Paper ID #23124

Capturing Narratives of Graduate Engineering Attrition through Online Forum Mining

Carey Whitehair
Dr. Catherine G.P. Berdanier, Pennsylvania State University, University Park

Catherine G.P. Berdanier is an Assistant Professor in the Department of Mechanical and Nuclear Engineering at Pennsylvania State University. She earned her B.S. in Chemistry from The University of South Dakota, her M.S. in Aeronautical and Astronautical Engineering and Ph.D. in Engineering Education from Purdue University. Her research interests include graduate-level engineering education, including inter- and multidisciplinary graduate education, online engineering cognition and learning, and engineering communication.

Capturing Narratives of Graduate Engineering Attrition through Online Forum Mining

Abstract: This research paper presents methods by which researchers can harvest data from social media forums as a way to gain insight on sensitive issues or populations. In the present research, we are interested in studying doctoral attrition, which is a complex and multifaceted phenomenon that poses practical significance to funding agencies, advisors, and students themselves. Sampling non-completers is difficult, and researchers generally find it difficult to collect nationwide narratives of attrition. This paper presents a novel method for studying attrition using the publicly-available online forum Reddit.com to collect first-hand accounts and authentic narratives of attrition. These often- anonymous online discussions offer a unique view into the authentic thoughts of engineering graduate students considering leaving their program, throughout the decision-making process. This paper proposes a method to efficiently collect and parse open-source information into coherent narratives across "posts" or "threads" of conversation using data mining tools. The underlying methodology developed is based on achieving a holistic view of the discourse patterns and authentic narratives surrounding attrition, which in turn allows researchers to capture meaningful, authentic, and credible emergent themes unbiased by social response. We present a short summary of results to show the dominant narratives of attrition achieved through this method; however, the main focus of this paper is to present the method itself, which has the potential to be extended and modified to aid in other large data mining efforts to answer other research questions related to sensitive topics.

1. Introduction and Literature Review

According to the Council for Graduate Schools,[1] graduate attrition ranges between 24%-68% across disciplines. While engineering disciplines tend to be at the low end of the range, due to reliable funding and a low time to graduation relative to students in the humanities and social sciences, the low end of the range is still a remarkably high number. Attrition is monetarily costly for universities, colleges, departments, and research advisors—who often fully fund these students, and emotionally costly and time costly on the part of research advisors and the students themselves. Although attrition is often of interest to the higher education space and in the disciplines that represent the upper ends of the attrition range,[2] it is often difficult to study the experiences of non-completers or students considering leaving their programs because it is a sensitive topic embedded within the social dynamics of graduate school.[3]

Across disciplines, the academic achievements of students considering attrition does show that academic preparation is typically not one of the main reasons for attrition [4,5]. In other words, most students who leave academia choose to leave because of their own personal decision, not because they failed qualifying exams or are doing poorly in their courses [5–7]. Indeed, Barnes et al.'s [8,9] studies of graduate attrition showed that the attributions that professors give for their students that leave are different than the rationale that the corresponding non-completing students give for leaving. The misalignment, misunderstanding, or attribution bias that may exist (from both parties)

is worthy of study and is likely due to the issues that have arisen with sampling a sensitive population.

Further, most attrition literature takes a sociological view of attrition, noting the structural or "cultural" facets that cause attrition [10–14]. Fewer works approach attrition as a decision-making process from a psychological point of view to understand the complex process as a student decides to leave their program. Most attrition literature sees attrition as an "impulse function" where one moment a student is a student, and the next they have left the university. However, qualitative attrition research understands that attrition is nuanced and complicated, and that the decision to leave might stretch over a long period of time as the student experiences the ups and downs of graduate school, and so it is important moving forward for researchers to explore the oscillations the arise during the decision-making process [4,5,15,16].

