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

**Leila Abuabara[1,2], Maria Gabriela Valeriano[1,2], Giane Amorim Pessoa[1,2,3]**

[1] Universidade Federal de São Paulo (UNIFESP) – São José dos Campos (SP), Brazil

[2] Instituto Tecnológico Aeronáutico (ITA) – São José dos Campos (SP), Brazil

[3] Universidade Federal de Santa Catarina (UFSC) – Florianópolis (SC), Brazil

{leila.abuabara@unifesp.br, mariagrabrielavaleriano@gmail.com, giane.abp@gmail.com}

***Abstract.*** *In a democracy, the public sphere shall influence the State decision-making. The citizens participation especially those who contribute with opinions of experts is essential in this process. Media has a fundamental role in organizing an open and impartial dialogue. By using a collection of interviews organized by a public television in the pre-election period in October 2018, we used topic modeling to identify relevant issues and offer insights to assist public policies development in Education in Brazil. The findings indicate that attention to simple issues can contribute to a systemic improvement. Thus, the teacher is the actor that needs more attention towards an improvement process.*

***Keywords****: Text Mining, Topic Modeling, Latent Dirichlet Allocation, Education, Brazil*

***Resumo.*** *Em uma democracia, a esfera pública deve influenciar a tomada de decisão do Estado. A participação dos cidadãos, especialmente daqueles que contribuem com opiniões de especialistas, é essencial neste processo. A mídia tem um papel fundamental na organização de um diálogo aberto e imparcial. Por meio de uma coleção de entrevistas organizadas por uma televisão pública no período pré-eleitoral em outubro de 2018, usamos modelagem de tópicos para identificar questões relevantes e oferecer insights para auxiliar no desenvolvimento de políticas públicas na Educação no Brasil. Os resultados indicam que a atenção a questões simples pode contribuir para uma melhoria sistêmica. Neste sentido, o professor é o ator que mais precisa de atenção para um processo de melhoria.*

***Palavras-chave:*** *Mineração de Textos, Modelagem por Tópicos, Alocação Latente de Dirichlet, Educação, Brasil*

## 1. Introduction

The period leading up to a presidential election is one of intense discussion in any country. This is when candidates present their main proposals, and also a moment for citizens participation especially those who can contribute with their opinions of experts (Brady et al., 1995). At this time, we have the central attribution of the media, the journalists and the public television in developing the voice of these professionals and citizens in an impartial and organized manner (Graván, Mateos, & Broullón-Lozano, 2019). In a democracy, the public sphere shall influence the State decision-making (Habermas, 2012). In this sense the public journalism is an important resource for promoting citizens participation through an active and horizontal dialogue. This work focuses on discussing education plans in Brazil during the 2018 Brazilian presidential and governor's pre-election campaign. We intend to 'listen' to education professionals and assess their speech for the purpose of having insights to assist public policies development in this field aiming to promote a quality education plan for Brazilians.

Education is a broad issue that considers many aspects. We start mentioning *education levels*: preschool, elementary school, high school, under graduation, graduation, and technical school, where each phase comprises specific ages and challenges. On the other hand, we have a diversity of learning *configurations*. Some examples include distance learning, full-time learning, private and public learnings, inclusive education, literacy in the right time, teacher training and so on. Thus, when suggesting any education related agenda, these diversities shall be taken into account. That means we can easily lose focus when considering education as one single issue in a nation, especially in Brazil, which is a huge country summing up 45.4 million of people in basic school age (4-17) (IBGE, 2010). Brazil counts on many educational statistics that can reveal points to be dealt. Even so, we believe that some changes can benefit a larger number of education areas. *How to know which are these changes?* To have an idea of the answer to this important question is the objective of this work. We intend to raise central aspects to be the focus of change from a collection of interviews with education experts in Brazil by using a text mining technique. We believe that, beyond educational statistics, people who are directly involved in the educational system can richly contribute by providing their opinions and perceptions to a better educational planning.

The major purpose of this research is to unfold relevant topics and use them to support general future public policies in the area. In a more detail level, from a set of discourses with education experts that were organized by journalists in a public media, we used probabilistic topic modeling via Latent Dirichlet Allocation (LDA) for Portuguese language, which is an unsupervised text-mining technique, to extract relevant topics.

Our article is organized in four sections including this introduction which contains the context of the research and the objective. Section 2 outlines an overview of the Brazilian education system; and provides text mining techniques details. Section 3 presents our material and research methodology. Section 4 presents some results and discussions. Finally, we conclude our work in Section 5 presenting some limitations and recommendations for future studies.

## 2. Literature Review

In this section two main issues were explored. First, we provide general information about how the educational system is organized in Brazil including the means of its assessment.

Education is our main research topic. Second, we discourse on text mining, our main tool in this research.

*An overview on the Brazilian Education*

In Brazil, school attendance is mandatory for children from 4 to 17 years old (Constitutional Amendment n.59/2009). However, about 8.5% of this group are not enrolled in any educational institution (IBGE, 2010). Education in Brazil is basically organized as shown in *Table **1***:

*Table 1: Education structure in Brazil.*

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

In the years 1970s, the main government concern was to expand the educational system in order to promote access to all students. Only in the years 1980s, an interest with the quality of education became part of the agenda and evaluation systems began to be implemented. Beyond being a way to facilitate educational system management, it was also a mechanism for transparency in the use of public resources. In 1990, the Brazilian Ministry of Education established the **Basic Education Assessment System** (SAEB – *Sistema de Avaliação da Educação Básica*) to evaluate students who has just concluded elementary school and high school. In 1999, the **National High School Exam** (ENEM - *Exame Nacional do Ensino Médio*) was instituted in the country as a voluntary exam which is also used to access the under-graduation level. *Educational statistics* through census data, and *educational performance assessments* are essential information for understanding, managing, and improving education in all levels (Cotta, 2014).

