Social Data and College Statistics

Sean Choi
Stanford University
yo2seol@stanford.edu

Elena Grewal
Stanford University
etgrewal@stanford.edu

Kai Wen
Stanford University
kaiwen@stanford.edu

ABSTRACT

We correlate aspects of data from the online micro-blogging service, Twitter.com, with college statistics, in order to understand whether the sentiment of discourse on Twitter can predict important outcomes like the number of applicants to a college or the graduation rate. We also test different methods of sentiment analysis of Tweets. We find that the amount of "buzz" about a college on Twitter predicts the number of applicants to the college, even when controlling for the number of applicants in the previous year. We also explore various methods to classify the sentiment of tweets about a school. We find that the sentiment of insiders at a college predicts the freshman retention rate, but that this result is explained by average SAT score and school size. The sentiment of Twitter messages about a college does not predict the number of applicants, the acceptance rate, or the graduation rate. In addition, we find high variance in the sentiment classifications from our three methods, pointing out the importance of being mindful of the classification method and the way in which the method chosen might influence results. The paper adds to the growing literature on the predictive power of social data, documenting its strengths and limitations, and applies these techniques to a novel set of outcomes.

General Terms

Experimentation, Measurement

Keywords

Social data, college statistics, sentiment analysis

1. INTRODUCTION

Social data has been used to predict a number of different outcomes including political opinion, stock prices, and movie revenues [1, 2, 3, 4]. However, such data has never been used to predict important outcomes for colleges such as the number of applicants, the acceptance rate, the US News and World Report rank, the freshman retention rate, and the graduation rate. If social data can predict these measures, then it can be used by college administrators to allocate correct resources to admissions or to respond to a predicted increase or decrease in the measures. It could be used by applicants to understand whether or not a college is actually a positive place. College ranking publications such as U.S. News and World Report could add a "social data" factor- a new ranking of colleges based on how people feel about the colleges rather than just the average SAT score or average GPA of students in attendance.

We hypothesize that the "buzz" about a school as measured by the number of tweets that mention a school will be a positive predictor of the number of applicants to the school and the acceptance rate of applicants at the school. In addition, we hypothesize that the sentiment of tweets about a school will predict the number of applicants, as well as other outcomes such as the freshman retention rate and the graduation rate. The sentiments of Twitter users who know more about a school and possibly attend the school should be an even better indicator of measures such as the freshman retention rate and the graduation rate. When those individuals express positive sentiments about a school, we predict that the freshman retention rate and the graduation rate will be higher. There are limitations in that the individuals who tweet may not be representative of the broader population, and the sentiments of those on twitter are different from those not on twitter; prior research finds that twitter users are not representative of the broader population with regard to geography, gender, and race [5]. However, it may be that the population of potential college students are better represented on Twitter.

We find that the count of mentions of a school is a positive predictor of the number of applicants, controlling for other factors. In addition, we contribute to the literature on sentiment classification of tweets by comparing multiple measures of the sentiment of tweets, namely a classifier using a lexicon of positive and negative words from OpinionFinder and then a multi-variate naive Bayes classifier. Though some of our measures using the sentiment classifications do predict college outcomes, we find that the sentiment of the tweets does not provide additional predictive power when matched against the average SAT score of students at the school and the size of the school. In addition, we find that our sentiment classifiers do not identify the positive to negative ratio well on our hand-labeled test data, providing further evidence that better methods are needed for categorizing the senti-

#Influence12 - Symposium on Measuring Influence on Social Media, September 28-29, 2012, Halifax, NS, Canada.

ment of tweets on topics such as colleges. Multiple prior papers have used the OpinionFinder lexicons but did not rigorously investigate the accuracy of the measure on test data. Our paper provides evidence to call into question the use of this sentiment classification method when using social data to predict other outcomes.

