David P. Stonehouse

University of Maryland University College

DATA 650 – Big Data Analytics

Fall 2016, Section 9040, Professor Gortcheva

Assignment 2: Analyze Kombucha Twitter Feed

## Introduction/Purpose

The Southside Medical Clinic prides itself on taking a balanced approach concerning advice to patients who are considering new trends in nutrition. Many of their patients get excited about a new trend that has many followers, but for which there are limited number of clinical studies. In these cases, they seek to understand the patient’s point of view. One new topic is the consumption of kombucha, a fermented drink deemed to have benefits related to “healthy bacteria”. By studying twitter feeds that relate to kombucha, they hope to be able to more effectively engage and influence patients following the kombucha topic on social media.

## Data Exploration

The data for this project was retrieved from IBM’s Insights for Twitter service. This service allows the user to search for a term used in tweets (in our case, “#kombucha”) and pull back the results into database tables. The source data is a subset (random 10% sample) of the full set of tweets someone could download directly from Twitter. The benefit of using this subset is that IBM enriches the data with insights gained from natural language processing (NLP), accomplished with IBM Social Media Analytics. Insights for Twitter attempts to determine the sentiment of the message (positive, negative, etc.) and gender of the sender. The returned data also includes the term used to determine the sentiment. This data (over 8000 tweets) was loaded into IBM’s dashDb database service.

After loading the data, dashDB presents some basic statistics about the tweet set. From Figure 1, it is seen that there has been a steady stream of tweets with #kombucha over the past two years. The monthly tally has not fallen below 250 from the end of 2014 through last month.

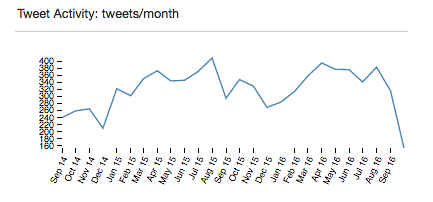


Figure . Kombucha tweets per month in Insights repository.

Looking at Figure 2, many tweets did not indicate the user’s country. For those that did, the majority were from the USA, followed by Canada.

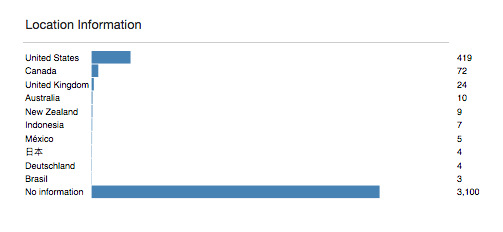


Figure . Locations for senders of kombucha tweets.

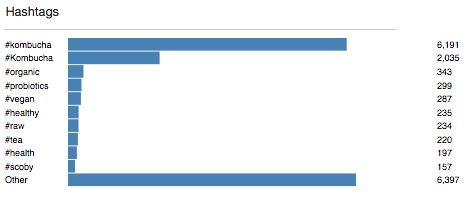
The hashtag information from Figure 3 reflect the relationship between kombucha and some other trends. For example, kombucha is part of a family of products that have probiotics (healthy bacteria). Also, some subset of kombucha tweets are also related to words like “vegan”, “organic”, “raw”, and “healthy”.

Figure . Hashtags in kombucha tweets

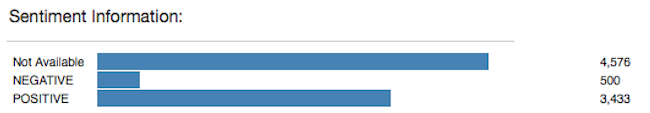
Figure 4 shows the last graph displayed after loading, concerning the sentiment polarities determined by Insights For Twitter. No sentiment was determined for over half the tweets. For those that had a sentiment polarity value, positive sentiments outweighed negative ones 7-to-1.

Figure . Sentiments in kombucha tweets.

The loading process produced a series of tables with the base name “KOMBUCHA” (see Table 1).

