Social Media Analysis for Higher Education

Anamaria Berea, William Rand, Kevin Wittmer
University of Maryland

Gerard Wall
vibeffect

Abstract

The educational system involves a complex set of actors, including learners, parents, teachers, and administrators. However, we now have more data than ever to analyze this system, which could result in a quick understanding and evaluation of public policies in this complex policy area. This paper explores a new area of data about the educational experience, namely social media data. This paper outlines an exploratory analysis of the Twitter discussions regarding higher education in the USA. Based on a collection of more than 1.5 million tweets over a period of 4 months, we identify a few key issues in the current higher education discourse on social media. We also identify the effect of the expressed feelings of the social media users when it comes to college applications, decisions and completion. We conclude that policies in higher education can be better tailored if they are informed by social media discussions.

Introduction

THE INCREASING AMOUNT OF DATA, the decreasing cost of computational power, and the improving state of analytics has revolutionized fields from stock trading to social analytics, but somehow higher education has not received as much attention. The technology that has transformed many for-profit businesses and governments can be applied at various colleges and universities.

One obvious place that analytics could be useful is in the classroom, but currently instructors at many universities are using outdated and inefficient methods to grade assignments and compile these scores into self-generated databases. In fact, Darrell West argues that "many of the typical pedagogies provide little immediate feedback to students, require teachers to spend hours grading routine assignments, are not very proactive about showing students how to improve comprehension, and fail to take advantage of digital resources that can improve the learning process" (West 2012). Data mining and analytics provide the capabilities necessary to circumvent the traditionally cumbersome grading processes and glean

insights from student data about performance, learning approaches, and other metrics. For example, Leah Macfadyen and Shane Dawson developed an "early warning system" which correctly identified 81% of students who failed an online course by creating a regression model that analyzed such variables as total number of discussion messages posted and total number of assignments completed (Macfadyen and Dawson 2010).

Big data analytics within education could also be used to monitor student progression through various course sequences for specific majors, online courses that change activities by measuring everything from individual clicks to aggregate performance and algorithms that suggest courses a student should take by analyzing her past grades in similar courses (Bienkowski *et al*. 2012). While traditional in-person classrooms may allow for the collection of big data for these applications, Anthony Picciano notes "to move into the more extensive and especially time-sensitive learning analytics applications, it is important that instructional transactions are collected as they occur" (Picciano 2012). This rapid collection of data is most likely to be facilitated by course management/learning management system architectures and online and blended learning course structures (Worsley 2012).

There is little work that has looked at how to use analytics methods outside the classroom to improve the overall educational ecosystem, as well as educational policy. However, insights produced by the previously described learning analytics systems can also be used to inform policy decisions. According to van Barneveld, Arnold, and Campbell (2012), "Like business, higher education is adopting practices to ensure organizational success at all levels by addressing questions about retention, admissions, fund raising, and operational efficiency". Michael Horn and Katherine Mackey (2011) suggest that education analytics can be used to shift the focus from inputs to outputs when measuring academic institutional success. Instead of using seat-time, faculty-student ratios, and dollars spent as a measure of success, analytics software can provide information on more appropriate metrics such as student performance and retention rates. The biggest obstacles to establishing more such systems are building data sharing networks where these myriad metrics can be aggregated, holistically analyzed, and shared among different institutions (West 2012). A recent paper proposes a model and algorithm that would

help prospective students make better informed decisions about the best fit and best college eco-system based on their unique personalities and behaviors (Berea *et al*. 2015).

Text mining, social media, or sentiment analysis on the college decision process has generally not been discussed in education analytics literature and therefore presents an interesting opportunity to further advance research in this area. A recent survey by Piper Jaffray found that teens are abandoning Facebook in favor of Instagram; 76% of teens are on Instagram and they are using it to gain an unfiltered look at colleges (Stampler 2015).

Data Analysis

We collected data for this education analytics project for a period of 4 months, between March 4th and July 1st, 2015. For this collection we used TwEater, an original and proprietary collection tool developed at the University of Maryland (TwEater 2015). Originally, the collection was based on 57 keywords and hashtags, such as: "igotin", "college", "campus", "acceptanceletter", and many more, and the original data set comprised more than 10 million tweets. Since most of these keywords were not necessarily related to the idea of higher education and college admissions and applications, we selected a list of 25 hashtags pertaining exclusively to college, high school and higher education. Out of these, only 20 rendered more than a tweet, with a minimum of one tweet for the hashtag #choosingacollege and a maximum of 1,153,618 tweets for the hashtag #college followed by 282,139 tweets for the hashtag #campus (see Table 1).