In engineering in particular, graduate education is highly understudied, and relies heavily on outside disciplines to study graduate attrition. While it is likely that theories of socialization still hold, and can be interpreted across disciplines, there are contextual differences in disciplinary academic culture that do not align well[9,12,17]. For example, reliable funding is one of the primary causes for attrition in the humanities, although, as Crede and Borrego [18] note, this reasoning does not typically apply to graduate engineering students, who are upwards of 80% fully funded. Advisor relationships do still play a strong role in the attrition process, as does the laboratory culture, since a student's laboratory is like a family and plays the role of peer mentoring and socialization in most graduate departments [19–21]. Other research at the graduate level has hinted at the role that non-technical competencies have in the ability to complete, such as academic engineering writing [22]. However, the psychological decision-making processes by which students decide to leave their programs is still unknown and represents an enormous gap in the scholarship. Furthermore, it is important to employ creative sampling methods in order to study students who are actually considering leaving or who have left their programs, but this has proven to be quite difficult.

The explicit objective of a broader project this paper represents is to capture and analyze the narratives of engineering graduate student attrition in ways that are unbiased by researcher selection, recruitment, and the inherent bias that comes when students answer questions from researchers. This paper is a methods paper representing the first part of this broader study, presenting an approach we have developed to begin to understand the sensitive topic of attrition and the decision-making process. We demonstrate the use of "scraping" social networks—in this case, the online forum Reddit.com—to gather students' real thoughts on the attrition process without interference by researchers or bias that comes with sampling. As social media forums are open on the internet, no IRB or explicit participant recruitment is necessary to gather and analyze this data. The remainder of this paper outlines the development of the method, discussing our decisions by which technological and methodological issues were overcame. We posit that this method, and similar methods that future researchers will develop, will help the engineering education methods and research community think creatively and strategically about bodies of data that might help to answer research questions that fall outside traditional interview and survey methods, as a means of studying sensitive topics and populations.

2. Method Development

We decided on the use of online forums as a source of data for a number of reasons. The exploration of online forums has been used in a wide array of research studies ranging from political engagement to socialization to uses in distance-education [23–28]. While these methods have not been applied to graduate education and attrition, the same benefits gained in other studies may afforded to engineering education. By using an anonymous online forum, we are offered the opportunity to observe first-hand accounts and the decision-making process of engineering graduate students as they perceive the internal and external factors affecting their decision to persists or not. Because of the anonymous nature of these forums without the risk of social or academic repercussions this allows us to capture meaningful and authentic themes unbiased by social response bias.

A. Introduction to Reddit and Programming of a Data Mining "Bot"

This data set was collected from a public, online forum called Reddit. Reddit is an overarching website that houses multiple "subreddits." These subreddits are effectively different pages within Reddit that host forums and discussions that can be posted by individual users, about a specific topic. For example, the subreddit named "LadiesofScience" is intended as a platform where women in science may go to discuss anything related to being a female in a science field. These forms can range from venting, to asking questions or gaining advice from other women. For each subreddit, there is at least one individual (typically the founder of the subreddit) who acts as a moderator. These moderators design the layout for their subreddit's specific page, curate content on the subreddit, and designate appropriate "associated subreddits" that are linked from their subreddit.

Data collection was accomplished through an automated web-scraping bot. The "bot" (short for robot) is effectively a program written to perform a certain task, in this case, to gather forum threads with specific criteria through Reddit. This bot searched within a researcher-chosen set of subreddits, where discussion for particular groups or on particular topics occur, such as "graduate school" or "women in science". The subreddits were defined from the "popular" set of subreddits as determined by Reddit (e.g. AskReddit). We included the "associated subreddits" as defined by each of these as well as subreddits mentioned within posts about graduate school, and through this process, we gathered a list of 15 relevant subreddits. Any subreddit without posts within the past 6 months related to graduate attrition were excluded from the set of subreddits. The code for this bot can be found by contacting the authors of this paper.