While these statistics and assessments measurements reflect the educational realities of this large country, it does not capture any qualitative information from the people directly involved in the educational context (Alves & Da Silva, 2013). Involving education specialists in an open and participatory dialogue is an opportunity to reduce the gap between citizens and decision makers towards important and broad impact decisions, beyond being a legitimate model of participatory democracy (Lindell & Ehrström, 2020).

*Text Mining*

Text mining is a set of methods used to discover useful knowledge, get patterns, understanding and analyzing text data and thus support the decision making (Aggarwal & Zhai, 2013). Different databases have distinct structure, features and contents, thus depending on the objectives to be reached, each database requires specific treatment and modeling techniques (Romero & Ventura, 2013). In this research, we use topic modeling as our main tool to get useful information from texts. This non-supervisioned machine learning technique identifies groups of words that occur frequently in a text through statistical methods. Thus, this technique allows understanding relations between words inside a text and finding abstract topics that represent the main document themes. For each word in a text is assigned a topic belonging probability and the entire document is a set of topics each one with a ratio of presence in the text (Aggarwal & Zhai, 2013). By using unsupervised text-categorization, the human intervention required during a pre-analysis is reduced. Moreover, we do not need to have previous substantive knowledge in the subject treated (Debortoli, Müller, Junglas, & vom Brocke, 2016).

## 3. Materials and Methods

### a. Materials and research context

The research material is composed of a set of 37 interviews with experts in education in many fields dated September and October 2018, the Brazilian presidential and governors' pre-electoral period, and were organized by a public tv. The complete playlist can be found at: https://tvcultura.com.br/playlists/221_de-olho-na-educacao-de-olho-na-educacao.html. The context included discussions on current problems in the Brazilian education system and educational projects for the future governments to be considered in their campaign proposal. The title (and so the main question covered) of each interview, the name of the interviewee and date of the interview, and the professional position at that moment are listed in ***Table 4*** in the Appendix A.

All the material is primary available in video which had to be transcribed into text for the purpose of our study. To perform caption extraction from the *YouTube* videos, we used the *YouTube* handler *PyTube*, a *Python* compatible library. By using a *YouTube* object (from *PyTube*), we got the playlist links which were stored in a *DataFrame* from *Pandas* library. Following, by using *Caption* object, we performed auto-generate subtitles extraction in Portuguese language. We included a column with the interview's title, an optional task. Alternatively (and due to some instability in *PyTube* library), we extracted the subtitles through the DownSub website (https://downsub.com/) to *.txt backup files.

### b. Methods

The methodology that we followed was initially proposed by Debortoli, Müller, Junglas, & vom Brocke (2016) in a tutorial in topic modeling application, particularly the extraction of topics by applying Latent Dirichlet Allocation (LDA). This work provides directions to conduct a text mining study that is going to be used as a start to the present project. In LDA, the relation between documents and topics and, topics and words are represented as statistical distributions. The authors selected LDA technique as it has been widely used in the literature and it could be implemented through several open-source software. *Figure **1*** illustrates a step by step of the proposed methodology.

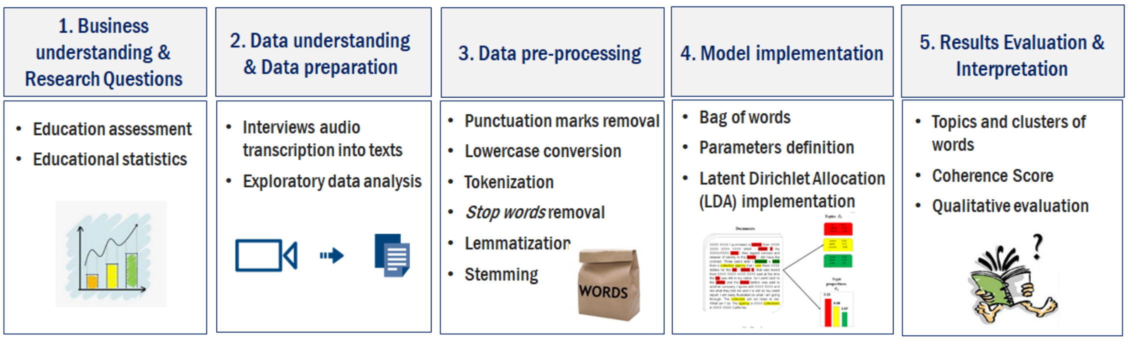
1. Business understanding & Research Questions
Education assessment
Educational statistics
2. Data understanding & Data preparation
Interviews audio transcription into texts
Exploratory data analysis
3. Data pre-processing
Punctuation marks removal
Lowercase conversion
Tokenization
Stop words removal
Lemmatization
Stemming
WORDS
4. Model implementation
Bag of words
Parameters definition
Latent Dirichlet Allocation (LDA) implementation
5. Results Evaluation & Interpretation
Topics and clusters of words
Coherence Score
Qualitative evaluation

*Figure 1: Proposed methodology (adapted from Debortoli, Müller, Junglas, & vom Brocke, 2016).*

Firstly, we conducted an exploratory analysis of the data set to understand its potential and identify possible data quality problems. By using statistical analysis, which includes the number of documents and average of words per document and visualization of word frequency, just to mention some examples, we followed by cleaning the data and preprocessing it as needed. Preprocessing techniques develop an important place in text mining models performance (Vijayarani, Ilamathi, & Nithya, 2015) since the quality of the result depends on it.

