Taking the Pulse of US College Campuses with Location-Based Anonymous Mobile Apps

YANQIU WU, New York University Shanghai, Shanghai, China
Tehila Minkus, New York University, New York, USA
Keith W. Ross, New York University Shanghai, Shanghai, China; New York University, New York, USA

We deploy GPS hacking in conjunction with location-based mobile apps to passively survey users in targeted geographical regions. Specifically, we investigate surveying students at different college campuses with Yik Yak, an anonymous mobile app that is popular on US college campuses. In addition to being campus-centric, Yik Yak's anonymity allows students to express themselves candidly without self-censorship.

We collect nearly 1.6 million Yik Yak messages ("yaks") from a diverse set of 45 college campuses in the United States. We use natural language processing to determine the sentiment (positive, negative, or neutral) of all of the yaks. We employ supervised machine learning to predict the gender of the authors of the yaks, and then analyze how sentiment differs among the two genders on college campuses. We also use supervised machine learning to classify all the yaks into nine topics, and then investigate which topics are most popular throughout the US, and how topic popularity varies on the different campuses. The results in this paper provide significant insight into how campus culture and student's thinking varies among US colleges and universities.

1. INTRODUCTION

I find that the three major administrative problems on a campus are sex for the students, athletics for the alumni, and parking for the faculty.

Clark Kerr, first chancellor of UC Berkeley

What topics do college students discuss on campuses across the United States? What is the general sentiment — positive, negative, or neutral - on a college campus on any given day or throughout an academic year? Though researchers have traditionally explored students' views and sentiment through surveys and focus groups [Kiesa et al. 2007][Ng et al. 2010], these approaches are expensive and hard to scale. Additionally, in a non-anonymous focus-group format, students may be reluctant to disclose their views about polarizing or taboo issues.

In this paper, we explore how location-based anonymous apps can be leveraged to take the pulse of college campuses across the United States. Today, many of the most popular mobile apps — such as Waze, Yelp, Tinder — are location based, that is, they provide information relevant to the user's current geographical position. The user's current geographical position is determined by the smartphone using GPS (with its constellation of satellites), AGPS, cellular networks, and Wifi. A user's current geographical position, simply referred to as the GPS position, is then made available to mobile app software through the smartphone's API.

It is well known, however, that a smartphone's GPS location can easily be faked using GPS hacking. With GPS hacking, a user physically in Paris can set its smartphone GPS coordinates to a specific street corner in Brooklyn. In this manner, the user in Paris can use Yelp to browse reviews of restaurants in Brooklyn, or use Waze to see current traffic conditions in Brooklyn. In this paper, in lieu of expensive and labor-intensive surveys and focus groups, we explore how GPS hacking can be used in conjunction with the mobile app Yik Yak to take the pulse of college campuses.

This work was supported in part by the NSF grant CNS-1318659.
Author's emails: Yanqiu Wu, yanqiu.wu@nyu.edu; Tehila Minkus, tehila@nyu.edu; Keith W. Ross, keith-wross@nyu.edu.
© YYYY
2157-6904/YYYY/01-ARTA $15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000
Yik Yak was founded in late 2013 and has since gained significant traction on college campuses across the USA [Mahler 2015]. Yik Yak provides a simple message board to which users post short messages, called yaks, which are typically less than 200 characters. Yik Yak differs from traditional social networks in two respects. First, using the smartphone’s GPS location, posts on Yik Yak are shared only to nearby users. In particular, a Yik Yak user only sees messages in a rectangular region centered at the user’s current location, with the rectangular region typically covering a few square kilometers [Xue et al. 2016]. The region is large enough to cover most US college campuses. The second way Yik Yak differs from traditional social networks is that it is anonymous. Up until recently, all Yik Yak posts were fully anonymous. More recently, Yik Yak modified its service so that posts now carry pseudonymous usernames, which users may change as frequently as they like.

In this paper we use Yik Yak and GPS hacking to collect student posts from campuses across the US. We first develop an environment that carries out this data collection in a fully automated fashion. We then employ this environment to collect nearly 1.6 million yaks from 45 US colleges and universities. We also classify each of the 45 colleges into the following categories: Christian universities; liberal arts colleges; top-ranked universities; big-ten campuses; public universities; women’s colleges; men’s colleges; historically black colleges; and two-year college campuses. We then build a pipeline to categorize, compare, and analyze the data at each campus and in each category. We seek to answer the following questions:

— What is the general sentiment (positive or negative) on each of the 45 campuses. Does campus sentiment differ significantly among campuses or among campus categories? We use natural language processing to address these questions.
— Is it possible to predict the gender of the author of an anonymous yak with good precision? If so, how does the general sentiment compare for males and females on the 45 campuses? Is it possible to use gender prediction to partially de-anonymize Yik Yak posts?
— What are the most popular topics being discussed over Yik Yak? Do any topics dominate the discussions? What words are used within the topics? Are certain topics discussed more frequently for some campus categories? Are certain topics discussed more frequently in some US geographical regions?
— Are there correlations between certain topics? Are there correlations between certain topics and campus enrollments or campus admission rates? We examine these issues for the topics of dating and sex, academics, substance abuse, and politics and religion.

Yik Yak’s anonymity allows students to express themselves candidly without self-censorship. The methodology in this paper can provide significant insight into how student’s thinking and campus culture varies among US campuses.

