Abstract

Factors influencing secondary educational attainment (staying in high school) are well characterized, but several shortcomings of current educational attainment theory preclude a real-time and wide-scope understanding of the issue. Further insight could be provided by text analytic technology. In this article, the authors review the current status of secondary educational attainment and its associated theories, and identify areas of weakness which could be reinforced by adopting text analytic research methods. The authors conclude with a discussion of potential applications of text analytic research within school systems aimed at equipping educators and counselors with additional resources to more accurately pinpoint intervention efforts.

Keywords

Text analytics, Educational attainment, interventions, predictive modeling

Introduction

Achieving a high school diploma is associated with favorable late-life social and economic outcomes, but a significant portion of United States students do not complete high school. Dropping out of high school is theorized as a process of gradual loss of engagement with school, beginning in students' years in elementary school (Alexander, Entwisle, and Horsey 1997). This engagement with academic success relies on interactions between many social, individual, cultural, family, and school factors (Janosz, Archambault, Morizot, and Pagani 2008) including educational performance, individual behaviors such
as absenteeism and alcohol use, school-related attitudes, and background characteristics such as socio-economic status (Rumberger and Lim 2008).

However, the picture is not complete, as some students who drop out of high school actually make adequate grades (Alexander, Entwisle, and Dauber 2003). Further insight into the issue could be provided by text analytic technology, which uses computer algorithms to efficiently extract meaningful data from text sources such as essays, emails, and social media profiles. Such data could potentially be used to more strongly predict those students at the highest risk of dropping out, and facilitate more accurately targeted intervention efforts. The present article will briefly review the current state of secondary educational attainment, then provide an overview of text analytic / text mining technology and its present uses. We will conclude with a discussion of how text analytic software could be implemented in school systems to equip educators and counselors with additional tools to identify those students at the greatest risk of dropping out.

Educational Attainment in the United States

According to a report from the U.S. Department of Education, 8.1% of individuals in the United States between ages 18 and 24 do not possess a high school degree or GED (Chapman, Laird, and KewalRamani 2011). This is a pressing issue for a number of extremely important reasons. First, individuals who drop out of high school result in an estimated $200,000 in increased government spending and lost taxes (Alliance for Excellence in Education 2011). Second, they are at greater risk of negative health outcomes (Pleis, Ward, and Lucas 2010). Third, they are more likely to be periodically unemployed or cycling in and out of the prison system (Alliance for Excellence in Education 2011). In fact, if those who had dropped out of the class of 2011 had graduated, they would have generated approximately $154 billion in additional income for the U.S. economy (Alliance for Excellence in Education 2011).

Scholars have used models rooted in diverse theory domains to determine why students drop out of high school. Recent findings suggests that staying in school is strongly a function of motivation, and that students’ motivation to remain in school is influenced by a variety of sources including teacher expectations, peer aspirations, and their perceptions of the value of education (Cham, Hughes, West, and Im 2014). Other models paint a more complex picture and hypothesize that student perceptions of teacher & parent support (social context) predict students’ identification with the school (self-perceptions), which then predicts academic engagement and dropout rates (Fall and Roberts 2012). Others suggest that models using only two possible outcomes – “still in school” or “dropped out” are missing an important third outcome, being pulled out due to family/financial obligations (Bradley and Renzulli 2011) and painting an incomplete picture.

While these explanations are theoretically sound, their application to real educational attainment situations is not as successful as we may expect. This may be due to several reasons. First, the variables of interest in the leading educational attainment theories aren’t measured in a timely fashion. The day-to-day, week-to-week, and cumulative effect of some of these influences are both difficult to assess and even more difficult to determine their relationship with academic success. Second, the number of variables that previous research studies can assess is limited. Thus, all of these theories might work better together to be an effective predictor of engagement. Text analytics could help provide a solution to these issue by (1) providing more up to the minute measurement of important variables and continually examine the relationship between these variables and important outcomes of interest, an (2) providing a more effective way to assess a larger number of variables of interest and test multiple or combined theories.
Text Analytics and its Uses

Technological limitations have, until recently, relegated educational attainment researchers (and perhaps social science researchers in general) to the realm of the easily measurable and quantifiable. Writing and free-response-based data collection, and the subsequent interpretation and coding, are exceedingly time consuming with large studies. Due to the sheer amount of work hours previously required, vast sources of student data, such as writing samples and online profile content, have not been quantified and remained unexamined. Text mining, or text analytics, represents a new tool at researchers’ and educators’ disposal that could allow them to efficiently utilize this untapped resource, test and develop theories of academic engagement, and build superior prediction models for educational attainment.

The term “text analytics” refers to the process of deriving data from text, essentially allowing text to be used as a substrate for statistical analysis (Yang et al. 2013). The process is carried out by sophisticated computer software designed to use lexical analysis, pattern recognition, concept/entity extraction, sentiment analysis, and link and association analysis in order to convert text and writing into structured data that can be statistically analyzed. This technology, available from companies such as SAS (Zhao, Albright, Cox, and Bieringer 2013) and IBM (Yang et al. 2013) allows users to fine-tune the software to search for and analyze unstructured text data from specified sources, including online databases, local networks, and social media profiles. Adopting text analytics has allowed modern businesses a first-time view into the vast amount of unstructured data that is being produced at an unprecedented rate. For example, Twitter recently announced that its users generate over half a billion tweets daily (Zhao et al., 2013), and without text analytics, it would be an impossibly difficult task for companies to analyze, monitor, and predict what their customers are saying about them in real time.