Each of the (audios transcribed into) texts went through the preprocessing activities that are described below:

i. Punctuation, time stamps and other special characters removal.
ii. Conversion of all uppercase to lowercase to avoid the risk of the same word to be considered as two different words.
iii. Tokenization that means to split a string, text into a list of tokens (a character group delimited by blank spaces).
iv. *Stop words* removal which are words without semantic content such as prepositions, articles, and conjunctions. This process needed to be complemented with manual removal as well.
v. Lemmatization that means a process to reduce each word to its lemma. The lemma is the word without flexions. E.g.: The lemma of the 'written' is 'write'. Remark that as lemmatization and stemming have the same performance in language model techniques (Balakrishnan & Lloyd-Yemoh, 2014), we have just chosen one of them (the first one) without comparing the techniques. As the previous item, we performed additional manual lemmatization.

Following Debortoli, Müller, Junglas, and vom Brocke's (2016) tutorial, next step is to implement topics extraction model. Before, the parameter related to the number of topics to be extracted shall be determined. By choosing a high value the algorithm could discover an infinity of topics that do not differ significantly among each other. On the other hand, selecting a low number of topics could limit the modeling potential. A recommended practice is to test different values for the number of extracted topics and evaluate the quality of each model resultant. This process has already been used in similar works (Chen, Yu, Zhang, & Yu, 2016). However, we need to keep in mind that the topics will be interpreted and evaluated qualitatively by humans (the authors), thus the number of topics chosen is usually between 10 and 50 (Debortoli, Müller, Junglas, & vom Brocke, 2016).

Due to the non-supervisioned topic generation task nature, its evaluation generally is measured through the performance when they are employed in a subsequent task, usually predictive (Debortoli, Müller, Junglas, & vom Brocke, 2016). According to those authors, topic modeling is descriptive by itself, that is, they represent qualitative summary of large documents collections. These topics can be grouped, classified and their evolution could be monitored over time. In this work, the documents initially do not belong to any class or group, so this predictive task could not be used to evaluate the topic generated. Thus, the topic is going to be evaluated only by qualitative analysis, based on the Boyd-Graber, Mimno, & Newman (2014)’s work who proposed criteria that include the topic interpretability, coherence and utility. In this research, the topics generated are useful if they could be used to support the development of educational public policy. Besides that, a contribution with Portuguese text treatment is also intended, which we consider challenging once the major works in the literature use English *corpora*. Topic modeling process was performed using the *Mallet* wrapper for *Gensim* (Řehůřek, 2009). *Mallet* is a *Java* program, but the *Gensim* wrapper allows *Mallet*'s LDA implementation to be used in *Python*.

## 4. Results and discussions

We start this section by presenting some statistical results in relation to the interviews. We follow with a graphical representation of knowledge through the Word Cloud, the coherence score assisting the definition of the number of topics and, finally the topics and the respective key terms.

### 4.1. Descriptive statistics of the interviews

The statistical summary of the interviews is present in *Table 2*. A total of 37 interviews involving 28 experts in education composed our research base. Each interview had a duration of 19 minutes (± 1.6) in average showing a balance among the themes discussed. The number of words per interview was also quite uniform among the videos (2,722 ± 293 words). This was due to the organization of the TV station which allocated a similar and limited time for each interview. Additionally, this was a way of ensuring that none of the themes was discoursed more than the other, avoiding bias. This was also reflected in the number of valid words, which is the set of words after the *stop-words* filter (1,414 ± 158 words).

*Table 2: Statistical summary*

| | **Total** | **Average** | **Standard Deviation** |
|---|---|---|---|
| **Number of views on YouTube** (up to Nov 16, 2020) | 197,112 | 5,327 | 8,406 |
| **Video duration** (hour:min:sec) | 11:45:06 | 0:19:03 | 0:01:37 |
| **Total word count** (entire text) | 100,709 | 2,722 | 293 |
| **Valid word count** (excluding *stop words*) | 52,311 | 1,414 | 158 |
| **Valid words** (%) | - | 52% | 2% |
| **Number of professionals interviewed** | 28 | - | - |

Some videos are more popular than others. This can be verified by the standard deviation of the number of views on YouTube (dated Nov 16, 2020). We listed the five most viewed videos:

- 1st: *What is the Common National Curriculum Base?* 38,259 views.
- 2nd: *How is Education in Brazil nowadays?* Also, the debut program: 33,519 views.
- 3rd: *What is the concept of full-time education?* 19,037 views.
- 4th: *Where does FUNDEB's money come from?* 17,838 views.
- 5th: *Are there many functional illiterates in Brazil?* 11,432 views.

All five themes are quite broad. Additional detailed information in relation to each individual interview is available in the Appendix A.

## 4.2. Word Cloud

Word Clouds are generated when sizes and colors (or weight) are attributed to the text to represent a feature (herein the frequency) of the associated terms and therefore its relevance. It is useful as an initial screening tool for assessment by checking basic concepts (or lack of them) (DePaolo & Wilkinson, 2014). It is an additional appeal to make data easy-to-read and comprehend (DePaolo & Wilkinson, 2014) and the first impression of a problem (Katre, 2019). Even being in the original language (Portuguese), the word cloud produced in this article and showed in ***Figure 2*** provides us a quick and intuitive sense for our initial conclusions.

Four words take our attention: education (‘*educação*’), teacher (‘*professor*’), student (‘*aluno*’) and school (‘*escola*’). We can conclude that despite all possible educational configurations and levels that were herein discussed, the main actors that emerge from all interviews are in the teacher-student relationship within the educational space.

