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 Twitter data with college statistics. 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. 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. Our paper contributes to the growing literature on the predictive power of social data by investigating these relationships. 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 other measures. College ranking publications such as U.S. News and Word 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. If those people express positive sentiments about a school then we predict that the freshman retention rate and the graduation rate will be higher. There are limitations in that the people 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, net of 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 sentiment 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. DATA

In this section, we briefly describe the two sets of data used in the paper: social data from Twitter, and college statistics from IPEDS (Integrated Postsecondary Education Data system) and the U.S. News and World Report.

2.1 Twitter data

Twitter is one of the most widely used social networks nowadays, and the messages people wrote every day, called "tweets", are of huge value for social data analysis. One month of tweets are provided from October 7th, 2011 to November 7th, 2011, in the format of JSON-based logs. The tweet logs contain more than 3 billions entries, which total 2.1 terabytes after compression stored on Amazon Simple Storage Service (S3), and around 20 terabytes after decompression. A notable feature of a tweet is that it usually very short: the number of characters are limited to 140 characters, and most tweets are even much shorter than 140 characters. Each log consists of various information about a tweet, including:

- tweet text,
- time stamp of generating the tweet,
- geolocation where the tweet is generated if available,
- how many times the tweet is forwarded by other people (number of retweets),
- and information of the author and his/her followers and following statistics.

2.2 College statistics

The primary statistics for colleges are obtain from IPEDS, which has been conducted yearly by the U.S. Education Departments National Center for Education Statistics since 1980. We look at the following variables:

- Number of applicants in 2010 and 2011
- Acceptance rate in 2010
- Graduation rate in 2011
- Freshman retention in 2011

We also obtain the US News and World Report College Ranking, in which we extract the list of top 100 US national university and top 100 US liberal arts colleges. Of these, 183 also have data on the other variables of interest. These are the main schools in our sample. 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.

3. EXPERIMENT PREPARATION

The first task to complete in order to correlate Twitter data with college statistics is to identify tweets that mentioned a college, which is not a trivial problem. The accuracy of the tweets identification will affect the final result significantly. Following the identification of tweets about a college, we also need to prepare training data for sentiment analysis.

3.1 Tweet identification

We employ two different methods to identify tweets that are related to a college. One is to use only college names to match the tweet text, the other one is to use the geo-location of the tweets to identify tweets made by "insiders."

3.1.1 Term match method

The term match method uses college names to identify tweets about a college. For example, we can say that the tweets containing the phrase "stanford university" are related to Stanford University. Also there are abbreviated names like "wfu" for Wake Forest University which are used in tweets to mention colleges.

In processing tweets using term match method, we first collect full names and abbreviated names for the 183 colleges and universities. Then we run a Map/Reduce job on the given tweets log, which first convert the tweet text to lower case and then match them with the college names. We make no attempt to limit user location or message language when matching college names to tweet text. The results of 4 example universities and colleges are shown in Table 1. The numbers of tweets matched college names may not always be proportional to the size of the colleges, due to several reasons including the popularity of Twitter, and the popularity of the schools in public media, etc.

We match both the full name of the colleges and the abbreviated names. However, there are a number of issues with matching abbreviated names rather than full names. If a college's abbreviated name is also a common location name, or another common term, there will be many erroneous matches. We might only consider abbreviated names that we think would have a good chance of just referring to the college. For example, "MIT" would not be considered, but "UCSB" is considered. However, at the same time, there are problems with selectively using abbreviations. For most of our models we use the full names of the schools. Using the full name ensures that we can include schools like "rice university" in which the abbreviation "rice" would certainly not identify the college. We believe that each college would have a similar decreased number of tweets when looking at the full name, because most schools include "college" or "university" after the name. Nonetheless the recall is probably low since many tweets might mention to a college but not include any full or abbreviate college name.

For some of the colleges, the abbreviated name is used much more than the full name - for example, "pitzer" is mentioned in 571 tweets, but "pitzer college" is mentioned in only 69 tweets. We did try an alternate model where we used the abbreviated names of eight schools with long and little used full names, such as "ucsb" instead of "university of california-santa barbara," and our main results did not change. For the purposes of this paper, however, we do not include information from any of the other abbreviated names. It may be that the choice to not use abbreviated names changes our results and it would be helpful to check this in the future.

