Of Oxen and Birds: Is Yik Yak a useful new data source in the geosocial zoo or just another Twitter?

Grant McKenzie
STKO Lab, University of California, Santa Barbara
grant.mckenzie@geog.ucsb.edu

Benjamin Adams
Centre for eResearch, The University of Auckland
b.adams@auckland.ac.nz

Krzysztof Janowicz
STKO Lab, University of California, Santa Barbara
janowicz@ucsb.edu

ABSTRACT

The landscape of social media applications is littered with novel approaches to using location information. The latest platform to emerge in this geosocial media realm is Yik Yak, an application that allows users to share geo-tagged, (currently) text-based, and most importantly, anonymous content. The fast adoption of this platform by college students as well as the recent availability of data offers a unique research opportunity. This work takes a first step in exploring this novel type of data through a range of textual, topical, and spatial data exploration methods. We are particularly interested in the question of whether Yik Yak differs from other geosocial data sources such as Twitter. Is it just another location-based social network or does it differ from existing social networks, establishing itself as a valuable resource for feature extraction?

Keywords

yik yak, twitter, anonymous, geosocial, social media

1. INTRODUCTION

In 2013, the social media application *Yik Yak* was launched out of Greenville, South Carolina. In the years since its release, it has quickly established itself as a popular, and controversial [10], location-based social media application produced for mobile platforms. Yik Yak differentiates itself from other location-based or geosocial media applications in a number of ways. First, posts (referred to as *yaks*) are published anonymously in that all identifiable information on the person posting, other than a rough location, is hidden from users of the application. Second, yaks are tied to the posting user's location[1] and are only accessible by users within a 1.5 mile radius of the yak. In combination, these two features offer a unique experience for users, and, in turn, influence the way users interact with each other. In this work we examine data produced via Yik Yak in the Los Angeles area and compare with data from the same geographic area published through the more established (and much more highly researched) social platform of Twitter.

In comparison to geotagged content from Twitter, we hypothesize that Yik Yak posts differs in two key respects. First, because Yik Yak posts are only locally visible to a cohort of other users within the same geographic area, we expect the thematic contents of posts to be more localized than Twitter *tweets*. This difference in audience means that we expect to see themes that are only relevant for few locations, e.g., about mid-term exams at university. In contrast, Twitter users (even georeferenced ones) are communicating with users that span a much wider geographic area, and, thus, we expect that themes will be less location-specific on average. The second way in which we hypothesize Yik Yak content to differ is due to the fact that it is an anonymous medium, and, thereby, users are likely to discuss many themes that they otherwise would not (or at least not by using the same terms).

Both of these assumptions are important as they determine for what and how Yik Yak data can be used. Here, we are specifically interested in whether Yik Yak can be used to extract different features[2] that would not be available from other data sources. To give a concrete example, while yaks cannot be precisely geo-located, the fact that they are only locally visible and only users within the range of a yak can comment on it, may enable the creation of place-centric features.

The research contributions of this work are as follows.

RQ1 Do the themes discussed on Yik Yak differ from those on Twitter and if so, are these differences merely caused by a few selected terms (e.g., place names) or are they prevalent? To answer this research question, we will learn an LDA [2] model for yaks and tweets and compare the similarity of the resulting topics using multidimensional scaling. Intuitively, if yaks and tweets are similar with respect to the discussed themes and audience, or if they only differ by a few prominent place names, say UCLA, the learned LDA topics should be similar (and thus close in an MDS scatterplot).

RQ2 Are the themes discussed in yaks more local in na-

[1] Geographic coordinates (latitude and longitude) are provided via the user's mobile device and are stored with a precision of two decimal places.

[2] We use the term *feature* here in its machine learning interpretation, e.g., for the construction of models, and will use the term *place* for geographic features such as universities.

SIGSPATIAL'15, November 03–06, 2015, Bellevue, WA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3975-9/15/11...$15.00
DOI: http://dx.doi.org/10.1145/2830657.2830659

ture, i.e., are they more specific to certain places or regions? To answer this question, we will perform marked point pattern analysis for all aggregated yaks and tweets for each LDA topic and then compare how many of these topics show a strong spatial dependency. Intuitively, if one would plot all yaks about a specific LDA topic on a map, these plots would show a clustering pattern more often than for data from Twitter.

