Enhancing diagonal comprehension with advanced topic modeling technique: DIAG-LDA

Fatima-Zahrae Sifi[1], Wafae Sabbar[1], Amal El Mzabi[2]
[1]Laboratory of Machine Intelligence (LIM), Faculty of Sciences and Technology, Hassan-II University, Mohammedia, Morocco
[2]Laboratory of Economic Performances and Logistics (PEL), Faculty of Law, Economic and Social Sciences, Hassan-II University, Mohammedia, Morocco

Article Info

Article history:

Received May 7, 2024
Revised Aug 6, 2024
Accepted Aug 11, 2024

Keywords:

Convolutional neural network
Diagonal reading
Latent dirichlet allocation
Natural language processing
Topic modeling

ABSTRACT

With the speed increase of reviews or other forms of text, natural language has the ability to convey large and complex amounts of information in relatively small communications. This capability is being leveraged by the machine-learning algorithm known as latent dirichlet allocation (LDA), which can be utilized to discover latent topics within documents. LDA can be also used to generate summaries or abstracts from a given set of documents. However, LDA can struggle to identify topics in short documents or in data with high levels of noise. This article will introduce a new method for topic modeling with LDA based on diagonal reading for sentences (DIAG-LDA). Primarily, the features are selected using the TF-IDF algorithm, and the highest relevant features are extracted using the confidence value. Besides, the classification step is executed utilizing the LDA classifier. Ultimately, we evaluate our model using the convolutional neural network algorithm. The experiment results show that DIAG-LDA performs well in identifying features from text data, achieving a 94.4%, and 89.5% in accuracy for the datasets on international economics and the political economy.

Corresponding Author:

Fatima-Zahrae Sifi
Department of Machine Intelligence, Faculty of Sciences and Technology, Hassan-II University
Mohammedia 28806, Morocco
Email: fatimazahrae.sifi@univh2c.ma

1. INTRODUCTION

The extraction of meaningful information from large datasets is a critical challenge in fields such as electronics, telecommunications and computer science. Efficient and accurate data analysis methods are necessary to handle the ever-growing volume of textual information [1]. To address this challenge, advanced techniques in natural language processing (NLP) [2] and information retrieval are employed. Specifically, combining latent dirichlet allocation (LDA) for topic modeling [3] with effective reading techniques such as Diagonal reading enhances the capability to extract and comprehend information from extensive text corpora. Furthermore, incorporating convolutional neural networks (CNN) aids in identifying trends within these large datasets.

The problem at hand involves the need for efficient methods to extract key themes and subtopics from vast amounts of text data [4]. This is addressed by integrating diagonal reading techniques with LDA [5], which automatically uncovers hidden topics within documents [6]. Diagonal reading helps in swiftly identifying crucial elements like topic sentences and headings [7], thus enhancing the topic modeling process. Previous works have utilized LDA to discover topics across various domains, including economy and politics

[8], by learning topic-word distributions through word co-occurrences [9]. However, these methods were usually insufficient or of limited accuracy/precision for large and complex datasets [10].

This paper proposes a novel approach that surpasses existing LDA models in the literature, demonstrating superior accuracy and precision in feature identification. Key aspects of this approach include the implementation of numerous pretreatment techniques [11] such as text lowercasing, negation handling, uniform resource locator (URL) and number removal, stop word elimination, tokenization, and lemmatization to enhance data quality. The data is then segmented into sentences based on the average number of words per sentence. Term frequency-inverse document frequency (TF-IDF) is applied to determine significant fixing points. Association rules are integrated using confidence values to select the most pertinent features. LDA is employed as a classifier to categorize each feature into five labeled topics, ensuring the capture of primary topics without overloading or diminishing their significance. Our recent approach surpasses alternative LDA models found in the literature, demonstrating superior accuracy and precision in feature identification. Finally, our proposal utilizes a CNN to treat text as a sequence, capturing relationships within sentences. By combining diagonal reading with LDA and CNN, our approach facilitates rapid comprehension of main themes and subtopics within text, revealing trends from extensive textual datasets.

The paper is organized as follows : in section 2, related work is reviewed. In section 3, the problem is formally defined and the proposed approach is outlined. In section 4, the results of our investigations are presented and compared with other methodologies in the field. Finally, a conclusion is given.

2. RELATED WORKS

NLP involves analyzing vast amounts of data in natural language [12]. Challenges in NLP include sentence misunderstanding, varied word meanings, informal language, and incorrect segmentation. Text preprocessing before analysis is crucial for clarity and optimal results [13].

Despite the abundance of available books, there is limited literature on diagonal reading. Yu [14] conducted a study comparing reading speeds for horizontal and vertical text, discovering that diagonal reading is faster and more efficient, particularly for small font sizes. Winsler *et al.* [15] propose that diagonal text processing differs from horizontal text due to containing more high-spatial frequency information. Regarding speed reading, the article [16] emphasizes its impact on text comprehension and retention. While sacrificing some comprehension for speed may be suitable for tasks like skimming, many situations require a slower pace for adequate understanding. The study [17] examines the impact of different orientation conditions on reading performance, measuring response times and accuracy rates, providing insights into how spatial orientation influences reading processes and cognitive processing in reading tasks. These studies provide that diagonal reading can be more efficient and faster than horizontal reading, depending on the font size, and that diagonal reading may have a different processing mechanism than horizontal reading.

LDA is a widely used method for identifying thematic structures in large text datasets. Initially introduced by Blei *et al.* in 2003, LDA has become prominent in NLP, information retrieval, and computational social sciences. Blei *et al.* [1] introduced LDA as a generative model representing documents as combinations of hidden topics. Their seminal paper outlines LDA's fundamental concepts, mathematical framework, and inference techniques, setting a benchmark for subsequent research. Addressing the challenge of large-scale topic modeling, another study [18] proposes a flexible and scalable approach using variational inference. This statistical method efficiently estimates latent topics in vast document corpora, essential for discovering thematic patterns in natural language processing and text-mining tasks. In social media analysis, Rohani *et al.* [19] investigate topic modeling's application. Their study addresses challenges like noisy text and scalability for large datasets. Evaluation is critical for assessing LDA models. Wallach *et al.* [20] propose a comprehensive framework, including metrics like perplexity and topic coherence, to measure model quality and interpretability.

