Filtering twitter feed

Alexey Levin and Aaron Elliot
Boston University
(Dated: May 8, 2017)

We consider data from twitter, working with text contents of tweets. Our purpose is to obtain classification of tweets, which could be applied to user's feed to filter it. The classification is done using LSA and applying GMM. We further used k=6 clusters for GMM because of relatively small increases in explained variance for subsequent values of k. We also build network of user interactions, using mentions in tweets. Network structure then can be used to complement the classifier and improve relevance of a particular users' feed.

I. PROBLEM

Twitter is one of the most heavily used services today. It is a social media platform that lets users publish their thoughts in form of short 'tweets' - posts that are at most 140 characters long. The brief nature of communications on twitter make it a very 'real-time' platform, where large volumes of posts can appear in a very short time.

This poses a problem: a particular user's feed (tweets from everyone they are subscribed to) becomes extremely cluttered, to the extent of being useless. Also, different accounts post with frequencies varying by many orders of magnitude, so feed becomes dominated by most abundant, noisy channels, which makes the average importance of tweets in the feed low.

II. METHOD

Ultimately the solution is to offer some filtering method.

The methods that could be used to approach this problem could be:

a) Ranking tweets by 'importance' or 'relevance'.
Downside: highly subjective.

b) Classifying tweets into categories.
Downside: does not ultimately tune to user, user still has work to do, by choosing which categories to read. However, categories could be determined based on "importance" of information (as in 'casual chat' vs 'breaking news'), which solves original problem.

The problem with a) is that it highly depends on user, so needs a lot of complementary information (user features, ground truth labeling for each user).

Our strategy is to use b) to classify tweets. After that the filtering can be done by choosing particular slices from feed (eg. "updates from friends + politics").

We also analyze the network structure of user relations on twitter, where an edge connects users if one of them has mentioned the other in a tweet. That can be helpful in determining tweets from which users are relevant to a particular one.

We base our analysis solely on text data from tweets (ignoring metadata), because in this case the approach can be easily reapplied to any other blogging platform (facebook feed, etc).

III. OBTAINING DATA

We wrote a parser that accesses tweets from twitter API. We also found several academic datasets [1, 2] that are suitable for our task, which we used to enlarge the size of dataset we are working with. We were able to access the whole tweet object through API. The datasets contain very limited information about tweets, however it is more than enough for our purposes, because we are only interested in the text contents of tweets themselves. We merged data from all of these sources and used that as our dataset.

IV. DATA PREPARATION

Our analysis consisted of two components: analysis of network structure and NLP.

First thing we had to do was construct the network. We did this by parsing texts of tweets using regular expressions and adding an edge between users if one of them mentions the other.

For the document-term analysis, we used 'bag of words' approach and TfIdf technique to vectorize the data (using english stopwords dictionary and thresholding the term frequency on both sides to clean data of corpus specific stopwords or irrelevant terms that appear only in a few documents). Finally, we reduced the dimensionality of our data by performing PCA. We keep 15 largest singular values in our further analysis.

V. ANALYSIS

A. Basic properties of data.

Our final dataset had 1610200 tweets from 659775 users. The network of mentions, which also contains users from outside of dataset that were mentioned had

889334 nodes and 616462 edges. The largest connected component that we are working with has 339766 nodes. While, the most connected node was @mileycyrus with 3450 edges.

The only numerical data we had on users was date/time of tweeting, ranging from Apr 06 22:19:49 2009, Jun 16 08:40:49 2009 for our larger data set. As well as sparsely from 4/17/17 to 5/4/17, which was when our parser was running. I am satisfied with the dataset we used; however, the academic datasets [1, 2] ended up swamping the much smaller parsed data in the analysis.

B. Network

For the network data we efficiently found clusters (communities) using Kmeans++ on spectral representation of a network using 3 lowest non-zero eigenvectors. We chose k=6 for Kmeans in order to be comparable with our NLP analysis in the next section, as well as because of relatively small increases in explained variance for subsequent values of k.

