PAPER SUMMARY | TASK 1

How Community Feedback Shapes User Behaviour

Originally authored by J. Cheng, C. Danescu-Niculescu-Mizil and J. Leskovec

Summary

Background and Setting

Social media systems rely on user feedback and rating mechanisms for discovering trending content, personalisation, ranking and filtering. However, when users evaluate content contributed by a fellow user, they also implicitly evaluate the content's author, leading to complex social effects within the underlying communities.

While related work has analysed the effects of this feedback at the community level, this paper attempts to understand them at the individual author's level and determine whether the feedback is indeed beneficial to the community - which is its intended purpose. The paper makes this judgement against the *operant conditioning framework* from behavioural psychology, which predicts that community feedback should improve content quality, by discouraging poor-quality and encouraging high-quality content.

Verifying and questioning the applicability of fundamental and long-relied-on frameworks from psychology, philosophy and related fields in modern settings is a popular direction within research in Social Computing, and this paper is another example.

Problem Statement and Contributions

This paper tracks the changes in user behaviour after receiving positive or negative feedback, concerning the quality of their posts (perceived and textual), the feedback they provide to others, frequency of posting and retention rates.

The significant contributions are:

- Introducing and validating a matric for aggregating community feedback and using it within a propensity-score-matching-based framework to quantify the effects of community feedback on the users' posting behaviour.
- Discovering that the effects of community evaluations are generally detrimental to the community, contradicting the operant conditioning framework.
- Revealing an important asymmetry between positive and negative feedback, somewhat in line with the *negativity effect*.

Methodology and Discussion

The paper follows a data-driven methodology; the dataset is collected from four online social communities that allow users to comment on articles and up- and down-vote others' comments. The authors propose a post's *proportion of upvotes* as a single metric for quantifying the community feedback and find it to have a high correlation with how independent annotators perceive those numbers of up-/down-votes (even though it doesn't capture the sheer numbers of votes).

The authors find that negatively evaluated users end up getting more negatively evaluated in the future, either because of a decrease in post quality or because of the community's perception of them. To disentangle these two factors, the authors train a binomial regression model to measure post quality using only the post's textual content. They find that there is a significant drop in textual quality of the posts coming from negatively evaluated authors, as well as a community bias against them (measured by the difference in the textual quality and proportion of upvotes). The corresponding mirrored patterns are surprisingly not found among positively evaluated users. Further, negatively evaluated users are 30% more frequent than before in posting, as opposed to 20% for positively evaluated ones. Users who receive no feedback are likely to discontinue, indicating that ignoring undesired behaviour may be a good way to discourage it. The user-pairs chosen for all these experiments were similar before they received the positive and negative evaluations, respectively. The authors also notice tit-for-tat behaviour, where people tend to give the kind of feedback they receive, although it is not specific to a thread or person (and hence not indicative of a flame war).

Analysing the networks of users, the authors find coalitions of users who up- and down-vote a post, and these networks are especially polarised when the numbers of up- and down-votes are equal.

Critical Analysis

The paper does an excellent job of explaining the motive behind each experiment and the factors that were kept in mind. For instance, each propensity score matching experiment specifies the variables that were kept identical before the event and when it was not possible to do so (retention rate, for example). The figures and tables are also complete and remarkably easy to read. The paper also mentions several shortcomings of its analysis and gives directions for future work.

Although the paper analyses four different communities, doubts about the generalisability of this analysis remain. While the communities differ in the topics they discuss, they are all article-sharing websites. It would be interesting to know how well these results generalise to websites that only allow upvoting, as well as across websites with different *"unwritten norms"* of providing feedback (for example, up-/down-votes play a significant role in moderation on StackExchange, while users might upvote almost everything they see on Facebook).

Further, the quality prediction model, which is central to this analysis, completely ignores the context surrounding a post - which could be important in determining its *"usefulness to a discussion"*.

In my opinion, one crucial difference up-/down-votes have from the experiments that led to the operant conditioning framework which the paper does not touch upon is that this feedback is publicly visible and affects the user's social image. Whether or not feedback that's only visible to the author would have similar effects is a question worth considering.

Future Work

There are several questions in this field worth exploring, such as the effects of flame wars on the users and the community, or the importance of the feedback giver's relative authority or reputation. Improving the quality prediction model or analysing communities with different feedback rules as described above are also ways of improving our understanding of these social feedback loops.