In addition to the difficulty with choosing which names to use for the college term-match, we noticed that many of the tweets that mention a university are foursquare checkins or news announcements and so there may be a higher share of

Table 1: Term match results of 4 universities and colleges.

Institute name	# of students in 2011	Matched tweets
Agnes Scott College	228	123
George Washington University	2383	9985
Harvard University	1661	169239
Stanford University	1707	227185

neutral texts in this group of tweets than in all the tweets that actually refer to a college or to another random sample of tweets. This may have implications for our sentiment analysis.

3.1.2 *Insider method*

The other approach we use to identify tweets that relate to a college is based on the geolocation of the tweets. 1% of the given Twitter dataset are tagged with a location inside the US. These tweets enable us to find those Twitter users located within a college campus. However, such tweets are still a small portion of the total tweets that may mention a college. We take a further step to identify the users who ever composed a tweet inside a college campus, and we name these users as college insiders. Then we find all tweets from the users as a sample of tweets that are related to colleges.

The data processing of tweets with geolocation data is done using Locomatix[6], which provides spatial indexes to allow us to query tweets within a college campus given its campus polygon. For example, Figure 1 illustrates the process in which the polygon of the campus is drawn and sent to Locomatix, and the tweets are returned within the given polygon area.

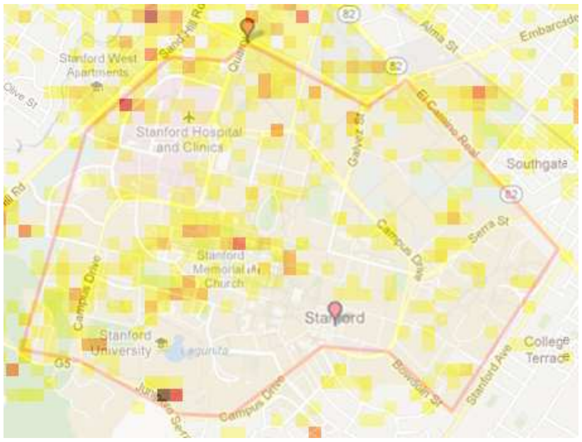

Figure 1: Number of tweets histogram overlay on the Stanford area map. The red polygon draws the boundary of the campus of Stanford University. The light yellow to dark red blocks visualize the number of tweets located within the blocks, and the darker the color, the more tweets in the block.

We extract the tweets located within the campuses of 30 colleges in US, and the authors of those tweets are identified as insiders. Finally we run an additional round of Map/Reduce jobs to fetch all tweets from these college insiders, and the results for four of the schools are shown in Table 2.

The insider match method is comparably more accurate in the sense that we are also interested in relating the insiders' opinions on the colleges and the colleges statistics. It also has the advantages over the term match method in finding the tweets that did not include any college name but still mentioned a college, increasing the recall rate. But it may contain visitors to a college campus who tweeted inside the campus during their visits, rather than true insiders. A easy way to remove those erroneous counts is to identify college insiders who tweeted at least a certain number of times. However, at this point, we are more interested in collecting the opinions from the users even including visitors, so we do not set any limitation on the number of times the user tweeted in the campus when extracting the tweets.

3.2 Training and test data for sentiment analysis

In order to train our model for sentiment analysis, a proper training dataset is crucial. We have tried using three different training data:

1. Training data from Ref. [7], which include around 500 tweets manually labeled with positive, negative, and neutral.
2. Amazon product review data labeled with positive and negative[8]. It has the negative and positive reviews for four different product categories, including books, dvd, electronics, kitchen). Each category has 1000 positive and 1000 negative reviews that are used for training.
3. Tweets labeled by positive emoticons and negative emoticons. For example, if a tweet text includes a smiley face :-), it is considered to be positive, while a sad face :-(identifies a negative tweet.

To perform the test in evaluating our model, we use part of the training dataset to do cross validation, and also employ around 500 manually labeled tweets as test data.

4. METHODS

4.1 Sentiment Analysis

We explored three different methods to determine the sentiment of tweets about a college, ranging from a simple word count algorithm to a machine learning based classification algorithms. First, we use the 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

Table 2: Insider match results of 4 universities and colleges.

