Project Description:

For an ever-growing range of decisions, individuals are explicitly or implicitly made to process others' responses before responding themselves. A judge on a multimember court may be required to account for not only the legal materials related to a case, but also the other judges' votes on that case (Posner and Vermeule 2016); online shoppers who want to make informed purchases are invariably exposed to others' product reviews when looking up product specifications; and an investor typically observes industry experts' opinions and other investors' actions while deciding on his or her own investments.

Given the prevalence of these group decision-making contexts, it is important to clarify how an individual's response to a decision could be affected by others' responses to the same decision. We therefore focus on the phenomenon of herding. Broadly, "herding" occurs when everyone does what everyone else is doing, even when private information suggests otherwise. More specifically, the classic herding model describes the following scenario: A group of people collectively chooses an outcome via sequential voting. Each member of this group has some chance of receiving private, though not necessarily correct, information about the possible outcomes they are voting for, so each member can vote according to both his or her private information and the sequence of preceding votes. As votes accumulate, the incentive of the remaining voters to rely on their private information decreases (Banerjee 1992). A better understanding of herding can improve existing guidance on group decision-making, because herding can lead to suboptimal outcomes.

There is evidence indicating that herding may be related to the difficulty, expertise, and or sentiment associated with the decision at hand (Spyrou 2013; Bobe and Piefke 2019). Intuitively, then, it is worthwhile to study herding with text, given that text can capture these and other factors. Moreover, many people are capable of training themselves to communicate in different styles; text therefore opens up new interventions for group decision-making, conditional on textual factors having a meaningful impact on herding. Our goal is to determine the relative importance of social and textual factors in guiding decisions within selected text data.

Thus, we focus on the Wikipedia Articles for Deletion (AfD) debates. As a community-driven encyclopedia, Wikipedia offers a unique platform for studying group decision-making, especially through language. Removing a Wikipedia article requires a Wikipedia user to post a nomination on AfD. Other users then post a "keep" or "delete" vote (among other options, which we omit for simplicity), along with a comment describing their rationale. They may also post non-voting comments, which are generally replies to the previous vote. The sequences of votes left by users on AfD pages, as well as their rationales, provide a rich data source for examining social decision-making. Each user's vote is likely influenced by the existing discussion, both of which are available to us. By examining the votes and comments, we can determine the extent to which users conform to the existing majority of the debate.

Several works support the existence of herding behavior in AfD. Taraborelli and Ciampaglia (2010), for example, used baseline probabilities to provide evidence of herding on early votes. Mayfield and Black (2019) later showed that language features in AfD could predict herding using a BERT-based model. We hope to further contribute by using a model to predict individual votes rather than debate outcomes in AfD; distinguishing between persuasion and herding; and partially addressing the issues of selection into debates and endogeneity of user preferences.

Data:

For this proposal, our data will come from the Wikipedia Articles for Deletion Corpus included in the Cornell Conversational Analysis Toolkit (ConvoKit). This corpus is a ConvoKit-formatted version of the data released by Mayfield and Black (2019), a collection of approximately 400,000 AfD debates that occurred between 1/1/2005 and 12/31/2018 on the English-language Wikipedia. For some exploratory analyses, we will also do additional cleaning and parsing with ConvoKit's built-in functions.

Hypotheses:

H1: Behavior that is consistent with herding is present in the Wikipedia AfD debates. Moreover, herding is more likely to occur after a threshold debate length.

We propose this hypothesis because we expect that herding-like behavior would manifest as agreement with the existing majority opinion in a given debate, or perhaps agreement with the most recent group of votes. We also suspected that this kind of behavior may only be observed after a sufficient mass of votes.

H2: Herding behavior is more pronounced for keep votes than for delete votes.

Since the deletion of an article is a more significant action than the keeping of an article, we hypothesized that herds might form more readily around majority "keep" debates. If less is at stake, users might be more likely to engage in social heuristics for decision-making rather than thinking through their opinions.

H3: Previous vote comments that are long and specific exert more influence on future votes than comments that are short, less content-rich, and more general.

We hypothesized that "stylistic" comments, as defined by the listed features, might be more influential than other comments. This is perhaps an initial distinction between persuasion and herding, as will be explained in subsequent sections.

Methods:

Going forward, we will define the sequence of previous votes that a given voter sees as the prefix of the debate.

We will investigate our hypotheses by, firstly, looking at activity statistics. Across debates, we will calculate the percentage of (k+1)th voting comments that contain a "delete" ("keep") vote given the percentage of "delete" ("keep") votes in a prefix of length k. If herding behavior exists, then we predict that a low percentage of "delete" ("keep") votes in the prefix would bias the percentage of (k+1)th votes that are also "delete" ("keep") toward zero, while a high percentage of "delete" ("keep") votes in the prefix would bias the percentage of (k+1)th votes that are also "delete" ("keep") votes toward one. This assumes that a setting without herding would exhibit an approximately one-to-one relationship between the percentage of (k+1)th votes that are “delete” (“keep”) and the percentage of “delete” (“keep”) votes in the prefix.

In a similar vein, we will also look at the probability that the (k+1)th vote agrees with the majority of the preceding k votes, as a function of k. If herding is present, we would expect to see the probability of agreeing with the majority increase with k.

Conditional on the activity statistics indicating behavior consistent with herding, we will then construct a binary logistic model of user votes using subsets of features of the preceding votes. Ideally, we will be able to use the results of this model to indicate whether early or recent vote sequences, or early or recent textual factors, have relatively stronger influences on subsequent voters.

