Reviewer’s report

Title: Mining Online E-Liquid Reviews for Opinion Polarities about E-Liquid Features

Version: 0 Date: 12 Feb 2017

Reviewer: Quinghua Yang

Reviewer's report:

The authors conducted a timely and important research on e-liquid reviews using text mining and sentiment analysis. This study could potentially be valuable for understanding e-liquid flavors and people's attitudes towards them, after several major and minor concerns being addressed.

Introduction

The authors made an argument that "there are no published studies that have systematically summarized and analyzed e-liquid flavors and other features based on online review websites". However, in a recently published study at JMIR, titled "Identifying Topics for E-Cigarette User-Generated Contents: A Case Study from Multiple Social Media Platforms", although authors didn't conduct a sentiment analysis per se, they did examine the flavors and sentiment across three social media platforms, including JuiceDB. Therefore, the authors may want to update their review of literature and make a more appropriate statement.

Methods

1. First, how do you know that the 17,000 reviews are all unbiased? It is highly possible that e-liquid companies also post reviews.

2. In the data analysis, the authors stated that "we identified some important features of e-liquids, including many flavors". "Many flavors" are very vague. Please specify or provide some examples. Similarly, the authors said "we added some flavors mentioned in the reviews but not in the previous literature". What flavors are referred to and what's the rationale of including these flavors above others? Are they the most popular ones in the reviews? Please clarify.

3. My major concern with the methods is the classification of sentiment. Is there a reason of not including the "neutral or mixed" category, which is typical in sentiment analysis? Also, just by reading those reviews on JuiceDB, many of the reviewers provided a neutral or mixed opinion of certain flavors. How did the authors take those reviews into consideration?
4. The authors said that they regarded the reviews with rating higher than 7 as positive and others as negative. What's the scale that is used in this rating and how did the authors come up with this criterion? Since the review scale on JuiceDB is 1-5, the scale used in the study is unclear.

5. More details are needed to explain the classifier training process. For instance, how did the authors train the classifier to obtain an accuracy of 82.04%. How did the authors differentiate long and short paragraphs, and why different classifiers are needed?

Results

1. No need to repeat the sampling in the methods section, which is already explained in the methods section.

2. The categorization is very confusing; several categories and subcategories are the same (e.g., fruit, cream, beverage). How did the authors come up with the subcategories and what's the differences between the categories and subcategories in the same name? Also, there're frequency and positive sentiment ratio for all other categories but not for ingredients and smoking feelings. Why is this the case?

3. Some of the results were misinterpreted or at least inconsistent. For instance, according to Table 1, tobacco, menthol, and beverages are the least favored, instead of "sweet, tobacco and beverage" in the manuscript.

4. Please move the definitions of cloud production and throat hit to the introduction section.

5. The conclusion that "the users pay little attention to nicotine and are not very pleased with the ingredients, while they are satisfied with the cloud production and throat hit of e-liquids" is not convincing, given that the positive ratios of ingredients and throat hit are very close and we don't whether they are significantly different from each other.

Discussion

1. Move the operational definitions of popularity and preference to the introduction section.

2. Again, the conclusion that "the favorite flavor list is different from the popular flavor list" is not convincing with only the descriptive statistics without significant testing or confidence intervals.
3. The authors concluded that "the most favorite flavor category is cream, followed by nuts", but the positive sentiment ratio of nuts (0.878) is higher than cream (0.872). Please revise to keep them consistent.

4. Also, the conclusion that "fruit, cream and nut categories are most often mixed with other flavors" is shaky. Did the authors conduct any statistical analysis to come up with this conclusion? Similarly, how did you come to the conclusion that "users generally were not concerned about nicotine" and "nicotine plays a less important role in the online e-liquid reviews" (compare to what)?

5. In general, the discussion section needs to be strengthened by drawing conclusions with rigorous testing and providing insightful discussion beyond the results.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal