LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

Extracting topics from Brazilian Education discourse: insights to support public policies

ABSTRACT

In a democracy, the public sphere must influence the State's decision-making process. To accomplish this, citizens' participation is essential. The media has a fundamental role in organizing an open and impartial dialogue that relies on experts' knowledge. In this article, we adopted topic modelling to identify relevant issues in a collection of interviews organized by public television during the Brazilian pre-election period in October 2018. The primary objective of this research was to identify and explore pertinent topics that can serve as a foundation for future public policies in education. At a more granular level, we employed probabilistic topic modelling using Latent Dirichlet Allocation (LDA) specifically designed for the Portuguese language. LDA is a non-supervised text-mining technique that allows us to extract relevant topics from a collection of discourses by education experts organized by journalists and conducted through public media channels. The extracted knowledge was then subjected to qualitative analysis. A minor objective is to try to follow up on whether the issues pointed out by the education experts were present in the government agenda of the following years. The graphical representation of all the transcript interviews provided by the word cloud left clear that we need to pay attention to the primary elements (and actors) of the educational system, that is, teacher and student, and their relationship, which is present regardless the level or configuration of learning. By analyzing the list of topics generated by our model more deeply, we confirm that the role of the teacher is hugely relevant to contribute systematically to a quality education. That includes their training, skills, and competencies. Financing and norms (common national curriculum base) are issues responsible for the State and supporting all education systems. Lastly, the family follows the student from childhood to youth in their different needs and levels, guaranteeing that the education cycle is not broken. Our findings offer insights to assist public policy development in Education in Brazil, indicating that attention to simple issues can contribute to a systemic improvement. We also compare these topics with those generated by the official news of the Ministry of Education in the subsequent years, evidencing the absence of relevant issues pointed out by education experts.

KEYWORDS. First keyword. Second keyword. Last keyword.

1. Introduction

The period leading up to a presidential election is characterized by intense discussions in any country. This is when candidates present their main proposals, which becomes crucial for citizens and also a moment for citizens' participation, especially from those who can contribute with their expert opinions [Cross, 2010]. In a democracy, the public sphere shall influence the State decision-making [Souza, 2012]. Therefore, public journalism is vital for promoting citizen participation through active and inclusive dialogue. This study discusses education plans in Brazil during the 2018 Brazilian presidential and governor's pre-election campaign. Our objective is to "listen" to education professionals and evaluate their perspectives to gain insights that can contribute to the development of effective public policies in education. The ultimate goal is to promote a comprehensive and high-quality education plan for Brazil.

Education encompasses a wide range of aspects, beginning with different levels of education, including preschool, elementary school, high school, undergraduate studies, graduation, and technical schools, each with its own age groups and unique challenges. Additionally, there are various learning configurations to consider, such as distance learning, full-time learning, private and public education, inclusive education, timely literacy programs, and teacher training. Therefore, when suggesting any education-related agenda, these diversities must be considered. Viewing education as a singular issue within a nation, particularly in a vast country with a basic school-age population of 50.5 million individuals (from 4 to 17 years old) [1], can lead to a loss of focus.

Brazil counts on many educational statistics that can reveal points requiring attention. Nonetheless, we believe that implementing certain changes can positively impact multiple educational domains. The main objective of this work is to gain insights and identify key aspects that should be the focus of change through the analysis of a collection of interviews with education experts in Brazil using text mining techniques. Besides relying on educational statistics, we firmly believe that individuals directly involved in the educational system can provide valuable contributions by sharing their opinions and perceptions to enhance educational planning.

The primary objective of this research is to identify and explore pertinent topics that can serve as a foundation for future public policies in education. At a more granular level, we employed probabilistic topic modelling using Latent Dirichlet Allocation (LDA) specifically designed for the Portuguese language. LDA is a non-supervised text-mining technique that allows us to extract relevant topics from a collection of discourses by education experts organized by journalists and conducted through public media channels. The extracted knowledge was then subjected to qualitative analysis. A minor objective is to try to follow up on whether the issues pointed out by the education experts will be present in the government agenda of the next few years.

The article is structured into four sections, starting with this introduction, which provides the research context and objective. Section 2 overviews the Brazilian education system and provides detailed information on the text-mining techniques employed. Section 3 presents the materials used and describes our research methodology. Section 4 presents the results and subsequent discussions. Finally, in Section 5, we conclude the work by highlighting the limitations encountered during the study and providing recommendations for future research endeavours.