RQ3 Given that there is no 140 character limit for yaks, limited hastag usage, and the audience is predominantly college students, is there a difference between the readability of yaks versus tweets? To address this questions, we will compare *reading ease* and *grade level.* Intuitively, yaks should have a higher grade level on average.

An in-depth analysis and comparison with other location-based social networks is left for future work.

2. DATA

Approximately 800,000 Yik Yak posts, or yaks, were collected from the greater Los Angeles area over a five month period starting in December of 2014. A grid with cells measuring 1 square mile was placed over the greater Los Angeles area (as defined by the 2013 US Census Urban Areas dataset) and the unstructured text content of all yaks were spatially aggregated into regions based on these grid cells. Given the 1.5 mile radius of influence and coordinate rounding (to two decimal places) of the yak location, if a unique yak identifier was discovered in more than one grid cell, the coordinates were averaged and the yak was assigned to the grid cell that contained the averaged location.

After collection, the data were cleaned to remove duplicate and empty yaks as well as those which only contained emoticons[3] or other special characters. After cleaning, 664,839 yaks were used for exploration and analysis. The average length of a yak in this dataset is 66.6 characters long with a standard deviation of 42.2 characters.

By way of comparison, a set of tweets were accessed for the same region during the same time frame. These tweets were aggregated to the same spatial grid of 4559 cells. After cleaning, removing replies and URLs (and hashtags for topic modeling), 684,793 tweets remained and were used for analysis in this research.

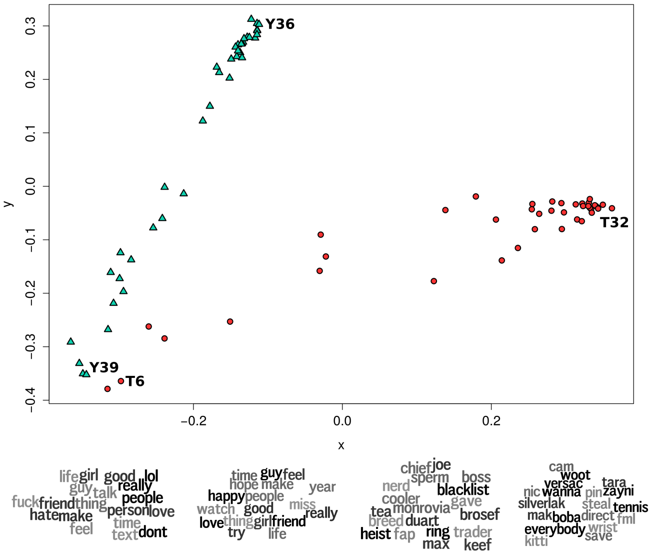
3. THEMES AND KEYWORDS

This section explores the themes that are found in both the Yik Yak and Twitter datasets. A topic modeling approach was applied to extract topics from both of the datasets independently. A *latent Dirichlet allocation (LDA)* topic model [2] was used with a number of 40 topics specified. LDA takes a bag-of-words approach to extracting topics from text by examining the co-occurrence of words in a document (aggregate of content within a grid cell j in our case) and describes a document as a probabilistic distribution, θ_j, across topics. Running two separate models for Yaks and Tweets produced two sets of 40 topics allowing us to describe each grid cell, j in the dataset as 1) a distribution of *Yak* topics (θ_j^Y) and 2) a distribution of *Tweet* topics (θ_j^T).

Multidimensional Scaling

Through the LDA process, each word in the corpus (set of tweets or yaks) is given a probability value for its occurrence in each topic. A exhaustive set of words was constructed by combining all words from both the Yik Yak (110,770) and Twitter (82,089) datasets. A total of 30,905 common words were found between both datasets, resulting in a total set of 161,954 words. An overall word-probability distribution for each topic in each of the two datasets was then constructed by taking the probability of the word in the given topic (in many cases this was 0 given that many Yak terms did not appear in the Tweet dataset and vice versa) and normalized so that the sum of each topic distribution equaled to 1.