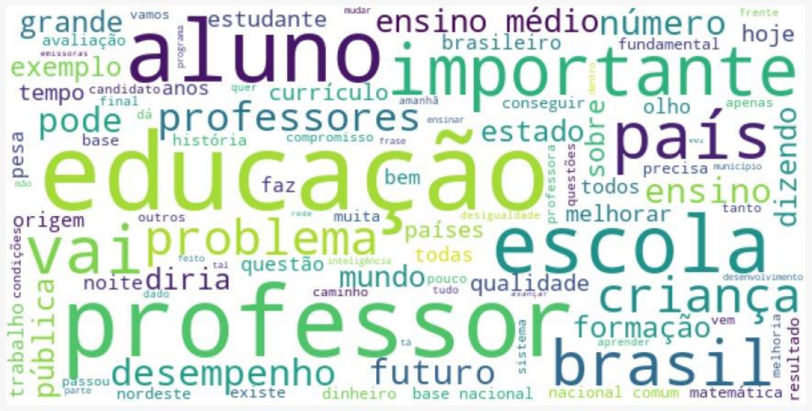

Figure 2: Word Cloud.

## 4.3. Topics

One important parameter to be defined is the **number of topics** (k-value) to be generated by the model. Our process included the combination of coherence measures and human judgement of the interpretability of topics (Röder, Both, & Hinneburg, 2015) to establish a reasonable k-value. *Figure* ***3*** shows the coherence score according to the number of topics. After evaluating a range which included values from 5 to 15 topics, the authors agreed at 11.

At this point we also realize an initial of a stability phase of the coherence score (values between 11 and 16).

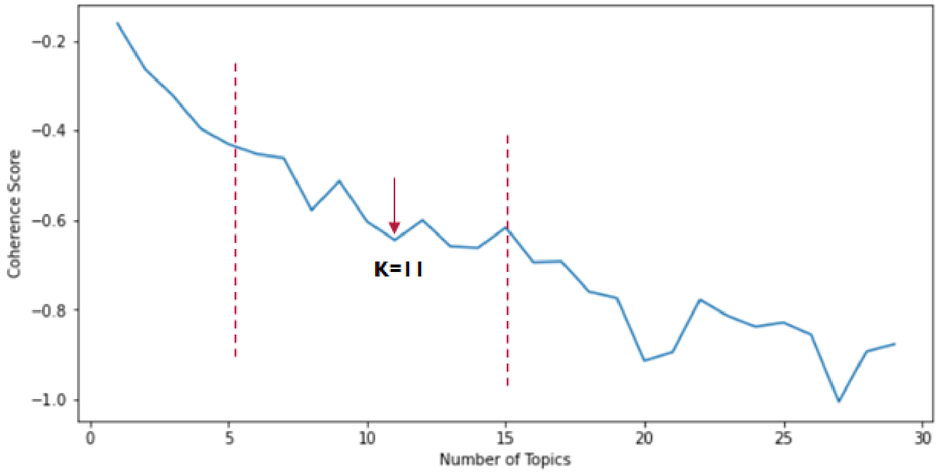

*Figure 3: Stability analysis of different topic solutions.*

Finally, we present in ***Table 3*** the eleven topic labels assigned by the authors and the eleven related keywords groups (topics generated by the model). The original outcome is presented in ***Figure 5*** in the Appendix B. It is important to mention that it was tricky to translate the topics at this stage of the work since the interpretation of free words (out of a sentence/context) is more difficult and can generate ambiguity.

*Table 3: Topic and the key terms.*

| Cluster # | Topic Label | Keywords |
|---|---|---|
| 0 | High school | teach_high_school, student, young, night, high school[1], discipline, today, choose, school, little |
| 1 | Nation | today, high school, pass, school, only, last, important, brazil, million, national |
| 2 | Undergraduate education | student, power, university, attend, program, person, problem, change, area, course |
| 3 | Study versus work | power, nation, teach, school, education, need, work, example, time, succeed |
| 4 | Common national curriculum base | base, student, curriculum, new, competence, knowledge, learn, common_national_curriculum_base, issue, Brazilian |
| 5 | Community participation | student, school, technology, power, beyond, involve, community, feel, participate, pedagogic |
| 6 | Teacher training | teacher, little, relation, classroom, academic_eduation, class, student, teacher_training, power, issue |
| 7 | Family & education | child, family, father, school, elementary, respect, son, put, educate, wait |
| 8 | Importance of preschool | child, day_care_center, example, important, need, condition, bring, middle, adult, son |

[1] Here, the original word '*médio*' was translated as '*high school*'.

| | | |
|---|---|---|
| 9 | Education financing | education, continue, resource, county, political, Brazilian, better, big, challenge, go |
| 10 | Literacy and reading | reading, read, book, need, literacy, word, language, join, library, world |

An additional way of evaluating and providing interpretation to our outcomes is through the *bar chart* (on the right) and the two-dimensional *intertopic distance map* (on the left) as presented in ***Figure 4***. The topic circles have similar areas since the numbers of words that belongs to each topic is the same (10 words). This means that all topics (circles) have the same prevalence in the *corpus*. The bar chart shows the 30 most salient terms indicated by the total frequency of the term across the entire *corpus*. The most prominent words in descending order are teacher (‘*professor*’), student (‘*aluno*’), school (‘*escola*’), power (‘*poder*’), education (‘*educação*’), and child (‘*criança*’). We can conclude that these are the main pillars to be considered in the future plans.

Although ***Figure 4*** is static, it provides information interactively. Once performed such command, we can select a topic in the *intertopic distance map* and the bar highlights their colors to display the most salient words included in that specific topic (SydneyF, 2020). The distance among circles shows how different they are. However, two of them are overlapping.

