Ruma Anand, Ryan Miller, Daniel Trielli

EECS 510: Social Media Mining / Alok Choudhary

6/13/2018

## Sentiment analysis of tweets from U.S. universities – final report

### Introduction

The goal of this project was to discover if it is possible to extract meaningful insights about quality of life on American university campuses using data from social media. To do that, we combined data extraction from Twitter and sentiment analysis to try to extract sentiment profiles from higher education institutions.

The motivation for this work comes from traditional university rankings. Students applying to college turn to these indexes to try to choose the best institution to attend. The rankings use a variety of different criteria, but most commonly they rely on academic excellence[[1]](#footnote-0), reputation[[2]](#footnote-1), or financial outcome of students after graduation[[3]](#footnote-2). Our motivation of this work is that applicants might also be interested in metrics of quality of life on campus.

Literature shows that sentiment analysis combined with social media data can provide deep insights about social and political issues. According to Paul et al. (2017), "analyzing and understanding societal response and crowd reaction to important and emerging social issues and events through social media data is increasingly an important problem."[[4]](#footnote-3) It is also understood that emotion measurements extracted from social media and user-generated content can be essential for brands to determine how consumers are reacting to their products and services.[[5]](#footnote-4)

Based on our goal and the literature available, we decided to do this pilot project to try to answer two questions:

**Q1: Is it possible to classify universities according to sentiment?**

**Q2: Does this classification provide robust insights about quality of life on campus?**

### Methods

For our pilot project to extract sentiment about universities from social media, we developed a method that is divided in three phases: first, we created a sample of American universities to analyze; second, we collected tweets related to these universities; and third, we ran sentiment analysis on these tweets, generating summary statistics related to each institution.

#### Sampling

We collected a sample of tweets from 25 American universities, one of them being our own, Northwestern University. Northwestern is part of the Association of American Universities, an organization that gathers 60 leading higher education universities in North America. We decide to extract our sample from that set, since all these institutions have common features, such as academic excellence, focus on research and high notoriety.

However, even within that set, there is a wide variety of institution size, from the California Institute of Technology (around 2,200 students) to the University of Toronto (84,000). This difference in total students imply different volumes of social media chatter, and therefore might create unwanted noise in the data. Therefore, instead of randomly selecting institutions that had from each other – and would produce incomparable results – we instead decided to focus on universities that were most similar to Northwestern, which has approximately 21,000 students. We therefore focused on all universities that had between 10,000 and 30,000 students. There are 25 universities in that subset (see Table 1), and fortunately they have distinct features that can produce interesting comparisons. First, they are a mix of private and public universities (16 private and nine public); and second, they represent all four regions of the U.S. (as defined by the Census Bureau), which eliminates geographical bias of sentiment. In total, there are four in the West, eight in the South, ten in the Northeast, and three in the Midwest.

**Table 1: Sample of U.S. universities**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Institution** | **Control** | **Total students** | **US\_Region** | **State** |
| Cornell University | Private | 21904 | Northeast | New York |
| Harvard University | Private | 21000 | Northeast | Massachusetts |
| Johns Hopkins University | Private | 23073 | South | Maryland |
| Northwestern University | Private | 21208 | Midwest | Illinois |
| Stanford University | Private | 15877 | West | California |
| University of Oregon | Public | 22980 | West | Oregon |
| University of Pennsylvania | Private | 24630 | Northeast | Pennsylvania |
| University of Virginia | Public | 22391 | South | Virginia |
| UCSB, Santa Barbara | Public | 24346 | West | California |
| Carnegie Mellon University | Private | 12,908 | Northeast | Pennsylvania |
| Columbia University | Private | 29,250 | Northeast | New York |
| Duke University | Private | 14,600 | South | North Carolina |
| Emory University | Private | 14,513 | South | Georgia |
| Georgia Tech, Atlanta | Public | 29,370 | South | Georgia |
| MIT, Cambridge, MA | Private | 11,301 | Northeast | Massachusetts |
| Stony Brook | Public | 25,272 | Northeast | New York |
| The University of Chicago | Private | 14,954 | Midwest | Illinois |
| The University of Kansas | Public | 27,983 | Midwest | Kansas |
| Tulane University | Private | 13,462 | South | Louisiana |
| UC Irvine | Public | 29,588 | West | California |
| UNC Chapel Hill | Public | 29,390 | South | North Carolina |
| University of Pittsburgh | Public | 28,649 | Northeast | Pennsylvania |
| University of Rochester | Private | 10,290 | Northeast | New York |
| Vanderbilt University | Private | 12,795 | South | Tennessee |
| Yale University | Private | 12,223 | Northeast | Connecticut |

#### Scraping

The social media data that we gathered focused on three types of tweets related to universities: 1) tweets that were geolocated within a radius of the selected colleges and universities; 2) mentions of the institution by tagging its Twitter account (e.g. @NorthwesternU); and 3) tweets that used a hashtag that refers to the institution (e.g. #northwestern). We did not collect tweets from official university accounts because they would be mostly positive.

