Mining Students' Funds of Knowledge from Textual Data to Inform Culturally Relevant Instruction

Amir Lopatin Fellowship
Raquel Coelho

September 8, 2020

Introduction

"Big data" is a buzz word that crops up increasingly often in education research, and while many researchers have made important inroads into the areas that big data opens to them, much potential remains untapped. Over the past two decades, two research communities have emerged looking to capitalize on the promises of big data for improving student learning: Education and Data Mining (EDM) and Learning Analytics (LAK). The former's focus is technological and the latter's pedagogical, but both communities share the goal of transforming data into forms that are digestible and utilizable by students, teachers, and other relevant stakeholders seeking to improve learning (Gaševic, Dawson, & Siemens, 2015; Wise, 2019). The current project aims to contribute to this goal, but to do so in a way that incorporates an emphasis on cultural diversity and inclusion, which have been given considerably little attention in LAK and EDM research, as evidenced by recent literature reviews (e.g., Mangaroska & Giannakos, 2018; Papamitsiou & Economides, 2014).

Largely unexplored in LAK and EDM research, for example, are students' "funds of knowledge" (Hogg, 2011; Moll & Greenberg, 1990; González, Moll, & Amanti, 2006) – that is, the knowledge, skills, and social practices that students develop from their participation in everyday, out-of-school activities and then bring with them as resources to formal learning environments. This is especially true for students who have been historically marginalized, as their strengths too often go largely or even entirely unrecognized (González et al., 2006).

Culturally Relevant Education (CRE) approaches (Aronson & Laughter, 2016) include a variety of theoretical frameworks that have been proposed over the past 30 years as ways of incorporating historically marginalized students' FoK into classroom instruction. Approaches that fall under this umbrella include, for example, Culturally Responsive Teaching (Gay, 2018), Culturally Relevant Pedagogy (Ladson-Billings, 1995), and Culturally Sustaining Pedagogy (Paris, 2012). A central tenet of CRE approaches is the belief that, by tapping

into students' FoK, teachers will be better able to build trusting and fruitful relationships with students and thereby engage those students in learning in ways that have more immediate relevance to students' lives.

Despite a growing body of empirical research that documents the connection between classroom practices that capitalize on historically marginalized students' lived experiences and their engagement and learning outcomes (Aronson & Laughter, 2016; Steele, 2011), a variety of factors complicate the implementation of CRE approaches in practice. These factors include things like deficit views of students and communities, simplistic conceptions of what FoK-informed instruction might look like, conflicts with school-level policies and/or limited support from colleagues and leadership, lack of practical know-how, and limits on time and resources. The current project seeks to draw on insights gained from big-data analysis to mitigate the impediments to implementing CRE approaches that arise as a result of teachers having limited time, resources, and materials.

While the original and most FoK-inspired approaches suggest that teachers engage in "ethnographic inquiry" (González et al., 2006), visiting students and families and documenting the knowledge and skills that students display in their household contexts, teachers are overburdened as it is and it is often unrealistic to expect them to regularly be involved in extensive ethnographic research (Hattam & Prosser, 2008). Further, students' FoK are unlikely to be stagnant. Given the dynamic nature of the data thus in question, it is imperative that our methods of mining and utilizing that data be dynamic as well. Thus, while ethnographic approaches can contribute greatly and are an excellent option when feasible, it could serve the field well to complement such approaches with a big-data-backed approach that is more easily adaptable to low-resource contexts.

While such an approach might not have been feasible previously, big data is becoming increasingly accessible. As data collection technology evolves, there are increasingly-many data sources available to researchers, and the insights drawn from such data can and should be used to inform instruction. One ripe source of dynamic data that is waiting to be mined is the vast array of student-generated texts, photographs, and other assignments. Researchers such as those reviewed by Llopart and Esteban-Guitart (2018) have proposed using such student-produced artefacts as ways to obtain empirical information about students and then craft connections between curricula and students' contexts. Such use has most often been proposed in the context of a single classroom, using artefacts collected by individual teachers, but a similar idea could also be leveraged on a broader scale.