On the basis of this collection, we assembled a data set of 1,523,817 tweets where most of them (73%) refer to the general idea of "college". Many of these tweets are quite general, but some of them focus on specific issues, such as: making college applications friendlier for the LGBT community, businesses supporting campuses, parent-student conflicts in college decision making, and hard college choices between various schools.

Text Mining

Based on this collection, we built a dictionary of about 470,000 unique words that are specific to the discourse about higher education in the

USA. This is a dictionary roughly half the size of the English language (the Oxford English Dictionary has over 600,000 words alone), with the caveat that some of the words in our dictionary are informal or abbreviations or pronouns that may not be currently recognized as being part of the formal English.

The most frequent words in the education discourse are “campus” and “college”, but if we leave these obvious terms aside, words such as “highschool”, “acceptance”, “life”, and “met” are highlighted as the most frequent ones that are not directly related to colleges. This gives us an indication that students do talk about college acceptance, life, and college related meetings on Twitter.

We also analyzed each of the 20 keywords separately and created a histogram of word frequency for each of the 20 keywords. After “college”, “campus”, “higher education” and “highschool”, the largest corpuses (indicated by the number of tweets) belong to hashtags such as #collegeopportunity, #collegetour and #collegebound. The second most frequent word in most corpuses is “student”. Some interesting words, which are sparse (low frequency) but appear more than once and are associated with the most frequent terms mentioned above, are terms such as: “success”, “community”, “hard”, “chip” and “app”. There is a very large gap between the most frequent words and the second most frequent words (showing the long tail distribution of the words) (see Table 1).

Twitter only allows for a fixed number of characters per tweet, therefore we also checked how many unique words are being used in a tweet in our data: #collegedecision, #collegechoice and #backtocollege have the most “rich” tweets (an average of ~7 words per tweet), while #collegeopportunity has the least number of unique words per tweet (an average of 0.2), probably due to a different type of content used in the tweet (*i.e.*, hyperlink or video) (see Table 1).

Table 1. The summary statistics for higher education Twitter data

	Number of tweets	Highest frequency	Second highest frequency	Absolute sentiment score	Corpus size (no. unique words)	Words per tweet
Min	6	2	2	-461647	17	0.2072
1st Quart	14.8	10.25	3.75	3.8	91.5	1.2242
Median	198	155.5	50.5	37.5	515	3.1876
Mean	76190.9	20985.15	3399.55	-34069.8	37949.7	3.2889
3rd Quart	2562.2	838.5	151.25	414.8	2727.2	4.5682
Max	1153618	193802	48383	1945	469748	7.3636

Sentiment Analysis

We matched the words in each of the 20 dictionaries with the AFINN standard sentiment dictionary (Nielsen 2011) and calculated the sentiment scores of the tweets in our data (see Figure 1). The AFINN dictionary uses a scale from -5 to +5 to rate the effect of approximately 2000 words. We calculated both the absolute and the weighted scores for each keyword. The absolute scores show that the first largest corpuses ("campus", "college" and "highschool") are also strongly negative, while all the rest are positive (with the exception of #backtocollege, where the absolute score is only -1, close to neutral, and #collegedecision, which is 0). However, raw sentiment scores do not take into account the volume of the tweets for each keyword. We therefore examine weighted sentiment scores – on corpus size and on tweet – since the distributions of the corpus sizes and number of tweets are quite skewed. The weighted sentiment scores show that #highschool is the most negative talk on Twitter, while #collegematch and #collegeopportunity are the most positive ones.

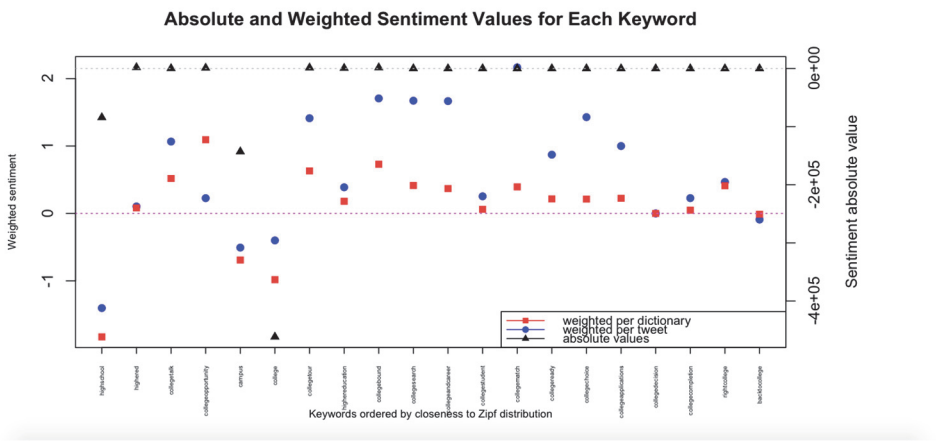

Figure 1. Sentiment values for each keyword.