2. RELATED WORK

A number of studies look at correlations between Twitter sentiment and other outcomes. O'Conner et al. correlate results from the Index of Consumer Sentiment (ICS), Gallup's daily presidential job approval tracking poll, and other tracking polls during the 2008 U.S. presidential election cycle to sentiment in contemporaneous Twitter messages. They identify tweets related to consumer confidence as those that contain the topic words "economy," "job," and "jobs", tweets about Obama as those that contain the word "obama" and tweets about the campaign as those with the words "obama" and "mccain." They identify tweet sentiment by looking for positive and negative words in the tweets, using the word lists from OpinionFinder, and find a strong correlation between sentiment of tweets and poll results. Similarly, Bollen et al. use Twitter data to predict stock market outcomes using both OpinionFinder word lists and the Google Profile of Mood States (GPOMS), which measures mood in terms of 6 dimensions (calm, alert, sure, vital, kind, and happy). They find that variations along the mood dimensions of "calm" predicted changes in stock market prices, but the positive and negative sentiments identified by the OpinionFinder word lists do not. Neither of these papers evaluate the accuracy of the OpinionFinder sentiment classification; O'Conner et al note that they found many errors in the sentiment classifications but do not document the error rate of the sentiment classification.

Asur and Huberman find that the count of mentions of a movie predicts box office success. Wong et al. collect data from Twitter between February-March 2012 and manually labeled 10K tweets (using MT) as training data, and then create a set of SVM classifiers, rather than using positive and negative word lists to classify Tweet sentiment. They then look at the ratio of positive tweets before a movie is released to positive tweets after the movie is released and see no correlation with IMBD and Rotten Tomatoes reviews. again they do not test the accuracy of their sentiment analysis.

Prior research in the field of education indicates that students take into account a number of factors when deciding where to apply to college, such as geographic proximity and price, and then a number of other factors determine whether a student chooses to stay in school. It is likely that students will tweet about colleges and that those tweets reflect their feelings about a school. No prior work has correlated Twitter sentiment with college outcomes, and the prior papers that do correlate Twitter sentiment with other real world phenomena do not assess the validity of the sentiment analysis.

3. DATA & METHODS

We use one month of tweets, provided from October 7th, 2011 to November 7th, 2011, in the format of JSON-based logs. The primary statistics for colleges are obtained from IPEDS (The Integrated Postsecondary Education Data System). We also obtain the US News and World Report College Ranking and extract the list of top 100 US national universities and top 100 US liberal arts colleges. These are the main schools in our sample, though the sample size of schools with complete information on all variables is smaller (see Tables 2 and 3 for the sample size in each model). In addition we include a measure indicating the average SAT score of students at the college, calculated by taking the midpoint between the 25th percentile score and the 75th percentile score, as well as the size of the school, measured by the number of full time enrolled students at the school.

We employ two different methods to identify tweets that are related to a college. One is to look for tweets that contain the full name of a college; the other is to use the geolocation information provided for some tweets to identify tweets from within the college campus. We consider tweets written within the college campus as those written by by "insiders" who either attend the college or visit the college. We find 4,617,923 tweets that mention the name of a college, and 3,558,905 tweets from the geographic location of the colleges.

We explored three different methods to determine the sentiment of tweets about a college. First, we use a list of 1,600 and 1,200 words marked as positive and negative, respectively, from OpinionFinder [9], and classify a tweet as positive if it contains a positive word and negative if it contains a negative word. In this schema, any word that is not positive or negative will be considered neutral and will not provide any addition to the sentiment score. A tweet can be both positive and negative and so the method is similar to counting the total number of negative and positive words as most tweets do not have many words. Second, we identify tweet sentiment by looking at whether the tweet contains positive or negative emoticons. The last method that we explore is a Naive Bayes classifier. We use both the words contained in tweets as features, and then also tried models where we removed stop words and used chi-squared feature selection. We found that the model that contained all words was best.