Taking up that task, in this project I present a systematic analysis of a large collection of

textual data, using computational techniques to automatically identify topics written about by a geographically diverse body of Brazilian high school students. I examine not only the topics students choose to write about, but also whether and how those topics differ between states, cohorts (year), and rural and urban students. I argue that the topics students choose to write about provide an important window into those students' FoK.

Background

Brazil's educational system is (unofficially) segregated along both racial and class lines. Public school students make up roughly 90% of the students in Brazil (BRASIL, Inep, 2019), and the 10% of students who attend private schools are not at all a random cross section of the high-school-aged population. High-quality private schools remain largely the purview of white and relatively affluent students, while most non-white, low-income students attend public schools. As such, research on public school students in Brazil can provide insights into the education of students who have historically been marginalized in Brazilian society.

While public schools in Brazil are often under resourced, there have been some government efforts to increase opportunities for students in public schools. One such opportunity that began as an initiative of civil organizations and later, through partnership with the government, became public policy is the Portuguese Language Olympics (PLO), a biyearly writing contest started in 2002 with the goal of improving public-school students' writing and reading abilities, and with a strong orientation toward cultural competence and social justice. The competition invites 5th to 12th graders nationwide to write poems, literary memoirs, chronicles, or opinion articles (or, as of recently, to create a short documentary) about the places they live. Students' teachers submit students' submissions, and thereafter the set of submissions is narrowed down in several stages (school, municipality, state, regional, and national) until the winners are eventually selected. The PLO has a very broad reach; for example, in 2019, 85,908 teachers from 42,086 schools across Brazil participated, representing 4,876 of Brazil's 5,564 municipalities. Further, the reach is not skewed geographically; about 90% of students in the PLO come from urban areas, and both the urban/rural makeup and the regional distribution of PLO students are fairly similar to wider Brazilian demographics.

Given the PLO's reach (geographically, demographically, and over time), student submissions to the PLO serve as an excellent source of large-scale textual data that could inform educational practice. Capitalizing on this potential, the current project makes use of this dataset. Specifically, the dataset used in this study consists of the collection of all opinion articles written by 11th- and 12th-grade students participating in the PLO from 2012 to

2016 who reached the municipal level of the competition. In total, this includes 14,526 essays (4,525 from 2012, 5,312 from 2014, and 4,750 from 2016), for each of which a student was asked to write (in Portuguese) a persuasive essay articulating their view on a polemic topic in the place they live. The essays have been digitized, and the data, provided to the researcher by Brazil's Center for Studies and Research in Education, Culture, and Community Action (CENPEC), also includes relevant demographic information about students, teachers, and schools.

Data Analysis

The data were analyzed using Structural Topic Modeling (STM), an unsupervised machine learning algorithm (Roberts, Stewart, & Tingley, 2014) that takes in a set of textual data and, one, identifies latent topics within each text, and, two, estimates the relationship between those topics and information about each text (topical prevalence covariates). Each text is defined as a mixture over topics, meaning that a single text can be composed of multiple topics, and the sum of the proportions of all topics across a text is always one. In the case of this study, STM is useful for discovering what students write about in their essays and for estimating the relationship between students' characteristics (such as where they live) and how much attention they give to a particular topic in their essays.

Before fitting a structural topic model, I cleaned the data in accordance with the literature in computational linguistics (e.g., Jurafsky & Martin, 2009). This involved performing a set of text normalization tasks whereby the texts were put into a form more convenient for analysis. As part of this normalization, punctuation was removed from the texts, as were stop words (recurrent, non-content-heavy words such as the articles "a" and "o" in Portuguese, equivalent to English "the"), numbers, and a few high-frequency words such as "vida," "cidade," "município," and "viver" ("life," "city," "municipality," and the verb "to live," respectively). Subsequently, all remaining words were lowercased and stemmed using an algorithm designed for Portuguese.

After estimating a structural topic model, I determined that there were 14 topics, and labeled them as follows: "School Quality and Policies," "Public and Private Investment," "Economy and Unemployment," "Tourism and Natural Attractions," "Life in the Countryside," "Violence, Crime, and Abuse in Youths' Lives," "Transportation Issues," "Land Rights Issues," "Politics and Elections," "Water Scarcity and Pollution," "Public Health Services," "Agriculture," "Cultural Heritage and Festivities," and "Local community."