3. METHOD

In diagonal reading using NLP, several issues were addressed while executing system processes and extracting words. Firstly, there's a risk of losing important information, resulting in a superficial comprehension of the text. In addition, quick reading aims to cover more text in less time, highlighting keywords for efficiency. Retaining information and enhancing retention are crucial. Our proposed approach addresses these concerns, as shown in Figure 1.

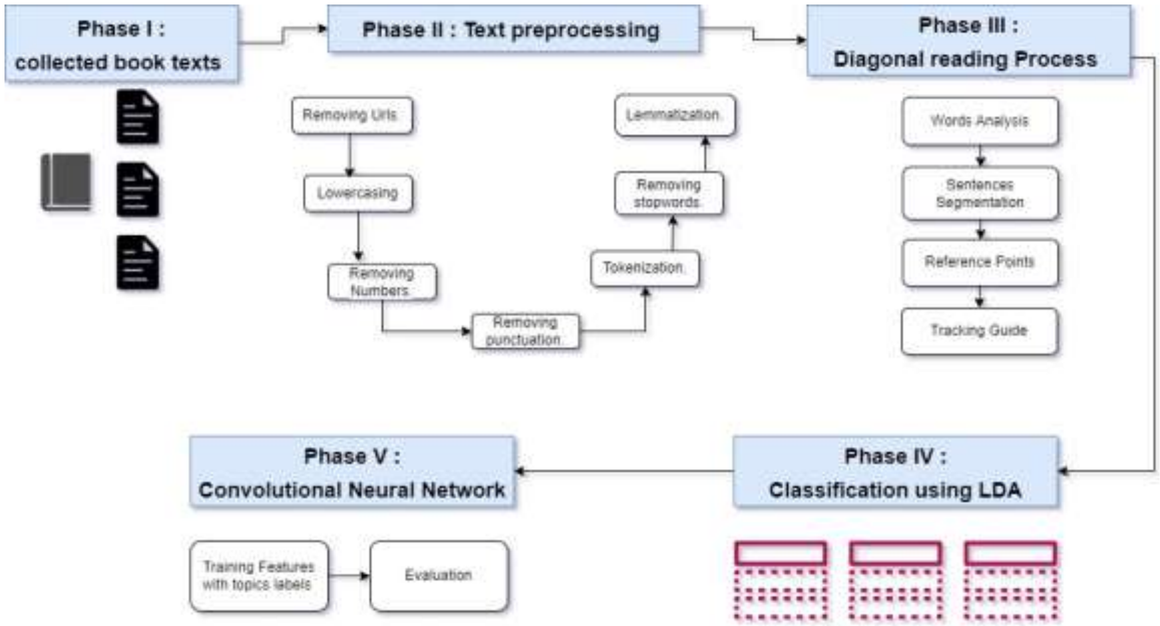

Figure 1. Global architecture of our DIAG-LDA method

3.1. Data pretreatment step

In the digital age, online books have become popular sources of reading material worldwide. However, non-preprocessed sentences in these books can hinder comprehension. To address this issue, publishers and platforms must prioritize robust pre-processing strategies for accuracy and readability.

Upon analyzing the document, we observed empty words and excessive punctuation. Our approach involved noise reduction for sentence normalization, including URL removal and punctuation elimination. Stop words lacking meaningful context were also removed, and all text was transformed to lowercase for consistency. Additionally, stemming was applied to unify word forms. Further details on these preprocessing steps are provided as displayed in Figure 2.

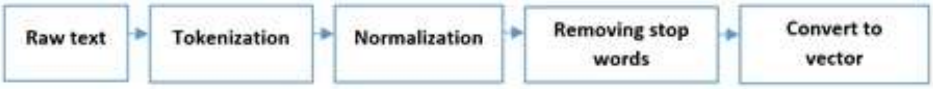

Figure 2. Text preprocessing process

3.2. Word analysis

Diagonal reading involves understanding words presented diagonally, hypothesized to rely on visual perception rather than linguistic processing. Research suggests diagonal reading can be faster and more efficient, especially with small text sizes [21]. In literature, shorter words are more frequently skipped than longer ones. For instance, three-letter words are omitted approximately 67% of the time, while 7–8 letter words are excluded only around 20% of the time [22]. In our approach, we adjust the word length threshold for removal based on dataset characteristics and research goals. By filtering out short words, we enhance the quality and efficiency of subsequent diagonal and NLP analyses.

3.3. Sentence segmentation

Previous studies have used sentence length to assess syntactic complexity [23], [24]. The study [23] introduces a computer-based approach that evaluates text readability by considering linguistic and structural features like sentence length, vocabulary complexity, and coherence. Forti *et al.* [24], examine language models producing longer or shorter sentences and evaluate their effect on narrative coherence and engagement. For our proposal, sentence segmentation involves calculating the mean word count per sentence in the processed dataset:

$$\mathrm{Average(w/s)} = \frac{\mathrm{Sum(w)}}{\mathrm{Sum(s)}} \quad (1)$$

Average(w/s): average words per sentence; sum(w): overall count of words in the text after preprocessing; sum(s): overall count of sentences in the raw text. Once the average word count per sentence is calculated, it can be used as a threshold to segment the text. If the average is calculated to be 10, our dataset is divided into a sequence of words that contains 10 words in each sentence.

