Predicting Student Dropout by Mining Advisor Notes

J.D Jayaraman
New Jersey City University and Teachers College, Columbia University
jjayaraman@njcu.edu

ABSTRACT

More Americans are attending college than ever before but almost half of them do not complete college. Thus, early detection of students at risk of dropping out of college is of paramount importance. This study describes a novel attempt at using notes made by student advisors to predict student dropout. We use a Natural Language Processing (NLP) technique called sentiment analysis to analyze unstructured textual data to extract the positive or negative sentiment contained in the advisor's notes. We then use the sentiment extracted from the notes as features to train a random forest model to predict student dropout. We achieve 73% accuracy in predicting student dropout. Thus, our study demonstrates the value of unstructured data held in institutional databases for identifying at-risk students.

Keywords

Dropout prediction, Sentiment analysis, Machine learning models, At-risk students, Natural language processing, Text mining

1. INTRODUCTION

Student retention is a major challenge at American universities with the average 6 year graduation rate hovering around 59% [12]. Graduation rates vary with institutional selectivity [19]; the situation being particularly grave at institutions with open admission policies where the 6 year average graduation rate is a meager 32% [12]. Low retention rates not only impact the financial well-being of individuals but the economy as a whole, since it is a well- established fact that income level rises with a college degree. Median income levels for young adults with a bachelor's degree are 64% higher than those with only a high school diploma [12]. Low retention rates also adversely affect the reputation of the educational institution and could lead to potential loss of funding and inability to compete for quality students. Thus, improving student retention is of paramount importance at institutions of higher education.

A critical factor in increasing student retention is the ability to accurately identify at-risk students, so that relevant interventions can be provided. Much of the prior research has been devoted to modeling the factors that impact student retention using traditional statistical methods. But, machine learning and data mining techniques have started becoming actively employed in student retention research in the recent past. Most research articles, though, have been focused on using structured data, such as GPA, SAT scores etc., that are readily available in institutional databases. To the best of our knowledge, as of this writing, there is no literature that tries to use unstructured data (e.g. free form text, images etc.) in predicting student dropout. Roughly 80% of the data generated in the world today is unstructured. Large amounts of unstructured data are generated by universities and colleges. Examples include advisor notes, discussion forum postings, online chats, emails etc. This is a treasure trove of information that has not been adequately exploited to help predict student dropout.

This paper describes a novel approach to predicting college student dropout using the information contained in free form notes recorded by student advisors on a student advising platform (e.g. EAB). We use Natural Language Processing (NLP) techniques to unearth the information contained in these advisor notes and use it to predict student dropout. To the best of the author's knowledge this study is one of the first to employ NLP techniques to predict student dropout. Thus, our study contributes to the literature by introducing an additional novel approach to predicting student dropout by using NLP techniques to analyze unstructured textual data in the form of advisor notes.

2. LITERATURE REVIEW

Research on student attrition has traditionally been based on surveying student cohorts and following them to assess dropout. These surveys contributed to the building of theoretical models of student retention, the most famous of them being the Tinto model [16]. Survey based research have been criticized for being too specific to an institution and hence not generalizable [1]. Also, these large scale surveys are not cost-effective to conduct. An alternative to survey based research is to use the data that most higher education institutions routinely collect about their students. This type of research based on institutional databases has been shown to be comparable to survey based research [2].

Prior research has also been mostly focused on identifying various factors that impact student dropout. Tinto [17] highlights academic difficulty, adjustment problems, lack of clear academic goals, lack of commitment, inability to integrate with the college community, uncertainty, incongruence and isolation as factors involved in student dropout. Tinto's theory of student integration posits that past and current academic success are crucial factors in determining student attrition and many studies have found high school GPA and SAT scores to have a strong effect on student retention [13]. Declaration of major and number of credit hours taken during the first semester have been used as proxies for institutional and goal commitment and have been found to be significant predictors of student attrition [1]. There have been many studies that have investigated the effect of financial aid on student retention [8, 9, 14]. These studies found that the type of financial aid that the student received had an impact on student retention.

Students receiving aid based on academic achievement had higher retention rates, while student loans had a negative effect on retention. Also, if students lost a scholarship or grant due to poor grades, it had a negative impact on retention. Thus, as evidenced above, almost all the studies have focused on factors that are part

J.D Jayaraman "Predicting Student Dropout by Mining Advisor Notes" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 629 - 632

of structured data collected by educational institutions. Factors such as emotions and sentiments that are embedded in unstructured data, have not been considered much in the literature.