I selected students' state, school location (urban or rural), and essay year as topical

prevalence covariates to be studied, and estimated expected proportions of essays belonging to each of the 14 topics as a function of these covariates. Preliminary results are discussed below.

Preliminary Findings

Figure 1, below, shows the proportions of the corpus that the model showed belong to each topic.

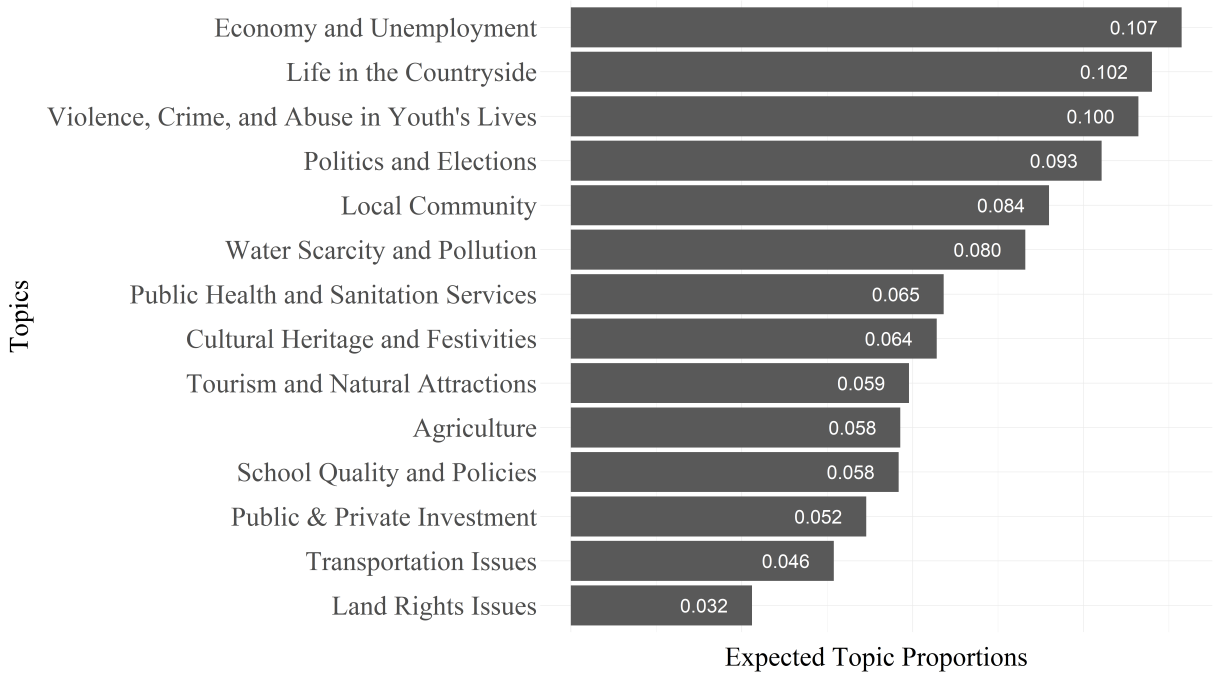

Figure 1: Top Topics in order of prevalence for the corpus as a whole

Figure 2, below, shows the proportions of the corpus that belong to each topic by state.