3.4. Reference point

During reading, our eyes don't continuously move along the text, relying instead on stable positions for visual perception [25]. To optimize reading, we aim to expand fixational gaze and minimize fixations, representing the information within a fixation. Researchers link eye fixation duration with deeper cognitive engagement [26]. Our method uses TF-IDF to identify reference points:

$$\mathrm{TF-IDF(m,s)} = \mathrm{TF(m,s)} \times \log\left(\frac{\mathrm{T}}{\mathrm{DF(m)}}\right) \quad (2)$$

m: word; s: sentence; TF(m, s): term frequency of m within sentence s; DF(m): number of sentences where the m word appears; T: total count of sentences. Distinctive features in our diagonal algorithm are identified using relevant words with high TF-IDF rates, aiding cognitive processing. Inspired by diagonal reading methods, our study selects two reference points per sentence based on their high TF-IDF scores.

3.5. Tracking guide

Untrained readers often experience regressions, consciously rereading text [27], [28]. Utilizing visual guides aims to ease analysis by aiding text tracking skipping crucial words [29], ultimately enhancing reading speed [30]. After identifying two reference points per sentence, association rules are used to reveal word relationships. This iterative process involves eliminating words that fall below a certain minimum threshold and analyzing associations. Formally, the confidence value for an association rule (word1→word2) can be defined as:

$$\mathrm{Confidence(m1} \rightarrow \mathrm{m2)} = \frac{\mathrm{N(m1,m2)}}{\mathrm{N(m1)}} \quad (3)$$

N(m1, m2): number of transactions containing both word1 and word2; N(m1): number of transactions containing w1; confidence assesses the conditional probability of word1 given word2, ranging from 0 to 1. Higher confidence scores signify stronger associations between words within the same sentence. For instance, 'stock-exchange' and 'financial-market' exhibit a strong association, resulting in high confidence values.

LDA, first presented in 2003 [2], stands as one of the pioneering topic models. Operating as a 'bag of words' model, LDA focuses solely on term-document frequency, disregarding text organization [31]. The visual depiction of LDA is illustrated in Figure 3.

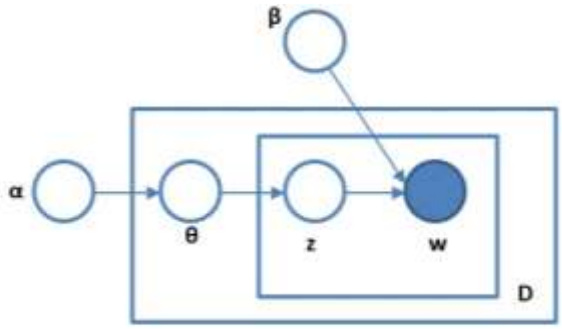

Figure 3. Graphical depiction of LDA

D: count of documents; α: distribution of document-topic θ; M: words present in the document; β: distribution of words within each topic; w: word; θ: topic distribution per document; z: topics to which the word is associated.

Our algorithm computes the confidence threshold by considering confidence values for each word pair and a percentile threshold. Rules surpassing the confidence threshold are considered strong and practical. Figure 4 outlines the process of determining the appropriate confidence threshold value.

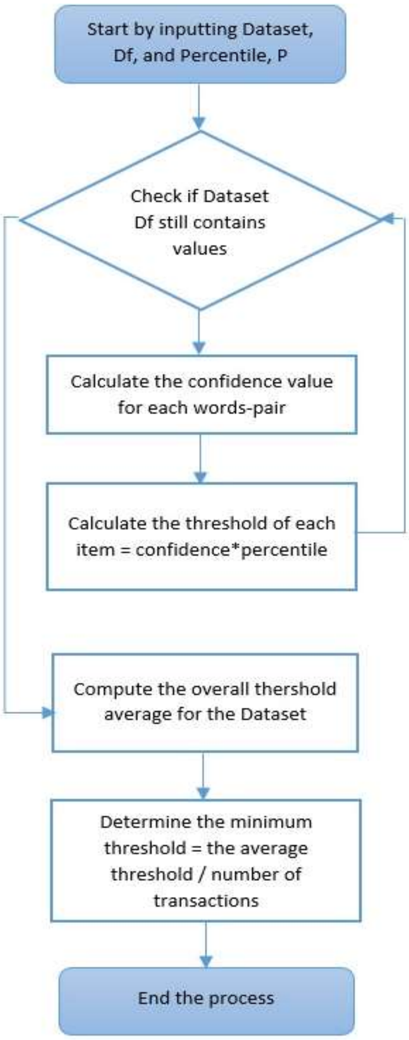
Start by inputting Dataset, Df, and Percentile, P
Check if Dataset Df still contains values
Calculate the confidence value for each words-pair
Calculate the threshold of each item = confidence*percentile
Compute the overall thershold average for the Dataset
Determine the minimum threshold = the average threshold / number of transactions
End the process

Figure 4. Confidence threshold flow

To evaluate results, we used coherence and perplexity as evaluation metrics:
Coherence: measures the coherence among words within a topic, indicating the interpretability of topics generated by LDA. Coherence can be computed using pointwise mutual information (PMI). We calculate the PMI scores for all word pairs within a topic. We take the average to obtain the coherence score for each topic. We compute eventually the overall coherence score by averaging the coherence scores across all topics. The formula for word pairs (m1, m2) is:

$$\mathrm{PMI}(\mathrm{m}_1, \mathrm{m}_2)=\log_2\left(\frac{\mathrm{P}(\mathrm{m}_1,\mathrm{m}_2)}{\mathrm{P}(\mathrm{m}_1)\cdot\mathrm{P}(\mathrm{m}_2)}\right) \quad (4)$$

P(m1, m2): the likelihood of word co-occurrence m1 and m2; P(m1) and P(m2): the likelihood of individual words m1 and m2, correspondingly. Perplexity: is a widely used measure to evaluate the quality of LDA models. It assesses the model's predictive capability. It is calculated using (5).

$$\mathrm{Perplexity}(\mathrm{D})=\exp(-\textstyle\sum\log(\mathrm{p}(\mathrm{w}))) \quad (5)$$

D: the dataset; p(w): the probability assigned to word w.