Table . Insights for Twitter tables

|  |  |
| --- | --- |
| **Table Name** | **Description** |
| KOMBUCHA\_TWEETS | Main table: text, URL, |
| KOMBUCHA\_HASHTAGS | The hashtags found in the messages, one row per hashtag. |
| KOMBUCHA\_LINKS | Expansion of tinyURL links in messages to the full URL. |
| KOMBUCHA\_LOCATIONS | Lat/Lon location – user profile, message transmission. |
| KOMBUCHA\_MEDIA | Expanded URLs media locations |
| KOMBUCHA\_SENTIMENTS | Polarity and text that indicate sentiment |
| KOMBUCHA\_USERS | User Id, name and screen name |

All of the tables have the key MESSAGE\_ID (e.g., “tag:search.twitter.com, 2005: 600637624315813888”), so that the data from different tables can be joined together as needed.

An example set of tweet texts from the KOMBUCHA\_TWEETS table is shown in Table 2.

Table . Example tweet texts.

|  |
| --- |
| **MESSAGE\_BODY** |
| RT @craftyleftdee: Slumped brown #Kombucha bottle for spoon rest #ecofriendly #etsymntt CraftyleftDee http://t.co/Qt75VOMuY8 via @Etsy |
| #Kombucha: Thrill Hill, Ginger Berry, Thai Temple (Lemon Grass, Lime Leaf) |
| I'm really enjoying this Ginger #kombucha from @livekombucha - and the cap says "we're betterâ€¦ http://t.co/MEfCKBrNXX |
| Homemade #kombucha first taste. So freaking delicious! This is my fermented beverage for the #GreyCup |
| Burnt cookies, #buchi\_kombucha, and #bongojava coffee = good morning! #wakeup #kombucha #paleo #grainfree http://t.co/TPis5SlC8a |
| RT @dbzweier: #kombucha for breakfast, lunch and dinner gets expensive. So I decided to brew my own: http://t.co/GkXbWtC5vy |

Table 3 shows example user information from KOMBACHA\_TWEETS. Note again that USER\_GENDER is a field generated by IBM Social Media Analytics.

Table . Example rows of user information

|  |  |  |  |
| --- | --- | --- | --- |
| **USER\_ GENDER** | **USER\_SCREEN\_NAME** | **USER\_SUMMARY** | **USER\_LOCATION\_DISPLAY\_NAME** |
| female | adams\_madeline | An artist in Alabama. |  |
| unknown | AileenMcGraw | A little heart and a whole lotta floral. | Evanston, IL / Elsewhere |
| unknown | amphore\_oz | Producer of Sydney's finest Ginger Brew Kombucha & the super-probiotic, dairy-free, sugar-free Coco-Kefir. | Sydney, Australia |
| female | annedooner | minneapolis. athleta. W hotel. travel. fitness. fashion. food. Instagram: lizannedooner | Edina, MN |
| unknown | AquaPamela | Tillbaka till grundlÃ¤ggande | estocolmo |
| Female | bAdLadyVet | #Veteran #Student #Millennial #MotivatingTeenSpirit #Volunteer #Advocate #Feminist does NOT = man haterâœŒï¸ #Videographer #SocialEntrepreneur | California, USA |

Some examples of negative and positive sentiment words that Insights for Twitter detected are shown in Table 4.

Table . Kombucha sentiment information

|  |  |
| --- | --- |
| **SENTIMENT\_POLARITY** | **SENTIMENT\_TERM** |
| NEGATIVE | nightmares |
| NEGATIVE | problem |
| NEGATIVE | ran out of |
| NEGATIVE | miserable |
| POSITIVE | best |
| POSITIVE | Love |
| POSITIVE | tasty |
| POSITIVE | beneficial |

## SQL Queries

Several SQL queries were run to gain insights into the data. The first one (Figure 5), is the first in a series of three queries to understand the negative sentiment messages.

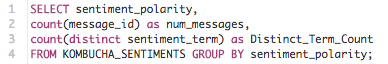
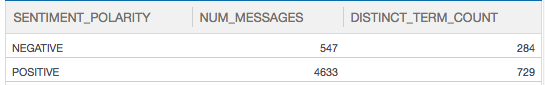


Figure . Query 1: Pattern of unique terms for each sentiment polarity.

The results (Table 5) indicate that over half (284 out of 547) of the negative tweets were analyzed with a term only used once. In contrast, almost 4000 tweets used repeated positive terms. That could make for a good word cloud.

Table . Results for unique sentiment terms.

The next query (Figure 6) found the top terms related to negative sentiment tweets.