Every topic j in each of the datasets was then compared to each other topic using *Jensen-Shannon distance (JSD)* [7, 8] to measure dissimilarity between topics. A distance matrix of these JSD values was produced as input to a metric multidimensional scaling. Among other features, multidimensional scaling (MDS) visually depicts the similarity between items in a dataset (topics in tweets and yaks) by reducing their dimensionality (2, in our case) while trying to preserve inter-object distance. Figure 1 shows a two-dimensional MDS scatterplot of Yik Yak and Twitter topics.



(a) Y39 (b) T6 (c) Y36 (d) T32

Figure 1: Multidimensional scaling applied to Jensen-Shannon distance values calculated from combined Twitter–Yik Yak weighted word vectors.

The results show a clear grouping of dataset-specific topics extracted from the *Twitter* data (red circles) and those pulled from the *Yik Yak* data (green triangles). Only a few topics are similar across these social networks. For example, *Yik Yak topic 39* is relatively similar to *Twitter topic 6.* In looking at the most prevalent terms within these topics (Figures 1a and 1b) one notices some common words. In contrast, *Yik Yak topic 36* and *Twitter topic 32* are highly dissimilar. This is supported through examination of the top words associated with the topics (Figures 1c and 1d). This is an interesting finding and important for the exploration of differences in the spatial distribution of patterns described

[3] A metacommunicative pictorial representation of a facial expression.

below. It means that Yik Yak topics differ from Twitter topics not just by a few prominent terms (particularly toponyms such as UCLA) but by the themes being discussed. This addresses the questions asked in **RQ1**. Unsurprisingly, Yik Yak topics are dominated by terms that matter to young adults (college students in specific) and contain terms that would typically not be used in a non-anonymous setting.

4. SPATIAL DISTRIBUTION OF TOPICS

In this section we compare the spatial distributions of topics in Yik Yak versus Twitter. Our methodology is to evaluate the spatial dependence between topic strengths in any two locations and the distance between them. As described above, for every location j we have a vector θ_j, which describes the distribution of topics associated with that location. For a given topic i we construct a marked point pattern, mpp_i, (i.e., a set of locations and value) where the mark values are set equal to $\theta_{j,i}$. The continuous valued Stoyan's marked correlation function shown in Equation 1 describes the relationship between the observed mark values and expected mark values given a spatially homogenous process [9].

$$\rho_f(r) = E[f(m1, m2)]/E[f(m, m')], f(m1, m2) = m1 * m2 \tag{1}$$

When marked correlation (not a true correlation in a statistical sense) is plotted against distance, r (units in decimal degrees), a mark pattern that has strong spatial dependence will start at a high value at a short distances and move toward 1.0 as distance increases. A spatially homogeneous topic will, in contrast, have a value near 1.0 at all distances.

We plotted all 40 topics from Yik Yak and Twitter against one another. As seen in Figure 2 there is a significant number of Yik Yak topics that show high spatial dependence in contrast to Twitter topics. The corresponding maps shown in Figures 2a and 2b depict the spatial distribution of these topics, and Figure 2c shows the highest spatially dependent Twitter topic while Figure 2d shows one of the least localized Twitter topic. The results of this analysis answers the question outlined in **RQ2** and supports our intuition that Yik Yak topics are more localized than Twitter topics.

5. READABILITY

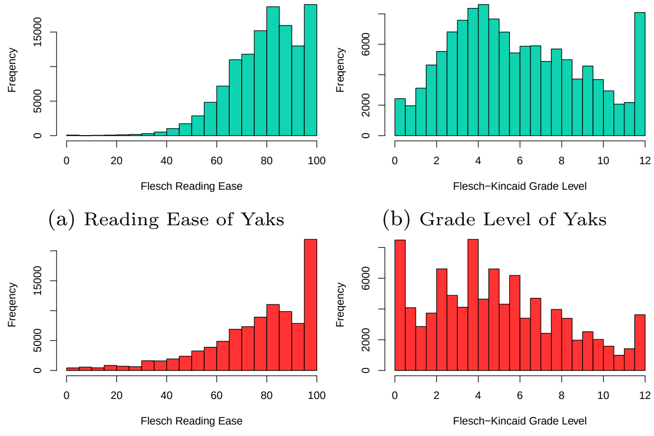
Existing research in social media analysis has employed *text statistics* as a means of describing and comparing datasets [1]. One of the most commonly used set of methods in textual analysis are the *Flesch-Kincaid readability tests*. The Flesch reading ease [4] and the Flesch-Kincaid grade level [6] tests have been used in previous work ranging from assessing the reading level of social media posts [3] to presidential speeches [5].