2. Literature Review

In this section, we explore two main issues. First, we provide an overview of the organizational structure of the Brazilian educational system, including the means of its assessment. Second, we discourse on text mining, the primary tool employed in this research.

[1] https://observatoriocrianca.org.br/cenario-infancia/temas/populacao/

2.1. An overview of the Brazilian Education System

In Brazil, compulsory education is mandatory for children between 4 and 17 years old[2]. The education system is structured into two primary categories: basic education, encompassing preschool through high school, and specialized education, which includes undergraduate, graduate, and technical schools. The organization of the education system is summarized in Table 1.

	Level	Age
Basic Education	Preschool	4-6 Years
	Elementary School	7-14 Years
	High School	15-17 Years
Specialized Education	Under Graduation	18+ Years
	Graduation	18+ Years
	Technical School	14+ Years

Table 1: Structure of the Education System in Brazil.

During the 1970s, the government's primary focus in Brazil was to expand the educational system, aiming to provide access to the education system for all students. However, it was not until the 1980s that the concern for the quality of education started gaining attention and evaluation systems began to be implemented. These evaluations served not only as a means to facilitate educational system management but also as a mechanism to ensure transparency about public resources employment. In 1990, the Brazilian Ministry of Education established the Basic Education Assessment System (SAEB - Sistema de AvaliaÃ§Ã£o da EducaÃ§Ã£o BÃ¡sica) to assess students who had recently completed elementary and high school. Subsequently, in 1999, the National High School Exam (ENEM - Exame Nacional do Ensino MÃ©dio) was introduced as a voluntary exam that offered undergraduate education access. Educational statistics derived from census data and educational performance assessments play a vital role in comprehending, managing, and enhancing education at all levels [Cotta, 2001].

While educational statistics and assessments provide valuable insights into the educational landscape of a large country like Brazil, they often lack the qualitative information and perspectives of individuals directly involved in the educational context [Alves and Silva, 2013]. To bridge this gap and foster a more inclusive and participatory decision-making process, involving education specialists in open and participatory dialogues is crucial. Such dialogues provide an opportunity to incorporate diverse perspectives and reduce the disconnection between citizens and decision-makers. Moreover, this approach aligns with the principles of participatory democracy, where deliberative processes contribute to more informed and legitimate decision-making [Lindell and Ehrström, 2020].

As an example of analyzing expert interviews in the educational context, we cite the study conducted by Seechaliao [2017]. Their research aimed to identify institutional strategies that could support educational innovation. To achieve this, the author collected interviews with specialists in the field and categorized them into key issues and themes. A descriptive analysis was then conducted to explore data and extract insights. This analysis showed effectiveness in identifying strategies that have the potential to foster innovation in learning.

Another study analysing interviews of education experts was carried out by Rapanta et al. [2021]. The authors interviewed four specialists to understand and compare their perceptions and knowledge about teaching in a post-pandemic scenario. The work presents full answers and develops only a qualitative analysis. It is important to note that conducting a manual analysis of large

[2]Constitutional Amendment n.59/2009

collections of interviews can be time-consuming and even impractical. To address this challenge, researchers can rely on automated text-mining strategies to analyze data. One such strategy is topic modelling, which will be discussed in the next section. By leveraging automated techniques, meaningful information can be extracted from a large volume of interview data, facilitating the identification of key themes and patterns.

2.2. Text Mining

Text mining is a collection of methods to uncover insights, identify patterns, comprehend, and analyze textual data, thereby aiding the decision-making process [Aggarwal, 2018]. Since different databases possess unique structures, features, and content, specific treatments and modelling techniques are necessary depending on the desired objectives [Romero and Ventura, 2013]. In this study, we employ topic modelling as the primary tool to extract information from the available data. Topic modelling enables us to uncover latent themes and patterns within the data, facilitating an automatized understanding of the content.

Topic modelling is a text mining technique that operates in a non-supervision manner to identify groups of words that frequently occur within a given text. This strategy enables the exploration of relations between words inside a text and the discovery of abstract themes that represent the underlying content of a document. Each word in the text is assigned a probability of belonging to a specific topic, allowing for a comprehensive representation of the entire document in terms of topic proportions [Aggarwal, 2018].