Institute name	# of students in 2011	Matched tweets
Columbia University	1391	149536
University of Florida	6395	84200
Harvard University	1661	312362
Stanford University	1707	153841

of negative and positive words as most tweets do not have many words.

We define the sentiment score x to be the ratio of positive to negative tweets for a school s:

$$x_s = \frac{count(pos.word)}{count(neg.word)}$$

This method was susceptible to many problems and resulted in low accuracy on the test set (0.50). If the word in the tweet that contained negative sentiment did not appear in our corpus, it would be ignored and would result in false classification. The biggest problem we found was that tweets that were truly neutral were classified as positive or negative because of the presence of one of the words on the list. However, this method has been used by multiple prior papers identifying the sentiment of tweets [1, 2] and it serves as an additional comparison.

For each of our sentiment measures, we also calculate the ratio of the sum of the number of positive and negative tweets over the total number of tweets as a measure of "polarity."

Second we find sentiment using emoticons. An emoticon is a pictorial representation of a facial expression using punctuation marks and letters, usually written to express a person's mood. Given such a definition, we can see that the presence of emoticons would provide some information about the sentiment of a given tweet. Thus, using a method similar to above, we obtained a set of positive and negative emoticons as a corpus instead of a list of words. This method also was susceptible to similar problem as above. Another problem is the lack of tweets containing emoticons. We found that only 1% of the entire tweets contained an emoticons, and that many tweets that were positive or negative did not contain an emoticon and so were not found. In the hand labeled test data, we only found five tweets with emoticons. These were classified correctly for their sentiment, but there were then an additional 114 positive and negative tweets that were not found.

The last method that we have explored is the Naive Bayes classifier for sentiment analysis. To provide some background, the Naive Bayes classifier uses Bayes rule to estimate the probability of the sentiment of a tweet given the words in the tweet, using information from training data that is already classified. Assume that we have some document in the training data. We model each document as a set of independently occurring set of words $d = (x_1, x_2, ..., x_n)$. Given these set of words, the task is to classify a new document by

$$\begin{aligned} c_{MAP} &= \arg\max_{c_j \in C} P(c_j|x_1, x_2, ..., x_n) \\ &= \arg\max_{c_j \in C} \frac{P(x_1, x_2, ..., x_n|c_j)P(c_j)}{P(x_1, x_2, ..., x_n)} \\ &= \arg\max_{c_j \in C} P(x_1, x_2, ..., x_n|c_j)P(c_j) \end{aligned}$$

(MAP is an acronym for maximum a posteriori, meaning the most likely class). Notice that the second line comes from applying a basic Bayes rule. Now, for Naive Bayes to work easily, we need to make three different assumptions:

- $P(c_j)$ can be estimated from the frequencies of classes in the training examples
- $P(x_1, x_2, ..., x_n|c_j)$ cannot be estimated unless we have a very very large training example thus we assume that $P(x_1, x_2, ..., x_n|c_j) = P(x_1|c_j)P(x_2|c_j)...P(x_3|c_j)$.

Thus, we arrive at the final form:

$$c_{MAP} = \arg\max_{c_j \in C} P(c_j) \prod_i P(x_i|c_j)$$

We explore two models for estimating these probabilities. One is the multivariate Naive Bayes model, which estimates:

$$\hat{P}(x_i|c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

which represents the fraction of documents of class c_j that contain the term x_i. Notice that this model disregards term frequency information, so it is also called the binomial Naive Bayes. The second model we have explored is called the multinomial Naive Bayes model, which estimates:

$$\hat{P}(x_i|c_j) = \frac{n_{x_i}}{n_{c_j}}$$

where n_{x_i} represents the number of x_i in the entire training documents of class c_j and n_{c_j} is the total number of terms in the collection of documents with the class c_j. Notice that when these probabilities are zero, we can have zero probability for the entire product. So, we use of Laplace smoothing (add-one smoothing), in which we change each term estimate into:

$$\begin{aligned} \hat{P}(x_i|c_j) &= \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)} \\ &\Rightarrow \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k} \end{aligned}$$

where k is the number of possible values for X_i.