In order to train our model for sentiment analysis, we use training data from Ref. [7], which include around 500 tweets manually labeled with positive, negative, and neutral. We use 90% of these tweets for the training set and then 10% for the development set, training the model and then testing on the development set and using cross validation to identify the accuracy of our results. We set aside an additional 500 manually labeled tweets as the test set. Laplace smoothing was performed to reduce over fitting. The training and test set represent a small fraction of the identified tweets, and so it is is possible that we do not have enough data to train and test our models, a potential threat to the validity of our results.

We use a linear regression model to predict college outcomes using social data.

4. RESULTS

4.1 Descriptive statistics

The college term match results returned the counts of mentions of the colleges. Boston College is an outlier here, men-

Table 1: Correlations between the measures of the ratio of positive to negative tweets

	OpinionFinder	Emoticons	MV Bayes
OpinionFinder	1		
Emoticons	-0.169	1	
MV Bayes	0.025	0.239	1

tioned in 18,060 tweets. Figure 1 shows the distribution of counts, excluding Boston College.

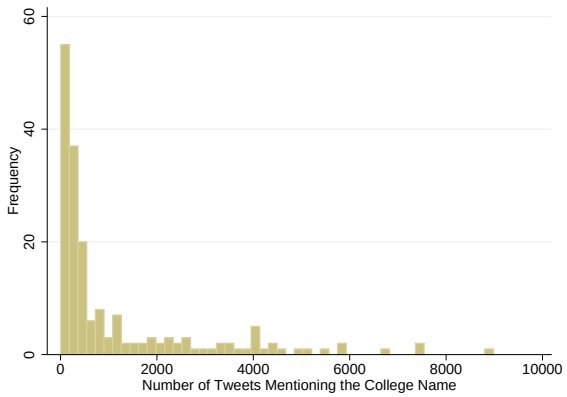

Figure 1: Distribution of the counts of mentions of colleges.

The count of mentions of a college's name is positively correlated with the number of applicants, as shown in Figure 2.

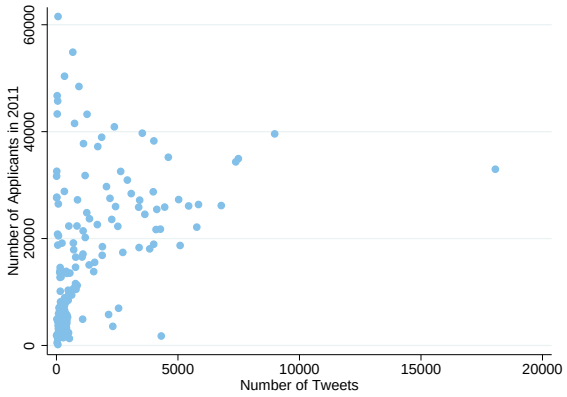

Figure 2: Number of tweets versus Number of Applicants in 2011.

The precision of the opinion finder, emoticon, and naive bayes classifiers was 0.50, 0.83, and 0.73 respectively. The recall was 0.29, 0.08, and 0.61. We calculate the balanced F1 measure (2PR/(P+R)), and find 0.54, 0.15, and 0.66. While the emoticon method was very precise, many positive and negative tweets did not contain emoticons, so recall was low.

As shown in Table 1, because the correlations are low among

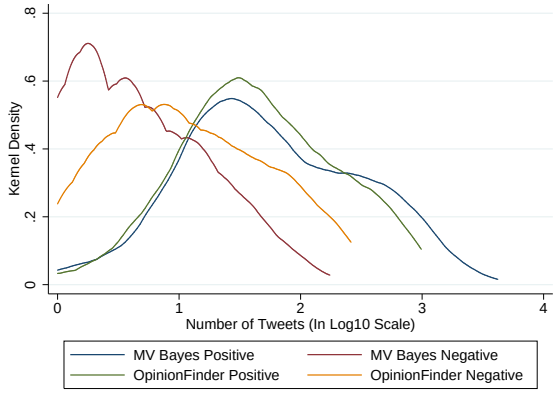

Figure 3: Kernel Density Plot of the Log of Number of Tweets Classified as Positive or Negative.