One of the key advantages of topic modelling is its ability to reduce the need for extensive human intervention and prior domain knowledge during the pre-analysis stage. Topic modelling does not require predefined categories or labelled data. It autonomously discovers patterns and structures within the text and can be adopted to various text types, including documents, articles and social media posts [Vayansky and Kumar, 2020].

In the field of topic modelling, various techniques have been developed. In this study, we employ a classic topic modelling technique called Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. LDA represents the relationship between documents and topics, as well as the relationship between topics and words, using statistical distributions. By leveraging these distributions, LDA uncovers the latent topics within the text and provides a quantitative measure of their presence and influence in the document corpus.

Topic modelling was already employed in bibliometric studies to identify emerging topics inside a collection of papers published in specific journals [Chen et al., 2020; Ozyurt and Ayaz, 2022]. Shrader et al. [2021] analysed interviews with students involved in cheating behaviours. The analysis allowed authors to identify primary topics emerging topics from students' reflections. Another application in education was the assessment of students' writings and reflections during a course conducted by Chen et al. [2016]. The authors adopt topic modelling to find themes present in these journals and automate journal assessment and grading.

Although we found different applications of topic modelling in education, they were not directly related to the educational system evaluation. To our knowledge, this is the first work employing this topic modelling to address this task. In fact, Manych et al. [2023] points out the scarce literature adopting topic modelling to analyse interviews. The authors extracted topics from interviews collected in a political analysis of the economy. They adopted 212 interviews from people in different countries about the political economy of coal, trying to understand the current investments in coal-fired power plants. The authors argue that topic models effectively supplement qualitative case studies, particularly when analysing large amounts of text.

In this work, we aim to analyse a collection of interviews with education experts conducted during a pre-electoral period. To accomplish the task, we rely on automatic strategies. Our

materials and methods are discussed in the next section.

3. Materials and Methods

This section describes the materials and methods used in the computational experiments conducted in this study. The research material comprises a collection of 37 interviews conducted with education experts from various fields. These interviews were recorded during the pre-electoral period of the Brazilian presidential and governors' campaigns in September and October 2018. They were organized by a public television channel[3]. The interviews focused on discussing current issues in the Brazilian education system and exploring educational projects that could be considered by future governments as part of their campaign proposals.

All the research material for this study was initially available in video format and needed to be transcribed into text. To accomplish this, we adopted the auto-generated subtitles provided by YouTube in Portuguese. For the extraction of captions, we utilized *YouTube Transcript/Subtitle API*, a Python package.

The methodology employed in this study was based on the tutorial proposed by Debortoli et al. [2016] on applying topic modelling, specifically extracting topics using Latent Dirichlet Allocation (LDA). This tutorial provided valuable guidance for conducting a text-mining study, serving as a foundation for our present work. The decision to employ the LDA technique was motivated by its widespread use in the literature and its availability in various open-source software tools [Sreenivas et al., 2023; Tabiaa and Madani, 2021; Farkhod et al., 2021].

Firstly, we conducted an exploratory dataset analysis to assess its potential and identify any possible data quality issues. This involved performing simple descriptive statistical analyses, such as determining the number of documents and calculating the average number of words per document. Additionally, we visualized word frequency to gain insights into the data. Subsequently, we proceeded to clean and preprocess data. Preprocessing techniques play a crucial role in text mining, as the quality of the analysis greatly relies on this step [Hickman et al., 2022]. The text of each interview underwent the following preprocessing activities:

1. Punctuation, time stamps and other special characters removal.
2. Conversion of all uppercase letters to lowercase to ensure consistency and prevent the same word from being considered as multiple different words.
3. Tokenization involves splitting the text string into a list of tokens (groups of characters delimited by blank spaces), typically corresponding to individual words.
4. *Stop words* removal, they are words without semantic content, such as prepositions, articles, and conjunctions.
5. Lemmatization, which aimed to reduce each word to its lemma, represents the word's base or dictionary form. Lemmatization helps to normalize the words by removing inflexions. For example, the lemma of the word "written" is "write". It is important to note that while both lemmatization and stemming techniques are effective in language modelling [Balakrishnan and Lloyd-Yemoh, 2014], we chose lemmatization for this study without comparing the two techniques.

