Predicting Students' College Drop Out and Departure Decisions by Analyzing their Campus-Based Social Network Text Messages

Rebecca Allen
University of Cincinnati
allenr0@mail.uc.edu

Alex Nakonechnyi
Mount St. Joseph University
alex.nakonechnyi@msj.edu

Abraham Seidmann
University of Rochester
avi.seidmann@simon.rochester.edu

Jacqueline Roberts
Mount St. Joseph University
jacqueline.roberts@msj.edu

Abstract

Undergraduate student retention is a key concern in the US higher education system. Having a scientific method for predicting undergraduate student departure would enable institutions to deploy targeted interventions with the goal of retaining a particular student who is at risk of dropping out. We explore the use of Latent Dirichlet Allocation (LDA), Systemic Functional Linguistics (SFL), and new techniques for Social Network Analytics addressing student communications within a novel campus-based closed social networking platform. Our research results indicate that students who were ultimately retained sent three times as many messages than those who were not, and analyzing the patterns of use of the closed social network in an academic setting reliably predicts undergraduate student dropouts and leads to a more effective deployment of retention resources over time.

1. Introduction

Higher education in the United States is a costly endeavor. Students are borrowing large sums of money to pay for their education. In recent years, student debt in the US has exceeded credit card debt [1]. While a college degree generally remains a sound investment, notwithstanding the cost, college dropout bears significant negative financial consequences for the student, their families as well as the institution [2]. For institutions of higher learning, student retention is both a moral as well as financial imperative. Student retention is key to continued viability of universities both in terms of tuition revenue as well as preserving the brand value. Unfortunately, nearly 50% of students who begin college will not graduate – this ratio has remained steady over the past four decades [4]. This dropout rate represents not only a social problem but a business problem as well. Kuh et al. [4] states that while scholars have long recognized the need to improve student retention rates in higher education there is an urgent need for institutions of higher learning to take concrete steps to improve the situation.

The branding that sells students on their initial college choice may not be what keeps them enrolled. Instead, within academic literature, the assertion has been that the more engaged students are, that is the degree to which they use and interact with the people, places, and events the university has to offer, the less likely they are to dropout [5]. One of the most common metrics of student engagement and retention is the National Survey of Student Engagement (NSSE). NSSE measures student engagement through a survey administered to graduating seniors [6].

NSSE findings have been widely used to formulate recommendations such as practices seen to benefit student retention [4]. Yet, NSSE, while a useful tool, has the limitations inherent to surveys and structured data. First there is response bias: it has been shown that students do tend to present a rosier picture of their own engagement than the actual reality [7]. Also, because NSSE is administered only during students' senior year, the ongoing development of relationships or factors that help students engage and the experiences of those who drop out are not captured. Hence, administrators and stakeholders may benefit from additional sources of data regarding student engagement so as to improve student retention.

In view of this need for supplemental data sources and analysis, we propose that using big data to understand student retention may provide additional useful insights. Thus, in this project, we analyze approximately 7,000 chat messages written using a university's closed social networking platform. The messages were written by students, some of whom were retained after the first year and some of whom dropped out. Herein, it is postulated that by using Latent Dirichlet Allocation (LDA) for sentiment analysis and Systemic Functional Linguistics with corpus linguistics (computer assisted linguistic analysis) tools, fundamental differences and trends can be seen in the development of language of students who disengage and drop out as compared to those who engage and stay. The specific questions asked by this research are as follows:

Q1: Do retained students communicate more frequently using the university's closed social

URI: https://hdl.handle.net/10125/64387
978-0-9981331-3-3

networking platform's messaging functionality?

Q2: Using LDA, do non-retained students less often express themes about university-related topics?

Q3: Using LDA with sentiment analysis, do retained students express a higher degree of positivity in their message than non-retained students?

Q4: Using SFL, how does retained and non-retained students' use of modality change over time?

Q5: Using SFL, how does retained and non-retained students' use of engagement language change over time?

All of these questions can be answered quantitatively based on the application of Latent Dirichlet Allocation (LDA) to sentiment analysis and Systemic Functional Linguistic (SFL) approaches.

2. Conceptual Framework

Again, it is important to understand that engagement has been widely seen as the antidote to dropout [5] . But just what is engagement? Even within the narrow area of student retention, engagement has been defined loosely or differently across various works [8]. Both fundamental and conceptual understandings of engagement are needed. The following sections discuss engagement from the interdisciplinary frameworks used within the paper, illustrating that the common thread running through each framework is that engagement is a function of involvement with a surrounding community.