the different ratios, it is clear that the classifiers are working quite differently. MV Bayes had the highest precision rate on our hand-labeled test data, but the ratio of positive to negative tweets identified in the test data was much higher than the actual ratio. The actual ratio was 0.95; the OpinionFinder classifier found a ratio of 1.68, and the MV Bayes found a ratio of 9.5. The reason that the MV Bayes classifier has such a high positive to negative ratio is because the classifier finds few negative tweets. Figure 3 shows a kernel density plot of the log of the counts of positive and negative tweets found by MV Bayes vs. OpinionFinder. The number of positive tweets has a similar distribution, but the number of negative tweets found by MV Bayes has a different distribution, shifted to the left indicating a higher density of counts of low numbers of tweets identified as negative.

4.2 Regression results

The regression results indicate that the count of the number of mentions of a college name is predictive of the number of applicants in 2011. The coefficient on the count is positive and statistically significant even when controlling for the number of applicants in 2010 and the size of the school and mean SAT score of the school. In contrast, the sentiment of the tweets is not predictive of the number of applicants when the number of applicants in the prior year is included as well as the other variables. The sample size changes because in some schools there were no negative tweets and so the ratio could not be calculated, and the location data only includes 30 schools; it may be that the results would be significant if there were a larger sample of schools in the models.

Though the count of the number of tweets does improve predictions of the number of applicants to a college, the total applicants in the prior year explains much of the variance. The R-squared increases from 0.179 to 0.987 (see columns (1) and (2) of Table 2) when the applicants from the year before, the size of the school, and the average SAT score are included in the model. Note that some schools did not report SAT scores so the sample size is smaller when that variable is included. The results are the same if we only include the total applicants from the year before as covariate (and in that case, the sample sizes remain the same for each pair of models).

Table 2: Predicting Number of Applicants.

	(1)	(2)	(3)	(4)	(5)	(6)
Count	2.666***	0.133*				
	(0.424)	(0.064)				
Total Apps in 2010		1.025***		1.031***		1.023***
		(0.014)		(0.017)		-0.016
Number of Enrolled Students (in 1000s)		25.194		103.887		98.615
		(97.724)		(109.075)		(98.004)
Average SAT Score		2.358+		1.805		1.988
		(1.276)		(1.608)		-1.306
MV Bayes Ratio			48.700**	-1.665		
			(17.150)	(2.222)		
OpinionFinder Ratio					-348.785+	-14.072
					(208.568)	(27.204)
Number of Schools	183	165	130	121	161	146
R-squared	0.179	0.987	0.059	0.986	0.017	0.986

Notes: Standard errors in parentheses. *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. The "ratio" refers to the ratio of positive to negative tweets.

We then turn to regressions predicting graduation rate and acceptance rate and find that none of our measures from Twitter predict these outcomes. We also check whether our measures predict the U.S. News and World Report ranking but did not find consistent results.

At first glance, it appeared that the higher the ratio of positive to negative tweets and the more emotional the tweets the lower the freshman retention rate - see columns (1), (3), and (5) in Table 3. In particular, the sentiment measures derived from the insiders data - the ratio of positive to negative tweets and the overall polarity of tweets - was negatively associated with the freshman retention rate - perhaps because the many positive tweets are due to excessive parties that might be associated with lower rates of retention. However, as shown in Table 3 this result does not hold when size of school and average SAT score are included in the model. Size of school and average SAT are positively correlated with freshman retention and negatively correlated with the ratio of positive to negative tweets, so the coefficients were biased downward when those variables were not included in the model. The location data only includes 30 schools, so it may be that if the sample size were larger, then the results would be significant for the sentiment of insiders. Interestingly, the coefficients on the sentiment scores are negative, indicating that the higher the ratio of positive to negative tweets the lower the freshman retention rate.