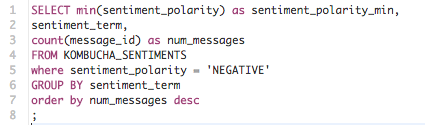
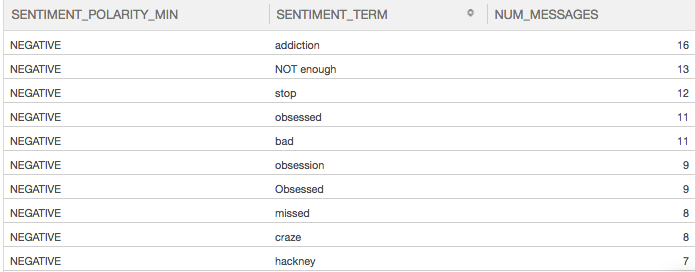


Figure . Query 2: top terms used to determine negative sentiment.

The results (Table 6) indicate that “obsessed”, “addiction”, “not enough” and “stop” are the top words. It is interesting to note that Insights for Twitter makes use of “not” as a contextual influence to determine sentiment.

Table . Results for message counts for the top negative terms.



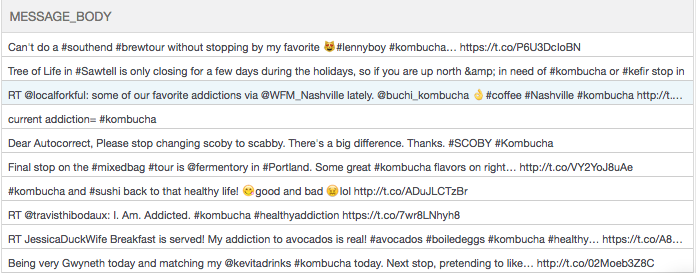
The third query (Figure 7) retrieves the messages that contain some of the top negative terms.



Figure . Query 3: Find matching messages for the top negative sentiment terms.

The results are enlightening. It appears that Social Media Analytics is either employing shallow NLP algorithms or still has a way to be trained for finding sentiment in tweets. As an example, “stop” was the negative sentiment, even though it is part of the phrase “next stop”, so it is a noun. The noun form of “stop” has no negative connotation. If this sample is any indication, there are likely very few negative sentiments in these tweets. The twitter crowd just loves their kombucha!

Table . Tweets containing the top negative sentiment terms.



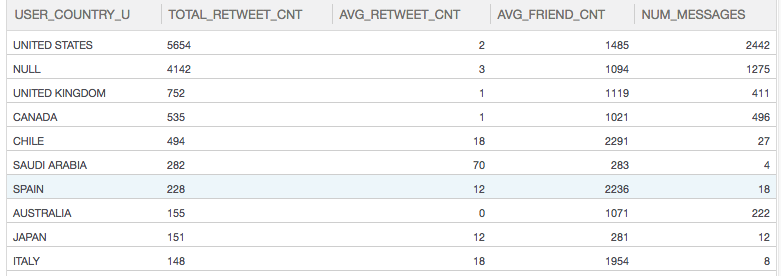
Switching topics, the next query (Figure 8) aims to find out which countries do the most retweeting. In support of this query, a view called SENTIMENT was created. The code to create that view is in the appendix.



Figure . Query 4: List the top countries for retweeting.

The results (Table 8) show the USA as the top country, which is not a surprise since Americans tweeted the most about kombucha. An interesting note is that the few messages from Saudi Arabia kombucha tweeters were retweeted 70 times. Similarly, influence is expanded in Chile and Spain.

Table . Top 10 countries for retweeting.



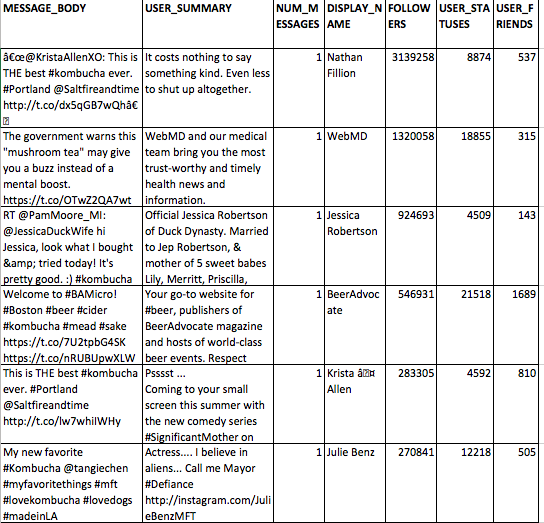
The last query (Figure 9) is an attempt to determine the set of influencers who are tweeting about kombucha. These are people who have potential for a very large influence based on their number of followers, friends or their number of status posts. The cutoff values were chosen based on basic statistical analysis in R. See the “Analysis in R” section for more details.