We also explored a method of feature selection called χ^2 feature selection. χ^2 feature selection is a method of scoring each term in the corpus for each class to see if it is a good indicator of the classification. We rank each term for each class:

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

where e_t is the indicator if the term is present and e_c is the indicator if the class is selected. For example $N_{e_1 c_1}$ will denote the number of term t that are in the class. After we rank them, we took the top k vocabulary to limit the features.

Lastly, given that Naive Bayes classifier is a linear classifier and we have 3 classes to classify into, we have explored different ways to formulate the problem that gives us the greatest accuracy on our test set. One is the method of One vs. All classification, where we test to see if the document is contained within one of the desired classes or not. In this schema, we create two classifiers that classifies between positive and not positive, and negative and not negative and picks the choice that occurs in the set. For example, if a class is not positive but negative, we consider it a negative document. If the document is both positive and negative, we consider their probabilities to see if they are more positive and negative. One other method is One vs. One classification, where we compare each class with all other classes to see where the document belongs to. Under this schema, we get the class with the highest probability out of the three and label it as such.

4.2 Regression Analysis

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

$$CollegeOutcome = \beta_0 + \beta_1 OtherData + \beta_2 TwitterData + \epsilon \quad (1)$$

The four college outcomes are each considered separately in regressions including features of the Twitter data such as the count of mentions of a college or the sentiment of tweets about a college. Then, additional control variables are included in the models to see if Twitter data can explain variation in the outcomes unexplained by other variables.

5. RESULTS

5.1 Selection of sentiment classifier

First, we choose the multivariate Naive Bayes method over multinomial Naive Bayes for the following reasons. Multinomial Naive Bayes relies heavily on the term frequencies for each document. However, each tweet is very short and rarely contain more than one important terms. Thus, it was not possible to classify the documents well. We obtained about 63% accuracy over our test set using multinomial Naive Bayes model. Since there are not many duplicates, we assumed that the presence of each term is a huge signaling factor for classification. Thus, we choice the multivariate Naive Bayes model. All of the results below is obtained from this model.

As discussed above, we obtained a corpus of manually labeled twitter training data, as well as some Amazon review data. We wanted to find out if the addition of Amazon training data helped our classification. Also, we wanted to find out if χ^2 feature selection helped our classification results. In order to obtain the following accuracy values, we have used k-fold cross validation on the training data to obtain consistent result. Also, Laplace smoothing was performed to reduce over fitting.

Table 3: Results of k-fold cross validation using different classifiers

	$k=1$	$k=5$	$k=10$	$k=20$
Amazon, χ^2	0.6574	0.6100	0.6295	0.6267
No Amazon, χ^2	0.8412	0.7270	0.7047	0.7298
Amazon	0.8886	0.7604	0.7632	0.7465
No Amazon	0.9694	0.7660	0.7660	0.7716

Table 4: Accuracy comparison for One vs. All and One vs. One approaches

One vs. All	0.7067
One vs. One	0.6311

Table 3 shows that adding no amazon training data nor doing any χ^2 feature selection resulted in the highest accuracy values. Our assumption for such result is due the size of the training data set as well as the difference in term distributions in different training data. Since the size of the training data set is small already, reducing the feature might have greatly reduced it's predicting power, thus χ^2 feature selection did not help. As for Amazon training data, since the training data was so different than the twitter data, it might have reduced our accuracy on the test set, which was a set of twitter documents.

Now, given that we decided to not include any Amazon data or do any feature selection, we evaluated the accuracy of two different methods of multi-class classification. The accuracy was obtained for the test set that we hand labeled, in order to get the final accuracy results. The accuracy is shown in Table 4. Notice that One vs. All method outperformed One vs. One method. Thus, this was our choice of Naive Bayes classifier to do the sentiment analysis on the set of tweets.

5.2 Descriptive statistics

The college term match results returned the counts of mentions of the colleges. Boston College is an outlier here, mentioned in 18,060 tweets. Figure 2 shows the distribution of counts, excluding Boston College.

The count of mentions of a colleges name is positively correlated with the number of applicants, as shown in Figure 3.