3.6. Training and evaluation dataset

For extensive economic data classification, deep learning models such as CNN is favored. CNN architecture typically includes an embedding layer, a convolutional layer, and a max-pooling layer for output, as depicted in Figure 5. The preprocessed data is converted into a sequence of words, which are mapped to corresponding word vectors indexes. These indexes serve as input to the CNN model. Our convolution utilizes 64 filters, with each filter group reducing dimensionality through 1-max pooling. A fully connected layer with a dense function and rectified linear unit (ReLU) activation computes probabilities for the 5 labels in our case. The model is prepared through training using 5 epochs and a batch size of 32. Evaluation metrics for our model include accuracy, precision, recall, and the F1-measure.

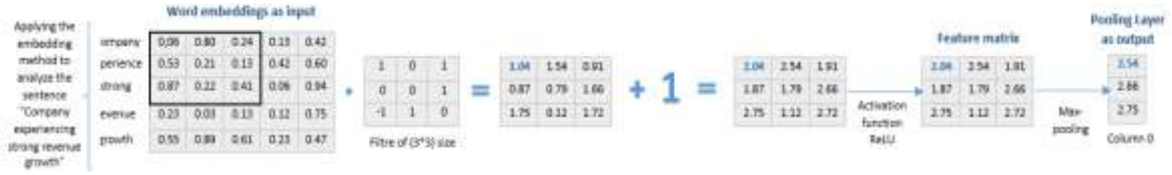

Figure 5. Example of one filter of CNN classifier in a words-sentence

4. RESULTS AND DISCUSSION

This study investigated the effects of DIAG-LDA on text comprehension and information retention. While earlier studies have explored the impact of various reading techniques on learning outcomes, they have not explicitly addressed its influence on the effectiveness of LDA in identifying thematic structures within texts. By focusing on how diagonal reading, a method that involves skimming and scanning text diagonally, interacts with LDA's topic modeling, this research aims to fill a gap in understanding how reading strategies can enhance or impede the algorithm's ability to uncover coherent themes in large textual datasets.

To assess the efficacy of the model, we tested our approach using metrics like accuracy, precision, recall, and execution time. We evaluated its performance with the DIAG-LDA classifier on economic books and analyzed results from the CNN classifier within our model. Then, we compared our framework's performance with existing methods in similar studies. The first dataset within this research is a Book [32] of 3,930 sentences in the English language about the political economy. Additionally, we have evaluated the proposed method on the international economics book [33] of 10,918 sentences in the English language to make a literature comparison. The results regarding coherence, and time execution of DIAG-LDA and LDA are presented in Tables 1 and 2.

Table 1. Coherence of different datasets

Model	Coherence (%) International economics	Coherence (%) Political economy
DIAG-LDA	75.65	74.86
LDA	41.32	36.46

Table 2. Execution time of different datasets

Model	Execution time (seconds) International economics	Execution time (seconds) Political economy
DIAG-LDA	1.395	1.087
LDA	106.256	42.403

The DIAG-LDA method achieved a coherence of 75.64% and a timing of 1.395 seconds toward the international economics dataset and a coherence of 74.86%, a timing of 1.087 seconds toward the political economy dataset. Our method outperforms other techniques in coherence and execution time. This superiority is due to leveraging LDA with diagonal approach, which incorporates word confidence to define topics and assign appropriate words to sentences. The harmony of LDA's topic classification within our contextual framework, aided by a CNN classifier, further enhances its effectiveness.

Regarding coherence, the presence and quality of semantic relationships between words greatly impact their value. Well-associated words contribute positively. Additionally, execution timing is determined by the complexity of the employed algorithm, the volume of data being processed, and the hardware resources available. Faster execution is often achieved through optimized algorithms, parallel processing, and efficient resource allocation. However, using the basic LDA for our datasets, we achieved a coherence of 41.32% and timing of 106.256 seconds for international economics dataset and a coherence of 36.46%, timing of 42.403 seconds toward political economy dataset. DIAG-LDA has the best scores in coherence and is level-headed in time execution.

Although the benefits of LDA, our approach surpassed the performance of these enhanced classification techniques. This can be attributed to the utilization of a diagonal method utilizing the LDA topic modeling approach. Our methodology incorporates contextual data in the form of word confidence, presenting confidence values for every word within a topic. Through considering prominent words with higher values, the topic is defined, and subsequently, appropriate words are assigned to represent specific sentences within the text. This situationally representative diagonal method yielded better outcomes in comparison to the aforementioned approaches. The superior effectiveness of the DIAG-LDA approach can also be attributed to its consideration of a set of harmonious procedures for topic classification within the contextual framework. This was accomplished by implementing a CNN classifier to evaluate the process.

Figure 6 display the accuracy rate of CNN classifier applied to our datasets, as a function of 5 epochs. In the instance of the CNN classifier utilized to DIAG-LDA for the international economics dataset, the accuracy score achieved of 94.4% attained after 5 epochs. The DIAG-LDA for the international economics dataset with CNN classifier achieved the highest accuracy result. Similarly, CNN classifier applied to DIAG-LDA for the political economy dataset achieved 89.5% of accuracy score after 5 epochs. The CNN measures of DIAG-LDA for the political economy dataset also goes up to a satisfactory level. In the other side, the CNN classifier applied to LDA for the international economics dataset exhibited the accuracy of 67% after 5 epochs. Furthermore, the CNN classifier applied to LDA for the political economy dataset revealed the accuracy of 52.9% reached after 5 epochs.

In terms of precision, the CNN classifier applied to DIAG-LDA on the international economics dataset achieved, in Figure 7, a precision of 94.7% after 5 epochs, while for the political economy dataset, the precision was 93.3% after the same number of epochs. These results show the model's high ability to accurately predict classes in both domains. Conversely, the CNN classifier applied to LDA for the international economics dataset had a precision of 66.7%, and for the political economy dataset, the precision was 54.3% after 5 epochs.