From there the bot used a designated set of search terms to find relevant post for discussion by individual users within each chosen subreddit. This search criteria were set to search for terms related to graduate school and attrition; if only one set of search terms was found the post was excluded. Additionally, a set of exclusionary search terms were used to prevent unrelated posts from being found. These search terms and exclusionary terms are presented in Table 1. For example, a post about "Leaving a game of dungeons and dragons because of the game master" would fall within the search if the exclusionary search terms were not used. These exclusionary terms were selected based on an initial trial run of the code, noting particular words or themes (e.g.

Disney and dungeon that shared many keywords in common with the search criteria terms for whatever reason.)

Table 1: Search criteria to direct "bot" data selection

Search Criteria		Exclusionary Search Terms
Set 1	**Set 2**	
Grad, graduate, PhD, doctoral, doctorate, masters, MS, Msc	Leave, leaving, dropping out, drop out, quit, quitting, mastering out, left, done, withdraw, withdrew	Disney, high school, highschool, dungeon

The search terms were designed to cast a wide net with the goal of not excluding searches that used different syntax and phrasing. For each subreddit, up to 500 posts that met the search criteria were collected, including username, date and time of submission, post title, and post submission text. Examples of the way one thread appears on the Reddit website and then by the Reddit scraper bot are shown in Figures 1 and 2, respectively.

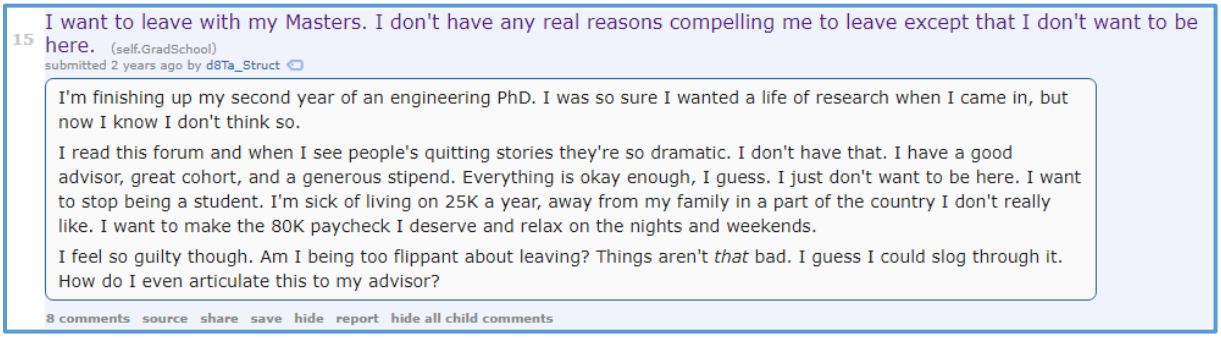

Figure 1: A sample Reddit post as viewed from the Reddit webpage.

AUTHOR: d8Ta_Struct

TITLE: I want to leave with my Masters. I don't have any real reasons compelling me to leave except that I don't want to be here.

SELFTEXT: I'm finishing up my second year of an engineering PhD. I was so sure I wanted a life of research when I came in, but now I know I don't think so.

I read this forum and when I see people's quitting stories they're so dramatic. I don't have that. I have a good advisor, great cohort, and a generous stipend. Everything is okay enough, I guess. I just don't want to be here. I want to stop being a student. I'm sick of living on 25K a year, away from my family in a part of the country I don't really like. I want to make the 80K paycheck I deserve and relax on the nights and weekends.

I feel so guilty though. Am I being too flippant about leaving? Things aren't *that* bad. I guess I could slog through it. How do I even articulate this to my advisor?

[POSTED]: Wed Apr 15 23:10:57 2015

Figure 2: A sample Reddit post as stored by the web-scraper bot.

The scraped threads related to graduate attrition were gathered from the bot, which outputs as a text file that can then be analyzed through qualitative analysis techniques, such as content analysis. For our purposes, this initial corpus was then sorted by hand to remove any post not related to graduate student attrition, in case any keywords returned unrelated items.