2.3. Zipf's and power law distributions

Zipf's law is a well-known statistical regularity observed in natural language (Zipf 1949) that states that the frequency of any word is inversely proportional with its' rank in the frequency table. We tested whether the Zipf law holds for each of the 20 corpuses and found that #highschool, #highereducation, and #collegetalk have distributions similar to the Zipf distribution (power of ~ -1), while #backtocollege, #rightcollege, and #collegecompletion show the farthest departures from the Zipf distribution (with power of ~ -0.3) (see Figure 2).

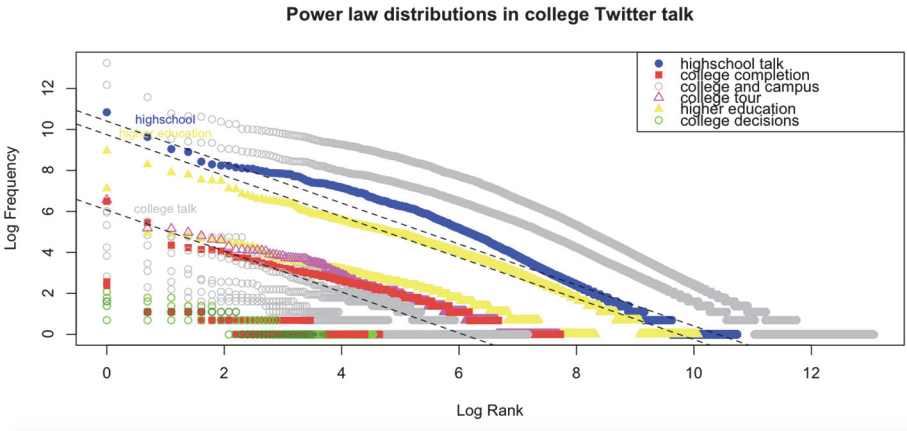

Figure 2. Power law and Zipf distributions of words.

One way to interpret this result (backed by the size of the corpuses, as well) is that there is more actual discussion involved in general topics about high school and college education as opposed to topics about college completion or matching, where the Twitter activity is more likely to inform with links and other types of information as opposed to offering opinions and personal insights and affect. There is no explanation today for why Zipf's law is characteristic to human language, but some prior research suggests that this distribution is more characteristic to natural language and the human memory of language (Cohen *et al*. 1997; Piantadosi 2014). Therefore tweets that contain other type of content than words are less likely to exhibit this pattern.

Retweets

Retweets in any Twitter data are one way to measure the degree of popularity of certain tweets. In our data the retweets to tweets ratio is quite high. The two keywords with the highest retweet to tweet ratio, "rightcollege" and "collegeopportunity", had retweeting activity for almost each and every tweet -- 0.933 and 0.903 respectively – but this is due to the majority of the tweets with "collegeopportunity" that are initiated by the users of @WhiteHouse and @BarackObama, which are popular and frequently retweeted.

Disregarding these outliers, the two keywords with the highest retweet to tweet ratio are "highered" and "collegebound" at 0.484 and 0.468, respectively – almost half of the tweets being retweeted. The high retweet to tweet ratio of "collegebound" provides an interesting insight in the context of this project. It indicates that many high school seniors revert to Twitter to broadcast their accomplishments to friends, who share the congratulatory experience. This conclusion is supported by a reading of the tweets. Many of the keywords with high absolute numbers of tweets also have moderately high retweet to tweet ratios, namely "highschool," "campus," and "college" at 0.443, 0.399, and 0.356 respectively.

Conclusion

Our analysis is constrained to only about 4 months of collection and a short list of keywords. But even so, our findings show the following: there is generally a negative sentiment regarding colleges, campuses, high school

and higher education; there is a tension between students and parents with respect to college decisions; campuses and colleges are being judged with respect to their inclusions (*i.e.*, LGBT); people are more interested in offering their opinions on general subjects (*i.e.*, "campus") than on specific ones (*i.e.*, "college tours", "back to school").