It is interesting to note in the results (Table 9) that the number two most followed user in this set of tweets is WebMD, which clearly has reservations about kombucha. Interesting in this case that these most influential people just tweeted once about kombucha. But with millions of followers, this is possible the introduction of the topic to a new set of people.



Figure . Query 5: Get the user summary and an example tweet for the biggest influencers. The biggest influencers are determined by their number of followers, statuses and friends.

Table . Example tweet and user summary text for the 6 users with the most followers.



## Analysis with R

In dashDB, an R script was created using template code created using an interactive tool. After selecting a table and the desired columns to import into R, dashDB produces R code to load it into a data frame. In order to validate this load, users should run the R summary command. The R summary command An important step before validation was to use the as.factor function to specify that fields related to gender, country and sentiment polarity were factors. After that, the summary results could be compared with the load results from when dashDB loaded the data from Insights for Twitter. Another good sanity check was that the tweets per month values (Figure 10) also validates that the distribution of tweets by date matches the summary statistics observed earlier.

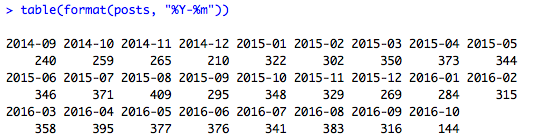


Figure 10. Table of number of kombucha posts by month.

After the data was loaded to R, there were a number of insights gained from doing basic statistical analysis of variables. For example, the summary information related to influencers in Figure 11 were used to set the cut off point for the SQL query in the previous section for retrieving influencers.

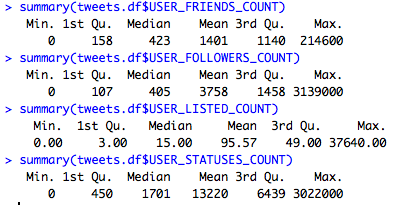


Figure . Summary information for variables that define influencers.

The clinicians were interested to understand whether there were any weekly or seasonal patterns to kombucha tweets. It does seem that Sunday is significantly quieter (Figure 12), while Tuesday to Friday seem to be pretty even.

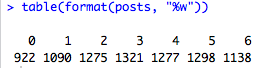


Figure . Number of kombucha tweets by day of the week

The day with the most kombucha tweets was on July 7th, 2015. More were sent on that day than in many whole months.

../../Desktop/Screen%20Shot%202016-10-16%20at%205.36.25%20PM.png

Figure 13. Busiest day for kombucha tweets.

Looking at Figure 13, there does not seem to be any interesting seasonal, quarterly or monthly patterns.

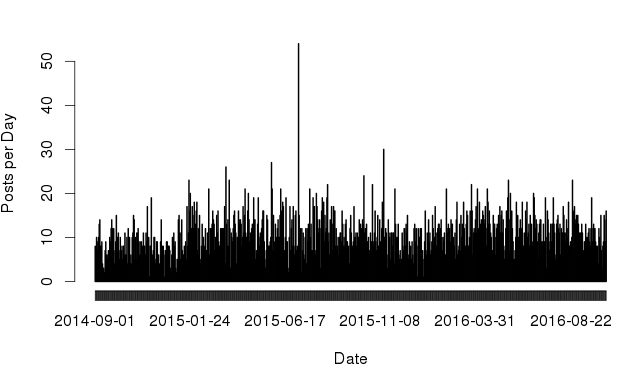


Figure . Plot of kombucha tweets for each day.

The clinicians were also interested in patterns related to gender. The analysts hoped to do analysis of variation (ANOVA) related to gender for the “influencer variables”: number of followers, number of friends, and number of statuses. Figure 15 and Figure 16 illustrate how a few outliers can dominate analysis of these variables. Just by removing the small number of outliers with more than 300,000 followers, the mean and standard deviations change from being drastically different between male, female and unknown gender classifications, to the numbers being quite similar. An attempt was made to do ANOVA on number of followers vs. gender, as well as number of friends, but there were no results that were shown to have significance.