For recall in Figure 8, the CNN classifier applied to DIAG-LDA on the international economics dataset achieved a recall score of 94.5% after 5 epochs, whereas for the political economy dataset, the recall was 89.4% after the same number of epochs. These results indicate the model's effectiveness in identifying positive instances within each dataset. In contrast, the CNN classifier applied to LDA for the international economics dataset had a recall of 66.8%, and for the political economy dataset, the recall was 52.5% after 5 epochs. Regarding the F1-measure, the CNN classifier applied to DIAG-LDA on the international economics dataset, in Figure 9, attained an F1-measure of 94.4% after 5 epochs, while for the political economy dataset, the F1-measure was 90.3% after the same number of epochs. These findings highlight an effective balance between precision and recall in classification for both domains. In contrast, the CNN classifier applied to LDA for the international economics dataset achieved an F1-measure of 66.7%, and for the political economy dataset, the F1-measure was 52.8% after 5 epochs.

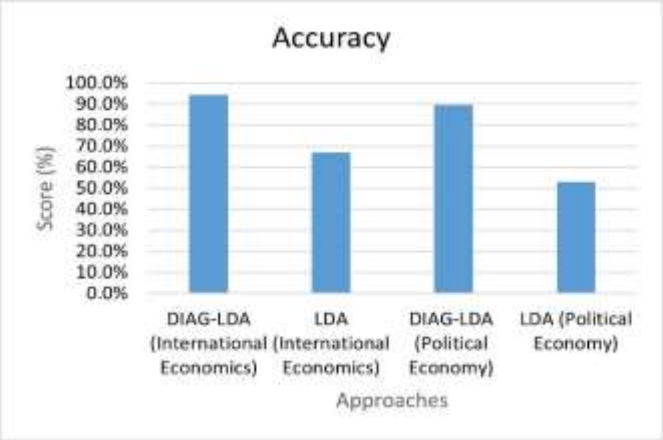

Figure 6. Accuracy comparison of CNN on DIAG-LDA and LDA of datasets

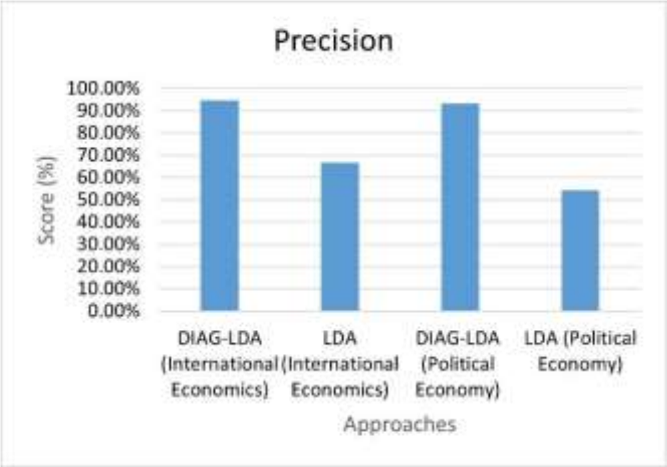

Figure 7. Precision comparison of CNN on DIAG-LDA and LDA of datasets

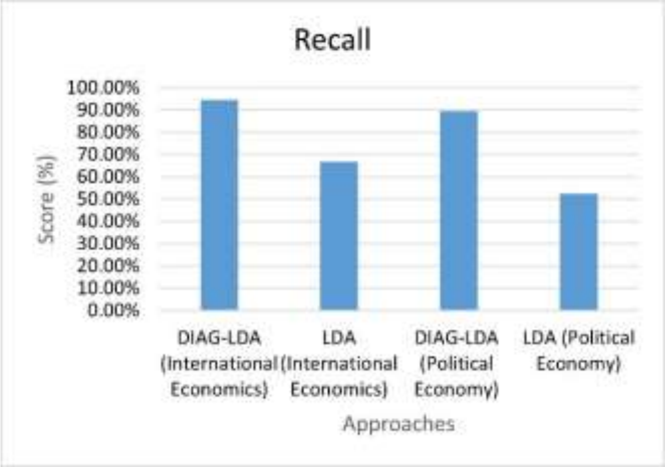

Figure 8. Recall comparison of CNN on DIAG-LDA and LDA of datasets

When training our datasets with more or less than 5 epochs, variations in results emerge due to the interplay between data volume and training duration. This potentially results in a risk of inadequate uncovering of diverse features found within the data. However, training with 5 epochs on a larger dataset offers extended learning opportunities, allowing the model to refine its understanding of data nuances and produce more accurate predictions. Thus, it is important to find a balance to achieve optimal performance.

The comparison results among the CNN classifier indicate that the DIAG-LDA for the dataset of international economics and DIAG-LDA for the dataset of political economy achieved higher results than LDA for the dataset of international economics and LDA for the dataset of political economy, which suggests that our classifier model is generally effective in distinguishing between classes, and it also indicates strong performance in categorization. The DIAG- LDA for the international economics dataset with CNN classifier achieved the highest classification results. Similarly, the CNN measures of DIAG-LDA for the political economy dataset also goes up to a satisfactory level.

Furthermore, in addition to our comparison of DIAG-LDA with traditional LDA, we compare it with three other LDA adaptations mentioned in the research: Jo and Oh [34] proposed SLDA for extracting aspects from user reviews. The objective of this methodology is to identify topics on a sentence level, as well as assign likelihoods of topics to words in every sentence. Yan *et al.* [35] suggested a modification of LDA for the short text that they designate the biterm topic model (BTM) for solving data sparsity that makes conventional topic models on short texts. The work of Ozyurt *et al.* [36] proposed an innovative modification of LDA algorithm for topic extraction based on sentence segmentation and sentiment analysis (SS-LDA).