At present, our baseline model will include the following features, though we anticipate adding additional features, such as politeness indexes, in order to make more informed decisions on distinguishing herding from persuasion and selection:

- Index in the debate of the vote to predict
- Two sets of the following features, where one set is computed for the first half of the prefix, and the other set is computed for the second half of the prefix:
    - Presence of previous delete votes
    - Presence of previous keep votes
    - Fraction of previous delete votes
    - Average values of the following features for prior “keep” and prior “delete” votes:
        - Length
        - Sentiment
        - Slang
        - External link references
        - Use of “per nom”

At this time, there are three key issues to address with respect to these methods. Firstly, we need to consider how our strategies can distinguish herding from persuasion. If our metrics show that users tend to vote according to the existing majority of AfD debates, and that our language features of interest mitigate that relationship, then we need to determine if our language features work by (dis)encouraging users to put aside their private information about the nominated article when that information conflicts with the majority. This mechanism is associated with herding, but not always with persuasion.

Persuasion is, broadly, communication where one agent has some interest in changing the behavior of another agent . Existing models of persuasion suggest that it can operate either through beliefs or independently of beliefs. The models describing the former mechanism, which DellaVigna and Gentzkow (2019) term as “belief-based models”, seem closely aligned with the concept of herding: Persuasive communications lead individuals to update their prior beliefs and act according to those updated beliefs, much like how the decision to herd and not use one’s private information is based on the probabilities individuals can infer from others’ actions. On the other hand, the models describing the latter mechanism, or “preference-based models” (DellaVigna and Gentzkow 2019), suggest that persuasive communications can alter behavior without conveying information (e.g. politeness, narrative structure, etc.).

One potential way to distinguish between herding and preference-based models of persuasion, then, may be to run our model separately on subsets of debates where voting comments' rationales offer no additional information, but vary in linguistic structure. (For example, voting comments that all vote "delete", but the accompanying rationales are all references to the same reference or previous voting comment worded in different ways.) We might expect that certain language features, like the inclusion of politeness, may better predict the probability that users vote with the majority in such subsets than in other subsets with more information in the voting comments.

The second key issue to address is selection, namely selection into debate participation and selection into vote type. Users may select into a given debate based on how much impact they believe their individual vote may have on the outcome. For example, users may be less inclined to vote at all if there is already an overwhelming majority in the debate. Conditional on participation, users may also select to vote for "keep" or "delete" based on some internal preference for "keep" or "delete", as opposed to the content of the debate itself.

To account for selection into debate participation, we can re-run our analyses on debates that begin and end within a short period of each other. (We would define a "short period" based on the distribution of timestamps in our dataset.) When voting comments are presented nearly simultaneously, it is reasonable to assume that participating voters will not have been able to base their choice to participate on an existing majority. Accounting for selection into vote type is more difficult. However, we believe we may be able to take advantage of usernames and IP address records across debates to generate a control for user features that we can incorporate into our model.

The last key issue to address is one of fundamental causality, for example, whether voting comments "aim to persuade" or "actually persuade". This, as well as the other issues, may be best addressed in a lab experiment, where we would have more control over the content of the voting comments. One potential laboratory manipulation would involve flipping the vote of (but retaining the original rationale) of a subset of voting comments, and then assessing whether or not lab participants vote differently in response. Similarly, another laboratory manipulation would present modified AfD debates to lab participants, where either the votes or the rationales are hidden from view, and then assess whether or not lab participants changed their stance towards the existing majority in those debates.

Impact:

We believe that this project may indirectly assist the administrators who determine the final outcomes of articles on AfD, especially since these outcomes are not necessarily equivalent to popular vote. The model output provides the greatest guidance in this regard, as the coefficients can point administrators towards the language features that more strongly predict behaviors consistent with herding, and therefore help the administrator attend to some rationales over others. This kind of indirect assistance may be of help in other projects involving debates, as well.

Our work on selection issues may also contribute to knowledge on new user participation and equality of participation.

More generally, our project may be useful to the 2030 Wikimedia Strategic Direction's recommendation for "improving user experience". To enable more participation in Wikimedia projects, it is necessary to understand the conditions under which users are willing to participate when they hold opinions that oppose those of existing users. Herding is an indirect contributor to those conditions.

Dissemination:

The results of this project would ideally be disseminated in conferences with OA proceedings, such as Open Access Week. We will comply with the WMF Open Access Policy and create project pages and reports as required by Wikimedia.

References:

1. Marie Christin Bobe and Martina Piefke. Why do we herd in financial contexts? Journal of Neuroscience, Psychology, and Economics, 12(2), 116–140 (2019).
2. Stefano DellaVigna and Matthew Gentzkow. Persuasion: Empirical Evidence. Annual Review of Economics, Vol. 2:643-669 (2019).
3. Elijah Mayfield and Alan W Black. Analyzing wikipedia deletion debates with a group decision-making forecast model. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–26, 2019.
4. Melissa Newham and Rune Midjord. Do Expert Panelists Herd? Evidence from FDA Committees. DIW Berlin Discussion Paper No. 1825 (2019).
5. Eric Posner and Adrian Vermeule. The Votes of Other Judges. Harvard Public Law Working Paper No. 16-04.
6. Dario Taraborelli and Giovanni Luca Ciampaglia. Beyond notability. collective deliberation on content inclusion in wikipedia. In 2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop, pages 122–125. IEEE, 2010.