We collected tweets from a year-long timeframe (from June 1, 2017 to May 31, 2018), in order to control for temporal biases and to gather a significant volume of data. To do that, we used a tool called GetOldTweets that can be found on github, created by Jefferson Henrique[[6]](#footnote-5). Since the Twitter search API prevents users from querying Twitter for tweets older than seven days (unless you pay for the premium API), we used GetOldTweets to circumvent this roadblock. GetOldTweets uses a technique that involves programatically emulating a web browser and pointing that emulated web browser to Twitter’s advanced search in order to trick it into thinking a human is viewing its web pages. Twitter then displays tweets without the seven-day-old restriction, which are scraped by the emulated web browser and returned to the calling program.

These tweets were then processed to a function that removes hyperlink and any special characters/unicode characters that add to the difficulty in processing the sentiment for a tweet.

#### Sentiment analysis

We chose to use Indico for our sentiment analysis. Indico[[7]](#footnote-6) is a multi-faceted toolbox full of utilities for machine learning, AI, NLP, and many more. We used their Emotion tool. The Indico Emotion tool is fairly simple; it takes in text and returns a breakdown of how much of each of five emotions the text snippet is. A very happy statement would have a very high score (almost one) in the joy category and very low scores (almost zero) in the sadness, fear, anger, and (possibly) surprise categories.

Using the almost 50,000 tweets scraped from Twitter by GetOldTweets, we passed those into the Indico Emotions API (separated by university) to get a five-emotion breakdown of each tweet. After averaging across emotion for every tweet in a university, we compare the averages for each university. Our findings are discussed below.