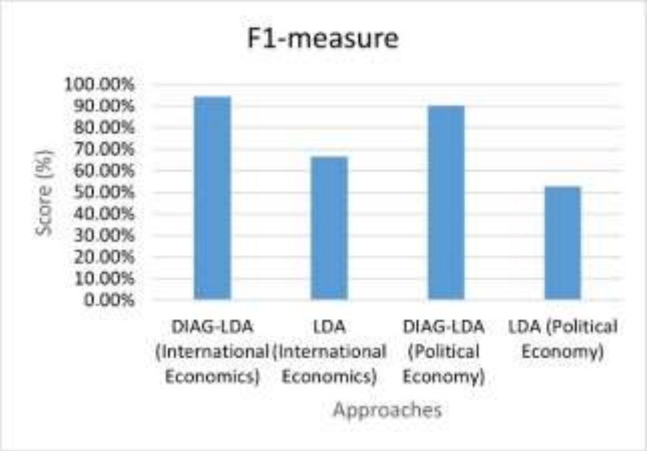

Figure 9. F1-measure comparison of CNN on DIAG-LDA and LDA of datasets

We implemented and validated three LDA adaptations on datasets of political economy [32] and international economics [33]. We made some adjustments to the code to incorporate DIAG-LDA and to ensure a just and equitable comparison with BTM, sentence-LDA, and SS-LDA. We executed these LDA adaptations on our two datasets and compared their success in topic modeling. According to the results shown in Figures 10 and 11, DIAG-LDA outperforms other methods with the highest scores.

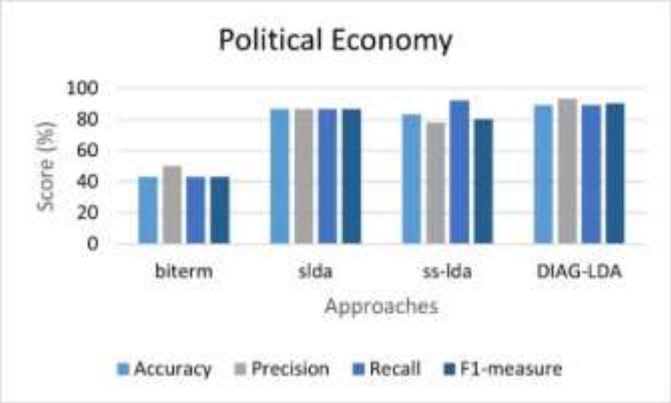

Figure 10. Evaluation metrics comparison of CNN classifier on DIAG-LDA, SS-LDA, SLDA, and biterm approaches applied to political economy dataset

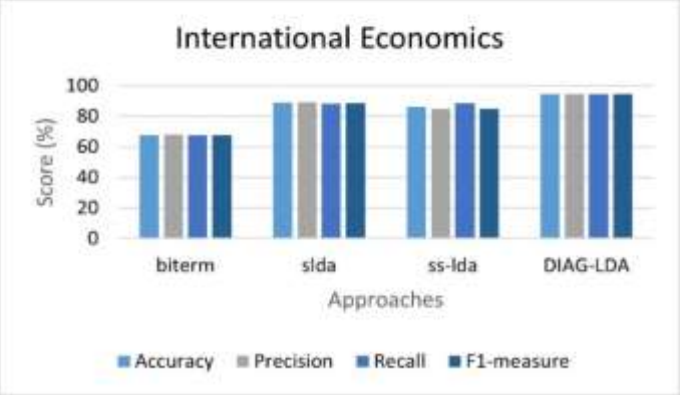

Figure 11. Evaluation metrics comparison of CNN on DIAG-LDA, SS-LDA, SLDA, and biterm approaches applied to international economics dataset

While SS-LDA demonstrates commendable recall, its precision value is comparatively lower. The SLDA method achieved satisfactory outcomes regarding accuracy, precision, recall, and F1-measure applied to the international economics, and the political economy datasets. Thus, an accuracy value of 86.76%, precision score of 86.75%, recall score of 86.68%, and F1-measure value of 86.71% for the political economy dataset, although DIAG-LDA achieved the accuracy value of 89.5%, precision score of 93.3%, recall score of 89.4%, and F1-measure value of 90.3% as well for the political economy dataset. Therefore, DIAG-LDA shows superior reaching in topic extraction. Nevertheless, the performance of the BTM in topic extraction is disappointing. This underperformance highlights the limitations of the BTM in effectively capturing and extracting topics from text data. It reaffirms the need for more sophisticated and advanced approaches, such as DIAG-LDA, to achieve better results in topic extraction tasks.

By running comprehensive literature comparisons, our study provides valuable perceptions of the achievement of DIAG-LDA to assess the impact of topic modeling on text analysis efficiency. The results clearly demonstrate the superiority of DIAG-LDA in topic extraction when compared to alternative approaches, especially concerning accuracy, precision, recall, and F1-measure. Our methodology also enriches the existing knowledge base and underscores the potential of DIAG-LDA as a powerful and puissant implement for topic extraction. However, further and in-depth studies may be needed to confirm its effectiveness in optimizing LDA’s performance, especially regarding the accuracy of thematic identification and the comprehensive interpretability of topic models generated from various reading strategies.

Our study demonstrates that DIAG-LDA offer more resilience in thematic extraction compared to other methods such as SS-LDA, SLDA, and BTMs. Unlike SS-LDA and SLDA, which often struggle with maintaining coherence across diverse textual datasets, and biterm, which can be sensitive to sparsity in large corpora, our method shows improved robustness in identifying and maintaining coherent topics. Future studies may explore integrating advanced preprocessing steps with diagonal reading methods to further enhance LDA’s performance. This could involve employing sophisticated text normalization procedures, such as semantic enrichment and syntactic parsing, to better prepare the data for analysis. By combining these advanced preprocessing strategies with DIAG-LDA approach, we could focus on feasible ways of producing more stable and interpretable topic models across varying types of textual data.