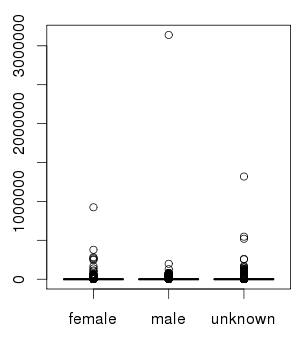
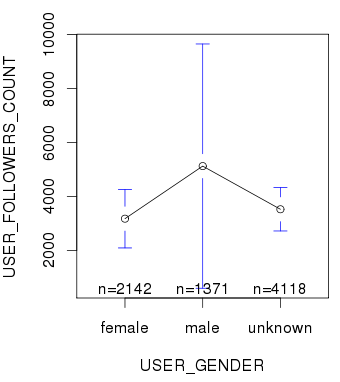


Figure . Plots of gender vs. number of followers: a) mean and standard deviation b) full distribution.

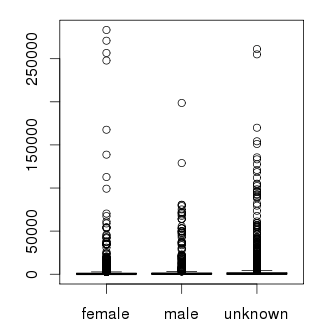
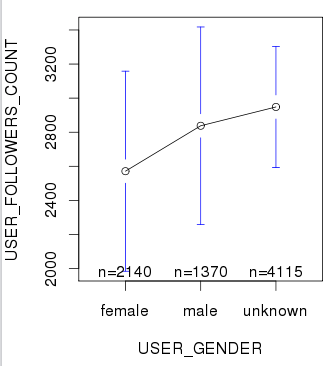


Figure . Plots of gender vs. number of followers with outliers removed: a) mean and standard deviation b) full distribution.

The clinicians were very interested to know how much these trends were taking globally, as that was likely to confirm to patients that viability of the claims. Figure 17 shows the distribution around the world by country and gender. It is interesting to see how the gender gap in interest seems to be significantly bigger in Canada.

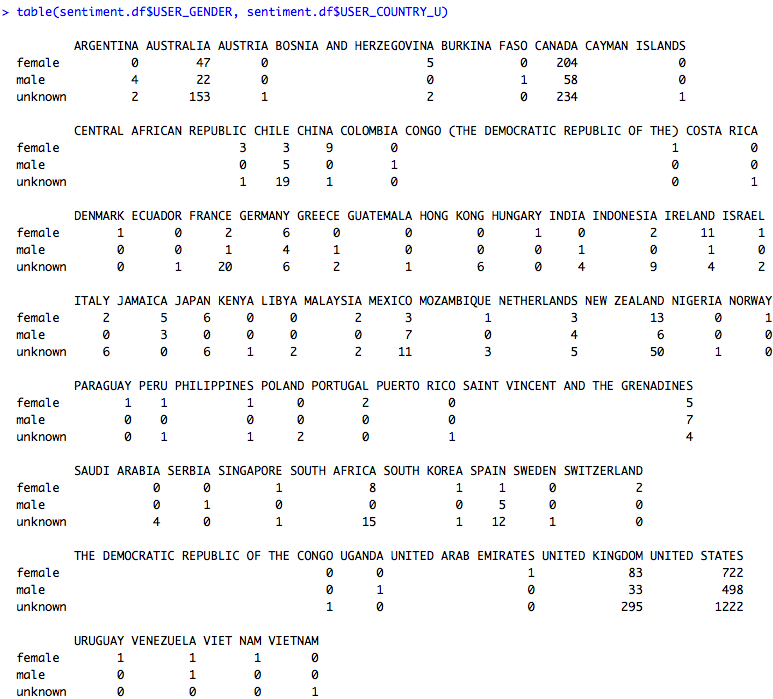
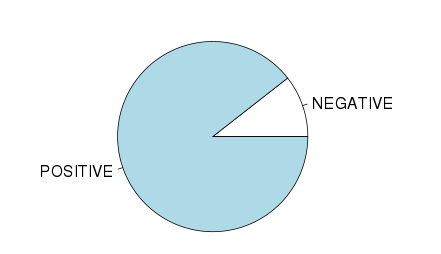


Figure . Distribution of messages by gender and country.