We then were able to analyze the resulting corpus through traditional qualitative and textual analysis techniques. First, we sorted the corpus into posts related to engineering, STM (science, technology and mathematics) and non-STM. Of these, there were 28 discussions explicitly related to engineering graduate school attrition. Because this study was primarily qualitative and exploratory in nature, these numbers were found to be appropriate [22,29–31]. We were prepared to perform another search with the web-scraping bot; however, after analyzing this corpus, saturation of the themes appeared to be reached and therefore it was felt that another search was not necessary [29,32]. A significant number of posts mentioned that the author was in a STEM field, however these were not included in the engineering categorization because they could not explicitly be identified as engineering. Because the method of collection was passive (collected from public online forums rather than seeking out individuals), no efforts were made for quota sampling by gender, engineering field, or ethnicity. This one-time collection of data resulted in post-dates ranging from 2010 to 2017. There was also a limited number of posts by the same users. If these posts were simply the same post within separate subreddits, one of them was excluded. Otherwise they were grouped with the previous posts by that user in chronological order.

This method of data collection is also easily repeatable and may be extended to other forums. Because the forum is publicly available, it doesn't require IRB approval. The search can also be extended to look at non-STEM programs, or even for different search criteria unrelated to graduate student attrition.

B. Overcoming Limitations

As with all methods, there are some limitations to the unique approach used in this work. The anonymous nature of these forums is a double-edged sword in that it may allow users to be more open and honest without fear of repercussions, but it also means that users are not guaranteed to be truthful and must be taken at their word for validity of all statements. In addition, this anonymity make results in negativity bias or in users simply using these forums to vent their frustrations. However, unlike other social forums such as Twitter, there is not a limited character space, so users have the opportunity to fully express themselves and their situation to whatever degree of detail they desire. Another limitation is that, as with any written exchange, certain emotions and sentiments may be difficult to convey without tone and social cues like facial expressions. Fortunately, some of this has been mitigated by certain online customs; for example, within Reddit users often add a "/s" at the end of a sentence or post to indicate sarcasm or they use characters to create emojis such as "=)" to indicate a friendly smile. These modes of discourse could be interesting to study in future work.

There may be self-selection bias present in this group of participants. Individuals who post on these forums are likely more willing to seek out help and discussion when facing a decision like graduate school attrition. However, some of the data points did indicate that the graduate students

really did not feel they had anyone else to talk to about these sensitive issues, and the only place they felt they could turn to was the internet community. These two factors indicate that researchers may still be missing important themes of attrition from those who leave their departments silently. The goal of this method is to capture at least some narratives of individuals who may not feel comfortable talking to their departments but are comfortable speaking to anonymous strangers, as is mentioned in several of the Reddit posts.

3. Results and Discussion

A. Brief Summary of Attrition Findings

The purpose of the present paper is to introduce the method of online forum mining as a way to collect anonymous and sensitive data from populations who may be feeling marginalized or uncomfortable discussing their situations. Therefore, the thorough results and discussion of findings will be left for future publications. Ultimately, the corpus resulted in a total of 28 Reddit posts from engineering graduate students considering leaving their programs that capture several themes that have emerged in other graduate attrition literature across all disciplines. For example, the role of the advisor, and financial support were common themes; however, these manifested differently in the engineering students than graduate students in other disciplines. Rather than complaints about being unsupported financially, graduate students considering leaving their programs understood that they could hold financially lucrative jobs even with a bachelor's degree and were more concerned about what employers would think of a gap in their resume. This financial cost was compared with other non-financial "costs" of staying in graduate school to their well-being.