Research on using machine learning techniques to predict student attrition is still in its infancy. Delen [6] used a dataset consisting of 39 variables such as SAT score, high school GPA, hours registered, hours earned etc. and several machine learning methods such as support vector machines and neural networks to model freshmen student attrition and found that support vector machines performed best, reaching a prediction accuracy close to 80%. Thammasiri, Delen, Meesad and Kasap [15] used data and techniques similar to Delen [6] to predict whether students would enroll for the second term. Lauria, Baron, Devireddy, Sundararaju, and Jayaprakash [10] used demographic and course related data to show that support vector machines performed better than decision trees at predicting at-risk students. Thus, almost all the research on student dropout prediction using machine learning and statistical techniques has focused on using structured data.

While there is not much literature using NLP techniques and unstructured data in predicting college student dropout, there is some recent literature in the related area of predicting student completion in Massive Open Online Courses (MOOCS). The most common NLP technique employed in these studies is sentiment analysis, which examines language in discussion forums and assignments to detect positive or negative emotion words and words that convey motivation, engagement etc. Wen, Yang, and Rose [18] examined students' opinion towards the course based on a sentiment analysis of discussion forum posts and used these opinions to predict course completion. Wen et al. [18] found that students who used words related to motivation were more likely to complete the course. Crossley, Paquette, Dascalu, McNamara, and Baker [4] used NLP techniques on MOOC forum posts and found that lexical sophistication, writing quality were predictive of student completion. Our study uses similar approaches to the literature described above but applies sentiment analysis to free form notes entered into an advisement system by the student's advisor, in order to predict student dropout.

3. METHODOLOGY

3.1 Data

The data consists of 19,562 notes entered over a period of four years (2015 - 2018) for 7343 undergraduate students at an urban university in the North Eastern United States which caters to a largely minority population. These notes are made by the student's advisor after each meeting with the student and are keyed into the student advisement system. These notes are free form and do not have any structure to them. Students typically meet with the advisor multiple times a semester to discuss enrollment, progress and any other issues. The notes the advisor makes documents the meeting in a reasonable amount of detail. Thus the notes are rich with information on any issues and difficulties students might be facing not only with respect to their academics but also with respect to their social and family life. We also compiled data on whether a student dropped out or not (a binary indicator variable). A student was considered to have dropped out if he or she did not enroll in any semester following the last semester of enrollment. Based on this definition we constructed a binary indicator variable to indicate whether a student has dropped out or not.

3.2 Analysis

3.2.1 Sentiment Analysis

Sentiment analysis is a NLP technique that attempts to categorize the emotions and sentiments in a block of text. Most sentiment analysis tools will categorize the sentiment as positive, negative or neutral and also provide indexes for affective states such as anger, sadness, happiness, etc. Sentiment analysis has been widely used to mine emotions from social media posts and has been effective in identifying depression, anxiety and other emotions [15].

There are two main approaches to extracting sentiment from text. The lexicon based approach uses a dictionary of words annotated with their sentiment polarities, while the text classification approach involves building classifiers from labelled instances of texts. Lexicon based approaches work well when there is insufficient human classified data or when human classification is time consuming and expensive. We use the lexicon based approach in this study as it would be very time consuming to hand classify the sentiment in the advisor notes to create a large enough training dataset. There are several sentiment lexicons available. We use a popular lexicon called the Bing lexicon [11] which consists of 6800 words, 2000 of which are positive and 4800 are negative. We also constructed a custom lexicon of 100 sentiment words relevant to the student retention domain and combined it with the Bing sentiment lexicon.

We preprocessed the data by removing stop words, punctuations, numbers, white spaces and other words such as will, student, etc. that would not be pertinent to conveying sentiment. The sentiment analysis was done on the preprocessed data. The output of the sentiment analysis is a list of words in each note tagged with a sentiment (positive or negative).

3.2.2 Imbalance

Data is said to be imbalanced if the number of instances in one class significantly outnumbers the number of instances in other classes. Since the number of dropouts is much smaller when compared to the number that don't dropout, student retention data sets are typically imbalanced. If the data is imbalanced the standard classifiers have a bias towards the larger majority class. One approach to correcting this imbalance is to preprocess the data in order to balance it out and then build the model. This approach uses various techniques to either oversample the minority class or undersample the majority class. Random oversampling attempts to balance the data by randomly sampling from the minority class and adding them to the training data set while random undersampling attempts to balance the data by removing data instances from the majority class. Undersampling has been shown to perform better than oversampling in some cases [7]. Synthetic Minority Oversampling Technique (SMOTE) is a popular and robust technique that uses a combination of oversampling the minority class and undersampling the majority class which results in better classifier performance than just oversampling or undersampling [3]. Our study uses SMOTE to correct the imbalance.