The next step involved implementing the topic extraction model. We adopted the *Mallet* wrapper implemented in the *Gensim* library [Řehůřek and Sojka, 2010]. Although *Mallet* is a Java

[3]https://tvcultura.com.br/playlists/221_de-olho-na-educacao-de-olho-na-educacao.html

program, the *Gensim* wrapper allows us to utilize *Mallet*'s LDA implementation within Python. The code used in this work and the datasets compiled in this study are available on Github [4].

Before extracting topics, it was necessary to specify the parameter related to the number of topics to be extracted. Choosing the appropriate number of topics is crucial as it impacts the quality and interpretability of the resulting model. Opting for a high number of topics may lead to the discovery of numerous topics that are not significantly distinct from each other. Conversely, selecting a low number of topics can limit the modelling potential. Typically, the number of topics chosen falls between 10 and 50 [Lindell and Ehrström, 2020]. It is recommended to test different values for the number of topics and evaluate the quality of each resulting model, as done in similar works [Chen et al., 2016].

The evaluation of topic quality in unsupervised topic generation tasks is commonly assessed based on their performance in subsequent tasks, often predictive ones [Debortoli et al., 2016]. However, when predefined classes or groups are absent in the initial documents, defining a predictive task for evaluating the generated topic structure becomes challenging. Therefore, this work will evaluate the topic structure mainly through qualitative analysis. To evaluate the quality of the topics, we can employ qualitative criteria proposed by Boyd-Graber et al. [2014], which include topic interpretability and utility. In the context of this research, the generated topics were considered valuable due to their potential to support the development of educational public policies. We highlight that the interpretability and evaluation of topics are qualitative tasks performed by humans (the researchers). Two of the authors listened to all the interviews and were able to perform this task.

We adopt the topic coherence measures to support the researchers' decision of the best number of topics, leading to a set with greater interpretability and utility. The *U_Mass* coherence score was adopted, which calculates the likelihood of two words appearing together in the corpus. This score is the logarithm of the ratio between the co-occurrence count of two words and the individual occurrence count of the first word. The lower the coherence score, the better the coherence of the topic. Coherence scores are data-dependent, and no universal threshold exists for determining their quality. Typically, the coherence score decreases with the number of topics, but the rate of decrease diminishes as the number of topics increases. The idea is to choose the lowest value of topics after the first curve decline.

To follow up on the most important events in education in the years after the election, we also extracted topics from the official news published by the Ministry of Education. The ministry's website has a news section where it can filter news articles within a specific time period [5]. We selected the four years of the president's mandate, and at the time of our last access, the earliest news article we retrieved was from May 2020, and the most recent one was from December 2022. We automatically extracted the news content using the libraries *Requests* and *beautifulsoup*. 1,480 news articles were collected, pre-processed, and analyzed using the same methodology. However, in this case, we relied solely on coherence measures to determine the number of topics since the authors did not have prior knowledge of the content.

4. Results

In this section, we begin by presenting statistical results from the interviews. We then provide a graphical representation of knowledge using a Word Cloud. Additionally, we utilize coherence scores to determine the optimal number of topics. Finally, we present the identified topics along with their corresponding key terms.

[4] https://anonymous.4open.science/r/brazilian-education-B748/
[5] http://portal.mec.gov.br/todas-as-noticias?view=noticias

4.1. Descriptive Statistics of the Interviews

The statistical summary of the interviews is presented in Table 2. Our research dataset consisted of 37 interviews involving 28 education experts. Each interview had an average duration of 19 minutes, with a standard deviation of less than two minutes, indicating a balanced distribution of the discussed themes. Furthermore, the number of words per interview exhibited uniformity across the videos, with an average of 2,722 words and a standard deviation of 293. This approach ensured that none of the themes were given more prominence than others, mitigating potential bias.

	Total	Average	Standard deviation
Video duration	11:45:06	0:19:03	0:01:37
Word count	100709	2722	293

Table 2: Summary statistics of the videos and corresponding texts extracted.

4.2. Word Cloud

Word Clouds are generated by attributing sizes and colours (or weight) to the text, representing the frequency and relevance of associated terms. They are an initial screening tool to assess basic concepts or their absence [DePaolo and Wilkinson, 2014]. Word Clouds make text data easy to read and comprehend, providing a quick and intuitive impression of a problem [Katre, 2019]. Despite being in the original language (Portuguese), the Word Cloud produced in this article (shown in Figure 1) allows us to draw initial conclusions.

