October 6, 2020

Dear editors and reviewers:

We present below our replies to the very helpful comments that you shared when reviewing the manuscript. We include both the text of the comments and (in bold) our replies. We are grateful for the constructive feedback that you've shared. Our manuscript is better because of it.

Sincerely,

Fred Boehm and Bret Hanlon

Editor's Comments to Author:

Associate Editor

Comments to the Author: The manuscript is a substantial revision and provides excellent information for JSE readers. The referees are generally quite happy with the work, and they offer some general guidelines for additional improvements.

In particular, the referees all spoke about the way that "research question" was addressed in the manuscript. Although your focus away from the student projects makes sense, bringing in their research question would be a way to address the referees suggestions that twitter data is hard to use for generating hypotheses. Indeed, you might address the point by speaking to the difference between "data science research" and "political science research". It seems as though the former was a learning goal and the latter would be difficult to do with such exploratory data as twitter. Can you expand on that point?

> **Thanks for raising this point. We've now added text in the introduction that addresses the interplay between data science research and discipline-specific research, like that in political science.**

In terms of George Floyd, my guess is that by May 26, the 1% of twitter hadn't get caught wind of the situation? You might make a comment to that effect.

> **This is an excellent point. We've now added text to specify that although our May 26, 2020 tweets occurred after George Floyd's murder, we didn't detect evidence of a high volume of tweets about George Floyd at 12pm Eastern Daylight Time on May 26. We suggest that the George Floyd Twitter activity likely increased after our collection time as more viewers watched the video.**

minor: * line 128 "Below is an example of Tweet JSON." Is that phrasing right? I might have said "Below is an example of one tweet in JSON format." Or "Below is information provided by the Tweet JSON script." I'm not sure. But I guess I don't know what it is an example *of*.

> **Thank you for the suggestion. We've tried to clarify this by rephrasing.**

Reviewer: 1

Comments to the Author General Comments:

The revised manuscript "What is happening on Twitter? A framework for student research projects with tweets" provides a case study for facilitating undergraduate research projects using twitter and topic modeling. The manuscript is well-written and is of interest to JSE readers, and is a strong "datasets and stories" article.

I have only two general suggestions: first, it would be nice to have a brief description of the two undergraduate projects within the manuscript. While I understand the data is no longer available, and think it was the right decision to focus on the case study for methods and results, including a short paragraph about each of the undergraduate projects you supervised would help to give the reader a sense of the range of possibilities of a twitter-based project. This is briefly mentioned at the end of the manuscript but it would help to see the motivations for such a project earlier.

Thank you for this suggestion! We've now added text that describes the two projects early in the manuscript.

Second, while I appreciate the concise and clear writing, adding transitions between the sections would help the reader anticipate what is coming. As one example, prior to Sec 5.2, there is no indication that the authors will walk through (a) querying twitter API (b) accessing full tweets via ID numbers (c) parsing JSON.

We appreciate this suggestion, too. We've worked to add transitions between sections with the goal of better guiding the reader's expectations.

Specific Comments:

- Page 2 Line 28: Should there be a reference for "rhetoric in recruiting political supporters"?

  Yes, we mistakenly omitted them. We've now added key references for this point.

- Page 2 Line 52: I'd recommend adding a statement along the lines of: "different statistics and data science researchers may categorize these skills differently, but our assignment is based on our program/specific projects." For example, I imagine some projects using twitter data may not categorize "Use text analysis tools to analyze tweets" as "Enduring Understanding", but would categorize "Use data visualization to clarify and inform quantitative analysis" in such a way.

  We've now added a statement like this to the manuscript to reflect this point.

- Page 4 Sec 4.2: I agree with the authors that differences in student research interests and goals should be an important consideration and it would be nice to see some examples.

  We've added two examples from our students.

- Page 4 Line 82: Would this type of project also be appropriate for a summer research experience?

  Yes, we've amended the text to reflect this.

- Page 4 Line 88-92: This wording could be a little more clear: is the case study essentially a replicate of a student project? It would be helpful to see an overview of what the original project was, and then introduce the case study as a replicate.

  We've reworded this to clarify the point. Yes, the case study essentially replicates a student project.

