Question1: Read and summarize the following paper Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 'How Community Feedback Shapes User Behavior'. In Proceedings of the Eighth International Conference on Weblogs and social media, ICWSM 2014.

A) Write a detailed summary of the paper's objectives, research questions/hypotheses, methodology, and results.

This paper investigates how rating on some author's content affects his/her future behavior.

If positive rating act as the reward and negative rating act as the punishment, then according to the "operant conditioning framework" the rewarded author must contribute more, and punished author should contribute less in the future.

In this paper they developed a method for comparing the effect of reward and punishment on the author's future behavior in the community and whether community feedback affects the quality and quantity of a user's future contribution.

They investigate four online news communities CNN.com, Beitbart.com, IGN.com and Allkpop.com

In these website users can comment(post) on the news articles and other user can up vote or down vote this comment(post).Now to aggregate the post's votes into a single number that corresponds to the magnitude of the reward/punishment that will be received by the author of the post they have used Amazon's mechanical Turk and asked users how they would perceive receiving up/down votes and found out that the proportion of up-votes is a very good measure of overall feedback of the community.They considered a post to be positively evaluated if proportion of up-votes is 75 percentile of all post and negatively evaluated if it is 25 percentile.

They build a machine learning model that predicts a post's quality by using textual features of the post. Using this model, they match pairs of users (A, B) that contributed post of the similar quality I.e., they select the pairs of the user that appear indistinguishable before the evaluation and one of the users from that pair is punished while the other one is rewarded. Then to see the effect of the evaluation on the behavior of the A and B their next three posts will be examined.

The results were in the opposite direction of the operant conditioning framework. Rather than increasing the post quality, negative evaluation decreases the quality. The effect of positive feedback is inconsistent and not significant.

Now to check community bias they again use the same setup and found out that community bias affects the negatively evaluated users because the perceived quality of their post was much lower than their textual quality.

Next to determine the effect on the quantity of the author's post by reward/ punishment they changed the variable of the interest from similar quality post users to the users with same posting frequency. They found out that the users who received no feedback they actually slow down, and they post with 15% less frequency while who received positive feedback write 20% more frequently and those who received negative feedback write 30% more frequently.

Also punished users not only chances their posting behaviors but also voting behavior by evaluating their fellow users negatively. Feedback effects become stronger if a user trusts the authority providing the feedback.

Thus, both types of evaluation suggest that providing negative feedback to the bad users is not a good strategy to avoid undesired behavior in the community, we should rather ignore undesired behavior and provide no feedback at all.

B) Express your opinions on the paper in terms of its strengths and weaknesses, what it should have done differently, possible extensions, and its utility.

In this paper we understood how feedback mechanisms are used in online systems and how they affect the underlying community. They relate their findings with operand conditioning theory which explains the mechanisms behind reinforcement learning.

If I talk about the strengths, then it is ability of the machine learning model to estimate the textual quality of the post since text quality is a very hard machine learning problem. This model can be further optimized by fitting the model with more validation and test data. Improving the model will allow further analysis and reveal even more relations between post quality and community feedback. This model only considers up and down votes of a post to evaluate the quality of the post while ignoring other facts like the comments on the post. If we consider the comments also then the model will be more accurate and can give better relation between the post quality and the community feedback. To do so we can train a machine learning model that counts and number of up votes, number of down votes, total number of comments that post has.

After that it will process each comment one by one and identify each comment as a positive comment or negative comment that can be done by preprocessing of the comment and linguistic analysis. Number of up votes, number of down votes, number of positive comments, number of negative comments will be the feature values for our model, and it will calculate a score for each post. We can use that score instead of the score used in this paper for better results.

Also, they have completely ignored the content of the discussion and the context in which the post appears, understanding the role of the contest can also reveal very deep interactions that occur in the communities.

C) The paper follows a common paradigm of verifying existing psychological theories (in this case, the operant conditioning framework) using large-scale data collected from social media and other sources. What do you think are the benefits and possible pitfalls of such analysis?

This paper follows operand conditioning theory from behavioral psychology which explains the underlying mechanisms behind reinforcement learning and finds the observed behaviors deviate significantly from what the theory predicts. This Method/framework can be used to find the behavior of an individual according to the feedback. But the problem is this framework was designed and developed to test and experiment on animal behavior. For example, cats and rats or pigeons.

The lack of experiment on human behavior with this framework can be attributed to the ethical and methodological issue. Because we found out in this experiment that user behavior is largely “tit for tat” that is different from what this framework suggest.

There are various reasons for that, we can say online feedback (Down votes) is much more different than the feedback in a laboratory (Electric shocks).

Question 3: Collecting redundant data is a useful way to assess the quality and consistency of distributed data collection. Read this paper by Windt et al.which develops and tests a system to collect reports of conflict events.

A) How does their design ensure redundancy?

The voix dese kivus system is a crowd seeding system that collects information via short message service (SMS) from the preidentified informants in a selected location. Under the crowd seeding approach anyone can send an SMS to a central platform. In this system they identified three reporters in each village, the chief of the village, head of the women association and one elected by the community.They were given a phone and trained how to send SMS, and a code sheet was also provided to them which contains the codes of the possible events that can take place in the village. Reporters can send SMS containing these codes or they can also send full text.

It is possible that one reporter sent an SMS containing code of the event while another sent a full text describing the whole event. It is also possible that some of the reporters send redundant SMS in another local language .To remain in the system they need to send at least one SMS per week (could be blank). So, the system might get a lot of blank SMS.