On average, the multi-variate naive bayes classifier classifies 15% of the tweets that mention a college name as either positive or negative, and the ratio of positive to negative tweets is 40. On average, the opinion finder classifier classifies 13% of the tweets that mention a college name as either positive or negative, and the ratio of positive to negative tweets is 6. On average, the emoticon classifier classifies 0.02% of the tweets that mention a college name as either positive

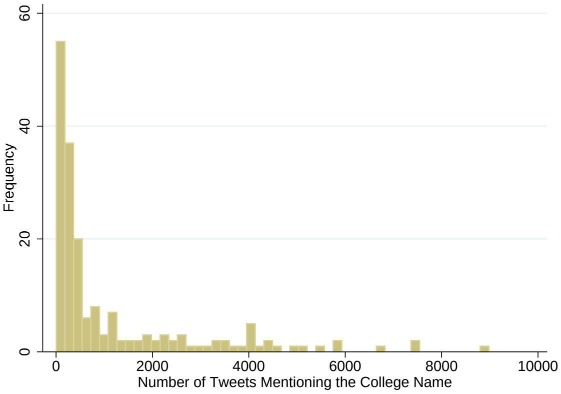

Figure 2: Distribution of the counts of mentions of colleges.

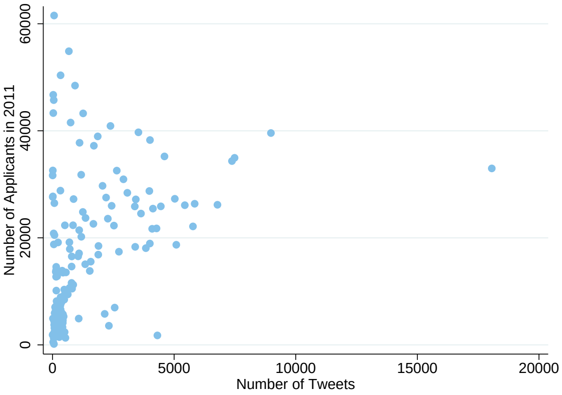

Figure 3: Number of tweets versus Number of Applicants in 2011.

or negative, and the ratio of positive to negative tweets is 9. The reason that the MV Bayes classifier has such a high positive to negative ratio is because the classifier finds few negative tweets. Figure 4 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.

As shown in Table 5, because the correlations are low among 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. Therefore we conclude that it is very likely that the MV Bayes is overestimating the number of positive tweets and that the OpinonFinder is also overestimating the ratio of positive to negative tweets.

Table 5: 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

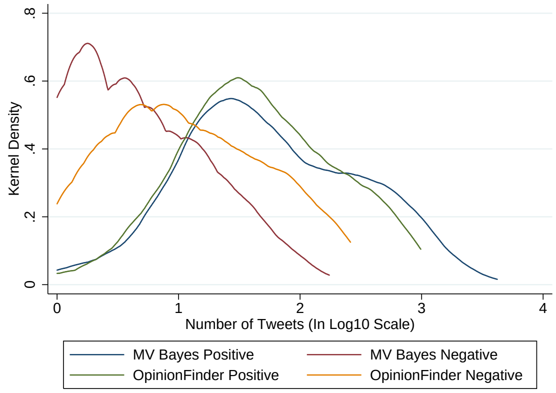

Figure 4: Kernel Density Plot of the Log of Number of Tweets Classified as Positive or Negative.

5.3 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 6) when the applicants from the year before, the size of the school, and the average SAT score are included in the model. Note that the sample size changes for the models including tweet sentiment because some schools had no tweets classified as negative, and so it was not possible to calculate a ratio of positive to negative tweets for those schools. In addition, some schools did not report SAT scores. 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).

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

Table 6: 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)		
Opinion Finder 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.

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 7. 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. However, as shown in Table 7 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.

6. 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 even when controlling for the number of applications in the previous year and the size of the school. 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 are results from the correlation of sentiments with college outcomes.

7. 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 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.

Table 7: Predicting Freshman Retention Rate.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
OF Location Ratio	-0.492*	-0.193						
	(0.188)	(0.139)						
OF Location Polarity			-29.572**	-10.308+				
			(8.317)	(5.646)				
MV Bayes Location Ratio					-0.328*	-0.127		
					(0.138)	(0.105)		
MV Bayes Location 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.