2.1. National Survey of Student Engagement

In the area of student retention, engagement is often understood quantitatively by measuring the frequency with which learners engage in various activities, as is the case with the NSSE [9]. Yet measuring what learners have done does not necessarily explain what precipitated their actions. Perhaps also being informed by socio-cultural and linguistics lenses would provide more insight into the narratives of student (dis)engagement.

2.2. Socio-Cultural Understandings of Engagement

With regards to socio-cultural aspects of student engagement, Rogoff [10] brings several significant points to bear. Rogoff suggests that learners learn by forming communities of participation [10], [11]. The knowledge to be, participate, and belong in such a community, is not necessarily transmitted through participation only; it is also transmitted more passively, through observation [12]. This project frames student engagement as successful sustained membership in the student community; students, impacted by their peers, have to learn to participate and stick with their academic community. Failure to join the community may be felt as student drop out.

2.3. Natural Language Processing, Linguistics, and Engagement

What shapes the learning communities and engagement in them in the first place? Language does. After all, humans create, declare, and shape their participation or lack thereof in their communities through language [13].

Latent Dirichlet Allocation with Sentiment Analysis: Unstructured data, such as the free-form chat messages found on our closed social networking platform, can present a challenge to analyze. Often, in humanities, such data is understood through thematic analysis [14], where data is manually categorized into themes. This sort of analysis, however, becomes untenable in a very large data set. Latent Dirichlet Allocation (LDA) offers a framework for uncovering the structure of data and categorizing the data based on its attributes, what would often be called themes in traditional qualitative research, using computational and statistical methods [15]. The particular application of LDA at hand is using LDA for sentiment analysis; namely using LDA as a tool to measure users' attitudes and totality through their online textual data [16]

Systemic Functional Linguistics: LDA proposes placing textual data into categories according to meaning, while Systemic Functional Linguistics (SFL) offers a distinct, but complimentary approach. SFL is not only interested in what a given data set means; SFL holds that how that meaning is created – the specific grammatical and word choices – represents conscious construction, and purpose on the part of the language creator [17], [18]. While SFL is not solely a big data tool, SFL implemented through computerized approaches to language analysis (corpus linguistics) have been cited as useful [19],

Mood and modality: In SFL, interpersonal meaning, that is, the relationship between speakers/writers, is realized through analysis of mood and modality which are collectively called interpersonal meaning [18]. Mood can be understood as the occurrence of types of clauses: declarative - clauses that tell about something; interrogative – clauses that ask about something; or imperative – clauses that mandate something [20]. Modality, the

expression of probability, is a complex area of English grammar [18], [21]. Although the set of modal verbs in English is "finite and familiar" [19, p. 69], and is comprised of *can, could, may, might, should, shall and would*, modality is expressed in ways additional to the use of just these modal verbs, such as *probably and maybe*.

Appraisal Theory: Another SFL area, appraisal theory in particular, has been understood as a way to understand engagement [22]. Martin and White's [22] appraisal theory has three components: attitude, graduation and engagement. Attitude consists of expressed feelings, judgement and/or appreciation (evaluation). In appraisal theory, the term graduation refers to the degree to which language amplifies or softens the meanings. Engagement is the extent to which other voices are incorporated into the discussion, and how social positions are created within the dialogue. If participants are to successfully build a community, these verbal resources – attitude, engagement, and graduation – are the linguistics tools afforded. Measuring the implementation of these aspects may give insight into community engagement and involvement.

Interestingly, in reflecting upon the theory which he largely crafted, White points out that are many registers and discourse domains to which the theory has yet to be applied. We may reasonably anticipate more breakthroughs in the mapping of this semantic domain as the number of researchers using the theory continues to increase. [23]. Thus, we have adapted this theory to the specific paradigm of student engagement.

3. Literature Review

To our knowledge, there has been no specific research to date that employs LDA or SFL analysis on student engagement social text message data. Student engagement apps are a relatively new phenomenon; though larger scale commercial apps do exist, most student engagement apps are localized initiatives [23]. However, there is abundant literature on LDA and SFL as it applies to user or student engagement in online and social media platforms. Herein, we review some key studies from which we model the methodology.

Our first, second, and third research questions measure simple frequencies and use LDA to facilitate prediction of student retention using on campus student platform text message data. Though no studies have been found using student campus app text messages, given findings from similarly generated student texts, the data and its analysis may be especially promising. Looking for predictive retention patterns, studies have analyzed admissions essays [24], Massive Open Online Course (MOOCs) [25], [26] open-ended freshman experience surveys [27], and online course postings [28].