- Page 5 Sec 5.2: would this process theoretically work for both windows and Mac machines? Is it necessary to use a crontab-like process, or could it be done entirely within R?

  Thanks for this question. We've added text to clarify that this process is possible from Windows, Mac, and linux operating systems. We've also included text to say that there are R packages that interact with operating system-specific scheduling software, so that a user would only interact with the R GUI.

- Page 6 Sec 5.4: This overview of JSON tweet data is very useful.

  Thank you.

- Page 7 Line 133: Include citations for rtweet and tidy text appear here, where they're mentioned first, instead of in the following section.

  We've now moved the citations to the first mention of these packages.

- Page 8 Line 152: It would be helpful to see a brief model equation.

  We agree! We've added the equations that describe Latent Dirichlet Allocation.

- Page 8 Figure 1:"beta" has not been mentioned before appearing in this graph (related to previous comment)

You make a good point. We think that the equations that we've added (to address the previous comment) also resolve this.

- Page 11 Line 162-163: "... our interest is in the transient appearance of a new topic" — this could be made more clear at the beginning of the case study. Another sentence or two motivating what to expect for results would be helpful.

  We agree with this suggestion, too. We have tried to clarify it by adding a sentence to this section.

Reviewer: 2

Comments to the Author Summary: This paper describes a case study using Twitter data and latent Dirichlet allocation to find topics in social media data. Overall, I felt like this edit of the paper was much stronger than the previous submission, much more useful to me personally and to a broad audience of statistical educators. I particularly appreciate the grading rubric!

As always with case studies, I think this is useful at a variety of different levels. First, it provides a template for doing a similar project with a student (i.e., a year-long independent research project). Then, it gives example code for a task that might otherwise be complex. And finally, if an instructor is just looking for an interesting example dataset, they could use your code to grab (and perhaps clean) data for their students to analyze, without exposing them to the entire pipeline.

Major Issues: N/A

Minor Issues:

I think sections 3 and 4 could be combined into one section that is perhaps "Structure of mentored research" or something like that. Perhaps there would be fewer subsections under that section header, but more details on how the research was structured.

We agree. We now use your suggested section title.

Throughout the existing sections, the verb tenses were strange. Since this is now a case study, I expected everything to be grounded in what you actually did. But, on page 3 it says, "These are our four learning objectives." Why not, "were"? Did you develop the learning objectives after the fact?

We didn't write out learning objectives at the time of working with these students, but we did informally discuss them. Back then (2015), we were new to research mentoring. We now are more explicit while designing projects with students.

At the bottom of the page there are some general statements about students in your statistics department, but not about the particular expertise of the two students who did the project. Did they have existing expertise in linux computing? It says that you guided students toward supplementary resources— did they not have previous R experience?

Great point! We now discuss this point, and we think that it strengthens the manuscript.

Section 4.2 says that "an initial brainstorming session may clarify." Why not "helped"? Is an initial brainstorming session just something you recommend/plan to do in the future, or did you actually do it with those students? You say that you anticipate that sharing completed student project reports will guide future students, but this is from the 2015-2016 school year. Since this project, has it become easier to talk to students about goals, grounded in examples like this? Back to the students in 2015-2016, what goals did you eventually land on?

Thank you for raising this issue. We agree that it needs clarification. We've added text to address this.

I assume that those discussions with students involved conversations about the learning objectives from the previous section, as well as the expectations for students (presentation at the end of the year, perhaps some amount of weekly work?).

Yes, although we were new research mentors when we began working with these two students. Regrettably, we didn't structure our conversations as we would now.

I found the case study pretty easy to follow starting at 5.3, but I personally have never used crontab before, so that was the most intimidating piece of this. I'd guess that my experience is pretty similar to many JSE readers, who may be familiar with R but not unix/linux commands. Can you insert a short description of what crontab is, and a reference to how to learn more? I'd guess that the instructions you provided would work on Mac, and probably on an RStudio server or RStudio Cloud, but not on a Windows computer. You may want to explicitly state this.

Yes, we completely agree. We've added text to explain what crontab is and what it does. We've also pointed readers to an R package for Windows

I like that you've shared the results from the analysis in the paper! One of my first observations was that there are a lot of swear words in the topics. You may want to address this directly in the paper. Real world data is messy! And may not be "safe for work." If faculty are going to supervise students in research like this, they need to be prepared for that.