3.2.3 Classification

From the output of the sentiment analysis we computed the number of positive sentiment words and number of negative sentiment words in a note. We then computed the ratio of the number of positive sentiment words to the total number of words in a note. This ratio and the number of positive sentiment words were used as a measure of positive sentiment in the advisor notes. Similarly the ratio of negative sentiment words to the total words in a note and the number of negative sentiment words were used as a measure of the negative sentiment contained in the advisor note. We then used the ratios and word counts as features to predict student dropout. We trained a random forest classifier on the features extracted from the sentiment analysis to classify a student as likely to dropout or not. The random forest model is a popular ensemble model that provides good performance. We used 75% of the data to train the model and the rest 25% to test the model. We used a tenfold cross validation to avoid overfitting.

4. RESULTS

Figure 1 shows a word cloud of the commonly used sentiment words in the notes and Figure 2 shows the top ten frequently occurring words by positive and negative sentiment.

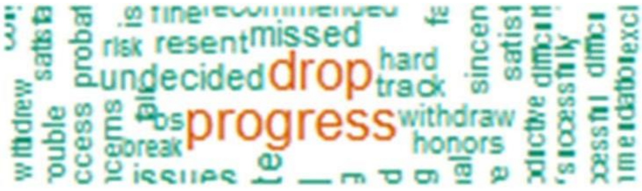

Figure 1: Word cloud of most used sentiment words in the advisor notes

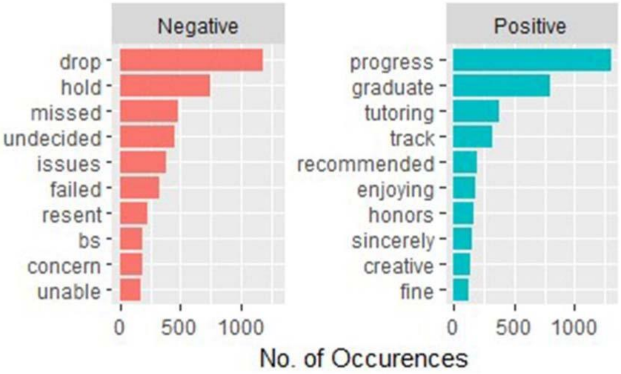

Figure 2: Top ten frequently occurring words by sentiment

The words "drop" and "progress" were the most used in the notes, with "drop" indicating a negative sentiment and "progress" a positive sentiment.

Table 1 shows some summary statistics of our NLP analysis of the advisor notes. After removing stop words and other words (e.g. appointment) that were used often in the notes but did not contribute to sentiment, we ended up with about 444,000 words that we used in the sentiment analysis. Some notes were long and detailed while others were short and cryptic. The maximum number of relevant words we found in a note was 92 and the minimum was just a couple of words.

Table 1: Summary statistics

Total number of words analyzed for sentiment	443,974
Maximum number of relevant words in a note	92
Maximum number of positive sentiment words in a note	42
Maximum number of negative sentiment words in a note	36

Since we corrected for the imbalance in the data (as described above) we report accuracy as the measure of classifier performance. The random forest classifier we trained produced a 73% accuracy predicting student dropout. The AUC metric was similar at 72%. We also trained three other classification models – Support Vector Machines (SVM), logistic regression and Classification and Regression Trees (CART) – to see if the accuracy could be improved upon. But, the logistic regression only achieved a 69% accuracy, while the SVM and the CART model achieved a 70% accuracy.

5. DISCUSSION

To the best of our knowledge this is the first study to use NLP techniques to mine advisor notes and use it to predict student dropout. The study demonstrates how information contained in unstructured data, such as advisor notes, can be automatically mined using machine learning techniques in a cost effective manner and used in early identification of at-risk students. The number of positive sentiment words and the number of negative sentiment words provide faculty, staff, administrators and advisors an additional indication of the risk of the student performing poorly or dropping out and can be used as a supplement to other traditional indicators of performance. Our analysis provides an indication of the likelihood of the student dropping out and thus helps in providing early intervention. The advantage of mining the advisor notes is that it picks up issues like family problems, stress etc. that cannot be picked up by simply looking at structured data such as grades, GPA, SAT scores, etc. Thus, what we have described is a powerful tool that can be used in addition to other techniques, such as predictive modeling using structured institutional data, to identify at-risk students.