We also noticed differences in dominant narratives based on student self-reported confidence levels. For example, some students used language to indicate high or low self-efficacy in their ability to succeed in graduate school, which weakly aligned with some of the facets of attrition. Of course, this study is a low-N qualitative study, and therefore, these correlations are anecdotal at best, but lay the groundwork for future attrition studies and research questions. These results will be best analyzed through attribution theory as well as other psycho-social theories of graduate attrition and persistence. These results will be presented in future publications and warrant an in-depth journal article to thoroughly interpret and discuss the findings.

B. Opportunities for Social Media Mining in Engineering Education Research

One of the main attributes of the present research is the opportunity to employ creative methods to study sensitive populations in engineering education. While survey and interview techniques still dominate the engineering education research methods landscape, it is important to consider alternative ways to recruit populations, especially those who are marginalized or may be feeling insecure. Although this study did not seek to reach out to these students, we anticipate that using social media forums to recruit participants for attrition-related research (for example) might be a lucrative approach. Similarly, we expect that other sensitive topics in education might be approached similarly through social media, whether that be Reddit or some other social media venue. A handful of engineering education researchers have dabbled in this area: For example,

Chen et al[33] analyzed tweets from one university in order to gain a deeper, unveiled understanding of the student learning experience than a more structured formal research campaign. However, these methods are not widespread in engineering education.

We recommend that the selection of an appropriate social media venue, the careful selection of inclusion and exclusion criteria, and then the appropriate selection of qualitative analysis techniques be central issues to researchers considering social media mining as a method of corpus collection. The limitations of such methods are important, such as the inability to gain any more context than simply the information given by the participant; however, for sensitive topics, these limitations might be overcome by the benefits of an increased understanding. As social media platforms become venues by which students can ask for- and gain advice, we as engineering educators and researchers should be willing to look to these venues to gain insight into new research questions and the issues facing students.

4. Conclusion

This paper proposes a method by which Reddit.com forums can be mined via a virtual bot in order to collect relevant forum threads related to graduate attrition. We prove the successful use of this method in an analysis of graduate engineering student attrition, by which an appropriate corpus manifested from the bot that could be analyzed by more traditional qualitative means. As graduate attrition is a sensitive topic, and therefore it is difficult to recruit non-completers or those considering leaving their programs to participate in research studies, analysis of data from online forums can provide insight on these sensitive topics in order to more effectively hone research questions before larger, more intensive studies are undertaken. We present recommendations for other researchers interested in employing social media forums and recommend that social media become a venue by which researchers and educators can gain insights on issues facing students.

References

1. Council of Graduate Schools. (20008). *Ph . D . Completion and Attrition: Analysis of Baseline Data*.
2. Gardner, S. K. (2009). Student and faculty attributions of attrition in high and low-completing doctoral programs in the United States. *High. Educ.* **58,** 97–112.
3. Lovitts, B. E. (2001). *Leaving the Ivory Tower*. Rowman & Littlefield Publishers, Inc.
4. Lovitts, B. E. & Nelson, C. (2000) The hidden crisis in graduate education: Attrition from Ph.D. programs. *Academe* **86,** 44–50.
5. Lovitts, B. E. (2001). Leaving the Ivory Tower: The Causes and Consequences of Departure from Doctoral Study. *Contemporary Sociology* **32**.
6. Ruud, C. M., Saclarides, E. S., George-Jackson, C. E. & Lubienski, S. T. (2016). Tipping Points: Doctoral Students and Consideration of Departure. *J. Coll. Student Retent. Res. Theory Pract.* **0,** 1–22.
7. Curtin, N., Stewart, A. J. & Ostrove, J. M. (2013). Fostering Academic Self-Concept: Advisor Support and Sense of Belonging Among International and Domestic Graduate Students. *Am. Educ. Res. J.* **50,** 108–137.
8. Barnes, B. J. (2010). The nature of exemplary doctoral advisors' expectations and the ways they may influence doctoral persistence. **11,** 323–343.
9. Barnes, B. J. & Randall, J. (2012). Doctoral Student Satisfaction : An Examination of Disciplinary , Enrollment , and Institutional Differences. 47–75.
10. Haydarov, R., Moxley, V. & Anderson, D. (2013). Counting chickens before they are hatched: An examination of student retention, graduation, attrition, and dropout measurement validity in an online master's environment.. *J. Coll. Student Retent.* **14,** 429–449.
11. Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Rev. Educ. Res.* **45,** 89–125.
12. Herzig, A. H. (2004). Becoming Mathematicians: Women and Students of Color Choosing and Leaving Doctoral Mathematics. *Rev. Educ. Res.* **74,** 171–214.
13. Weidman, J. C. & Stein, E. L. (2003). Socialization of Doctoral Students to Academic Norms. *Res. High. Educ.* **44,** 641–656.
14. Sweitzer, V. (Baker) (2009). Towards a Theory of Doctoral Student Professional Identity Development: A Developmental Networks Approach. *J. Higher Educ.* **80,** 1–33.
15. Lovitts, B. E. (1996). Who is Responsible for Graduate Student Attrition--The Individual or the Institution? Toward an Explanation of the High and Persistent Rate of Attrition. in *Annual Meeting of the American Education Research Association*.
16. Breckner, J. A. (2012). A Phenomenological Study of Doctoral Student Attrition in Counselor Education. University of Tennessee.