This paper is organized as follows. In Section 2 we provide a brief overview of Yik Yak. In Section 3 we describe our data collection methodology. In Section 4 we provide a sentiment analysis of the collected data. In Section 5 we carry out a gender analysis of the collected data. In Section 6 we explore topic analysis. In Section 7 we summarize related work and in Section 8 we conclude.

2. ABOUT YIK YAK

In addition to sharing yaks, Yik Yak users can view, upvote, downvote, and reply to others’ yaks. By default, a Yik Yak user views the yaks within a region of its geographical location.

Yik Yak employs community-driven moderation based on downvotes. Specifically, a yak that receives five negative ratings from other users is removed from the stream.
of yaks. Additionally, Yik Yak blocks posts within a certain radius of high schools and grade schools, an attempt to curtail digital bullying on Yik Yak[Graber 2014].

Yik Yak is particularly popular among students on college and university campuses. Much media coverage has focused on Yik Yak’s potential as a platform for bullying and hate-speech[Mahler 2015], and a handful of universities have blocked access to the app on their networks[Rubelke 2015]. However, some brands view Yik Yak as an avenue to engage with the coveted millennial demographic, which comprises an estimated 98% of Yik Yak’s user base[Lela 2015]. In particular, the BBC recently engaged with Yik Yak users during the 2015 Canadian presidential elections. BBC also solicited contributions from Yik Yak users during a special week of mental health coverage. Journalists found Yik Yak effective for eliciting honest responses on sensitive topics[Bilton 2016].

<table>
<thead>
<tr>
<th>Christian universities</th>
<th>Big Ten campuses</th>
<th>Women's colleges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brigham Young U, 39,919,</td>
<td>Indiana University, 44,647</td>
<td>Bryn Mawr College, 14,916</td>
</tr>
<tr>
<td>Liberty University, 8,072</td>
<td>Michigan State University, 103,037</td>
<td>Scripps College, 11,003</td>
</tr>
<tr>
<td>Grand Canyon University, 41,637</td>
<td>Northwestern University, 38,313</td>
<td>Wellesley College, 18,938</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Liberal arts college campuses</th>
<th>Top-ranked universities</th>
<th>Public universities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowdoin College, 7,130</td>
<td>Columbia University, 15,096</td>
<td>Arizona State U, Tempe, 45,209</td>
</tr>
<tr>
<td>Middlebury College, 5,474</td>
<td>Harvard University, 21,526</td>
<td>Miami Dade College, 6,056</td>
</tr>
<tr>
<td>Swarthmore College, 9,311</td>
<td>Princeton University, 28,058</td>
<td>Rutgers U, New Brunswick, 34,578</td>
</tr>
<tr>
<td>Williams College, 8,071</td>
<td>Stanford University, 19,042</td>
<td>Texas A&amp;M, College Station, 34,181</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Men's colleges</th>
<th>Historically black colleges</th>
<th>Public universities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hampden-Sydney College, 4,691</td>
<td>Florida A&amp;M University, 68,881</td>
<td>University of Minnesota, Twin Cities, 55,490</td>
</tr>
<tr>
<td>Saint John's University, 1,533</td>
<td>Howard University, 19,500</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>Wabash College, 1,464</td>
<td>Jackson State University, 3,642</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top-ranked universities</th>
<th>Public universities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ohio State U, Main Campus, 102,167</td>
<td>University of Minnesota, Twin Cities, 55,490</td>
</tr>
<tr>
<td>Penn State, Main Campus, 65,454</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>Purdue University, 65,815</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>U of Illinois, Urbana-Champaign, 61,470</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>University of Iowa, 42,053</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>U of Michigan, Ann Arbor, 74,330</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>U of Minnesota, Twin Cities, 55,490</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
<tr>
<td>University of Nebraska, Lincoln, 23,237</td>
<td>University of Nebraska, Lincoln, 23,237</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Women's colleges</th>
<th>Two year college campuses</th>
<th>Public universities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bryn Mawr College, 14,916</td>
<td>Austin Community College, 63,778</td>
<td>University of Florida, 34,848</td>
</tr>
<tr>
<td>Scripps College, 11,003</td>
<td>East Los Angeles College, 3,698</td>
<td>University of Texas at Austin, 65,462</td>
</tr>
<tr>
<td>Wellesley College, 18,938</td>
<td>Ivy Tech Community College, 3,282</td>
<td>University of Texas at Austin, 65,462</td>
</tr>
</tbody>
</table>

Table I: Categories, campuses, and number of yaks collected at each campus

3. DATA COLLECTION

3.1. Yak Retrieval

One approach to collecting yaks is to take screen shots and use OCR character recognition to extract the yaks. Although this process can be largely automated with existing task automation tools for smartphones, it is slow and prone to errors. We instead used the Yik Yak API to retrieve the yaks at the 45 campuses.

We set up a Python bot and employed GPS hacking to retrieve yaks. For each of the 45 campuses, we sent a request for recent yaks once per hour, yielding the 100 most recent yaks from that location. Because most yaks are visible for more than an hour, and thus were retrieved multiple times, we had to remove all duplicates from the dataset.