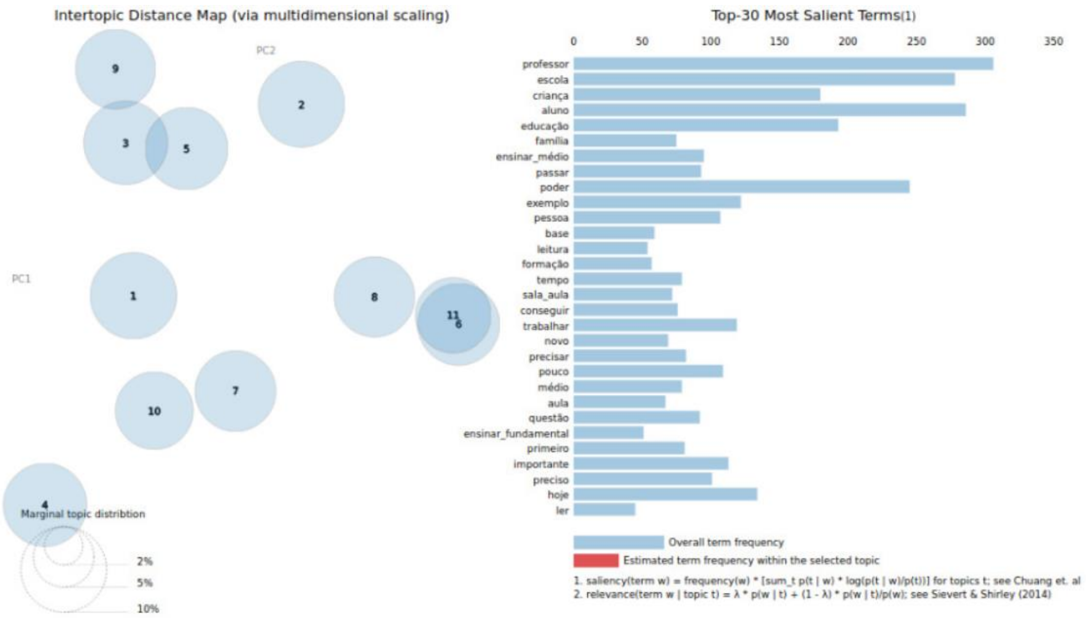

*Figure 4: Intertopic Distance Map and Bar Chart.*

## 5. Conclusions and future studies

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

*Conclusions*

Civic engagement and political participation shall be increasingly encouraged from the early ages (Yoon, 2020) even in developing countries (Kovalev, Stepanov, & Ilyushkina, 2019). Real democracy demands consistent and quality participation by all its citizens. And the media has a key role in establishing this active and horizontal dialogue.

In this work we proposed the use of topic modelling to automatically explore interviews in Education in discussion during a pre-electoral period in Brazil. Education per se generates large repositories of datasets. Applying data mining in education is an emerging interdisciplinary research field that has been frequently explored by researchers (Romero & Ventura, 2013). However, this work 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, Eden, & Brown, 2009), thus the statistical tools behind topic modeling 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 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 more deeply the list of topics generated by our model, we confirm that the role of teacher is hugely relevant to contribute systematically to a quality education. That includes their training, skills, and competences. Financing and norms (common national curriculum base) are issues that are responsible for the State and supports all education system. Lastly family plays the role of following the student from child to young in their different needs and levels guaranteeing that the education cycle is not broken.

*Limitations*

Working with a corpus in Portuguese language was challenging as we had to make many manual adjustments. We hope to have improvements in the next future aiming to explore the methodology in other contexts.

One limitation of our work is that most part of the experts are concentrated in São Paulo State and Brasília city. Brazil is very extensive, and we cannot consider a unique pattern as valid for the entire country.

*Recommendations*

The interviews considered took place before the Covid-19 pandemic. From this period many changes and challenges are happening in the educational field which include education at home, and so we believe that other topics would emerge such as the digital inclusion (for students and teachers), school dropout and other misfortunes. This is a subject that shall be explored in future studies.

## Referências

Ackermann, F., Eden, C., & Brown, I. (2009). *The Practice of Making Strategy: a step-by-step guide.* London (UK): SAGE.

Aggarwal, C. C. (2015). *Data Minig: The Textbook.* New York (USA): Springer.

Aggarwal, C. C., & Zhai, C. (2013). An Introduction to Text Mining. In C. C. Aggarwal, & C. Zhai, *Mining Text Data.* (pp. 1-10). Springer. doi:10.1007/978-1-4614-3223-4

Alves, T., & Da Silva, R. M. (2013). Estratificação das oportunidades educacionais Brasil: Contextos e desafios para a oferta de ensino em condições de qualidade para todos. *Educacao e Sociedade, 124*, pp. 851-879. doi:10.1590/S0101-7330201300

Balakrishnan, V., & Lloyd-Yemoh, E. (2014, August). Stemming and Lemmatization: A Comparison of Retrieval. *Lecture Notes on Software Engineering, 2*(3). doi:10.7763/LNSE.2014.V2.134

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. In. In E. M. Airoldi, D. Blei, E. A. Erosheva, & S. E. Fienberg, *Handbook of mixed membership models and their applications* (pp. 3-34). Boca Raton: CRC Press.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April 25-29). Topic modeling for evaluating students' reflective writing: A case study of pre-Service teachers' journals. *ACM International Conference Proceeding Series*, 1–5. doi:10.1145/2883851.2883

Cotta, T. C. (2014). Avaliação educacional e políticas públicas: a experiência do Sistema Nacional de Avaliação da Educação Básica (Saeb). *Revista Do Serviço Público, 4*, pp. 89-111. doi:10.21874/rsp.v52i4.316

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems, 39*(7), pp. 110-135.

