' numerical ratings? The underlying mechanisms of customers' decisions to rate products using latent dirichlet allocation (LDA)," *Journal of Consumer Marketing*, vol. 39, no. 2, pp. 230–241, Feb. 2022, doi: 10.1108/JCM-09-2020-4114.
- [12] T. Cvitanic, B. Lee, H. I. Song, K. Fu, and D. Rosen, "LDA v. LSA: a comparison of two computational text analysis tools for the functional categorization of patents," *In International Conference on Case-Based Reasoning*, 2016.
- [13] S. H. Mohammed and S. Al-Augby, "LSA & LDA topic modeling classification: comparison study on E-books," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 19, no. 1, pp. 353–362, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.
- [14] K. Henderson and T. Eliassi-Rad, "Applying latent dirichlet allocation to group discovery in large graphs," in *Proceedings of the ACM Symposium on Applied Computing*, Mar. 2009, pp. 1456–1461, doi: 10.1145/1529282.1529607.
- [15] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. Murty, "On finding the natural number of topics with latent dirichlet allocation: some observations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6118 LNAI, no. PART 1, 2010, pp. 391–402.
- [16] C. Chapman, P. Wang, and K. T. Stolee, "Exploring regular expression comprehension," in *ASE 2017 - Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, Oct. 2017, pp. 405–416, doi: 10.1109/ASE.2017.8115653.
- [17] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014, vol. 2014-June, pp. 55–60, doi: 10.3115/v1/p14-5010.
- [18] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent dirichlet allocation," *Journal of machine Learning research*, pp. 993–1022, 2003.
- [19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [20] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, Dec. 2020, doi: 10.1162/tacl_a_00349.
- [21] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [22] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Feb. 2015, pp. 399–408, doi: 10.1145/2684822.2685324.
- [23] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 952–961.
- [24] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *In Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100–108.
- [25] S. Duraivel, L. Lavanya, and A. Augustine, "Understanding vaccine hesitancy with application of latent dirichlet allocation to reddit corpora," *Indian Journal of Science And Technology*, vol. 15, no. 37, pp. 1868–1875, Jun. 2022, doi: 10.17485/ijst/v15i37.687.
- [26] L. Stracqualursi and P. Agati, "Tweet topics and sentiments relating to distance learning among Italian Twitter users," *Scientific Reports*, vol. 12, no. 1, Jun. 2022, doi: 10.1038/s41598-022-12915-w.
- [27] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," *In 2017 international conference on computer, communication and signal processing (ICCCSP)*, Jan. 2017, doi: 10.1109/ICCCSP.2017.7944061.

BIOGRAPHIES OF AUTHORS







Soufiane Aouichaty    is a Ph.D. student at the faculty of sciences and technology, part of Hassan 1st University, Morocco, since 2019. He works as data analyst at ARTEMIS Company in Casablanca, Morocco, since 2010. He obtained his master of sciences in computer sciences, from Hassan 1st University, since 2012 and a master of management in information systems, from Hassan 1st University, since 2013. His research interest includes machine learning, data science, database engineering and text mining of legal domains. He can be contacted at email: s.aouichaty@uhp.ac.ma.







Prof. Dr. Yassine Maleh     is an associate professor of cybersecurity and IT governance at Sultan Moulay Slimane University, Morocco, since 2019. He is a double Ph.D. in computer sciences and IT Management. He is the founding chair of IEEE Consultant Network Morocco and founding president of the African Research Center of Information Technology and Cybersecurity. He is a senior member of IEEE He has published over than 140 papers (international journals, book chapters and conferences/workshops), 27 edited books, and 5 authored books. He is the editor-in-chief of the International Journal of Information Security and Privacy. He can be contacted at email: y.maleh@usms.ma.



Prof. Dr. Abdelmajid Hajami     is a Ph.D. in informatics and telecommunications, Mohamed V-Souissi University Rabat-Morocco in 2011. Ex trainer in regional centre in teaching and training, assistant professor at the Faculty of Science and Technology of Settat in Morocco member of LAVETE Lab at Faculty of Science and Technology of Settat. Research interests: security and QoS in wireless networks, radio access networks, next generation networks and ILE: informatics learning environments elearning. He can be contacted at email: a.hajami@uhp.ac.ma.



Prof. Dr. Hakim Allali     was born in Morocco in 1966. He received his Ph.D. degree from Claude Bernard Lyon I University (France) in 1993 and the “Docteur d’Etat” degree from Hassan II-Mohamedia University, Casablanca (Morocco) in 1997. He is currently professor at Faculty of Sciences and Technologies of Hassan 1st University of Settat (Morocco) and director of LAVETE Laboratory. He is executive manager and founder of IT Learning Campus. His research interests include technology enhanced learning, modeling, image processing, computer networking and GIS. He can be contacted at email: hakim-allali@hotmail.fr.