5.1 Reading Ease

The Flesch reading ease approach uses a combination of the number of words, sentences and syllables to assign a readability value between 0 and 100 (Equation 2) with a value of 0-30 indicating a university graduate reading level and a value of 90-100 representing text that is easily understood by an average 11-year-old student.

$$RE = 206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}}\right) \tag{2}$$

In order to give the reading ease approach the highest chance of success, both the Yik Yak and Twitter datasets used in this analysis was restricted to posts with character counts above 100. This resulted in 123,546 yaks and 95,085 tweets being used in this analysis. The mean reading ease value for yaks in this dataset was 79.5 (median 80.8). For tweets, the mean reading ease was higher at 80.5 (median 76.53) indicating a lower reading level. It should be noted that Twitter handles (@ replies) were excluded from analysis but hashtag terms were kept. A histogram of the results are shown in Figure 3.



(a) Reading Ease of Yaks

(b) Grade Level of Yaks

(c) Reading Ease of Tweets

(d) Grade Level of Tweets

Figure 3: Histograms of Reading Ease and Grade Level based on yaks and tweets in the Greater Los Angeles area.

5.2 Grade Level

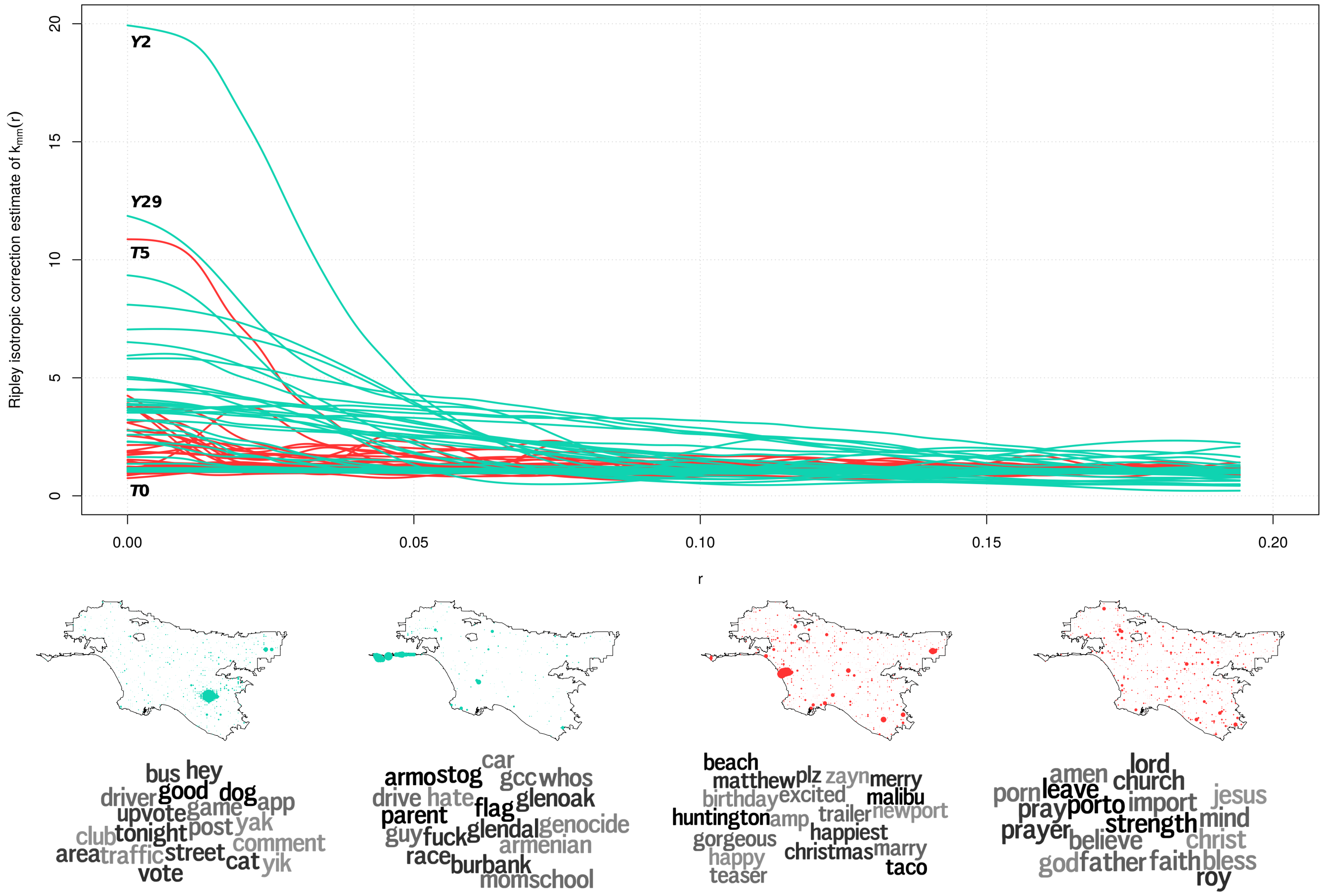
The Felsch-Kincaid grade level test is very similar to the reading ease test but with some adjustments in weights (Equation 3).