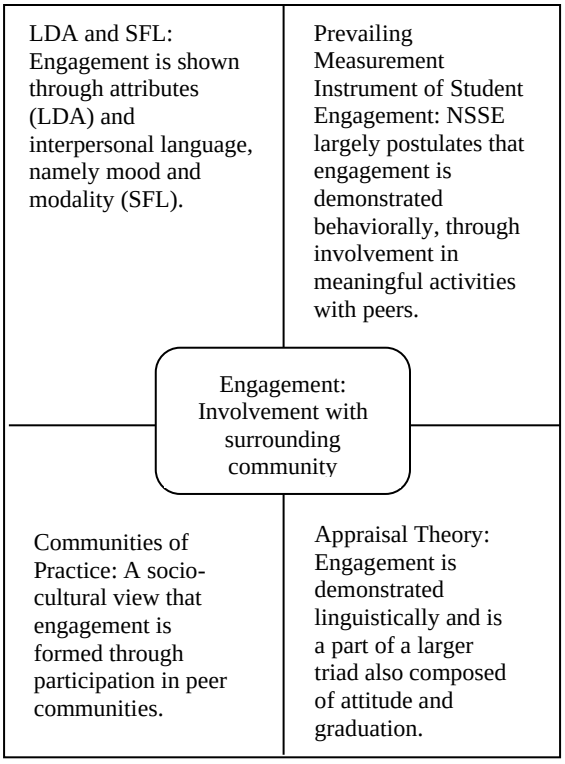

Figure 1. Engagement is defined differently in distinct disciplines and theories, but the commonality amongst the definitions is that engagement pertains to the involvement and investment with others.

LDA was identified as a useful predictive tool in all but one of the cases: the admissions essays [22]. For LDA to be a meaningful tool, it would seem that there needs to be some freedom or open-endedness and authenticity (arguably an admission essay is not open ended and is subject to third party involvement) in the data; ability to measure change over time in messages has also been cited as useful [28]. The platform data has both of these characteristics, so such data analysis may offer a strong contribution.

The fourth and fifth questions draw on SFL approaches to linguistically analyze mood, modality and engagement with the view of linking these linguistic characteristics to retention. While again, no SFL studies specifically on student retention on social network data have been done, past studies on online community building and engagement may offer methodological insight. McDonald and Woodward-

Kron [29] examine how users' interpersonal meaning, namely, mood and modality, evolved as they continued to engage in an online mental health support message board community. To carry out their study, McDonald and Woodward-Kron built a series of sub-corpora, with corpora one capturing a user's first posting, and corpora two capturing a user's second posting, and so on; this approach allowed them to note changes in language over time. The authors noted shifts in the interpersonal meaning employed by the message board users. Message board users began employing progressively more imperative statements with longer membership in the online community. Regarding modality, McDonald and Woodward-Kron reference the same difficulty that I previously outlined: modality is difficult to measure in the context of corpus linguistics because modal-like expressions are legion. Nevertheless, over the course of membership, they found that there is a marked increase in the use of will/would modals for the purpose of advice giving, in expressions such as "I would seriously consider".

The final fifth question also queries whether engagement measurably grew in the student text messages; Goertzen and Kristjánsson [30] use the SFL related area, appraisal theory, to measure engagement in student forum messages in an online course. We adopt their studied linguistic engagement aspects for our study: the aspects that they track that would be readily evaluable in a corpus driven SFL approach are primarily those that signal involvement. Such aspects include the use of: naming (the use of others' names), inclusive language (such as we or everyone), paralanguage (use of abbreviations and unconventional spellings such as TTYL), and use of emoticons (or emojis).

4. Setting, Participants, and Data Sources

The data set comes from a new student engagement and retention closed social media networking platform (comprised of a mobile platform, integrated digital signage, and administrative and reporting interfaces) which was developed at a private Midwestern University. This student engagement platform has been conceived and designed by a large team over a period of 18 months using a participatory action research approach; student opinions and needs were sought throughout the project development [31]. In response to student demand, the platform connects students with on campus activities, key university services, and the university help desks. The platform also allows students to see the names, pictures, and social media handles of their classmates (only of those students who chose to share this information). Students are also able to initiate chat messages directly with their peers through the platform itself. These 'student-to-student' text messages compose our research dataset. Over 6,000 internal chat messages (the database grows continually) have been collected into a database that links the messages with the first-year retention outcomes of the senders (i.e. if they dropped out or stayed). Though student usage of the platform was voluntary, it was widely adopted. We saw that 99.69% of freshman students installed it; 72.81% of freshmen sent at least one message. Hence, the platform saw extensive usage during 2018-2019 academic year. Key usage summary is provided in Table 1.