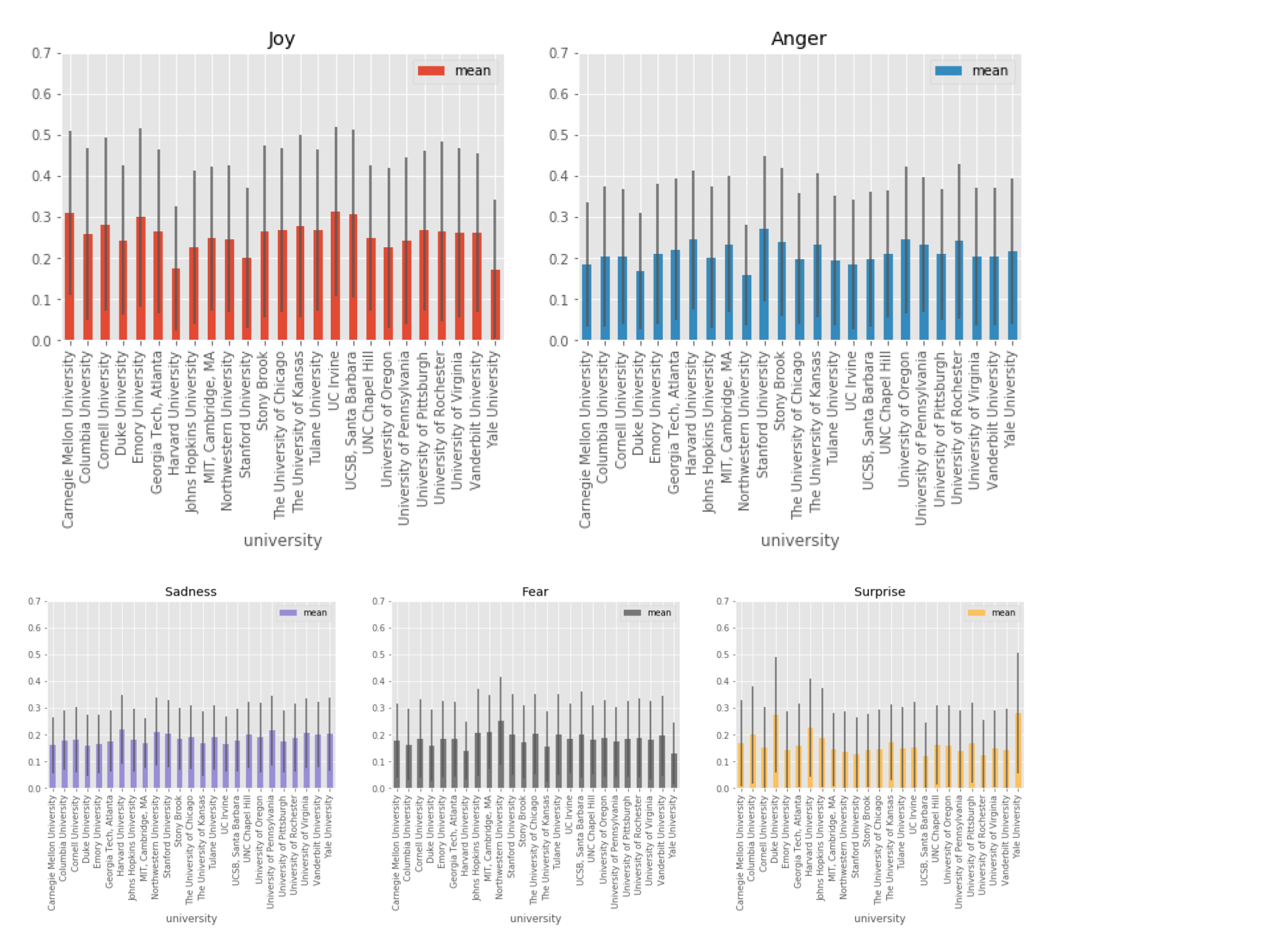
### Results

The results gathered from the data show promise in determining sentiment profiles from higher education institutions.

#### Comparing universities and regions

When looking at geographical tweets, that is tweets from a radius around the campuses, it is possible to discern some that have a higher degree of anger, joy, fear, sadness and surprise associated to each university (Figure 1).

**Figure 1: sentiments by university**



We also attempted to find distinctions from type of institution – private or public (Figure 2). Aside from a slight difference in joy and surprise (public universities have a bit more of the first and a bit less of the second), there are not any strong conclusions to be made.

**Figure 2: average sentiment per type of institution**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Joy** | | **Anger** | | **Sadness** | | **Fear** | | **Surprise** | |
| **Funding** | **mean** | **std** | **mean** | **std** | **mean** | **std** | **mean** | **std** | **mean** | **std** |
| **Private** | 0.243635 | 0.193796 | 0.213711 | 0.16777 | 0.187754 | 0.120591 | 0.185778 | 0.144489 | 0.169122 | 0.170309 |
| **Public** | 0.270735 | 0.20196 | 0.214042 | 0.168285 | 0.183692 | 0.120026 | 0.182303 | 0.141688 | 0.149228 | 0.147348 |

Looking at the summary of sentiment per region, it is also possible to see some general trends. The Midwest, for instance, has an average slightly higher joy and fear than the other areas; while campuses in the West have higher anger sentiment and the Northeast institutions have a higher surprise level.

**Table 3: average sentiment per region**

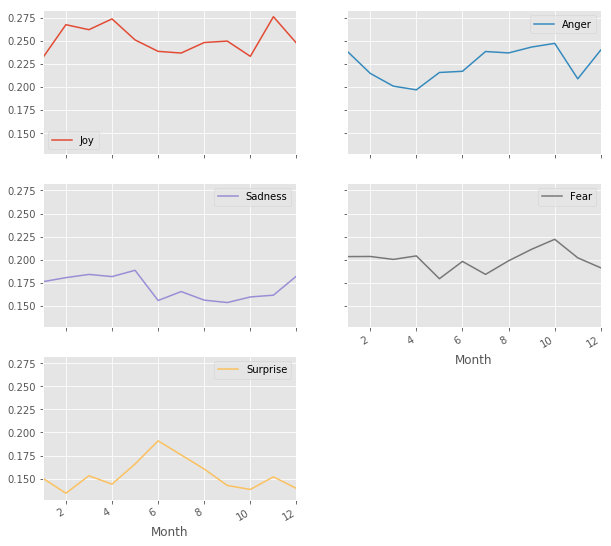
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Joy** | | **Anger** | | **Sadness** | | **Fear** | | **Surprise** | |
| **Region** | **mean** | **std** | **mean** | **std** | **mean** | **std** | **mean** | **std** | **mean** | **std** |
| **Midwest** | 0.263443 | 0.200229 | 0.195185 | 0.156779 | 0.189744 | 0.122713 | 0.203069 | 0.153854 | 0.148559 | 0.148036 |
| **Northeast** | 0.246959 | 0.197612 | 0.221086 | 0.16968 | 0.185963 | 0.119382 | 0.172566 | 0.135301 | 0.173427 | 0.168696 |
| **South** | 0.255494 | 0.19439 | 0.202936 | 0.164621 | 0.186514 | 0.121009 | 0.188723 | 0.146079 | 0.166333 | 0.16495 |
| **West** | 0.261061 | 0.200021 | 0.224979 | 0.172672 | 0.184892 | 0.120477 | 0.194462 | 0.148988 | 0.134606 | 0.145883 |

While these results show differences across universities and regions, the standard deviations are also quite high. But this means that there is internal variation of sentiment among students for each university. It is to be expected that not all students share the same sentiment; some are happier than others. That differentiation is what the standard deviation shows. Usually, a high standard deviation signifies a noise that diminishes the validity of the result. In this case, the noise is to be expected, since we do not expect that campuses that are "happier" only have happy people. They are aggregately happier or not, as shown by the average.