## Limitations and Future Research

.

## Conclusions

## References

IBM, 2016, “About Insights for Twitter”. Retrieved from <https://console.ng.bluemix.net/docs/services/Twitter/twitter_overview.html#about_twitter>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MESSAGE\_BODY** | **USER\_SUMMARY** | **NUM\_MESSAGES** | **FOLLOWERS\_COUNT** | **USER\_STATUSES\_COUNT** | **USER\_FRIENDS\_COUNT** |
| â€œ@KristaAllenXO: This is THE best #kombucha ever. #Portland @Saltfireandtime http://t.co/dx5qGB7wQhâ€   Agreed! | It costs nothing to say something kind. Even less to shut up altogether. | 1 | 3139258 | 8874 | 537 |
| The government warns this "mushroom tea" may give you a buzz instead of a mental boost. https://t.co/OTwZ2QA7wt #kombucha | WebMD and our medical team bring you the most trust-worthy and timely health news and information. | 1 | 1320058 | 18855 | 315 |
| RT @PamMoore\_MI: @JessicaDuckWife hi Jessica, look what I bought &amp; tried today! It's pretty good. :) #kombucha #gettinghealthy https://t.coâ€¦ | Official Jessica Robertson of Duck Dynasty. Married to Jep Robertson, & mother of 5 sweet babes Lily, Merritt, Priscilla, River, & Gus | 1 | 924693 | 4509 | 143 |
| KOMBUCHA ç¾Žå‘³ã—ã„ ãŠæ°—ã«å…¥ã‚Šã£ #kombucha https://t.co/vvnAlSl40S | SKE48ã®æ¾äº•çŽ²å¥ˆã§ã™ â– æ¯Žé€±æœˆæ›œ24:00ã€œãƒ‹ãƒƒãƒãƒ³æ”¾é€ã€ŒãƒŸãƒ¥ã€œã‚³ãƒŸï¼‹ã€å‡ºæ¼” å‡ºæ¼”æƒ…å ±ã¯ã“ã¡ã‚‰ã¸ æ¾äº•çŽ²å¥ˆã‚ªãƒ•ã‚£ã‚·ãƒ£ãƒ«ã‚¤ãƒ³ãƒ•ã‚© @rena\_info | 1 | 648862 | 5639 | 171 |
| Welcome to #BAMicro! #Boston #beer #cider #kombucha #mead #sake https://t.co/7U2tpbG4SK https://t.co/nRUBUpwXLW | Your go-to website for #beer, publishers of BeerAdvocate magazine and hosts of world-class beer events. Respect Beer. | 2 | 546931 | 21518 | 1689 |
| Y gracias Andreita @nutristetik por el #kombucha que me regalaste me cayo de perla. ðŸ˜‰ https://t.co/ooYRFp2i9q | Vivo en Guayaquil. Hago Radio y Revista. Tengo dos hijos; amo a mis chikis, a mi trabajo y sobre todo a Dios. | 1 | 378463 | 55094 | 1898 |
| This is THE best #kombucha ever. #Portland @Saltfireandtime http://t.co/lw7whiIWHy | Psssst ... Coming to your small screen this summer with the new comedy series #SignificantMother on The CW Network. Premieres August 3rd @ 9:30 | 1 | 283305 | 4592 | 810 |

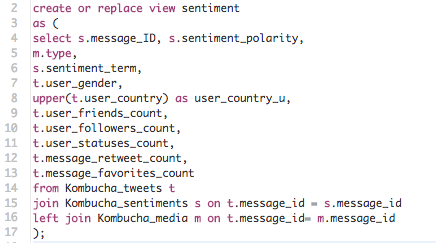
UCLA, “How can I subset a data set?”, retrieved from http://www.ats.ucla.edu/stat/r/faq/subset\_R.htm

IBM, 2016, “Getting started with dashDB”. Retrieved from https://console.ng.bluemix.net/docs/services/dashDB/dashDB.html

## Appendix: Supporting Information

**SQL code:**

The following code was used to create the sentiment view:

****

**R source code:**