DePaolo, C. A., & Wilkinson, K. (2014, May/June). Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. *TechTrends, 58*(3), 38-44.

Graván, M. L., Mateos, C., & Broullón-Lozano, M. A. (2019). The role of public service journalism and television in fostering public voice and the capacity to consent: An analysis of Spanish viewers' discourses. *Journalism*, 1-18. doi:10.1177/1464884919847593

Habermas, J. (2012). *Teoria do agir comunicativo - vol. 2: Sobre a crítica da razão funcionalista.* WMF Martins Fontes.

IBGE. (2010). *Instituto Brasileiro de Geografia e Estatística*. Retrieved from www.ibge.gov.br.

Katre, P. D. (2019, September). NLP based Text Analytics and Visualization of Political Speeches. *International Journal of Recent Technology and Engineering, 8*(3). doi:10.35940/ijrte.C6503.098319

Kovalev, Y., Stepanov, A. B., & Ilyushkina, M. (2019). Alternative Models of Political Participation of Population in Developed and Developing Countries: Cases of Switzerland, Germany, Brazil and Uruguay. In R. Bolgov, V. Atnashev, Y. Gladkiy, A. Leete, A. Tsyb, & S. Pogodin, *Proceedings of Topical Issues in International Political Geography* (pp. 204-227). St. Petersburg: Springer Geography. doi:10.1007/978-3-030-58263-0

Lindell, M., & Ehrström, P. (2020, january 20). Deliberative Walks: citizen participation in local-level planning processes. *European Political Science*. doi:10.1057/s41304-020-00243-4

Řehůřek, R. (2009). *Gensin: Topic Model for Humans*. Retrieved December 10, 2020, from https://radimrehurek.com/gensim/models/wrappers/ldamallet.html

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the Space of Topic Coherence Measures. *WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. doi:https://doi.org/10.1145/2684822.2685324

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*, pp. 12–27. doi:10.1002/widm.1075

SydneyF. (2020, August 05). *Getting to the Point with Topic Modeling | Part 3 - Interpreting the Visualization*. Retrieved December 10, 2020, from Alteryx Community: https://community.alteryx.com/t5/Data-Science/Getting-to-the-Point-with-Topic-Modeling-Part-3-Interpreting-the/ba-p/614992

Vijayarani, S., Ilamathi, J., & Nithya, M. (2015). Preprocessing Techniques for Text Mining. *International Journal of Computer Science & Communication Networks, 5*(1), 7–16.

Yoon, H. S. (2020). Critically Literate Citizenship: Moments and Movements in Second Grade. *Journal of Literacy Research*, 1-23. doi:10.1177/1086296X20939557