17. Gardner, S. K. (2010). Contrasting the socialization experiences of doctoral students in high- and low-completing departments: A qualitative analysis of disciplinary contexts at one institution. *J. Higher Educ.* **81,** 61–81.

18. Crede, E. & Borrego, M. (2014). Understanding retention in US graduate programs by student nationality. **39,** 1599–1616.

19. Crede, E. & Borrego, M. (2012). Learning in graduate engineering research groups of various sizes. *J. Eng. Educ.* **101,** 565–589.

20. Gardner, S. K. & Barnes, B. J. (2007). Graduate student involvement: Socialization for the professional role. *J. Coll. Stud. Dev.* **48,** 369–387.

21. Gardner, S. K. (2008). Fitting the mold of graduate school: A qualitative study of socialization in doctoral education. *Innov. High. Educ.* **33,** 125–138.

22. Berdanier, C. G. P. (2016). Learning the language of academic engineering. Dissertation: Purdue University.

23. Dunne, K. (2010). Can online forums address political disengagement for local government? *J. Inf. Technol. Polit.* **7,** 300–317.

24. Beuchot, A. & Bullen, M. (2005). Interaction and interpersonality in online discussion forums. *Distance Education* **26,** 67–87.

25. Li, N. & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis. Support Syst.* **48,** 354–368.

26. Fayard, A. L. & DeSanctis, G. (2010). Enacting language games: The development of a sense of 'we-ness' in online forums. *Inf. Syst. J.* **20,** 383–416.

27. Buckley, F. (2011). Online discussion forums. *European Political Science* **10,** 402–415.

28. Yassine, M. & Hajj, H. (2010). A framework for emotion mining from text in online social networks. in *Proceedings - IEEE International Conference on Data Mining, ICDM* 1136–1142.

29. Creswell, J. W. & Plano-Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research*. SAGE Publications, Inc.

30. Spaulding, L. S. & Rockinson-Szapkiw, A. J. (2012). Hearing their Voices: Factors Doctoral Candidates Attribute to their Persistence. *Int. J. Dr. Stud.* **7**.

31. Gopaul, B. (2014). Inequality and doctoral education : exploring the '"rules"' of doctoral study through Bourdieu's notion of field. *High. Educ.* **70,** 73–88.

32. Marshall, C. & Rossman, G. (2006). Designing qualitative research Sage. *Thousand Oaks, CA*.

33. Chen, X., Vorvoreanu, M. & Madhavan, K. P. C. (2014). Mining social media data for understanding students' learning experiences. *IEEE Trans. Learn. Technol.* **7,** 246–259.