Most text analytic software relies on machine learning to automate the process of evaluating bodies of text. This involves using sophisticated search and parsing algorithms to comb through the text, along with human-entered rules about how to interpret ambiguous segments. The software keeps track of the user’s patterns of categorization, and eventually “learns” the rules that the user relies on when categorizing ambiguous content. Eventually, the software is able to work largely independently and analyze information accurately.

For example, software from SAS (SAS 2011) can be used by businesses to automatically search the web for, and analyze the sentiment of, consumer-generated media (such as blog posts), in order to identify good and bad influencers for their corporate image or brand, in real time. Xiang, Schwartz, Gerdes, & Uysal (2015) used exactly this function to review large quantities of reviews from Expedia.com. They were able to deconstruct guests’ hotel experiences and identify the important dimensions of guest experience that correlated most strongly with satisfaction ratings (Xiang et al. 2015). Sifting through the sea of unstructured data known as consumer reviews, and extracting data that can be statistically analyzed in real time as it is generated, is obviously of great interest to businesses. Text analytics may represent the most efficient, cost-effective method of doing so.

However, text analytics isn’t just a tool for businesses. Text analytics represents a promising new method for predicting educational attainment due to the fact that the information which may facilitate a more accurate understanding of the issue and identify effective intervention efforts may be “locked up” in an unstructured form. In other words, students’ writing (e.g., papers, homework questions, or other assignments) could prove to be a strong predictor of risk of drop-out, but text analytic software is needed in order to “convert” such unstructured data into a form of data that can be statistically analyzed.
Some researchers have already begun to use text analytics to study students’ free-response, short answer exam questions. Leeman-Munk, Wiebe, and Lester (2014) recently designed software that analyzed fourth-grade students’ responses to short-answer questions, and found that the output grade from the software correlated with student performance. This type of application is not a novel phenomenon: Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter, and Person (2000) developed a fully automated online computer science tutoring system that is capable of evaluating college students’ answers to essay-based deep reasoning questions at the same level of accuracy as intermediate computer literacy experts. These are just two examples of how academia has already used text analytics to extract meaningful, analyzable, structured data from sources that previously would have demanded human evaluation.

**Combining Educational Attainment and Text Analytics**

The study of educational attainment stands to benefit from the adoption of text analytic research methods. On a very basic level, text analytics could be used to effortlessly evaluate students’ average sentence complexity and frequency of grammatical or spelling errors, factors which could plausibly predict students’ educational investment. More complex functions of text analytics, such as sentiment analysis, could allow easy screening of writing samples for salient elements – perhaps the presence of negative sentiments towards school or education, or positive attitudes towards social deviance, factors which are known to negatively predict educational attainment (Battin-Pearson et al. 2000).

Possible sources of analyzable material include writing samples, school emails, or other electronic school communications, as legal boundaries and privacy concerns permit. Public school systems often use online assignment submission systems where students turn in papers and other writing-based work electronically. Gaining access to such systems would afford us a rich data source that minimally imposes upon teaching professionals.

A key benefit of text-analytic based research over traditional educational attainment research methods is this lack of an intrusive element. Text analysis relies on a passive accumulation of data, rather than the active collection methods of surveys. In other words, the data is already being generated and does not require anyone to go out of their way to produce more of it. As an example, one potential application could be longitudinally tracking students’ writing samples, creating a standardized composite score based on sentence coherence/complexity and presence of negative attainment predictors such as negative or indifferent attitudes towards education, and tracking these scores over time. From such an analysis, we may discover that a sudden decrease in standardized scores during one academic year predicts dropout during the next – and the only data “collection” needed would be obtaining the electronic documents from teachers.

One other benefit to be found in adopting text-analytic research is the real-time element of text analytics. Text analytic programs constantly incorporate new data as they are created, alleviating the fear of outdated models. For example, what may have been a predictor for dropout among a given cohort in 2014 may not necessarily predict dropout in 2015. In this way, text analytics accounts for social and cultural shifts in a quick and adaptive way that other forms of research cannot match. Finally, the ultimate goal of applying text analytics to the educational attainment domain would be to enhance the quality and targeting of intervention programs. By using the vast amount of unstructured data available to us, we may be able to discover latent variables that both strongly predict dropout rates and are plausible intervention targets.
Concluding Remarks

In the knowledge economy of the 21st century, the demand for college-educated workers has never been higher. Adolescents who fail to complete high school are effectively barred from many positive social and economic outcomes, and face increased risk of negative health and wellness outcomes. In light of this, educational attainment researchers and practitioners should consider putting text analytics in their toolbox. Passive data collection, access to huge amounts of data, and real-time adaptability with the capacity to unlock massive, previously untapped sources of data make text analytics an extremely attractive research tool.

References