Four words capture our attention: education (*âeducaÃ§Ã£oâ*), teacher (*âprofessorâ*), student (*âalunoâ*), and school (*âescolaâ*). From all the discussions on various educational configurations and levels, it is evident that the main focus is on the teacher-student relationship within the educational space.

Figure 1: Word cloud of the videos transcriptions.

4.3. Topics of the interviews

One crucial aspect to determine is the number of topics (k-value) to be generated by the model. Our approach employed coherence measures and human judgment regarding topic interpretability and utility to arrive at a reasonable k-value [Romero and Ventura, 2013]. First, we undertook an evaluation of different sets of topics, with the number of topics in the range of 10 to 20. Since LDA has a random component, we set the seed value before generating these topics to ensure reproducibility. With previous knowledge about the interviews, the authors agreed on the set of 12 topics as the best description of the data.

Using Gensim library implementation of coherence metric, we calculated the coherence score in the range 1-30 topics, adopting three different random seeds. In all cases, we could observe an initial stability phase of the coherence score between the sets with 10 and 18 topics, which

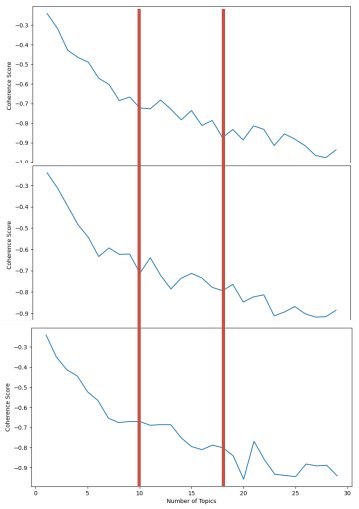

Figure 2: Coherence values according to the number of topics across three different random seeds.

supported our decision to choose 12 topics. The coherence score, depicted in Figure 2, illustrates the relationship between the number of topics and coherence.

Lastly, we provide Table 4, which displays the twelve topic labels assigned by the authors along with the corresponding groups of keywords. It is worth noting that the translation of the topics at this stage posed a challenge, as interpreting individual words without their sentence/context can be more complex and may lead to ambiguity.

We can observe the presence of different actors and levels of education in the extracted topics. The family appears in two topics, as well as the teachers, associated with training and remuneration. Other topics include literacy, the difference in resources between public and private schools, particularly regarding access to technology and the necessity for some students to balance work and study. Higher education is also mentioned as a more general cluster, identified in Topic 7. The everyday issue of classroom management is represented by Topic 2. Lastly, Topic 11 is associated with the integral development of students.

4.4. Topics of the news

To assess whether the issues presented by the experts were addressed in the following years, we also extracted topics from the official news published by the Ministry of Education during the president's term. Following the same methodology, we generate the coherence curves against the number of topics adopting three seeds. The curves suggested values between 12 and 20 topics, and the set with a minimum value of topics was chosen. For conciseness, we will not display all topic words here, but rather, we list the main identified topics 4. We refer the reader to our repository for a comprehensive analysis of the topics.

It is possible to see three elements that match the interviews: the family, the teacher and the basic education. However, the topics extracted from the news articles reveal a predominance of topics related to higher education. This can be understood based on the division of responsibilities outlined in the Brazilian Constitution: the municipal government is responsible for early childhood education and the first stage of primary education, while the state government and the Federal District prioritize secondary education, including the second stage of primary education. The federal government, on the other hand, is responsible for the financial and technical coordination of this educational framework and federal universities. However, the federal government continues to play a leading role in education, being able to decide over the budget, programs and policies.

Table 3: Topic labels and the corresponding keywords.

Cluster #	Topic Label	Keywords
0	Family & education	child, family, son/daughter, father, process, good, to look
1	Classroom discipline	student, problem, teacher, change, help, time, power, future, need
2	Family & school	father, always, teacher, essential, teaching, person, different
3	University education	university, program, Brazil, national, number, power, investiment, situation, university education
4	Study versus work	high school, today, work, few, young, happens, average
5	Teacher training	teacher, few, training, Brazil, need, knowledge, value, enjoy
6	Common national curriculum base	education, today, new, currilum, base, obligation, competence
7	Nation	education, school, example, world, Brazil, obligation, improvement, teaching, both
8	Public and private schools	class, technology, teacher, public school, study, knowledge, internet, use, study, private school
9	Literacy and reading	book, text, reading, student, always, literacy, mother, write, child, culture
10	Teacher's income	stay, night, now, student, example, big, resource, brazilian, money, FUNDEB (The Fund for the Maintenance and Development of Basic Education and the Valuation of Education Professionals)
11	Child & teenager development	child, work, example, development, important, space, create, environment, knowledge

Table 4: List of topics extracted from the news.