The Yik Yak API we used was based on the YakGrabber library[Gupta 2015], which we modified to accommodate for batch inputs of several target locations. Our library emulated an older version of the Yik Yak application for Android, building custom URLs to retrieve data for specific parameters. (The updated version of the Yik Yak
application employs additional measures making it more difficult to intercept and re-
play its communications, such as pinned SSL certificates and cryptographically hashed
timestamps. While older versions of the application may no longer post yaks, Yik Yak’s
servers still allow read-only access for older versions.) We sent HTTP GET requests,
and the Yik Yak endpoint returned the yaks in JSON format. We took care to rate-limit
our requests by sending only one request per minute, thus avoiding overloading Yik
Yak’s servers.

3.2. Campus Categories
As our goal is to get compare student’s thinking and campus culture across the United
States, we selected 45 universities are diverse in character and geographic location.
Specifically, we consider nine categories of colleges and universities and universities:
Christian universities; liberal arts colleges; top-ranked universities; big-ten campuses;
public universities; women’s colleges; men’s colleges; historically black colleges; and
two-year college campuses. We choose these categories because they are diverse yet
collectively include most of the universities in the US.

From each of these categories, we selected several colleges while aiming for geo-
graphic diversity. In selecting colleges from these categories, we prioritized campuses
with high enrollment, based on the figures provided by the National Center for Edu-
cation Statistics report[1]. For top-ranked universities, we
drew universities from the U.S. News and World Report annual college ranking. Table
I lists the categories and campuses.

3.3. Dataset
We hired workers on Amazon Mechanical Turk to quickly and cheaply determine the
GPS coordinates of each university. We collected a corpus of 1,579,733 yaks from the
45 U.S university campuses from 21 January 2016 to 20 June 2016. Each college in I is
followed by the total number of yaks collected at that college. Of these 1,579,733 yaks,
a little over 100,000 had handles (pseudonyms). We also mined publicly available data
from CollegeData about each campus.

3.4. Legal and Ethical Considerations
We applied for IRB approval and were found to be exempt. We also limited our crawlers
to ensure that the Yik Yak servers would not be overloaded.

4. SENTIMENT ANALYSIS
In this section we analyze and compare the sentiment of the yaks across all 45 univer-
sities. If one accepts that the candid and anonymous posts on Yik Yak reflect student’s
thinking and feelings, the results here can help determine at which types of schools
students are, in aggregate, happier. For both sentiment and gender analysis, we re-
move yaks that have only one or two words for both training and validation, since
these yaks are generally meaningless and irrelevant.

4.1. Lexicon-based Classification
There are two broad approaches that can be taken for sentiment analysis: supervised
machine learning approaches and un-supervised lexicon-based approaches. For super-
vised sentiment analysis, we first need to label a subset of the yaks as positive, nega-
tive, or neutral. Then using this labeled data, we use supervised machine learning al-
gorithms to classify the sentiment of the remaining yaks. In the un-supervised lexicon-
based approach, we use emotional dictionaries which classify words and phrases in
terms of emotional content. An emotional score is then given to each yak based on the
words and phrases in the yak. As the supervised approach requires significant effort
to create the labeled data set, and also suffers from the subjectivity of the labelers, we use the un-supervised lexicon-based approach in this paper.

Many sentiment analysis algorithms are focused on supervised, machine learning approaches. In this paper, we use an unsupervised, lexicon-based approach to measure the emotional intensity contained in yaks across the 45 university campuses. Specifically, we use a lexicon-based approach based on the LIWC program, which is publicly available with an open API [Paltoglou and Thelwall 2012]. The LIWC program, built with the aid of linguists and psychologists, classifies tokens in terms of their emotional content. For a given text, the algorithm extracts the polarity and intensity of words belonging to the emotional dictionary. Words five words before and after the emotion terms are checked for negators (e.g., “not”), intensifiers (e.g., “very”) or diminishers (e.g., “little”). Negators will reverse the emotion weights of emotion terms but reduce the absolute values by 1 (e.g., -3 becomes 2 and 3 become -2). Intensifiers will increase/decrease the emotion weights of positive/negative terms by its intensity weight. Diminishers, intuitively, do the opposite of intensifiers. The classifier outputs each yak as positive, negative, or neutral.

4.2. Sentiment results

Since the algorithm does not recognize emoticons, we transfer emoticons in the yaks into actual words based on a Python library called emoji [Kim and Wurster 2015]. Then we perform the lexicon-based sentiment analysis over all the yaks in our Yik Yak dataset. Table II shows examples of 5 positive yaks and 5 negative yaks.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Sample Yak</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>Handles are fucking amazing!!!!</td>
</tr>
<tr>
<td>positive</td>
<td>Talking to a match on tinder a million time out of my league, and it’s going great. Pretty successful day!</td>
</tr>
<tr>
<td>positive</td>
<td>Sitting on the patio. Glass of wine. Listening to a young girl talk about her upcoming wedding. Congrats!</td>
</tr>
<tr>
<td>positive</td>
<td>Congrats to all of the graduates today!!</td>
</tr>
<tr>
<td>positive</td>
<td>Good luck on your comps seniors!</td>
</tr>
<tr>
<td>negative</td>
<td>The fact that you can’t retake tests when you’re out sick is incredibly shitty</td>
</tr>
<tr>
<td>negative</td>
<td>Insomnia, heartbreak, anxiety....what do I do?</td>
</tr>
<tr>
<td>negative</td>
<td>The rate on dollars are so shitty rn I’m crying! Fuck America exchange</td>
</tr>
<tr>
<td>negative</td>
<td>I’ll be so sad if Hillary wins in PA. pensive face</td>
</tr>
<tr>
<td>negative</td>
<td>It’s so fucking annoying spending every weekend stressed out about doing my homework</td>
</tr>
</tbody>
</table>

Table II: Examples of five yaks classified as having positive sentiment, and five yaks classified as having negative sentiment.