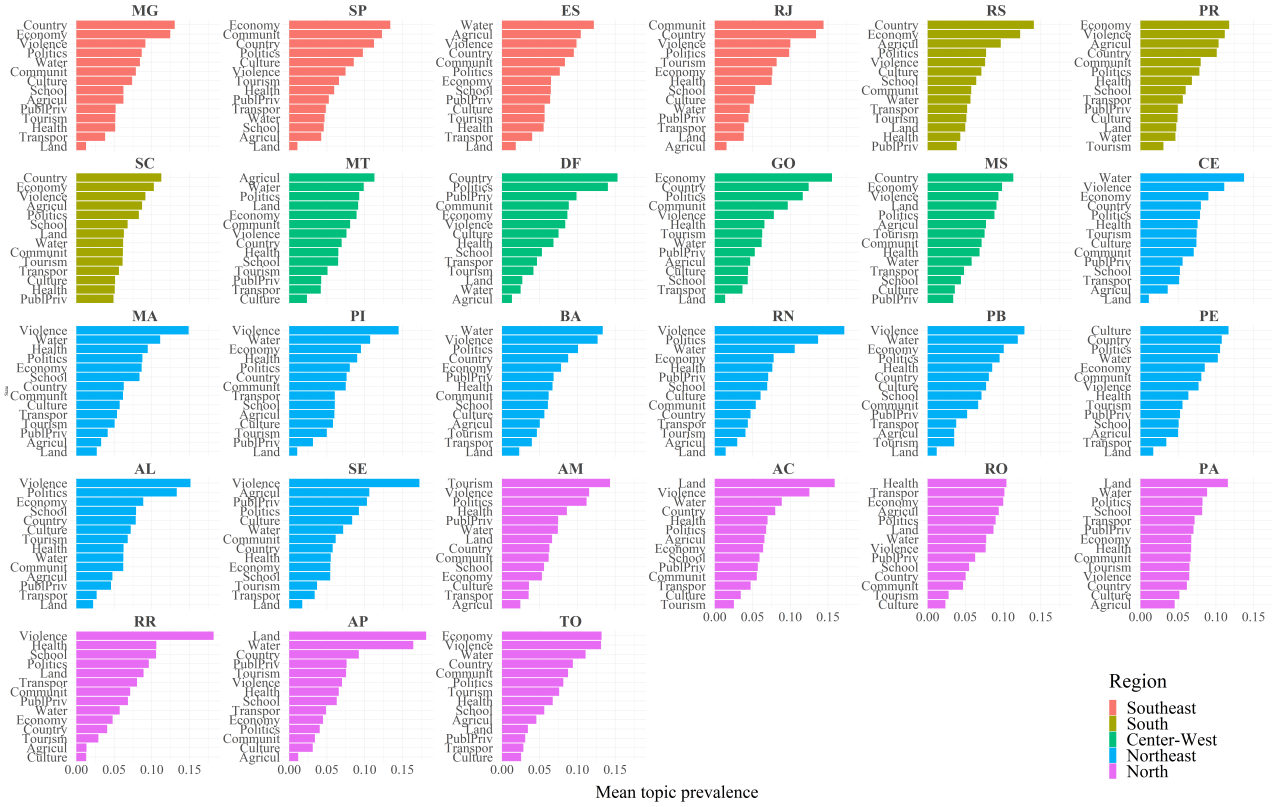

Figure 2: Top Topics in order of prevalence by state

As far as variation by state, for ten of the topics (all except "School Quality and Policies", "Transportation Issues", "Politics and Elections", and "Public Health Services"), there appears to be variation in the proportion dedicated to those topics across different states. Further, with few exceptions, the percentages of essays dedicated to each of the 14 topics appears more similar within regions than between them. São Paulo, which had the largest sample of essays, was used as a baseline. Differences were more pronounced and frequent between São Paulo (which is in the Southeast) and states from other regions, especially in the North and Northeast, than between São Paulo and other states in the Southeast.

As to topics that pattern along the rural/urban divide, certain topics (such as "Agriculture," "Environmental Issues," and "Local Community") were associated with rural areas, while others (such as "Public and Private Investment," "Cultural Heritage and Festivities") were strongly associated with urban areas. Other topics were observed across both urban and rural samples but were still more strongly associated with one type of area or the other;

this is the case, for example, for Violence, Crime, and Abuse in Youths' Lives, which was more strongly associated with urban areas, and Land Rights Issues, which was more strongly associated with rural areas. Still other topics, such as School Quality and Policies and Life in the Countryside, were associated more evenly with both urban and rural areas. Excepting three topics ("School Quality and Policies," "Life in the Countryside," and "Land Rights Issues"), further analysis confirmed statistically significant differences in topic proportion in student essays between rural and urban areas.

