Author’s response to reviews

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

Authors:

Zhipeng Chen (zhipengchen@email.arizona.edu)

Daniel Zeng (zeng@email.arizona.edu)

Version: 2  Date: 02 May 2017

Author’s response to reviews:

Hello Editor,

We appreciate the editor's comments and the reviewer's comments. We revised the manuscript according to the comments.

Reply to the editor:

1. Ethical consideration of the study

You said 'not applicable'. However, you need to write in the methods section how you think ethical requirement was not needed for your study (i.e. exemption).

All the data were collected from public online sources. We didn't collect any indentifiable information. The previous literature using social media data was not required for approval, such as, 'Li, Q., Zhan, Y., Wang, L., Leischow, S. J., & Zeng, D. D. (2016). Analysis of symptoms and their potential associations with e-liquids’ components: a social media study. BMC public health, 16(1), 674'. We explained this in the section Ethics Approval and Consent to Participate in line 2 of page 10. If the ethical approval is required, we will apply for it.

2. Validity of the response

I see this is an online forum used by registered users. How confident you are about responses were made by a unique user?

Would it possible for one person to post a response via different accounts? Is there any inclusion/exclusion criteria for the responses?
As the data source is social media, it is very difficult to confirm the true identity of the users. It is likely to happen that one person registers more than one accounts and posts multiple posts, but considering the large size of the dataset, this personal behavior has only minor influence. We added the discussion of this limitation in lines 16-22 on page 8.

3. Measurements

It is not clear what are actually looked (e.g. flavours, smoking feelings) and how relevant information was identified. Many are mixed up with Data analysis section. I would like you to separate measurements out from the Data analysis section.

To make the structure clear, we seperated feature text extraction and sentiment analysis from page 3 line 20 to page 5 line 15. Because we want to conduct sentiment analysis on the feature level, we first extracted the feature texts. Then we explained the classification of the feature texts at the beginning of Sentiment Analysis subsection: we classify the texts to two categories, positive and negative.

On page 4, line 11, you stated 'we manually identified two smoking feelings'. I would advise you to provide a reference for this justification.

Thank you for reminding. We added a reference on page 3 line 32.

4. Sentiment analysis

You need to describe the approach more. It is not clear what definition is applied to define 'positive' or 'negative' on feedback and whether this is done manually or empirically through key word search. Is this done on each flavour?

As we mentioned above, we revised the structure from page 3 line 20 to page 5 line 15. We first identified features, such as strawberry flavor, cream flavor VG and PG, and extracted feature texts by keyword search. Then we conducted sentiment analysis to classify the feature texts. We defined positive and negative at the beginning of the subsection Sentiment Analysis in lines 23-24 on page 4. If a text is in positive sentiment category, the writer likes the feature; otherwise if a text is in negative sentiment category, the writer dislikes the feature. The sentiment analysis was conducted by two steps: first, we trained the NBSVM model; second, we used the trained NBSVM model to analyze all the feature texts. Technique details can be found in the NBSVM paper 'Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90-94). Association for Computational Linguistics'.
5. Discussion

We do not use numbering in discussion. You could use subheadings if that would help. You could state the implications of your findings (i.e. significance) in relation to the aim of journal. What would your study contribute to health and well-being at population level? You mentioned some in conclusions, but these could be addressed under discussion. It is helpful to discuss your findings relevance to the safety of the products.

We extended the discussion in lines 1-15 on page 8. The implications focus on the policymakers can identify the attracting factors of e-cigarette and improve the regulation policies to leverage the tobacco control function of e-cigarettes. The safety and health issues should be further explored by clinical studies. Previous literature on mining symptoms from social media claims the VG and PG potentially affect users’ health, which is consistent with our result that the users are not satisfied with VG and PG. Therefore, our sentiment analysis can reflect potential health risks of e-cigarettes. We explained this point in lines 1-8 on page 8.

Reply to the reviewer:

1. I agree with the authors that the data set might be unbiased, but don't see convincing evidences, as the consistency between the popularity results and the analysis of Reddit data could happen by chance. A stronger argument is needed to conclude that "the bias is not strong and the conclusions can still hold".

As the data source is social media, it is difficult to check the bias. We further discussed this limitation in lines 18-22 on page 8.

2. What scale the authors used for review ratings is still unclear. Did the authors transform the JuiceDB 1-5 rating?

JuiceDB revised its user interface several times. At the time we collected data, the scale was 1-9. For example, one post says "pretty good! could probably hit a 9 after a good amount of steeping."

3. The statement that "The popularity and preference are not correlated (correlation=0.31), so the popularity of certain flavors doesn't suggest the e-cigarette users like them" is problematic. According to Cohen (1988), .3 is a moderate correlation coefficient. Please justify the statement.
Thank you for your advice. We have fixed it in line 32 on page 5.

Minor issues:

1. It should be "we identified and explored" at line 25 on page 3.

Thank you for pointing it out. We have fixed it in line 22 on page 3.

Reference