For each college, we define two sentiment metrics: *sentiment ratio*, which is the ratio of the number of negatively classified yaks to the number of positively classified yaks; and *emotion level*, which is the sum of the positively classified yaks and the negatively classified yaks, divided by the total number of yaks for that college. Figure 1 shows a scatter plot of the 45 universities with the sentiment ratio on the x-axis and the emotion level on the y-axis. From Figure 1 we see that all 45 schools have more negative posts than positive posts. We further see that the sentiment ratio varies greatly from campus to another, ranging from 1.22 to 1.66. The emotion level also varies significantly from school to school, ranging from 53% to 69%. However, we see from figure
4.2 Sentiment results

Fig. 1: Scatter plot of 45 universities with respect to sentiment and emotion

that most colleges have emotion level of 60% to 70%, thus for most colleges, 30% to 40% of the yaks are neutral. Figure 1 also shows the MSE regression line. We see that schools with higher sentiment ratios (i.e., more negative) tend to have a higher emotion level, although the correlation is not statistically significant.

Figure 2 (a) displays sentiment ratios by school category. For a given category there is one horizontal bar, with the left-side of the bar being the lowest sentiment ratio among all the schools in the category, and the right-side of the bar being the highest sentiment ratio among all the schools in the category. Each bar also has a vertical line indicating the average sentiment ratio in the category. Figure 2 (b) provides a similar visualization for emotion levels.

Surprisingly (at least to the authors), Figure 2 (a) clearly shows that the Christian universities are the most positive among the nine university categories. Moreover, all three Christian universities have almost the same sentiment ratio. The majority of the students at these universities are devout Christians. We can conjecture that these students have a relatively positive outlook on life, although we offer no means to validate this conjecture. One might conjecture that large universities would generally be relatively negative; however, the two Christian schools Brigham Young University and Grand Canyon University both have large student bodies, with about 30,000 and 20,000 on-campus students, respectively. Thus there does not seem to be a correlation between the size of a school and the sentiment of the students.

After the Christian schools, the liberal arts colleges are the most positive. In fact, all four of the liberal arts colleges are more positive than all of the schools in big ten, top-ranked, historically black, women’s college, and men’s college categories. The liberal arts colleges considered here are all small private schools with high tuition and high faculty-to-student and staff-to-student ratios. On the other hand, the big-ten schools have relatively low in-state tuition, and generally have larger classes and lower staff-to-student ratios. Again, without any ground truth to validate, we can conjecture that higher faculty and staff to student ratios lead to a happier student body in aggregate. Following the Christian and the liberal-arts schools, the top-ranked universities and women’s colleges are the most positive.

On average the two-year colleges are the most negative. Again without any attempt to validate the claim, this may be because students at two-year universities often come from less privileged economic backgrounds as compared to their four-year college counterparts. Alternatively, this may also be because the two-year colleges are commuter schools, whereas the four-year colleges considered in this study are primarily residential schools.
The emotion-level results, shown in Figure 2 (b), are not quite as striking. There is generally a lot of variety within each category. The men's colleges on average are the least emotional, and two-year colleges on average are the most emotional.

5. GENDER PREDICTION AND ANALYSIS

It is of interest to see if we can accurately predict the gender of the author of a yak for two reasons. First, if we can accurately classify the authors as male or female, then we can perform a gender-based analysis of sentiment and topics. Second, accurate gender classification would be one step further towards de-anonymizing the yaks. Indeed, in an earlier paper, it was shown that it is possible to determine the location from which a yak was posted to the granularity at a dorm building [Xue et al. 2016]. By combining known location with known gender, one can further narrow down the potential authors of a yak. De-anonymization puts into question the main feature of Yik Yak and other related anonymous mobile apps.

5.1. Labeling the gender of yaks

In order to classify the yaks as male or female using supervised learning, a subset of the yaks need to be labeled. Recall that a fraction of the yaks have handles (pseudonyms). Many of these handles reveal gender information, for example, handles like “Princessshh”, “missmango” and “crystalbeth” have a very high probability of being created by female users. Similarly, “SammyDavisJr”, “Blue Eyes White Guy” and “kingfkush” are very likely created by male users. We hired workers from Amazon Mechanical Turk (AMT) to quickly and cheaply classify handles as either female, male or not sure. Each yak was classified by 5 different workers on AMT. We declare its final gender label as not sure unless three or more workers vote the same gender and no contradictions occur with the remaining workers. For example, a handle would be labeled as male if and only if the 5 votes are MMMNN, MMMMN or MMMMM, with M representing male and N representing not sure. MMMMF would be labeled as not sure.
5.2 Preprocessing for prediction

We use two kinds of features, bag-of-words features and stylometric features, for gender prediction. Details about the stylometric features can be found in Table III. To ensure the quality and reliability of our feature choices, the following data pre-processing steps are applied:

— We group phrases and words with their corresponding abbreviations. For example, we group “bf” with “boyfriend”, “hw” with “homework”, “to be honest” with “tbh”, and “oh my god” with “omg”.

— We group contractions such as “will not” with “won’t”, and “does not” with “doesn’t”. We treat each grouping as one word.

— We classify punctuation into 3 classes: “?”, “!” and “.”. The “.” class covers everything else, including comma, period and punctuation strings.

— We stem every word.

After performing the above steps, we have 3,754 yaks labeled as female and 13,458 yaks labeled as male. To deal with this imbalanced data, we re-sample the minority class by duplicating entries. To avoid over-fitting, we over-sample after cross-validation, i.e., after we leave the validation set out of the training loop[Japkowicz 2000]. In this way, we can make sure that no duplicated entries will be both trained and validated. After testing a number of algorithms, we use Support Vector Machines (SVMs) and report averaged 10-fold cross-validation test results.

<table>
<thead>
<tr>
<th>Gender Classification</th>
<th>Approximate Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>59.1%</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>61.1%</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>64.9%</td>
<td></td>
</tr>
<tr>
<td>15%</td>
<td>68.2%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>74.4%</td>
<td></td>
</tr>
</tbody>
</table>

Table IV: Gender classification results with bag-of-words features

5.3 Gender prediction results

Table IV summarizes gender classification results when using only bag-of-words features. We control the value of recall by not predicting data points that are within a
5.4 Gender sentiment

Using only the yaks for which the author gender was manually labeled by AMT workers, we used the lexicon-based sentiment analysis tool to analyze the sentiment of the two genders. Female students and male students seem to have very similar sentiment ratios and emotion levels, as shown in figure 3. However, we must emphasize that this observation is being drawn from subset of yaks, namely, those which have revealed their genders through their handles. Interestingly, students who reveal their gender are on average more positive than students who do not reveal their gender.

6. TOPIC ANALYSIS

For the topic analysis, we decided to focus on a one-month period (January 21 to February 21, 2016) during which the presidential primaries were receiving intense media

---

Table V: Ten most weighted feature words

<table>
<thead>
<tr>
<th>Feature Words</th>
<th>Male Class</th>
<th>Female Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>girlfriend</td>
<td>boyfriend</td>
<td></td>
</tr>
<tr>
<td>man</td>
<td>tomorrow</td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>someone</td>
<td></td>
</tr>
<tr>
<td>fuck</td>
<td>cat</td>
<td></td>
</tr>
<tr>
<td>gal/girl</td>
<td>bed</td>
<td></td>
</tr>
<tr>
<td>yik</td>
<td>drink</td>
<td></td>
</tr>
<tr>
<td>shit</td>
<td>thing</td>
<td></td>
</tr>
<tr>
<td>know</td>
<td>🎉</td>
<td></td>
</tr>
<tr>
<td>point</td>
<td>stop</td>
<td></td>
</tr>
</tbody>
</table>

Table V: Ten most weighted feature words

certain distance to the optimal hyper plane. We see that we have nearly 59% precision. If we consider only the 15% of yaks that are furthest from the SVM hyperplane, then we can have nearly 71% precision. Thus, using word-usage alone, we can make reasonably-accurate predictions about whether the author of a yak is male or female.

Each gender class has its own set of most highly weighted feature words. Table V lists the top 10 feature words for each class. It is interesting that the most distinguishing predictor of whether a yak is authored by a male is the presence of the word “girlfriend” whereas symmetrically the most distinguishing predictor for a female is the presence of the word “boyfriend”. We also note that male yaks are also distinguished by the use of profanity, and female yaks by the use of emoticons.

We also apply SVM using stylometric features. Furthermore, we combine the normalized real-valued stylometric features with binary bag-of-words features. We normalize the stylometric feature values by dividing each value by the 2-norm of its stylometric feature array. In this way, every stylometric feature value is bounded by 0 and 1.

Table VI summarizes the classification results. In this case, we only compare precision for 100% recall. We can easily see that stylometric features work poorly. Women and men in aggregate appear to use similar styles for short messages like yaks.

Table VI: Gender classification results with different features (100% recall)

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-words features</td>
<td>59.1%</td>
</tr>
<tr>
<td>Stylometric</td>
<td>51.7%</td>
</tr>
<tr>
<td>Bag-of-words + stylometric</td>
<td>57.0%</td>
</tr>
</tbody>
</table>

5.4 Gender sentiment

Using only the yaks for which the author gender was manually labeled by AMT workers, we used the lexicon-based sentiment analysis tool to analyze the sentiment of the two genders. Female students and male students seem to have very similar sentiment ratios and emotion levels, as shown in figure 3. However, we must emphasize that this observation is being drawn from subset of yaks, namely, those which have revealed their genders through their handles. Interestingly, students who reveal their gender are on average more positive than students who do not reveal their gender.
This gives a sub-corpus from the dataset of nearly 500,000 yaks from the 45 schools. We manually examined 1000 yaks to develop a codebook of eight topics. Table VII lists the eight topics.

We randomly sampled 10,064 yaks from our dataset (roughly 2% overall). We then posted a task on Amazon Mechanical Turk, asking the workers to classify each yak into one of the eight topics shown in Table VII. We included a ninth topic for “I don’t know,” since some yaks are hard to understand. If at least two out of three workers chose the same topic for a single yak, we assigned the yak to that topic. There was consensus for 85% of the yaks. The frequency for each of the topics is shown in Table VII.

<table>
<thead>
<tr>
<th>Topic</th>
<th>% of sample</th>
<th>Sample Yak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dating and sex</td>
<td>23</td>
<td>Any girls looking for a late night make out session?</td>
</tr>
<tr>
<td>Local life, weather, announcements</td>
<td>15</td>
<td>Are there any bus routes to Reed Arena?</td>
</tr>
<tr>
<td>Culture, tech, and sports</td>
<td>12</td>
<td>Thomas Bryant is the only thing ESPN can talk about apparently</td>
</tr>
<tr>
<td>Academics and careers</td>
<td>11</td>
<td>Does anyone wanna work on the Stats project for Samantha Russels class together? Need some help</td>
</tr>
<tr>
<td>Health, drugs, alcohol</td>
<td>7</td>
<td>With the 2 beers I burned off, I can have another 2 beers!</td>
</tr>
<tr>
<td>Politics and religion</td>
<td>5</td>
<td>It’s just like the high schoolers on the news said. Donald Trump: make America hate again</td>
</tr>
<tr>
<td>Friends</td>
<td>4</td>
<td>I miss my college friends so much. Don’t take your friends for granted.</td>
</tr>
<tr>
<td>Diversity/discrimination</td>
<td>1</td>
<td>Some of these international students act like they’ve never lived in a civilized society before</td>
</tr>
<tr>
<td>I don’t know</td>
<td>6</td>
<td><em>trombone</em> <em>oven door slamming</em></td>
</tr>
</tbody>
</table>

Table VII: Topics and their occurrence within the sample set of 10,064 yaks. For 15% of the dataset, there was no consensus on a single category.

We can draw some preliminary conclusions based on this sample of 10,064 yaks. Notably, dating and sex comprise a large portion of conversations on Yik Yak. Popular culture and academics are widely discussed, as are local inquiries and announcements. There is a fair bit of discussion about health and substance issues, and political yaks have a consistent presence as well. However, counter to claims by many media outlets and university administrators, discriminatory and racist yaks are not very prevalent in our sample. This would seem to bolster claims by Yik Yak that their moderation and downvoting policies are effective.

Figure 4 shows the frequent terms from some of the topics. These word clouds help to understand the content in the yaks in these topics.
6.1 Scaling Up the Analysis

In order to learn more broadly about students’ yaks across the United States, we employ machine-learned classifiers to label all the yaks in the one-month dataset. We used the ground truth topics established above to train machine-learning classifiers for the most interesting topics: dating and sex; academics; politics and religion; and alcohol and substances. (We were unable to train a classifier for racist/bullying yaks since they were so rare in our ground-truth training set.)

After testing a number of algorithms, we used Random Forest and Decision Tree algorithms for the different categories. To determine each classifier’s performance, we ran the selected algorithm with 10-fold cross-validation. Average precision ranged from 86% to 92%, and average recall ranged from 87% to 97%. For features, we removed stopwords and generated word-count vectors and TF-IDF vectors. We did not use n-grams since they degraded precision and recall on our test sets.

6.2 Analyze Yaks at Scale

After developing a pipeline to categorize yaks, we analyze the dataset to answer a few important questions. How do campuses differ in yakking behaviors? Do certain campus characteristics (e.g. enrollment, selective admissions rates, demographics, geographic region) correlate to topic popularity?

In this section, we analyze yaks labeled by our machine-learned classifiers. We note that all ANOVA analyses mentioned in this section are significant at the $p \leq 0.05$ level, and are followed by post-hoc Tukey HSD tests. All correlations mentioned in this section are significant at the $p \leq 0.05$ level after a Holm-Sidak correction for multiple comparisons, unless otherwise specified. In figure 6, the $r$-value is the Pearson’s correlation coefficient, the blue line is the regression line, and the $p$-value is the 2-tailed $p$-value. The Pearson correlation coefficient measures the linear relationship between two datasets.

6.3 Dating and sex

The most popular topic in our dataset is dating and sex. How does its incidence differ across colleges, regions, and campus types?

We see from Figure that men’s colleges yak the least frequently about dating and sex. Women’s colleges yak more frequently but still relatively less than most of the campus categories. One possible interpretation is that single-gender campuses yak about dating and sex at lower rates because dating prospects are greatly limited for heterosexual students at such campuses. Historically black colleges, two-year colleges, public universities, and Christian universities tend to yak the most about dating and sex. The campus with the highest rate of yakking about dating and sex, Jackson State
6.3 Dating and sex

TOPIC ANALYSIS

(a) By institution type

(b) By region

Fig. 5: Topic Analysis: By Campus Category and by Region
6 TOPIC ANALYSIS

6.4 Academics

(a) Dating and sex vs. campus enrollment

(b) Alcohol vs. campus admission rate

(c) Dating and sex yaks vs. academic yaks

Fig. 6: The popularity of topics when compared with campus characteristics.

University, had 39.1% of its total yaks labeled with this topic. In contrast, the campus that yakked least about dating and sex (Hampden-Sydney College) had only 6.6% of its total yaks labeled with this topic.

Regional differences were also evident, as backed up by an ANOVA analysis. Campuses located in the west had a statistically significant higher rate of mentioning dating and sex (see Figure 5), with 27.1% of their yaks labeled with this topic. Meanwhile, only 18.6% of yaks from northeastern campuses were related to dating and sex. The south and midwest fell in the middle, with respective dating-and-sex yak rates of 24.7% and 21.5%.

Another trend we observe is, in general, students yak more about dating and sex at campuses with higher enrollment. Figure 6a demonstrates this relationship (though it is not statistically significant). It is possible that students who plan to engage in a vibrant dating scene choose to attend larger campuses, which would explain the difference.

6.4. Academics

All campus types (e.g. public, christian, two-year, etc.) have academic-yak rates between 5% and 10%. Interestingly, the two school categories with the most positive sentiment ratio – Christian schools and liberal arts colleges – are also the two categories with the lowest yaking rates about academics. One can conjecture that the schools at which students are the least stressed about academics (for whatever reason) are also the schools where the students are the happiest. All regions (northeast, south, midwest, and west) yakked about academics at rates between 7% and 9%, with no statistically significant differences between regions. Interestingly, we also find that campuses with higher rates of yacking about dating and sex had lower rates of academic-related yaks at a statistically significant level (see Figure 6c). Presumably, students who are more involved in the dating scene maintain lower levels of academic investment.

We also examined relationships between academic yak rates and university characteristics, such as average SAT scores, acceptance rates, and demographic composition of student bodies. However, we found no evidence for statistically-significant correlations.

6.5. Substance use

The widespread use and abuse of alcohol and drugs on U.S. college campuses is an important issue. Is this behavior reflected in yaks? And which campuses talk about alcohol and/or drugs most? In order to measure this more carefully, we submitted each yak that had been automatically labeled as substance-related to workers on Mechan-
6.6 Politics and Religion

Recall that for topic analysis we are considering the yaks for a one-month period (January 21 to February 21, 2016) during which the presidential primaries were receiving intense media coverage. As we shown in Table VII, politics and religion are fairly popular on Yik Yak, with approximately 5% of all yaks in our sample set falling into the politics/religion topic. Which campuses yak the most about politics during the time spanned by our dataset? The campus with the most political chatter was University of Iowa, which is located in the state with the first caucus or primary. In general, the campuses with higher rates of political chatter are located in states with earlier primaries. An ANOVA analysis of regions with respect to politics-and-religion discussion rates reveal that midwestern campuses were most likely to discuss politics and religion. It is likely that intense campaigning in early-primary midwestern states leads to higher rates of yakking about politics.

Which political candidates garnered the most discussion on Yik Yak? By searching for candidates using their names as keywords, we found that Bernie Sanders and Donald Trump each were mentioned in approximately 1% of the yaks; Hillary Clinton was mentioned in 0.5% of yaks; and the rest of the primary candidates were each mentioned in less than 0.2% of yaks.

Interestingly, the proportion of yaks about each candidate on campuses did not correlate with state primary results. This was because virtually all campuses’ discussions about candidates were dominated by chatter about Bernie Sanders and Donald Trump. Even in states where another candidate won the majority of the popular vote and/or delegate count, Trump and Sanders were still discussed at higher rates than the actual winner.

7. RELATED WORK

7.1 Sentiment Analysis

Sentiment analysis has previously been applied to different types of social media such as blogs, Twitter, and Facebook. Pang et al.[Pang et al. 2002] performed sentiment analysis on reviews based on machine learning approaches. Agarwal et al.[Agarwal et al. 2011] introduced POS-specific prior polarity features and tree representation features for sentiment analysis on tweets based on Dictionary of Affect in Language[Whissel 1989] and extended it using WordNet. However, their supervised algorithm requires labor-intensive manual labeling of the Twitter data.
Paltoglou et al. [Paltoglou and Thelwall 2012] implemented an unsupervised sentiment classifier based on the LIWC software [Pennebaker et al. 2001]. In their experiments on the Digg and MySpace dataset, the lexicon-based classifier outperforms other supervised approaches (SVMs and Naive Bayes). However, their sentiment analysis did not target US college campuses. To our knowledge, this is the only study that uses social media to investigate the sentiment on different college campuses.

### 7.2. Gender Prediction

Predicting genders of blog authors and social media users from text has been studied by several researchers. The problem is very important in the field of author profiling and de-anonymization. For long texts, such as fiction books and blogs, 76%-80% accuracy has been achieved [Koppel et al. 2003] [Argamon et al. 2009]. Because the texts are much longer than a yak, they were able to obtain relatively high accuracy rates.

Gender prediction has also been performed for Twitter, which is similar to Yik Yak in post length. Rao et al. [Rao et al. 2010] input Ngram-features and sociolinguistic features, such as emoticon usage and character repetitions, into SVMs for gender classification. They reported an accuracy of 71.8% using sociolinguistic features, an accuracy of 68.7% when using Ngram features and an accuracy of 72.3% when combining Ngram-features with sociolinguistic features. However, they did not make their predictions based on a single tweet, but instead on the aggregation of tweets from the same author handle. Our gender prediction study differs in that we study the more challenging problem of predicting gender based on a single short yak, which typically ranges from 50 to 200 characters in length.

### 7.3. Topic Analysis in Yik Yak

Though Yik Yak is relatively new, it has gained some attention from the research community. McKenzie et al. [McKenzie et al. 2015] compared Yik Yak to Twitter. They found that Yik Yak topics are different and more localized than Twitter topics. However, they only explored one square mile in Los Angeles, so their conclusions are not generalizable and do not explore differences among campuses.

Northcut [Northcut 2015] manually examined screenshots of 319 yaks and categorized them into four intentions: shock, joke, inquire, and emote. Heston et al. [Heston and Birnholtz] collected yaks from 35 universities for multiple months, and categorized the intent of the posts. By hand-coding 1800 yaks, they found that yaks were posted in eight intent categories: personal admission, observation, information/advice, opinion, venting/complaining, invitation, favor, and joke. This sets a framework for the intentions of Yik Yak users, but it does not elaborate on the popular topics discussed. Additionally, this approach does not scale to a larger dataset.

Black et al. [Black et al. 2016] manually categorized 4000 yik yak posts from 42 campuses over 3 days. They found that campus life, announcements, and sex were popular topics. They also found many profanities and rhetorical questions. However, the dataset was quite limited temporally, and their analysis is not scalable.

Koratana et al. [Koratana et al. 2016] examined public health evidence on Yik Yak. Using Latent Dirichlet Allocation (LDA), they modeled topics to find health-related yaks. While their work relates to our research, they do not compare the prevalence of such behaviors across different campuses; they also do not examine any non-health topics. Additionally, we found that automatic clustering methods did not generalize to other yak topics, instead producing noisy and imprecise categorizations.

Using surveys, Ma et al. [Ma et al. 2016] found that people are more likely to share information that is intimate and/or negative when they are assured anonymity in online social networks. See also Peddinti et al. [Peddinti et al. 2014] and Correia et al. [Correa et al. 2015] for a similar observation. In contrast with tweets, Whisper posts were gen-
erally more sensitive and covered a different set of topics. In another analysis of Whisper, Wang et al. [Wang et al. 2014] found that anonymous online social communities show high dispersion and low clustering, with little evidence of long-term relationships between users. Xue et al. [Xue et al. 2016] studied location privacy in Yik Yak, showing how an attacker can determine the location from where a yak was posted; such an approach could potentially allow an attacker to correlate yaks to specific users.

Our topic analysis approach has two main advantages over the previously published literature. Firstly, other approaches have primarily used labor-intensive manual labeling, which limited the extent of their inferences. Since our approach uses supervised machine-learning, it can be automatically scaled to larger datasets. Secondly, we are not aware of any other research that compared yakking topics across different settings. These two contributions (i.e. scalability and context-based comparison) are a significant and novel contribution to the body of research on anonymous location-based social networks.

8. CONCLUSION

In this paper we investigated a potentially powerful new methodology for making passive regional surveys, namely, combining location-based mobile apps with GPS hacking. Although in this paper we focused our study on using Yik Yak to survey student sentiment and interest at US college campuses, the general methodology can be employed with other location-based mobile apps. For example, ethic-group centric apps (such as WeChat for Chinese users) can potentially be used to survey diaspora and immigration trends; traffic apps such as Waze can potentially be used to estimate urban development.

We combined Yik Yak with GPS hacking to develop a novel platform for surveying college student interests and sentiment. Using this platform, we collected nearly 1.6 million yaks from a diverse set of 45 college campuses in the United States. We employed an NLP tool to determine the sentiment (positive, negative, or neutral) of all of the yaks. We also used supervised machine learning to classify all the yaks into nine topics, and investigate which topics are most popular throughout the US, and how topic popularity varies on the different campuses. We have provided examples of yaks that have high positive and high negative sentiment values, examples of the most distinguishing features for gender prediction, and examples of yaks from each of the topics. Some of more interesting findings are:

— At all 45 universities, there are more negative yaks than positive yaks.
— Sentiment ratios vary significantly from campus to campus. However, for many college categories — including Christian schools, liberal arts colleges, big-ten schools, and top-ranked schools — Yik Yak sentiment levels are remarkably similar for the colleges within the category.
— Student sentiment is most positive on Yik Yak at Christian universities, followed by liberal arts colleges. Yik Yak sentiment is least positive in two-year colleges.
— Based on the word content of a single anonymous yak, it is possible to predict to gender with a reasonable amount accuracy; some yaks may be predicted with a higher level of confidence. Stylometry, either alone or combined with the bag-of-words approach, does not improve gender prediction precision when using a single yak.
— Aggregated across all the universities, male sentiment ratio and emotion level are about the same as female sentiment ratio and emotion level.
— Dating and sex comprise a large portion of the conversations on Yik Yak. Popular culture and academics are also widely discussed.
— Campuses with higher rates of yakking about dating and sex have lower rates of academically-related yaks.
— Schools with lower admission rates and higher SAT scores have lower rates of alcohol and drug-related yaks. Southern schools yak more about alcohol than those from other regions.

Yik Yak’s anonymity allows students to express themselves candidly without self-censorship. The methodology in this paper can provide significant insight into how student thinking and campus culture varies among US campuses.

In future work, it would be of interest to do a similar study using Twitter and/or Instagram and compare the results with Yik Yak. Twitter and Instagram are not anonymous, so it would be of interest to see if people are being more open and candid in Yik Yak than in other social networks, where users may employ self-censorship. To carry out a similar analysis with Twitter and Instagram, for each university under investigation, one would first have to determine users who are current students at the university, which may lead to some complications.

Acknowledgements
This work was supported in part by the NSF grant CNS-1318659. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any of the sponsors.

REFERENCES


Ricardo Bilton, February 18, 2016. How BBC is using Yik Yak to talk to millennials (and get them to talk back). Digiday (February 18, 2016).


Matthew Heston and Jeremy Birnholtz. (In)visible Cities: An Exploration of Social Identity, Anonymity and Location-Based Filtering on Yik Yak. In iConference. iSchools.


Taehoon Kim and Kevin Wurster. 2015. Emoji for Python. Available at https://github.com/carpedm20/emoji/