Table 1. Key platform statistics for 2018-19 academic year

Undergraduate enrollment	1,205
Students using the app	1,090
Sessions	48,963
Messages Exchanged	15,913
Freshman Messages	6,515
Connections Formed	2,096
Social Connections Involving Freshmen	1,221

5. Methodology

The first and second questions deal with what students talk about, how often they talk about it, and if these findings are linked with retention. LDA was used to look at attributes of students' chat messages. Using statistical methodologies, we were able to explore whether the topics of retained students' conversations vary in comparison with non-retained students.

The third, fourth, and fifth questions look at SFL mood, modality, and engagement in relation to student retention outcomes, and implement time series analyses. Following McDonald and Woodward Kron's [29] approach, sub-corpora of the student chat messages that occurred each week or month has been developed to measure for the passage of time. The items that were identified and measured, include imperative statements, modal verbs and modal-like expressions, naming, inclusive language, paralanguage, and emoticons. Part of speech tagging can be used to identify imperatives [32]; and all of the other proposed categories constitute closed sets, and thus identifying related lexical items is trivial. Useful taxonomies for carrying out this comparison include Biber, Johanson, Leech, Conrad, and Finegan's [33] list of modal and modal-like expressions, lists of university student names and widely available list of emoticons [34].

6. Findings

Out of 15,913 messages on the system exchanged during 2018-2019 academic year, the research team isolated 6,515 messages written by freshman for whom retention status is known. On average, students who were ultimately retained sent three times as many messages than those who were not during each of the six months in our sample *(t = 2.8316, df = 6, p-value = 0.02989)*. Using LDA topic model [35], no significant differences were found between the retained and non-retained groups in communication related to school and university related topics. However, the overall positivity sentiment of those messages was higher amongst the retained group as compared to the non-retained group *(t = 2.3829, df = 734.14, p-value = 0.01743)*. The use of modality changed over time, yet no significant differences were found in the use of modality between the retained and non-retained groups.

Based on these results, we decided to explore the structure of the freshmen social network. The graphical representation of this network is provided in Figure 2. The size of circles in Figure 2 is proportional to the 'Between Centrality' measure of that student; it measures the students' capacity to broker social contacts among other students – to extract "service charges", to isolate other students or to prevent contacts between them. Hence, the more other students depend on that particular student to make connections with others, the more "social brokerage" power that in-between student will have. The results indicate that the ***Between Centrality*** of the retained students is nearly double the between centrality of the non-retained students (*t = 2.466, df = 245.03,* ***p-value = 0.01435***).

SFL findings regarding modality yielded promising trends. Although, the use of a speech tagger showed no significant differences in either group's use of imperatives, there was evidence arising from LDA topic-term analysis [35] to suggest that the retained group expressed more certainty than the non-retained group. As previously was mentioned, ways other than direct use of imperatives, so-called modal-like expressions, to give advice and direction--are of significance. Interestingly, within the data we can see that the retained students use significantly more punctuation than the non-retained group. Both groups use questions marks—that is, ask questions—with the same relative frequency. The difference in the retained group's increased relative frequency of punctuation is their more frequent use of commas and periods. In short, as a relative frequency, more of the non-retained students' utterances are non-interrogative; retained persons express more certainty.

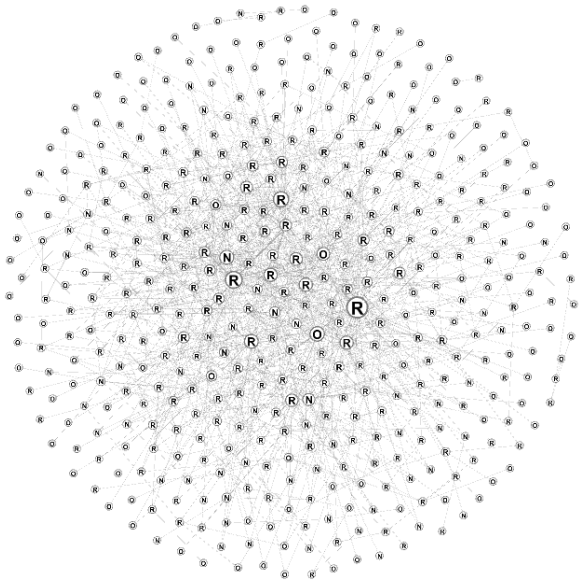

Figure 2. Social network connections of the freshmen on the platform. (R denotes retained freshmen students, N denotes non-retained freshmen, O denotes upperclassmen)

Retained students also used more language of engagement [23]. Although retained students did not use significantly more names, we could see a very significant uptick in the retained students' use of emoticons. Retained students used 48 smile face emoticons, while non-retained students used none. This significant difference also supports our finding that retained students express more positivity in their messages.