#### Analyzing time frames

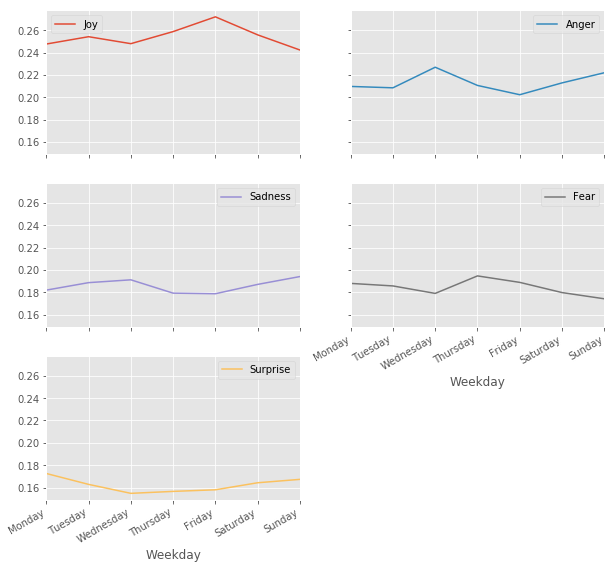
Looking at the evolution of sentiments across time also provides interesting insights. For instance, looking at the month by month change, it is possible to detect peaks of joy in the end of the calendar year and in the Spring, possibly because of Holidays and Spring Break. Fear drops around May, at the end of the school year, but increases steadily over the Fall, the same as Anger. Sadness drops sharply at the beginning of the Summer.

**Figure 2: sentiments across the year**



Analyzing sentiment by day of the week, unsurprisingly, there is a spike in Joy on Friday. The saddest and angriest day is Wednesday. Fear fluctuates across the week.

**Figure 3: sentiments across the year**



### 

### Discussion

The initial data that we processed and analyzed shows that the idea of using Twitter to provide a glimpse into the quality of life on U.S. campuses is promising. However, there are limitations to be considered in this type of work.

We began this work by posing two questions:

**Q1: Is it possible to classify universities according to sentiment?**

**Q2: Does this classification provide robust insights about quality of life on campus?**

The answer to the first question is, yes, but with caveats. Classification can be done by looking at sentiment contextually. While we might be inclined to disregard data with high standard deviation, this type of dispersion in the data is to be expected when dealing with thousands of tweets from thousands of students. In this regard, the combined averages of all tweets across dimensions of study can provide insights to the quality of life of college campuses, but that has to be tempered with the notion that campus life is, in itself, varied.

That answer also helps answer the second question. The robustness of insights is dependent on the context in which the data is presented, and the volume of data processed. It was possible to see, for instance, clear trends on time analysis of the sentiment of tweets. For instance, there is more joy on tweets posted on Friday and more fear and anger on tweets in the Fall.

A final point has to be made about an additional challenge in this type of work, which is the reliance on limited output of data provided by the platforms we aim to study. Twitter's API provides a very limited scope of data, which requires researchers to circumvent those limitations by using unorthodox means. Other social media platforms, such as Facebook, are even more restrictive on the type of information it provides, which helps explain the relative low volume of social media analysis work that is done with that platform.

In short, in the project we have learned the social media mining presents technical challenges, but also social and contextual implications.

1. https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings [↑](#footnote-ref-0)
2. https://www.topuniversities.com/qs-world-university-rankings/methodology [↑](#footnote-ref-1)
3. https://www.topuniversities.com/employability-rankings/methodology [↑](#footnote-ref-2)
4. Paul, D., Li, F., Teja, M. K., Yu, X., & Frost, R. (2017, August). Compass: Spatio Temporal Sentiment Analysis of US Election What Twitter Says! [↑](#footnote-ref-3)
5. Krebs, F., Lubascher, B., Moers, T., Schaap, P., & Spanakis, G. (2017). Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. [↑](#footnote-ref-4)
6. https://github.com/Jefferson-Henrique/GetOldTweets-python [↑](#footnote-ref-5)
7. https://indico.io/ [↑](#footnote-ref-6)