Our current research, although exploratory, points towards a few general conclusions when using social media or Big Data for education research. First, the selection of keywords and hashtags is essential, as these are going to determine the constraints for the data that are going to inform any analysis. Second, while there is considerable discussion on Twitter with respect to higher education, most of this discussion is negative. Third, social media is a great resource of information for education policy, as it gives in real time the opinions of the parents and prospective students when it comes to college applications, college acceptance, or college campuses.

Acknowledgements

The authors wish to thank Mrs. Ellie Cox for support and partnership in initiating and conducting this research. The work has been entirely supported by vibeffect.

References

Berea, Anamaria, Maksim Tsvetovat, Nathan Daun-Barnett, Mathew Greenwald, and Elena Cox. 2015. "A new multi-dimensional conceptualization of individual achievement in college." *Decision Analytics* 2(1): 1-15.

Bienkowski, Marie, Mingyu Feng, and Barbara Means. 2012. "Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics." *United States Department of Education Briefs* [https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf (accessed 6 September 2015)]

Cohen, Albert, Rosario, N. Mantegna, and Shlomo. Havlin. 1997. "Numerical analysis of word frequencies in artificial and natural language texts". *Fractals*, 5(01): 95-104.

Horn, Michael B. and Katherine Mackey. 2011. "Moving from Inputs to Outputs to Outcomes." *Innosight Institute*. [http://www.christenseninstitute.org/wp-

content/uploads/2013/04/Moving-from-Inputs-to-Outputs-to-Outcomes.pdf (accessed 7 September 2015)]

Macfadyen, Leah P. and Shane Dawson. 2010. "Mining LMS data to develop an 'early warning system' for educators: A proof of concept." *Computers & Education* 54(2): 588-599.

Nielsen, Fenn. 2011. "a new ANEW: Evaluation of a word list for sentiment analysis in microblogs" in *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages* 718 in {CEUR} Workshop Proceedings: 93-98. [http://arxiv.org/abs/1103.2903 (accessed 7 September 2015)]

Oxford English Dictionary. 2015. [http://public.oed.com/about/, (accessed 7 September 2015)]

Piantadosi, Steven T. 2014. "Zipf's word frequency law in natural language: A critical review and future directions". *Psychonomic bulletin & review*, 21(5): 1112-1130.

Picciano, Anthony G. 2012. "The Evolution of Big Data and Learning Analytics in American Higher Education." *Journal of Asynchronous Learning Networks* 16(3): 9-20.

Oxford English Dictionary. 2015. [http://public.oed.com/about/, (accessed 7 September 2015)]

Rand, William, David Darmon, and Radu Machedon, 2015. *TwEater.* [https://github.com/CenterForComplexityInBusiness/ (accessed 7 September 2015)]

Stampler, Laura. 2015. "How High School Students Use Instagram to Help Pick a College", Time Magazine, [http://time.com/3762067/college-acceptance-instagram-high-school/ (accessed 7 September 2015)]

Van Barneveld, Angela, Kimberly E. Arnold, and John P. Campbell. 2012. "Analytics in Higher Education: Establishing a Common Language." *Educause Learning Initiative*: 2-11.

West, Darrell M. 2012. "Big Data for Education: Data Mining, Data Analytics, and Web Dashboards." *Governance Studies at Brookings*: 1-10.

Worsley, Marcelo. 2012. "Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces." in *Proceedings of the 14th ACM international conference on Multimodal interaction (ICMI '12). ACM, New York, NY,* 353-356. DOI=http://dx.doi.org/10.1145/2388676.2388755

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley.

Bios

Anamaria Berea is a postdoctoral researcher in the Center for Complexity in Business, Robert H. Smith School of Business, University of Maryland. She is researching social phenomena using various computational methods.

William Rand is an assistant professor of Marketing and Computer Science at the University of Maryland. William Rand serves as the director at the Center for Complexity in Business. His work examines the use of computational modeling techniques, like agent-based modeling, geographic information systems, social network analysis, and machine learning to help understand and analyze complex systems, such as the diffusion of innovation, organizational learning, and economic markets.

Kevin Wittmer is an undergraduate researcher in the Robert H. Smith School of Business, University of Maryland. He is researching various aspects of qualitative and quantitative methods for data analysis.

Gerard Wall is the Solutions Architect at vibeffect, pioneering the investigation of how the Higher Education decision and its impact on families can become more transparent and relevant for the “consumer” as family.