The methodology we have described can easily be implemented in practice at any educational institution. Almost every educational institution uses some type of system that student advisors use to keep track of appointments and make notes. In fact, many institutions require advisors to document their discussion with the student. Thus, access to data is not an impediment. The NLP algorithm can be implemented using any open source tool such as R or Python and free libraries like NLTK for Python. Thus the approach used in this study can be cost effectively implemented at educational institutions and deployed via the advisement system.

There are several limitations and extensions to our study that will be addressed in future research. We used a general purpose sentiment lexicon that is more designed towards detecting sentiment in text like product reviews, etc. Most sentiment lexicons are general purpose ones, which is the reason why we augmented the Bing lexicon that we used with a 100 word custom lexicon. But, the custom lexicon we used is quite small. If we make our custom lexicon more comprehensive by increasing the number of words specific to the student retention domain, we would expect to get better results. Another limitation is that we have just used unigrams (single words) to detect sentiment. Using N-grams (multiple words) in our analysis should improve the sentiment detection and prediction accuracy. We could also take a non- lexicon based approach and hand classify the advisor notes to create a training dataset to predict dropout. It would be interesting to compare this type of approach and the lexicon based approach to determine if the expense of hand curating a training dataset is worth it. Further, we could combine the features extracted from the advisor notes with other traditional features such as GPA, SAT scores etc. to improve our prediction accuracy.

6. CONCLUSION

Unstructured data captured in various databases across the educational institution, including in online learning platforms (e.g. Blackboard), are a treasure trove of information that has not been adequately exploited to help the student in improving performance and avoiding dropout. Our study was an attempt at utilizing a small part of this unstructured data to help in the early identification of at-risk students. The fairly high level of prediction accuracy obtained in our study, even without much performance tuning, demonstrates the value of unstructured textual data in institutional databases for detecting at-risk students by predicting student dropout.

Future research should focus on unlocking the potential of unstructured data in institutional databases in helping the student. Other forms of unstructured data such as images, videos, audio clips, illustrations etc. that are created by students for different courses should also be used to extract information that could help provide early intervention and improve student retention.

7. REFERENCES

[1] Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. The Journal of Higher Education, 64(2), 123-139.

[2] Caison, A. L. (2007). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. Research in Higher Education, 48(4), 435-451.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[4] Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016, April). Combining click- stream data with NLP tools to better understand MOOC completion. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 6-14). ACM.

[5] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. ICWSM, 13, 1-10.

[6] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

[7] Drummond, C., & Holte, R. C. (2003, August). C4. 5, class imbalance, and cost sensitivity: why under- sampling beats over-sampling. In Workshop on Learning from Imbalanced Datasets II (Vol. 11). Washington DC: Citeseer.

[8] Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to- second year analysis of new freshmen. Research in Higher Education, 46(8), 883-928.

[9] Hochstein, S. K., & Butler, R. R. (1983). The Effects of the Composition of a Financial Aids Package on Student Retention. Journal of Student Financial Aid, 13(1), 21-26.

[10] Lauría, E. J., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012, April). Mining academic data to improve college student retention: An open source perspective. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 139-142). ACM.

[11] Liu, B. (2010). Sentiment Analysis and Subjectivity. Handbook of natural language processing, 2, 627- 666.

[12] McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., ... & Bullock Mann, F. (2017). The Condition of Education 2017. NCES 2017-144. National Center for Education Statistics.

[13] Porter, K. B. (2008). Current trends in student retention: A literature review. Teaching and Learning in Nursing, 3(1), 3-5.

[14] Stampen, J. O., & Cabrera, A. F. (1986). Exploring the Effects of Student Aid on Attrition. Journal of Student Financial Aid, 16(2), 28-40.

[15] Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications, 41(2), 321-330.

[16] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89-125.

[17] Tinto, V. (1993). Building community. Liberal Education, 79(4), 16-21.

[18] Wen, M., Yang, D., & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In Proceedings of the 7th International Conference on Educational Data Mining (EDM2014).

[19] Wetzel, J. N., O’Toole, D., & Peterson, S. (1999). Factors affecting student retention probabilities: A case study. Journal of Economics and Finance, 23(1), 45-55.Journal of Economics and Finance, 23(1), 45-55.