Finally, with respect to year, ten topics (all except "Lands Rights Issues," "Water Scarcity and Pollution," "Public Health Services," and "Local Community") were found to vary statistically significantly by year. Five topics' prevalence increased from 2012 to 2014 and 2016: "School Quality and Policies;" "Public and Private Investment;" "Violence, Crime, and Abuse among Youth;" "Transportation Issues;" and "Cultural Heritage." Conversely, five topics' prevalence decreased from 2012 to 2014 and/or 2016: "Economy and Unemployment," "Tourism," "Life in the Countryside," "Politics," and "Agriculture."

Further Research

Building on these preliminary results, I will perform further analysis on this data with the goal of extracting insights that teachers can apply to their curricula and instruction.

Moving forward, my goals are threefold: First, I aim to examine the relationship between what students write and textual genre. To that end, I am currently enlarging and diversifying my corpus to include the remaining Portuguese Language Olympics genres: memoirs, poems, chronicles, and documentaries. Here too students have the liberty to choose their own topics. My second goal is to study vocabulary differences in the writing of students from different regions, so as to understand what kinds of insights can be gained not only from what students write about, but from how they write about it. Thirdly, I will work with science, math, and language arts teachers in Brazil to understand how this information could be used to relate local curricula and school activities to students' lives outside of school. More specifically, my goal is to study how teachers use these unique insights about students in the process of designing learning experiences that are closer and more meaningful for their students.

Acknowledgements

I am deeply grateful to the Amir Lopatin Fellowship for funding this study. Without your generous support, the next stage of this study – the work with teachers – would not be

possible. I also want to extend my thanks to the Portuguese Language Olympics team at CENPEC, in particular Maria Cida Laginestra and Fabiana Cortello, for sharing the data and their expertise with me. I am also grateful to my friends and colleagues Alden McCollum, Lynne Zummo, and Emma Gargroetzi for their invaluable insights. A special thank you to Paulo Blikstein, for his incredible generosity with his time, expertise, and encouragement. Finally, I want to express my most sincere gratitude for my advisor, Roy Pea, for being the advisor that most could only dream of.

References

Aronson, B., & Laughter, J. (2016). The theory and practice of culturally relevant education: A synthesis of research across content areas. Review of Educational Research, 86(1), 163-206.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2019). Censo da Educação Básica 2019: Resumo Técnico. Brasília.

Gay, G. (2018). Culturally responsive teaching: Theory, research, and practice. Teachers College Press.

Gaševic, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64-71.

González, N., Moll, L. C., & Amanti, C. (Eds.). (2006). Funds of knowledge: Theorizing practices in households, communities, and classrooms. Routledge.

Hattam, R. & Prosser, B. (2008). Unsettling Deficit Views of Students and their Communities. Australian Educational Researcher, 35, 89-106.

Hogg, L. (2011). Funds of knowledge: An investigation of coherence within the literature. Teaching and Teacher Education, 27(3), 666-677.

Jurafsky, D. & J. H. Martin. 2009. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (2nd ed.). Pearson Education Inc., Upper Saddle River, NJ.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. American Educational Research Journal, 32, 465–491.

Llopart, M. & Esteban-Guitart, M. (2018). Funds of knowledge in 21st century societies: Inclusive educational practices for under-represented students. A literature review. Journal of Curriculum Studies, 50(2), 145-161.

Mangaroska, K. & Giannakos, M. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. IEEE Transactions on Learning Technologies.

Moll, L. C. & Greenberg, J. (1990). Creating zones of possibilities: Combining social contexts for instruction. In L. C. Moll (Ed.), Vygotsky and education: Instructional implications and applications of sociohistorical psychology (pp. 319-348). Cambridge: Cambridge University.

Papamitsiou, Z. & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. Educational Technology & Society, 17, 49-64.

Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. Educational researcher, 41(3), 93-97.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R package for structural topic models. Journal of Statistical Software, 10(2), 1-40.

Sleeter, C. E. (2012). Confronting the marginalization of culturally responsive pedagogy. Urban Education, 47(3), 562-584.

Wise, A. F. (2019). Learning analytics: Using data-informed decision-making to improve teaching and learning. In Contemporary technologies in education (pp. 119-143). Palgrave Macmillan, Cham.