7. Discussion

Our results strongly suggest that while the content of the communication authored by the retained and non-retained students on the institutional closed social network was very similar, differences exist in several key areas, such as the communications volume, positivity of the message sentiment as well as the use of positive emoticons. However, even more significant was the relationship between the numbers of messages authored and the retention status. And, the retention status is positively correlated with the number of social connections formed on the platform. This finding is aligned with the existing student engagement and retention theory which postulates that the students' social integration is one of the key drivers of student retention [36].

8. Conclusions, Limitations and Future Directions

The present study suggests that the use of closed social network technology on campus presents institutions with an opportunity to identify individual students who are at risk of dropping out before they leave. Network analysis shows that the 'between centrality' of the retained students is nearly double the between centrality of the non-retained students, and retained students express more positive sentiment. These findings indicate that students at risk of dropping out could be identified through measurement of the frequency of the use of their closed social network platform as well as by the analyses of the social networks formed. SFL derived constructs may also be useful— since the use of declarative modalities and emoticons seems to suggest that retained students express more certainty and positivity.

A larger scale study at multiple campuses is needed to validate and enhance our general model for identifying students who appear to be disengaged and thus are at risk of dropout. In addition, our study only examined the messages of freshmen students. Future research with larger sample sizes using students from all undergraduate years is needed to move towards developing a more robust mathematical prediction model; the current findings offer useful insights for our next larger scale research effort.

9. References

[1] W. J. Bennett and D. Wilezol, *Is college worth it?: A former United States Secretary of Education and a liberal arts graduate expose the broken promise of higher education*. Nashville, TN: Thomas Nelson, 2013.

[2] M. Schneider and L. Yin, "The high cost of low graduation rates: How much does dropping out of college really cost?," *American Institutes for Research*, Aug. 2011.

[3] A. L. Stephenson, A. Heckert, and D. B. Yerger, "College choice and the university brand: Exploring the onsumer decision framework," *Higher Education*, vol. 71, no. 4, pp. 489–503, 2016.

[4] G. D. Kuh, J. Kinzie, J. H. Schuh, and E. J. Whitt, *Student Success in College: Creating Conditions That Matter*. John Wiley & Sons, 2010.

[5] Kuh, "The national survey of student engagement: Conceptual and empirical foundations," *New Directions for Institutional Research*, vol. 2009, no. 141, pp. 5–20, Dec. 2009.

[6] "About NSEE," *National Survey of Student Engagement*, 2019. [Online]. Available: http://nsse.indiana.edu/html/about.cfm.

[7] K. A. Fuller, N. S. Karunaratne, S. Naidu, B. Exintaris, J. L. Short, M. D. Wolcott, S. Singleton, and P. J. White, "Development of a self-report instrument for measuring in-class student engagement reveals that pretending to engage is a significant unrecognized problem," *PLoS ONE*, vol. 13, no. 10, pp. 1–22, 2018.

[8] N. Zepke and L. Leach, "Improving student engagement: Ten proposals for action," *Active Learning in Higher Education*, vol. 11, no. 3, pp. 167–177, Nov. 2010.

[9] Center for Postsecondary Research, "Survey Findings on the Quality of Undergraduate Education Engagement," 2017.

[10] B. Rogoff, *Apprenticeship in thinking: Cognitive development in social context*. New York City, NY: Oxford University Press, 1990.

[11] B. Rogoff, "Developing understanding of the idea of communities of learners," *Mind, Culture, and Activity*, vol. 1, no. 4, pp. 209–229, Sep. 1994.

[12] B. Rogoff, R. Paradise, R. Mejía Arauz, M. Correa-Chávez, and C. Angelillo, "Firsthand learning through intent participation," *Annual Review of Psychology*, vol. 22, no. 1, pp. 11–31, 2014.