We completely agree. We've now added text to prepare readers for this finding.

I also audibly gasped when I saw the days you were collecting data. It doesn't look like George Floyd's death has made it into the topics, unless it is part of topic 10 from May 26, but again, may be worth an explicit mention. He was killed May 25, and protests began in Minnesota on the 26th.

We agree. We've added an explicit mention of this issue in the body of the text.

I really liked the fully-reproducible analysis, but I think it would also be worthwhile to discuss the projects your students actually did. You talked about those projects in the previous version of the paper, and while I know the data is no longer available, perhaps you could incorporate the ideas into the paper somewhere. Either when you talk about question generation, or at the end about other topics. Having a variety of ideas for using tweets might help instructors see that they could either follow your code, or do something related but different.

Thank you. We've now added text about the student projects. We've also added more references to point readers to other researchers' work with Twitter.

Typos/copyediting:

p. 3 I think Table 1 would benefit from rules on the left and right to close off the table.

We've added rules as you suggested.

p. 3 "We translated our prioritized list of skills [. . .] into learning objectives" rather than "we translated into learning objectives our prioritized list of skills"

We completely agree. The original wording was awkward, so we've fixed this as you suggested.

p. 4 I think "R for Data Science" should have headline case

We agree, and we've corrected this.

p. 13 - 14 many references need adjustment of capitalization. It looks like this was generated using bibTeX, so just in case this is useful— to force a capital letter you can use curly braces to surround something like {ACM SIGKDD} in the bibtex entry. I noted that Statistical need to be capitalized in the Blei reference, probably Gibbs and Dirichlet should be capitalized in the Porteous reference (as well as the conference name), R in the first Wickham reference.

Thank you for catching this error. We've now corrected the capitalization issues by using curly braces as you suggested.

Reviewer: 3

Comments to the Author The authors describe a framework for student research using data available using the twitter API. This is an interesting and relevant paper since many other uses of Twitter data have proven to be somewhat superficial in terms of the richness of data. The paper is grounded in research on projects.

My main issue relates to the question being answered here: what actionable insights are being extracted from these data? Is it really just that Memorial Day words show up on Memorial Day? Are there other questions that could be considered (even if the authors don't answer them)? I'm left with the concern that Twitter data is cumbersome to deal with (in terms of scale and restrictions on use of the API) and not particularly information rich.

Thank you. We've attempted to address these concerns by writing more about the original two student projects.

Suggestions:

1. Table 1 is a nice approach to classify project skills. I wonder whether it might be productive to connect these skills with other aspects of data science (e.g., the components of data acumen from the NASEM 2018 "Data Science for Undergraduates" report, https://nas.edu/envisioningds)?

   This is a really nice suggestion! Thank you! we've now added text to connect Table 1 with the data acumen components that you reference.

2. Need to add a ref for LDA (page 4): I would suggest capitalizing all words on first use, then using LDA.

   We've now added the original 2003 Blei, et al. article as the reference on page 4. We've also capitalized all three words in the first mention, and used LDA subsequently.

3. Page 5: I was confused by this sentence: "Additionally, on repeated querying of the API, different sets of tweet identification numbers return data." Can you clarify/rewrite?

   Thanks for pointing this out. We've now rewritten the sentence in efforts to clarify the point that querying the Twitter API is not fully reproducible.

4. It will be important to note early and in section 6/figure 1 that some tweets may not be suitable for work in terms of inappropriate language. This is a general issue that might be important for instructors to discuss.

   Thank you. We've added text to warn readers about this.

5. Please provide a link to the Amy Cooper Central Park incident.

   We've added a link to a New York Times article.

6. I really like the rubric. How closely tied is it to using Twitter data? What would be different if some other data source were provided?

   Thank you! This is a great point that you make, and, as you seem to suggest, much would not differ with other data sources. We've added text to the manuscript to expand on this point.

7. I wonder whether the title might be reconsidered. Perhaps "A framework for student research projects in data science courses"? The application to Twitter feels secondary to me.

   We appreciate the suggestion. We chose the title with hopes of promoting the use of tweets and text analysis in student research projects. With all of our supplementary materials explicitly dealing with tweets, we feel more comfortable with the current title.