# Appendix A – Material Information

*Table 4: Information and statistics of the interviews.*

| Item # | Main question of the interview *[original title in Portuguese]* | Interviewer (date) | Interviewer Position | Number of views on YouTube (up to Nov 16, 2020) | Video duration (min:sec) | Total word count (entire text) | Valid word count (excluding stop words) | Valid words (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | How is education in Brazil nowadays?<br>[*Como está o ensino no Brasil hoje?*] | Guiomar Namo de Mello<br>(Sep 3rd, 2018) | Professor, educator, and member of the São Paulo State Board of Education [*Conselho Estadual de Educação de São Paulo – CEE/SP*] | 33,519 | 18:20 | 2,567 | 1,378 | 54% |
| 2 | What is creative class?<br>[*O que é aula criativa?*] | Luis Carlos de Menezes<br>(Sep 6, 2018) | Senior professor at USP (University of São Paulo) and member of the São Paulo State Board of Education [*Conselho Estadual de Educação de São Paulo – CEE/SP*] | 3,678 | 18:23 | 2,410 | 1,245 | 52% |
| 3 | Is it essential that teachers know how to "play"?<br>[*É fundamental que os professores saibam "brincar"?*] | Gisela Wajskop<br>(Sep 7, 2018) | Sociologist, PhD, and researcher in Education | 3,781 | 18:35 | 2,690 | 1,430 | 53% |
| 4 | Is private education better than public education?<br>[*O ensino particular é melhor que o ensino público?*] | Mauro de Salles Aguiar<br>(Sep 18, 2018) | Principal of *Colégio Bandeirantes/SP*. | 4,758 | 17:56 | 2,472 | 1,318 | 53% |
| 5 | What are the pros and cons of participatory management?<br>[*Quais são os prós e contras de uma gestão participativa?*] | Débora Gonzalez Costa Blanco<br>(Sep 20, 2018) | Regional Education Director - São Carlos/SP region. | 3,441 | 18:20 | 2,582 | 1,344 | 52% |
| 6 | What is the concept of full time Education?<br>[*Qual o conceito de Educação Integral?*] | Simone André<br>(Sep 26, 2018) | Education executive manager at the Ayrton Senna Institute | 19,037 | 17:50 | 2,444 | 1,250 | 51% |
| 7 | Are schools educating their students well?<br>[*As escolas estão formando bem seus alunos?*] | Fernando Botelho<br>(Sep 28, 2018) | Professor of Economics at USP (University of São Paulo) and collaborator of education movement *Todos pela Educação* | 1,885 | 18:08 | 2,735 | 1,442 | 53% |
| 8 | Can any course be done remotely?<br>[*Qualquer curso pode ser feito à distância?*] | Guiomar Namo de Mello<br>(Oct 3rd, 2018) | Professor, educator, and member of the São Paulo State Board of Education [*Conselho Estadual de Educação de São Paulo – CEE/SP*] | 912 | 18:34 | 2,497 | 1,421 | 57% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | Education as a priority<br>[*A Educação como prioridade*] | Olavo Nogueira<br>(Oct 4, 2018) | Director of educational policies for the non-profit organization *Todos pela Educação* | 1,705 | 18:36 | 2,943 | 1,557 | 53% |
| 10 | What is an inclusive school?<br>[*O que é uma escola inclusiva?*] | Sylvia Figueiredo Gouvea<br>(Oct 5, 2018) | Educator | 7,134 | 18:47 | 2,643 | 1,199 | 45% |
| 11 | What is the continuous progression system?<br>[*O que é o sistema de progressão continuada?]* | Rose Neubauer<br>(Oct 8, 2018) | São Paulo State Academy of Education vice-president | 3,391 | 18:28 | 2,549 | 1,317 | 52% |
| 12 | Why is it essential to invest in digital technology in schools?<br>[*Por que é essencial colocar tecnologia digital dentro das escolas?*] | Maria Alice Carraturi Pereira<br>(Oct 10, 2018) | Ministry of Education and Culture (MEC) education director | 6,236 | 18:29 | 2,608 | 1,345 | 52% |
| 13 | Literacy in Brazil: what to improve? *[Alfabetização no Brasil: o que melhorar?]* | Flávia Yuri Oshima<br>(Sep 4, 2018) | Journalist specialized in Education and twice winner of the ‘Journalism in Education’ award. | 7,427 | 18:40 | 2,722 | 1,403 | 52% |
| 14 | Is it important to talk about teacher’s preparation?<br>[*É importante falar da preparação dos professores*?] \| | Guiomar Namo de Mello<br>(Sep 5, 2018) | Professor, educator, and member of the São Paulo State Board of Education [*Conselho Estadual de Educação de São Paulo – CEE/SP*] | 2,529 | 18:27 | 2,631 | 1,339 | 51% |
| 15 | Higher education incentive programs<br>*[Programas de incentivo à educação superior]* | Renato Pedrosa<br>(Sep 10, 2018) | Professor at UNICAMP (University of Campinas) | 1,821 | 17:56 | 2,722 | 1,433 | 53% |
| 16 | Are there many functional illiterates in Brazil?<br>*[Existem muitos analfabetos funcionais no Brasil?]* | Silvia Gasparian Colello<br>(Sep 11, 2018) | Pedagogue and professor in the Faculty of Education at USP (University of São Paulo) | 10,432 | 18:34 | 2,540 | 1,310 | 52% |
| 17 | How is the situation in high school?<br>*[Como está a situação do ensino médio?]* | Ilona Becskeházy<br>(Sep 12, 2018) | Education Consultant | 1,820 | 18:31 | 2,821 | 1,463 | 52% |
| 18 | Why is it important for children to attend day care?<br>*[Por que é importante para as crianças frequentar a creche?]* | Beatriz Cardoso<br>(Sep 13, 2018) | PhD in Education and President of the Laboratory of Education | 1,775 | 18:12 | 2,826 | 1,403 | 50% |
| 19 | Are colleges preparing our teachers well?<br>*[As faculdades estão preparando bem nossos professores?]* | Guiomar Namo de Mello<br>(Sep 14, 2018) | Professor, educator, and member of the São Paulo State Board of Education [*Conselho Estadual de Educação de São Paulo – CEE/SP*] | 1,828 | 18:33 | 2,737 | 1,408 | 51% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 20 | What is the Common National Curriculum Base?<br>*[O que é a Base Nacional Comum Curricular?]* | Valéria de Souza<br>(Sep 17, 2018) | PhD in Education and advisor to the São Paulo State Secretariat of Education | 38,259 | 18:16 | 2,772 | 1,440 | 52% |
| 21 | What are the assessments that are Inep responsibility?<br>*[Quais as avaliações que são de responsabilidade do Inep?]* | Maria Inês Fini<br>(Sep 19, 2018) | President of Inep (National Institute of Educational Studies and Research Anísio Teixeira) | 757 | 18:41 | 2,577 | 1,336 | 52% |
| 22 | The difficulties of implementing the new National Common Curricular Base<br>*[As dificuldades de implementação da nova Base Nacional Comum Curricular]* | Ghisleine Trigo Silveira<br>(Sep 21, 2018) | Member of the São Paulo State Board of Education [*Conselho Estadual de Educação de São Paulo – CEE/SP*] | 3,112 | 18:18 | 2,500 | 1,319 | 53% |
| 23 | What will change in high school?<br>*[O que vai mudar no ensino médio?]* | Simon Schwartzman<br>(Sep 24, 2018) | Sociologist and member of the Brazilian Academy of Sciences | 1,246 | 17:27 | 2,815 | 1,472 | 52% |
| 24 | The importance of the new National Common Curricular Base<br>*[A importância da nova Base Nacional Comum Curricular]* | Maria Helena Guimarães de Castro<br>(Sep 25, 2018) | Sociologist and Member of the National Board of Education | 1,221 | 18:37 | 2,336 | 1,247 | 53% |
| 25 | Will the reform of high school change the scenario of technical courses?<br>*[A reforma do Ensino Médio irá mudar o cenário dos cursos técnicos?]* | Almério Melquíades de Araújo<br>(Sep 27, 2018) | Coordinator of the Paula Souza Center | 2,880 | 18:17 | 2,509 | 1,314 | 52% |
| 26 | Why is the transition from Elementary I to Elementary II so delicate?<br>*[Por que a transição do Fundamental I para o Fundamental II é tão delicada?]* | Francisco Carbonari<br>(Oct 1st, 2018) | Philosophy teacher | 1,339 | 18:05 | 2,364 | 1,244 | 53% |
| 27 | Why is it important to read from an early age?<br>*[Por que é importante ler desde cedo?]* | Marisa Lajolo<br>(Oct 2nd, 2018) | Professor of literature at Mackenzie | 1,125 | 17:54 | 2,608 | 1,344 | 52% |
| 28 | Are Brazilian teachers qualified to teach English?<br>*[Os professores brasileiros estão habilitados para ensinar inglês?]* | Silvia Donnini<br>(Oct 9, 2018) | Secretary of Education of São Bernardo do Campo/SP | 1,552 | 18:14 | 2,398 | 1,199 | 50% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 29 | What resources are essential for the teaching of the Portuguese language?<br>*[Quais recursos são fundamentais para o ensino da língua portuguesa?]* | Pasquale Cipro Neto<br>(Oct 11, 2018) | Professor | 920 | 18:21 | 2,471 | 1,279 | 52% |
| 30 | Do our teachers receive good training?<br>[Nossos professores recebem uma boa formação?] | Mariza Abreu<br>(Oct 15, 2018) | Education Consultant | 1,247 | 19:52 | 2,839 | 1,479 | 52% |
| 31 | What are the problems of night teaching today?<br>*[Quais os problemas do ensino noturno atualmente?]* | Rose Neubauer<br>(Oct 16, 2018) | Educator and former Secretary of Education of the State of São Paulo | 760 | 18:09 | 2,693 | 1,360 | 51% |
| 32 | Does education start at home?<br>*[A educação começa em casa?]* | Guiomar Namo de Mello<br>(Oct 17, 2018) | Professor, educator, and member of the São Paulo State Board of Education [Conselho Estadual de Educação de São Paulo – CEE/SP] | 1,774 | 21:29 | 3,267 | 1,654 | 51% |
| 33 | What is literacy?<br>*[O que é alfabetizar?]* | Ilona Becskeházy<br>(Oct 18, 2018) | Education Consultant | 1,909 | 22:35 | 3,464 | 1,804 | 52% |
| 34 | What are the new skills to be developed by our students?<br>*[Quais são as novas competências a serem desenvolvidas por nossos alunos?]* | Maria Inês Fini<br>(Oct 19, 2018) | President of INEP | 1,622 | 22:54 | 3,187 | 1,683 | 53% |
| 35 | Where does Fundeb's money come from?<br>*[De onde vem o dinheiro do Fundeb?]* | Mariza Abreu<br>(Oct 23, 2018) | Education Consultant | 17,838 | 23:22 | 3,573 | 1,886 | 53% |
| 36 | What is the solution to the lack of places in public daycare centers?<br>*[Qual a solução para a falta de vagas nas creches públicas?]* | Ana Teresa Gavião<br>(Oct 24, 2018) | PhD in Education and principal at the Antonio Antonieta Cintra Gordinho Fundation. | 1,103 | 22:11 | 3,042 | 1,599 | 53% |
| 37 | What is the main problem of public higher education?<br>*[Qual o principal problema do ensino superior público?]* | Maria Helena Guimarães de Castro,<br>(Oct 26, 2018) | Conselho Nacional de Educação. | 1,339 | 23:05 | 3,155 | 1,647 | 52% |