Recent observations suggest that DIAG-LDA can lead to more stable and coherent topic models, which represents a significant advancement in the field of computational linguistics. Our findings provide conclusive evidence that this phenomenon is associated with improvements in thematic coherence and interpretability, rather than being driven by elevated numbers of topics or increased computational resources. This suggests that the enhanced performance observed with our approach is due to the effectiveness of the reading strategy in preprocessing and structuring the data for LDA, rather than simply increasing the complexity or scale of the model. Moreover, our approach DIAG-LDA stands out for its ability to effectively manage and interpret large volumes of textual data. By capturing the essence of the original content, our methodology enables comprehensive and rapid understanding of the entire dataset.

5. CONCLUSION

This study introduces DIAG-LDA, a novel topic modeling approach, utilizing diagonal reading. While previous methods often relied on manual labeling, our framework combines diagonal reading and LDA for extracting meaningful topics. Using English language datasets of international economics and political economics books, we address challenges posed by text length and complexity. While tailored for English, DIAG-LDA's adaptability extends to other languages. Additionally, we enhance classification performance by leveraging sentence averages and TF-IDF for feature extraction. We evaluate our proposed framework, which achieves an accuracy of 94.4%, a precision of 94.7%, and an F1-measure of 94.4% using the CNN classifier applied to the DIAG-LDA approach. Comparative analysis against the traditional LDA approach further highlights the superiority of our proposed method. While our method shows promise in economic books, it could extend to other domains like customer feedback reviews, educational texts, and scientific papers. In future studies, we aim to investigate additional methods for topic modeling for analyzing topics for labeling and enhancing classification performance. A possible avenue for future work is to ameliorate the sentence segmentation task, such as utilizing new measures or refining existing approaches, we will explore incorporating median-based techniques for enhanced accuracy and robustness. Accurate sentence segmentation is crucial for the overall success of the DIAG-LDA method.

The development of our summarization DIAG-LDA method represents a significant advancement in the field of computational linguistics, particularly in managing and comprehending large-scale textual data. The elements of our methodology encapsulate the essence of the original content, facilitating a comprehensive yet expedited understanding of the entire dataset. By bridging the gap between extensive content and accessible comprehension, our approach promises to revolutionize how researchers approach textual analysis, paving the way for new avenues of inquiry and discovery in the era of big data.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.7551/mitpress/1120.003.0082.

[2] A. Alarfaj and Y. Alshumaimeri, “The effect of a suggested training program on reading speed and comprehension of Saudi female university students,” *Procedia - Social and Behavioral Sciences*, vol. 31, pp. 612–628, 2012, doi: 10.1016/j.sbspro.2011.12.114.

[3] W. Wagner, “Steven Bird, Ewan Klein and Edward Loper: natural language processing with Python, analyzing text with the natural language Toolkit,” *Language Resources and Evaluation*, vol. 44, no. 4, pp. 421–424, Dec. 2010, doi: 10.1007/s10579-010-9124-x.

[4] S.-T. Park and C. Liu, “A study on topic models using LDA and Word2Vec in travel route recommendation: focus on convergence travel and tours reviews,” *Personal and Ubiquitous Computing*, vol. 26, no. 2, pp. 429–445, Apr. 2022, doi: 10.1007/s00779-020-01476-2.

[5] Z. Shi, “Perceptual intelligence,” in *Intelligence Science*, Elsevier, 2021, pp. 151–213.

[6] D. S. McNamara, “Speed reading,” in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2001, pp. 14887–14890.

[7] K. Rayner, “Eye movements in reading: models and data,” *Journal of Eye Movement Research*, vol. 2, no. 5, Mar. 2009, doi: 10.16910/jemr.2.5.2.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, “Probabilistic methods,” in *Data Mining*, Elsevier, 2017, pp. 335–416.

[9] D. M. Blei, “Probabilistic topic models,” *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.

[10] S. Alam and N. Yao, “The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis,” *Computational and Mathematical Organization Theory*, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8.

[11] C. P. Chai, “Comparison of text preprocessing methods,” *Natural Language Engineering*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[12] S. Sun, C. Luo, and J. Chen, “A review of natural language processing techniques for opinion mining systems,” *Information Fusion*, vol. 36, pp. 10–25, Jul. 2017, doi: 10.1016/j.inffus.2016.10.004.

[13] A. Kulkarni and A. Shivananda, “Advanced natural language processing,” in *Natural Language Processing Recipes*, Berkeley, CA: Apress, 2021, pp. 107–133.

[14] D. Yu, “Comparing reading speed for horizontal and vertical English text,” *Journal of Vision*, vol. 10, no. 2, pp. 1–17, 2010, doi: 10.1167/10.2.21.

[15] K. Winsler, P. J. Holcomb, K. J. Midgley, and J. Grainger, “Evidence for separate contributions of high and low spatial frequencies during visual word recognition,” *Frontiers in Human Neuroscience*, vol. 11, Jun. 2017, doi: 10.3389/fnhum.2017.00324.

[16] K. Rayner, E. R. Schotter, M. E. J. Masson, M. C. Potter, and R. Treiman, “So much to read, so little time,” *Psychological Science in the Public Interest*, vol. 17, no. 1, pp. 4–34, May 2016, doi: 10.1177/1529100615623267.

[17] N. Davidenko and A. Ambard, “Reading sideways: effects of egocentric and environmental orientation in a lexical decision task,” *Vision Research*, vol. 153, pp. 7–12, Dec. 2018, doi: 10.1016/j.visres.2018.08.006.

[18] K. Zhai, J. Boyd-Graber, N. Asadi, and M. L. Alkhouja, “Mr. LDA: a flexible large scale topic modeling package using variational inference in mapreduce,” in *Proceedings of the 21st international conference on World Wide Web*, Apr. 2012, pp. 879–888, doi: 10.1145/2187836.2187955.

