TEXT AS DATA METHODS FOR EDUCATION RESEARCH

Otamurodova Sabinur
Employee of academic lyceum of Tashkent pharmaceutical institute

ABSTRACT

Recent advances in computational linguistics and the social sciences have created new opportunities for the education research community to analyze relevant large-scale text data. However, the take-up of these advances in education research is still nascent. In this paper, we review the recent automated text methods relevant to educational processes and determinants

Key words: technique; text; untapped data; social media; scale;

ANNOTATSIYA

Hisoblash tilshunosligi va ijtimoiy fanlar sohasidagi so'ngi yutuqlar ta'lim taraqqiyot hamjamiyatiga tegishli keng ko'lamli matim ma'limotlarn tahlil qilish uchun yangi imkoniyatlar yaratdi. Biroq, ta'lim sohasidagi taraqqiyotlardagi bu yutuqlarni qo'llash hali ham yangi emas. Ushbu maqolada biz ta'lim jarayonlari va determinantlarga tegishli so'ngi avtomatlashtirilgan matn usullarini ko'rib chiqamiz

Kalit so'zlar: texnika; matn; foydalanilmagan ma'lumotlar; ijtimoiy tarmoqlar; masshtab

INTRODUCTION

The education-research community is in the middle of two advances that are changing the way we can analyze and understand educational processes. First, more than ever in the past, we have access to broad, rich, educationally-relevant text data from sources such as online discussion forums, transcribed video and audio of face-to-face classes, digital essays, social media, and emails and texts between parents and schools. Second, computational linguists and social scientists have recently developed more advanced and nuanced tools that can be applied to quantitatively analyze large-scale text data (Gentzkow, Kelly, & Taddy, 2017; Grimmer & Stewart, 2013; Jurafsky & Martin, 2018). Together, these advances in data availability and analytic techniques can dramatically expand our capacity for discovering new patterns and testing new theories in education. They can expand the types of research questions that researchers can ask, improve the external validity and representativeness of research, and reduce the cost to complete these types of studies.

Researchers can now ask questions that rely on previously untapped data sources, utilize more real- time data, and focus on mechanisms in addition to impacts. For example, researchers can utilize the discussion board forums of online classes to study classroom interactions, teachers' application essays to understand teacher motivation (Penner, Rochmes, Liu, Solanki, & Loeb, 2019), district web sites to catalog school improvement plans, or newspapers to study how educational policies are discussed in the media (Mockler, 2018). Instead of asking teachers to recall any professional development opportunities they have utilized, researchers can directly study the topics that teachers discuss in online professional development communities in real time (Anderson, 2018).

An Overview of Automated Text Analysis Methods

The existing methods for analyzing text generally cluster into three categories: lexical-based methods, supervised machine learning, and unsupervised machine learning. Lexical-based methods rely on simple document-level[1] word counts to accomplish tasks as diverse as measuring sentiment in a student essay, determining whether an online discussion post was directed at a particular student (Bettinger, Liu, & Loeb, 2016), and measuring policy uncertainty in newspaper articles (S. R. Baker, Bloom, & Davis, 2016). Supervised machine learning relies on researchers' hand-coded documents to train a model and predict the codes for an unlimited number of un-coded documents, which can dramatically expand the number of documents that a researcher can code. For example, Kelly, Olney, Donnelly, Nystrand and D'Mello (2018) use supervised machine learning to predict when teachers ask questions without predetermined answers by building a model based on a set of hand-coded questions. Similarly, Gegenheimer, Hunter, Koedel and Springer (2018) hand-code teacher observation feedback into seven major domains for a subset of teachers, creating a foundation for using supervised machine learning to predict the domains for all teachers in Tennessee.

Unsupervised machine learning procedures instead identify groupings in the data without relying on prior hand-coding, which is a less structured approach that can be particularly appropriate for researchers in a more descriptive and hypothesis-generating posture. For example, Beattie, Laliberte and Oreopoulos (2018) use unsupervised machine learning to examine open survey responses of students who perform significantly better or worse than predicted in their first year of college. They find that students who perform lower than expected are more likely to discuss topics like "getting rich" quickly, and students who perform better are more likely to discuss more philanthropic goals.

Applications of Text Analysis Techniques to an Experimental Study of Gender Bias in Education

In this section, we demonstrate both how to use lexical-based methods, supervised machine learning, and unsupervised machine learning and how to understand their affordances in the context of studying gender bias in an educational setting. Specifically, we rely on the text data from a field experiment conducted in the discussion board forums of 124 MOOCs. In this experiment, we created student profiles with names that sounded female or male, and randomized those fictitious students to post one of 32 different generic comments. We collected all of the written responses real instructors and students sent tothese fictitious posters, which allows us to use automated text analysis to examine how the written responses differ when they are sent to the fictitious female students as opposed to the fictitious male students.

Conclusion

In this study, we provided a comparative overview of a variety of techniques that have recently emerged for the study of text as data in the social sciences, including lexical-based methods, supervised machine learning, and unsupervised machine learning. We illustrated the use of these techniques (and their comparative properties) through the empirical example of gender bias in online classrooms using data from a field experiment in which the identities of discussion forum posters were randomly manipulated. We show that respondents provide 7 percentage points less assistance and 7 percentage points more individual attention to (the black and white) female posters. We see no significant differences in positive and negative sentiment, and small differences in topics discussed in response to male and female posters.

Each of these methods has its promises and pitfalls. Lexical-based methods are generally faster and easier to use than supervised machine learning methods but can perform poorly when applied to a different sample or domain and require validation. In our empirical application, we showed that less than half of the instances that our two lexical-based methods predicted as being positive were actually positive when compared to human coding. However, the relatively poor prediction performance did not induce much bias in our experimental analysis. Supervised machine learning requires a higher start-up cost of hand-coding, but it can perform well when it is trained in a similar sample with large enough samples.

References

Anderson, R. (2018). Beyond copy room collaboration: A case study of online informal teacher professional learning. In *Rethinking Learning in the Digital Age: Making the Learning SciencesCount, 13th International Conference of the Learning Sciences (ICLS)* (Vol. 3, pp. 1511–1512).London, UK: International Society of the Learning Sciences.

Baker, R., Dee, T., Evans, B., & John, J. (2018). *Bias in Online Classes: Evidence from a Field Experiment* (CEPA Working Paper No. 18–03). Retrieved from http://cepa.stanford.edu/wp18-03

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The QuarterlyJournal of Economics, 131*(4), 1593–1636.

Beattie, G., Laliberté, J. W. P., & Oreopoulos, P. (2018). Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, *62*(January 2017),170–182.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable

Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4),991–1013.

Bettinger, E., Liu, J., & Loeb, S. (2016). Connections Matter: How Interactive Peers Affect Students in Online College Courses. *Journal of Policy Analysis and Management*. https://doi.org/10.1002/pam

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine LearningResearch*, *3*, 993–1022.

Bohren, J. A., Imas, A., & Rosenberg, M. (2017). *The Dynamics of Discrimination: Theory and Evidence*