5. CONCLUSIONS

We conclude that the "buzz" about a school, as measured by the number of tweets that mention of the name of the school, is a significant predictor of the number of applicants to a college. The result holds when controlling for the number of applications in the previous year and the size of the school; the number of applications in the previous year explains most of the variance in the number of applicants in the current year, but the number of tweets still explains the residual variance. Twitter data does not predict graduation rates or acceptance rates or rank of the school. A number of measures predict the freshman retention rate, but the result does not hold when controlling for size of the school and mean achievement of the school.

The results would be improved by the inclusion of more training data and perhaps by trying different classifiers (such as maxent and svm). Both papers that looked at movie sentiment and revenue used training data consisting of thousands of hand labeled tweets, in contrast to our training data of around five hundred tweets. Our concern about the quality of our training data and sentiment models is further justified by the evidence that our current classifiers are overestimating the share of positive tweets in a way that might bias our results from the correlation of sentiments with college outcomes. Another consideration is that while we do not find that the current sentiment classifiers accurately classify the sentiment of individual tweets, it is possible that in aggregate the sentiment of the tweets is accurate. O'conner et al find that the sentiment from OpinionFinder did spike on national holidays, indicating that there was some signal in the sentiment classifications in aggregate.

Finally, one limitation of our analysis has to do with the timing of the tweet sentiment classification. We only have Twitter data from the month of October, and most college applications for the regular admissions cycle are due in December or January. We may find different results for predicting application numbers if we look at tweets during those months.

There are a number of possible extensions of this work. The first is to use these results to improve the sentiment classification ot tweets, in addition to gathering more training data. In addition, it would be interesting to examine the content of the negative or positive tweets and see whether certain types of negativity or positivity are related to college outcomes. Finally, it may be the case that certain members are more representative of the sentiments of students; we might check whether the sentiment of influential members of the Twitter network better predict college outcomes.

6. REFERENCES

[1] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on*

Table 3: Predicting Freshman Retention Rate.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
OpinionFinder Insiders Ratio	-0.492*	-0.193						
	(0.188)	(0.139)						
OpinionFinder Insiders Polarity			-29.572**	-10.308+				
			(8.317)	(5.646)				
MV Bayes Insiders Ratio					-0.328*	-0.127		
					(0.138)	(0.105)		
MV Bayes Insiders Polarity							-34.121**	-5.922
							(12.202)	(8.719)
Average SAT Score		0.017**		0.021***		0.017**		0.020***
		(0.005)		(0.004)		(0.006)		(0.005)
Number of Enrolled Students (in 1000s)		-0.092		0.153		-0.094		0.048
		(0.246)		(0.195)		(0.261)		(0.223)
Number of Schools	28	28	28	28	28	28	28	28
R-squared	0.209	0.776	0.327	0.787	0.178	0.771	0.231	0.762

Notes: Standard errors in parentheses. *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. The "ratio" refers to the ratio of positive to negative tweets. The "polarity" refers to the ratio of positive plus negative tweets to all tweets.

Weblogs and Social Media, pages 122-129, Washinton, DC, May 2010.

[2] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1-8 March 2011.

[3] S. Asur, and B. A. Huberman. Predicting the Future with Social Media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 1:492âĂŞ499, Lyon, France, August 2011.

[4] F. M. F. Wong, S. Sen, and M. Chiang. Why Watching Movie Tweets Won't Tell the Whole Story? arXiv:1203.4642v1, 2012.

[5] A. Mislove, S. Lehmann, Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the Demographics of Twitter Users. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, pages 554-557, Barcelona, Spain, July 2011.

[6] http://www.locomatix.com

[7] http://www.sentiment140.com

[8] N. Jindal, and B. Liu. Opinion Spam and Analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219-230, Palo Alto, CA, February 2008.

[9] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Joint Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, pages 347-354, Vancouver, Canada, October 2005.