## Appendix B – Graphical outputs

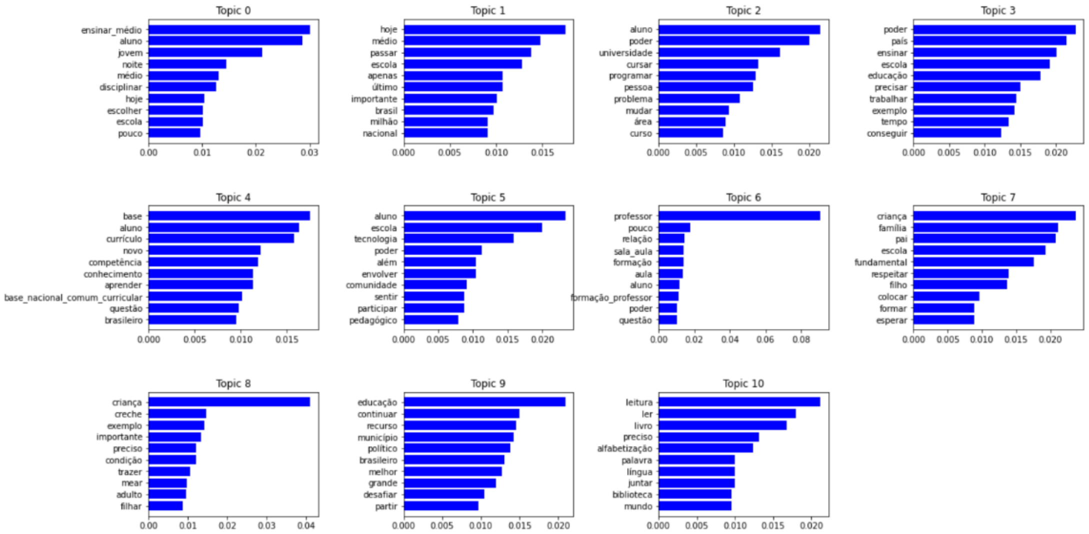

*Figure 5: Output for k-value 11 topics.*

## Appendix C – Notebook

The code commented in detail which was implemented in Python language at **Google Colab** is attached to this article in a file named PO240_CODIGO.ipynb.