Topic	Topic
Basic education	The National High School Exam (ENEM)
Federal institutes	The University for All Program (ProUni)
Technical education	*Higher Education Personnel Improvement Coordination* (CAPES)
Federal university	Teacher and knowledge
University Hospital	*National Institute of Educational Studies and Research* assessments (INEP)

We recognise and emphasize that the comparison that can be made between the topics generated is severely limited since, we are adopting data sources completely diverse in their proposes and discourse. To complement this discussion, an extensive analysis would be needed, including other sources of information and experts' presence in the analysis. A deeper study is needed to investigate how much the government's agenda and actions align with the expert's views and knowledge. Although our results give insights into the government's approach to addressing the challenges and priorities within the education system. We evidence the absence of crucial points that, according to experts, should be addressed to improve education in the country.

5. Discussions and conclusions

In this final section, we conclude our work and point out limitations and recommendations that can be used as avenues for future studies.

5.1. Contributions

Civic engagement and political participation shall be increasingly encouraged from the early ages [Yoon, 2020] even in developing countries [Kovalev et al., 2021]. Real democracy de-

mands consistent and quality participation by all of its citizens. Media has a key role in establishing this active and horizontal dialogue. In this work, we proposed topic modelling to automatically explore interviews in a discussion about Education during a pre-electoral period in Brazil. Applying data mining in education is an emerging interdisciplinary research field frequently explored by researchers [Romero and Ventura, 2013]. However, this work is considered more than a set of numbers in education. Instead, we used open discourse from experts in the education field. Topic modelling facilitated the analysis of the large text collections by extracting common themes discussed in the corpora. As in a conversation, the emphasized subjects are repeated several times by the interviewee [Ackermann et al., 2004], thus the statistical tools behind topic modelling assisted us in discovering these relevant issues.

The graphical representation of all the transcript interviews provided by the word cloud left clear that we need to pay attention in the primary elements (and actors) of the educational system, that is, teacher and student, and their relationship, which is present regardless the level or configuration of learning. By analyzing the list of topics generated by our model more deeply, we confirm that the role of the teacher is hugely relevant to contribute systematically to a quality education. That includes their training, skills, and competencies. Financing and norms (common national curriculum base) are issues responsible for the State and supporting all education systems. Lastly, the family follows the student from childhood to youth in their different needs and levels, guaranteeing that the education cycle is not broken.

It is worth noting that this work addresses the specific challenge of dealing with Portuguese text, which adds complexity as most existing literature focuses on English corpora.

5.2. Limitations and Recomendations

Working with a corpus in the Portuguese language presented various challenges, requiring numerous manual adjustments. We anticipate future improvements so we can explore the methodology in different contexts further. One limitation of our work is the concentration of experts primarily in the SÃ£o Paulo State and BrasÃlia city, which restricts the generalization of our findings to the entire country of Brazil due to its vastness and regional variations.

It is important to note that the interviews were conducted before the COVID-19 pandemic, which brought significant changes and challenges to education. Aspects like remote learning, digital inclusion for students and teachers, school dropout rates, and other related issues have gained prominence. These emerging topics warrant exploration in future studies.

In addition to the limitations mentioned, it is crucial to acknowledge other constraints. First, we relied on transcriptions that lack the nuances of natural speech, including intonation. Second, the subtitles were generated automatically and may contain errors or inaccuracies. Third, we did not compare our topic modelling approach with other state-of-the-art techniques like BERT. Incorporating such comparisons could be a valuable avenue for future research to enhance the depth of our analysis and provide a broader perspective.

Through our comparison between two sources of topics, we have uncovered a significant absence of crucial points in the official news released by the Ministry of Education during the elected presidency. We reinforce that a more comprehensive study would be needed to assess the extent of the gap between the agenda proposed by experts and the programs implemented by the government. In future research, it would be valuable to analyze data from other levels of government to determine whether the highlighted issues are present. This approach would provide a more holistic understanding of the education landscape and its alignment with expert recommendations.

