Yik Yak Proposal - EECS 349

Matthew Heston
Northwestern University

Jeremy Foote
Northwestern University

TASK

We plan to study the new anonymous mobile social media application Yik Yak. Yik Yak is a mobile application in which users can anonymously post short messages called yaks. These yaks are visible to other users who are within 1.5 miles of where the yak was posted. Yaks can be upvoted and downvoted. Users can also leave comments on yaks. These comments in turn can also be upvoted or downvoted. If a yak receives a score of -5, it is automatically deleted. We are interested in determining the features of yaks that are predictive of the score it receives.

Yik Yak is primarily used by students on college campuses. A campuss yik yak messages may be thought of, then, as a way to understand local culture. By understanding what features are predictive of different classes of yaks, we can gain insight into this culture. We are also interested in comparing these features across campuses. We might expect, for example, the types of yaks which are highly upvoted by large state universities to be different from smaller, liberal arts schools. As social scientists, machine learning here acts as a methodology in which to use social media data to understand differences in different localities. In addition, we believe our method (described below) can help us get insight into what types of yaks go viral. Again, comparative studies across campuses help us understand social expectations of different campuses.

DATA

We have collected on the order of 2 million yaks from 35 different campuses for another research project examining identity in anonymous social media. The 35 campuses were chosen based on previous work which examined anonymous social media at those locations. A script was written to collect data every hour from each of these 35 campuses. Every hour, the script runs and retrieves the last 100 yaks posted to that campus. If we collect a yak we had previously retrieved, we update the score in our database and store any new comments. For each yak we have: the time it was posted, the location it came from, any handle associated with the yak (this is a feature of Yik Yak that we can think of as a type of tagging), the score at the time we retrieved the yak, comments associated with the yak as well as the score of those comments. Features: Determining which features we plan to use for our task is part of our exploratory work. Ideas we have considered include bag of words representations (and related word vector approaches such as TF-IDF representation) and LIWC categories. Linguistic Inquiry and Word Count (LIWC), is a lexicon based psychometric tool. It consists of categories and dictionaries of words associated with those categories. An example of categories and associated words include positive affect (nice, sweet) or leisure (cook, chat). We can imagine using similar lexicons to create feature representations, or other linguistic attributes of the yak itself (average word length, etc.) Since we are thinking of this project as a way to identify which features are most discriminative, we are interested in generating many different features to test. Other features might include the time of day that a Yak is posted, the distance from the center of campus, the length of the text, the number of comments, and the time until the Nth comment is posted.

INITIAL APPROACH

Yak scores follow what we can think of as similar to a power law distribution. We plan to split the data, following the technique used by Cheng et al [1]. We first find the median number of votes for a post, and look at the differences between the posts which reach that level, and those which don't (i.e., a binary classification task). We then remove all of the items which did not reach the median number of votes, and recurse - for this new set, we again look at the differences between those which reached the median number of votes, and those which did not. This process continues until there are too few examples left to get meaningful results. We will then train various machine learning classifiers based on these different "classes of success". We will focus on classifiers that provide interpretable feature importance scores, and use feature selection techniques to understand what features are most predictive of these classes.

REFERENCES

1. Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. Can Cascades Be Predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, ACM (New York, NY, USA, 2014), 925–936.