$$GL = 0.39 \left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \tag{3}$$

The mean grade level value for yaks (Figure 3b) in this dataset was 5.9 (median 5.4) while the mean grade level value for tweets (Figure 3d) was 4.9 (median 4.7). This answers the question asked in **RQ3** and indicates that the contents of yaks are at a higher grade level than tweets on average. Again, the inclusion of hashtags may have had an impact on this readability analysis as hashtag prevalence is higher in tweets than in yaks. As hashtags are one of the foundational features of Twitter it is arguably inappropriate to remove them this analysis.

6. NEXT STEPS

This work represents an initial step in the data-driven exploration and analysis of anonymous geosocial content contributed via the social media application *Yik Yak*. While this location-based social network has rapidly increased in usage, it has not yet been widely studied by the research community. Thus, our goal was to describe and compare



(a) Y2

(b) Y29

(c) T5

(d) T0

Figure 2: Marked correlation for 40 Twitter topics (shown in red) and 40 Yik Yak topics (shown in green). The preponderance of topics with higher mark correlation curves are Yik Yak topics indicating a stronger spatial dependence in Yik Yak content versus Twitter content.

the dataset through a variety of textual, topical, and spatial data exploration techniques to study whether Yik Yak is an interesting complementary data source for feature extraction or just yet another clone of Twitter (or other social networks). Our initial results suggest that Yik Yak differs substantially from Twitter with respect to the themes being discussed and their local character. It also differs in terms of readability. Consequently, as the anonymity and the visibility scope create a distinct community, we expect that Yik Yak can be used to mine different sets of features that cannot readily be extracted from other data sources. Consequently, Yik Yak is a useful complementary data source worthy of future investigation.

Next steps in this area will involve comparing Yik Yak data with established social media platforms beyond Twitter, understanding how the local visibility of yaks can be exploited to characterize area, e.g., neighborhoods, and analyzing how data published anonymously reveals behavioral patterns that could not be studied using other types of geosocial data.

7. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] J. R. Davenport and R. DeLine. The readability of tweets and their geographic correlation with education. *arXiv preprint arXiv:1401.6058*, 2014.

[4] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[5] R. P. Hart. The language of the modern presidency. *Presidential Studies Quarterly*, pages 249–264, 1984.

[6] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[7] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[8] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[9] D. Stoyan and H. Stoyan. *Fractals, random shapes and point fields: methods of geometrical statistics*. John Wiley and Sons, 1994.

[10] E. Whittaker and R. M. Kowalski. Cyberbullying via social media. *Journal of School Violence*, 14(1):11–29, 2015.