References

Ackermann, F., Eden, C., and Brown, I. (2004). *The practice of making strategy: a step-by-step guide*. Sage.

Aggarwal, C. C. (2018). *Machine learning for text*, volume 848. Springer.

Alves, T. and Silva, R. M. d. (2013). Estratificação das oportunidades educacionais no brasil: contextos e desafios para a oferta de ensino em condições de qualidade para todos. *Educação & Sociedade*, 34:851–879.

Balakrishnan, V. and Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. `https://eprints.um.edu.my/13423/1/rp030_I3007.pdf`.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255.

Chen, X., Zou, D., and Xie, H. (2020). Fifty years of british journal of educational technology: A topic modeling based bibliometric perspective. *British Journal of Educational Technology*, 51 (3):692–708.

Chen, Y., Yu, B., Zhang, X., and Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the sixth international conference on learning analytics & knowledge*, p. 1–5.

Cotta, T. C. (2001). Avaliação educacional e políticas públicas: a experiência do sistema nacional de avaliação da educação básica (saeb). *Revista do Serviço Público*, 52(4):89–111.

Cross, K. A. (2010). Experts in the news: The differential use of sources in election television news. *Canadian Journal of Communication*, 35(3).

Debortoli, S., Müller, O., Junglas, I., and Vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, 39(1):7.

DePaolo, C. A. and Wilkinson, K. (2014). Get your head into the clouds: Using word clouds for analyzing qualitative assessment data. *TechTrends*, 58(3):38–44.

Farkhod, A., Abdusalomov, A., Makhmudov, F., and Cho, Y. I. (2021). Lda-based topic modeling sentiment analysis using topic/document/sentence (tds) model. *Applied Sciences*, 11(23):11091.

Hickman, L., Thapa, S., Tay, L., Cao, M., and Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146.

Katre, P. D. (2019). Nlp based text analytics and visualization of political speeches. *International Journal of Recent Technology and Engineering*, 8(3):8574–8579.

Kovalev, Y., Burnasov, A., Stepanov, A., and Ilyushkina, M. (2021). Alternative models of political participation of population in developed and developing countries: Cases of switzerland, germany, brazil and uruguay. In *Proceedings of Topical Issues in International Political Geography*, p. 204–216. Springer.

Lindell, M. and Ehrström, P. (2020). Deliberative walks: citizen participation in local-level planning processes. *European Political Science*, 19(3):478–501.

Manych, N., Müller-Hansen, F., and Steckel, J. C. (2023). The political economy of coal across 12 countries: Analysing qualitative interviews with topic models. *Energy Research & Social Science*, 101:103137.

Ozyurt, O. and Ayaz, A. (2022). Twenty-five years of education and information technologies: Insights from a topic modeling based bibliometric analysis. *Education and Information Technologies*, 27(8):11025–11054.

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., and Koole, M. (2021). Balancing technology, pedagogy and the new normal: Post-pandemic challenges for higher education. *Postdigital Science and Education*, 3(3):715–742.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Seechaliao, T. (2017). Instructional strategies to support creativity and innovation in education. *Journal of education and learning*, 6(4):201–208.

Shrader, C. B., Ravenscroft, S. P., Kaufmann, J. B., and Hansen, K. (2021). Collusion among accounting students: Data visualization and topic modeling of student interviews. *Decision Sciences Journal of Innovative Education*, 19(1):40–62.

Souza, T. A. S. (2012). *O comunicado da razão: crítica da razão funcionalista na Teoria do Agir Comunicativo*. PhD thesis, Universidade de São Paulo.

Sreenivas, G., Murthy, K. M., Gopali, K. P., Eedula, N., and Mamatha, H. (2023). Sentiment analysis of hotel reviews-a comparative study. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, p. 1–9. IEEE.

Tabiaa, M. and Madani, A. (2021). Analyzing the voice of customer through online user reviews using lda: Case of moroccan mobile banking applications. *International Journal*, 10(1).

Vayansky, I. and Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94:101582.

Yoon, H. S. (2020). Critically literate citizenship: Moments and movements in second grade. *Journal of Literacy Research*, 52(3):293–315.