In this project reporters sent around 4,783 nonempty SMS of 5,081 Events of which 4,623 were unique.

Identical messages sent by the same reporter within thirty minutes are treated as a single event, also identical messages sent by different reporters in the same village within 24 hours of each other are treated as single events.

That's how this system works and ensures redundancy!

B) They offer several approaches to validate the collected data. Summarize them. Which one do you find the most convincing?

Voix des kivus was launched in august 2009 and operated in only 4 villages then from august 2010 it was expended to 18 villages

Approaches to validate the data include investigating time, reporter type and event type and whether they affect if an event is reported or not.

First, Voix des kivus employed a coordinator to visit a village at least once every week and visit each of the reporters to assess the quality of the SMS. For Ex. The reporters employed the correct code or not, to verify whether an event had actually taken place or not and is there any event that took place and were not reported.

Another approach is the measure of internal validation because they distributed phones to three people per village, they show the share of events that were reported by at least two reporters out of all events reported.

Apart from that they conducted a survey in the same region in a similar period and compared the event information that was received from the system and from the survey.

I found out the first approach of employing a coordinator to go and ask the reporter and validate an event is most convincing. Since nineteen of the fifty-four reporters had only primary level education and only two had received education beyond secondary school so it is possible that they might not able to understand or misinterpret the code-sheet so there is a high chance of error there.

C) Propose a new way of validating the collected data. Try to increase confidence in the data while keeping the collection cost-effective and ethical.

To validate the collected data Voix des kivus proposed various ways. A new way might be to install the hidden Cameras in the village to keep an eye on whether an event is taking place or not. When the system gets SMS about any event from a village we can manually check and monitor whether it is true or not or we can get an idea about the event. To monitor the camera, we might need an operator who continuously keeps an eye and to reduce the cost we could install only (1-2) cameras only in any village.

Question 4: Amazon Mechanical Turk is a common crowdsourcing platform employed by many computational scientists for their research. Researchers use MTurk to access thousands of on-demand workers for different tasks ranging from annotations to data processing etc. Imagine the experience of an MTurker and then assess the design, quality, and ethics of human computation projects.

A) Critically evaluate human computation experiments on overall research quality, experiment design, and ethical grounds.

Mechanical Turk is Amazon's online labor market where Researchers/requesters post jobs and workers choose which jobs to do to get paid. MTurk is a flexible platform capable of supporting many kinds of human computation.

Human computation is an area where a computational problem is solved by using the power of humans that a computer alone cannot yet solve. Human computation includes a variety of data processing work such as translation, content generation, assessment, transcription, AI hard content analysis (Understanding text, images or multimedia), creating new content, making prediction, information seeking monitoring, digitizing

information of the physical world etc. Human computation experiments depend on the quality of work workers provide. To motivate different people to do tasks correctly different mechanisms include pay, fun, prestige, learning.

On AMT (Amazon's Mechanical Turk), a person who provides work is known as requester and a person who performs a given task is known as worker and the unit of work to be performed is called Human intelligence task (HIT). Distribution of HITs follows a power- law distribution I.e., few workers doing a lot of work while many workers doing a little work. It can't be said that it is a fault of online crowd work or AMTs, or it is because of different types of task/Experiment posted there.

Human computation experiments are affected by human factors as it is a human centric enterprise. Here poor quality of data often blames the workers for being lazy or ignorant. But some other factors like Experiment/Task instruction and/or interface may also be the reason for the same, I.e., if Experiment's/ Task's instructions are unclear or the interface is not designed properly then it affects the quality. A very Important factor of any such experiment is to ask the right questions, it seems very trivial but if not followed properly then it can produce undesirable results, since here a human design a task that will be completed by another human. The requester of the experiment must ensure that all the workers have a common knowledge of the question's meaning. Specific terminology should be avoided. Also, A task must be divided into many subtasks.

Human computing may appear unethical for some individuals due to relatively low wages, asymmetric power relationships and depersonalized work, for example you are doing some job and get nothing. I can take example of online game where people/workers play my game and it is generating some product as output of which workers may not be aware and they do not get any profit from it, but it is generating revenue for the requester.

B) Describe in detail how crowdsourcing human computation efforts help in improving data quality but does not help in removing bias.

Data quality is one of the most important things that may be reduced due to the noise in the responses provided by the users (Humans). One of the solutions to avoid this noise is to test the users. A qualification test in which we ask the workers some questions and qualify the workers who pass this test. This is a very good way to avoid those who are performing badly or avoiding the spammers. At any given day, it is true that requesters look for the workers so dividing a task into small and simpler is always appropriate since crowdsourcing these microtask helps the workers to achieve the goal faster and they do it more efficiently and accurately, thus helps in improving the work/data quality.

In crowdsourcing for a given HIT many workers from the globe will work on it. Ability to access a large and diverse population is a benefit of crowdsourcing. Decomposing work into very fine units helps in completing the work efficiently. Even though crowdsourcing helps in improving data quality through advance quality detection schemes and test patterns there are still issues that affect the performance of the workers.

For example, workers want to do jobs which they think 'interesting' or 'fun' to do. If there Is a point-based leaderboards, then workers mainly focus on improving their scores rather than doing their job properly and hence affects the quality of work so data quality might be compromised. Market design issue is another issue that affects workers since AMT does not allow workers to build an identity easily so for the other party all the users/workers are same whether they are low quality workers or high-quality workers. Thus, everyone ends up receiving low salaries.