

## DS8006: Lab 1 "Let's Talk"

Student's name: NAJLIS, BERNARDO

### 1. Briefly describe what search criteria did you use to collect data from each of the three platforms and why:

*I used the hashtag officially proposed by the campaign; #BellLetsTalk for Twitter and Instagram, and the Facebook page name BellLetsTalk.*

### 2. Did you consider any other search criteria for Twitter and why were they rejected?

*I purposely did not want to use any other search criteria than the main campaign keyword as I think the target audience has used the proposed hashtag. This is because the campaign has an inherent incentive to participate because Bell is donating money (CAD\$ 0.05) to the cause for every hashtag mention and I think that is the main reason why most people get engaged. Also, I think that adding additional search terms will limit the number of results returned, potentially skewing the data available for analysis.*

### 3. Summarize your datasets in terms of their size and any other attributes that you feel are important to mention here.

*When measured in number of interactions over time, the campaign was much more successful on Twitter (277.78 tweets/min) compared to Facebook (4.57 posts/min) and Instagram; when considering number of unique users Instagram comes first, Twitter in a very close second and Facebook in a third distant position. Also, as in this dataset every Twitter user has on average more followers than their Facebook counterparts, the audience reach on Twitter is 3x (200% more) when compared with Facebook. Notably, Instagram users were much more inclined to 'Like' posts from this campaign when compared to Facebook users: for every Facebook Like, there were 5 Instagram Likes.*

*A couple of minor side notes:*

- *as expected most of tweets and Facebook posts were in English and French*
- *most of Facebook post types were Status updates (the typical post in users' Wall)*
- *majority of Instagram posts didn't use any photo filter at all*

### 4. What are the major differences between the three datasets (Twitter, Instagram and Facebook) in terms of the types of data that they contain? Hint: When you answer this question, keep in mind the types of analysis that you can perform with the various data types.

*Every dataset has a basic set common fields shared among the three platforms (i.e.: pubdate, author, description) and fields that are specific for each platform ('like count' in Facebook, 'user\_count\_followers' in Twitter and 'filter' in Instagram). This suggests that the number of comparative analysis is reduced to only a couple of metrics, but the analysis that can be provided per platform is more detailed.*

### 5. If you are asked to analyze the effectiveness of this social media campaign, what other data fields (that Netlytic did not retrieve) would you want to collect using the available APIs to help you with the analysis? Hint: Refer to the API documentation from Step 5 above.

**6. Based on your research, what is the main difference(s) between Twitter's REST/Search API and its Streaming API? In terms of data collection for Bell Let's Talk campaign, which Twitter API is more appropriate and Why?**

*The main difference between the REST/Search API and the Streaming API is their time span reach: the REST/Search API goes back in time and return past results vs. the Streaming API only returns results after the call was made. Other difference is based purely in its technical implementation: the REST API returns a result dataset based on the time when the API call was done ("one shot query"), whereas the Streaming API keeps delivering data continuously after the first call was done ("continuous query"). The number of results that can be expected is therefore much higher with the Streaming API than with REST. Another main difference is the REST/Search API allows for more complex*

*For this campaign, considering its mainstream public reach and with the assumption to only use the #BellLetsTalk hashtag as a simple search term I think the streaming API is more appropriate. The streaming API allows to collect more data volume and the data returned is very similar to the REST/Search API results. This the implication that the data collection platform would have had to be prepared in advance of the campaign launch day.*

**7. Next week, you will start writing your own R scripts for data collection using Twitter Search API. Considering the rate limit imposed by Twitter, what is the maximum number of tweets would you be able to collect in a 60-min time window? Explain your answer.**

*Considering the REST/Search API allows a maximum of 450 requests per 15-minute time window (source: <https://dev.twitter.com/rest/public/search>), this gives a maximum of 1,800 requests per hour. If each query result returns up to 1,500 tweets (source: <https://twittercommunity.com/t/get-all-tweets-having-a-specific-word/73410>) , the theoretical maximum number is  $1,800 * 1,500 = 2,700,000$  tweets per hour.*