[13] J. Gee, *An Introduction to Discourse Analysis: Theory and Method*, 2nd ed. New York, NY: Routledge, 2005.

[14] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qulalitative Research in Pscyhology*, vol. 3, pp. 77–101, 2006.

[15] N. Bendle and X. Wang, "Uncovering the message for the mess of big data," *Business Horizons*, vol. 59, pp. 115–124, 2016.

[16] I. R. Putri and R. Kusumaningrum, "Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia," in *Journal of Physics: Conference Series*, 2017, vol. 801, no. 1, p. 12073.

[17] F. Christie, "Ongoing dialogue: Functional linguistics and Bernsteinian sociological perspectives in education," in *Language, Knowledge, and Pedagogy: Functional Linguistic and Sociological Perspectives*, F. Christie and J. R. Martin, Eds. London: Continuum, 2007.

[18] S. Eggins, *An Introduction to Systemic Functional Linguistics*. London, UK: Bloomsbury, 2013.

[19] S. Hunston, "Systemic functional linguistics, corpus linguistics, and the ideology of science," *Text and Talk*, vol. 33, no. 4–5, pp. 617–640, 2013.

[20] S. Eggins, "An overview of systemic functional linguistics," *An Introduction to Systemic Functional Linguistics*, p. 20, 2004.

[21] S. Hunston, *Corpus Appraoches to Evaluation*. New York, NY: Routledge, 2011.

[22] J. R. Martin and P. R. R. White, *Language of Engagement*. Basingstoke, Hampshire, UK: Palbrave Macmillian, 2005.

[23] P. White, "Appraisal Theory: The language of attitude, arguability and intepersonal positioning," 2015. [Online]. Available: https://www.grammatics.com/appraisal/index.html.

[24] M. Ogihara and G. Ren, "Student retention pattern prediction employing linguistic features extracted from admission application essays," in *16th IEEE*

International Conference on Machine Learning and Applications, 2017, pp. 532–539.

[25] S. Crossley, M. Dascalu, D. S. McNamara, R. Baker, and S. Trausan-Matu, "Predicing success in massive open online courses (MOOCs) using cohession network analysis," in *Repository of the Interntional Society of Learning Sciences*, 2017.

[26] G. Nanda, N. M. Hicks, D. R. Waller, D. Goldwasser, and K. A. Douglas, "Understanding Learners' Opinion about Participation Certificates in Online Courses using Topic Modeling," in *Eleventh International Conference on Educational Data Mining*, 2018, pp. 376–382.

[27] L. Aulck, J. Malters, C. Lee, G. Mancinelli, M. Sun, and J. West, "Helping Students FIG-ure It Out: A mixed-methods look at freshmen seminars via first-year interest groups (FIGs)," 2018.

[28] N. C. Ming and V. L. Ming, "Predicting student outcomes from unstructured data," *CEUR Workshop Proceedings*, vol. 872, pp. 11–16, 2012.

[29] D. McDonald and R. Woodward-Kron, "Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics," *Discourse and Communication*, vol. 10, no. 2, pp. 157–175, 2016.

[30] P. Goertzen and C. Kristjánsson, "Interpersonal dimensions of community in graduate online learning: Exploring social presence through the lens of Systemic Functional Linguistics," *Internet and Higher Education*, vol. 10, no. 3, pp. 212–230, 2007.

[31] S. Kemmis and R. Mctaggart, "Participatory action research: Communicative action and public sphere," in *Strategies of qualitative inquiry*, N. Deznin and Y. Lincoln, Eds. Thousand Oak, CA: SAGE Publications, Inc., 2013, pp. 271–329.

[32] M. El-Falaky, "Vote for me! A corpus lingustiics anaylsis of American presidential debates using functional grammar," *Arts and Social Sciences Journal*, vol. 6, no. 123, pp. 1–13, 2015.

[33] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English*. London, UK: Longman, 1999.

[34] "Full Emoji List, v12.0," *Unicode*, 2019. [Online]. Available: https://unicode.org/emoji/charts/full-emoji-list.html. [Accessed: 23-Apr-2019].

[35] K. Welbers, W. Van Atteveldt, and K. Benoit, "Text analysis in R," *Communication Methods and Measures*, vol. 11, no. 4, pp. 245–265, 2017.

[36] P. T. Terenzini and E. T. Pascarella, "Voluntary freshman attrition and patterns of social and academic integration in a university: A test of a conceptual model," *Research in Higher Education*, vol. 6, no. 1, pp. 25–43, 1977.