[19] V. A. Rohani, S. Shayaa, and G. Babanejaddehaki, “Topic modeling for social media content: A practical approach,” in *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, Aug. 2016, pp. 397–402, doi: 10.1109/ICCOINS.2016.7783248.

[20] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, Jun. 2009, pp. 1105–1112, doi: 10.1145/1553374.1553515.

[21] K. Pittrich and S. Schroeder, “Reading vertically and horizontally mirrored text: an eye movement investigation,” *Quarterly Journal of Experimental Psychology*, vol. 76, no. 2, pp. 271–283, Feb. 2023, doi: 10.1177/17470218221085943.

[22] K. Rayner, T. J. Slattery, D. Drieghe, and S. P. Liversedge, “Eye movements and word skipping during reading: effects of word length and predictability.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 2, pp. 514–528, Apr. 2011, doi: 10.1037/a0020990.

[23] C. Sun and W. Yang, “Research on text readability based on computer multi-level comprehensive evaluation model and algorithm optimization,” in *2022 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, Aug. 2022, pp. 325–332, doi: 10.1109/CIPAE55637.2022.00074.

[24] L. Forti, A. Milani, L. Piersanti, F. Santarelli, V. Santucci, and S. Spina, “Measuring text complexity for italian as a second language learning purposes,” in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 360–368, doi: 10.18653/v1/W19-4438.

[25] H. Moshtael, A. Nuthmann, I. Underwood, and B. Dhillon, “Saccadic scrolling: speed reading strategy based on natural eye movements,” in *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Aug. 2016, pp. 596–600, doi: 10.1109/IHMSC.2016.120.

[26] J. Hautala, S. Hawelka, and M. Ronimus, “An eye movement study on the mechanisms of reading fluency development,” *Cognitive Development*, vol. 69, p. 101395, Jan. 2024, doi: 10.1016/j.cogdev.2023.101395.

[27] C.-T. Hsu, R. Clariana, B. Schloss, and P. Li, “Neurocognitive signatures of naturalistic reading of scientific texts: a fixation-related fMRI study,” *Scientific Reports*, vol. 9, no. 1, p. 10678, Jul. 2019, doi: 10.1038/s41598-019-47176-7.

[28] F. Dary, M. Petit, and A. Nasr, “Dependency parsing with backtracking using deep reinforcement learning,” *Transactions of the Association for Computational Linguistics*, vol. 10, no. 6, pp. 888–903, Sep. 2022, doi: 10.1162/tacl_a_00496.

[29] K. Rayner, A. D. Well, and A. Pollatsek, “Asymmetry of the effective visual field in reading,” *Perception & Psychophysics*, vol. 27, no. 6, pp. 537–544, Nov. 1980, doi: 10.3758/BF03198682.

[30] Y. Zheng, Y. Que, X. Hu, and J. H. Hsiao, “Predicting reading performance based on eye movement analysis with hidden markov models,” in *2022 International Conference on Advanced Learning Technologies (ICALT)*, Jul. 2022, pp. 172–176, doi: 10.1109/ICALT55010.2022.00058.

[31] J. Xuan, J. Lu, G. Zhang, and X. Luo, “Release ‘Bag-of-Words’ assumption of latent dirichlet allocation,” in *Advances in Intelligent Systems and Computing*, vol. 277, 2014, pp. 83–92.

[32] D. Lederman, *The political economy of protection: theory and the chilean experience (social science history)*. California: Standford University Press, 2005.

[33] K. A. Reinert, *An introduction to international economics : new perspectives on the world economy*, 2nd ed. Mexico: Cambridge University Press, 2012.

[34] Y. Jo and A. H. Oh, “Aspect and sentiment unification model for online review analysis,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, Feb. 2011, pp. 815–824, doi: 10.1145/1935826.1935932.

[35] X. Yan, J. Guo, Y. Lan, and X. Cheng, “A biterm topic model for short texts,” in *Proceedings of the 22nd international conference on World Wide Web*, May 2013, pp. 1445–1456, doi: 10.1145/2488388.2488514.

[36] B. Ozyurt and M. A. Akcayol, “A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA,” *Expert Systems with Applications*, vol. 168, p. 114231, Apr. 2021, doi: 10.1016/j.eswa.2020.114231.

BIOGRAPHIES OF AUTHORS

Fatima-Zahrae Sifi is a Knowledge and Data Science Engineer from School of Information Sciences (ESI), Rabat, Morocco, 2016. She is member of the Laboratory of Machine Intelligence (LIM). Completed Higher School Preparatory Classes specializing in Mathematics in 2013 at Meknes, Morocco. She has published several research papers in international conferences and journals, including E3S Web of Conferences in 2021 and the 14th International Conference on Intelligent Systems: Theories and Applications (SITA) in 2023. Her research interests span the areas of text mining, topic modeling, machine learning, deep learning, and graph theory. She can be contacted at email: fatimazahrae.sifi@univh2c.ma.

Wafae Sabbar received the Ph.D. degree from Faculty of Science and Technology, Hassan 2 University, Mohammedia, in 2006. She is currently Professor in Higher Education at Faculty of Law, Economic and Social Sciences of Ain Sebaa (FSJESA), Morocco. She is member of the Laboratory of Machine Intelligence (LIM) and a team Leader of Learning and Data Science. Her research interests include information and computer science technologies and artificial intelligence. She can be contacted at email: wafae.sabbar@univh2c.ma.

Amal El Mzabi received the Ph.D. degree from Faculty of Science and Technology, Hassan 2 University, Mohammedia, in 2005. She is currently Professor in Higher Education at Faculty of Law, Economic and Social Sciences of Mohammedia (FSJESM), Morocco. She is member of the Laboratory of Economic Performances and Logistics (PEL). Her research interests include information and computer science technologies and artificial intelligence. She can be contacted at email: amal.elmzabi@univh2c.ma.