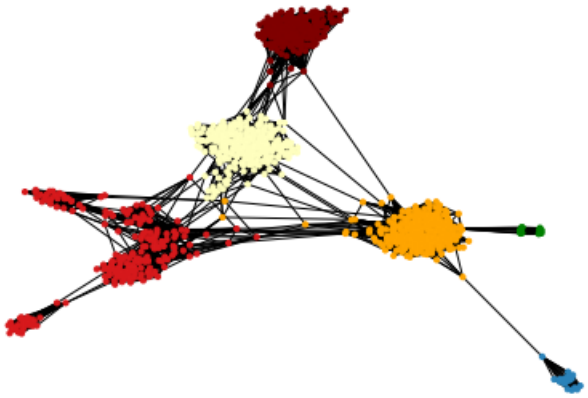

FIG. 1. Communities in the network of mentions on twitter, using spectral decomposition and kmeans++.

We performed basic analysis of the network that we built out of user mentions. Indeed, the network exhibits all the expected properties of a real world-network. In particular, degrees are distributed as a power law. We knew this because the log-log plot of degree distribution was approximately linear.

Properties of relations between users in the network of mentions give methods of predicting relevance of a tweet to the user (using network-based similarity measures) that are complementary to the NLP analysis.

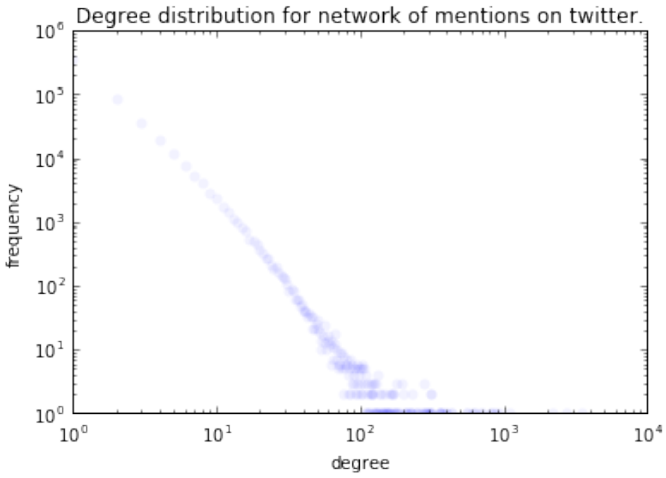

FIG. 2. Degree distribution.

C. NLP.

We used clustering with Gaussian Mixture Models and Kmeans++ to find topics that people talk about.

We analyzed the terms that were most popular in each cluster.

Almost every cluster had as it's most popular terms words related to time, especially to currents moment (such as *now, today, tomorrow, morning* and such), and words related to *work* or *school*.

The most obvious differentiator between clusters was, maybe surprisingly, the sentiment. The most defined discrepancies between clusters would be the proliferation of terms such as *sad, sick, tired, need*, or *happy, awesome, sunny, weekend*. The topics are not distinctly separated. We conclude that standard methods for clustering rather detect the mood of tweets, than the topic. Also we found that, as expected, twitter is mainly used to express emotions or feelings related to what is happening at the moment.

Comparing clusters to network data we observe the following. Mentioning communities on twitter are not interest-based. The distribution of topics that people tweet about is homogeneous over the network. So as an example, communities would be clustered as 'College 1', 'College 2', rather than 'divers', 'football players'; while the topics spoken about in College 1 and College 2 are similar.

[1] http://help.sentiment140.com.
[2] http://www.sananalytics.com/lab/twitter-sentiment/.

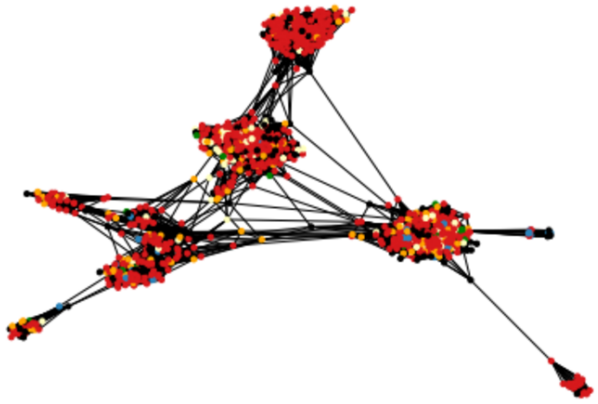

FIG. 3. The nodes are colored according to LSA clustering. Popular topics are not driven by belonging to a particular community.