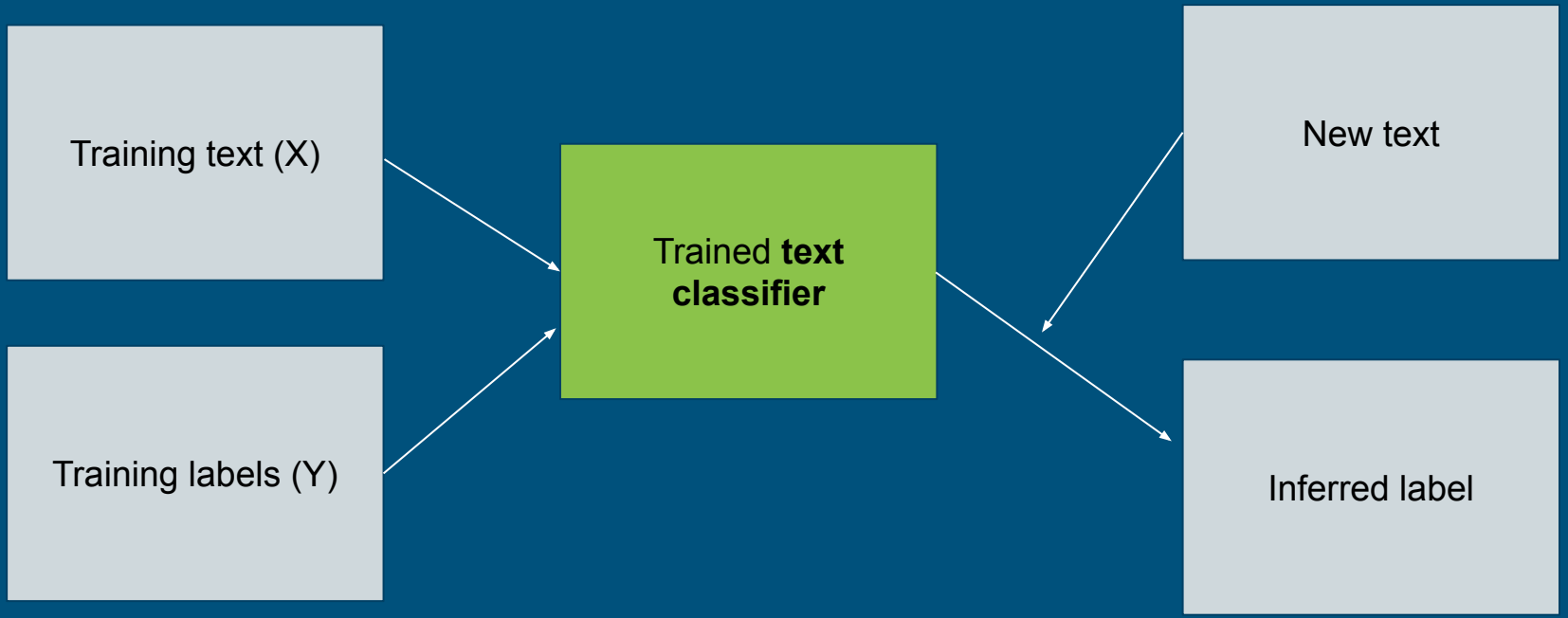
Aggregated Classification Pipelines: Propagating Probabilistic Assumptions from Start to Finish

Katherine A. Keith

NLP+CSS 201 Tutorial Series

March 30, 2022

Text classifiers



Text classifiers: two paradigms

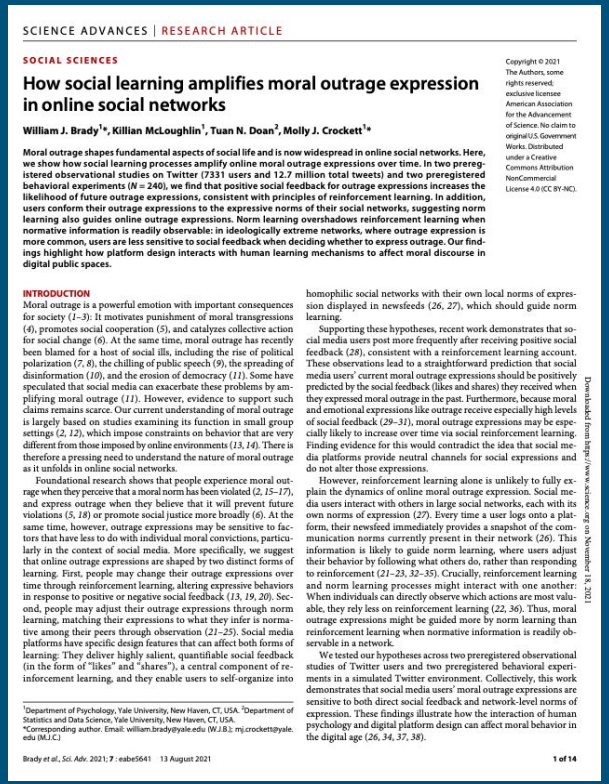
	“Conventional” NLP	**Social Sciences**
Example problem	20newsgroups label = type of text	Brady et al. 2021 label = moral outrage
Assumptions	One “true” label	• Rich social construct • Humans interpret texts different so may be natural disagreement on the labels.

Example of potentially ambiguous Tweet:
@SenBlumenthal You could be setting up the clean up of all the garbage along Route 91 instead. Wait that would take some effort!

One takeaway principle from this tutorial:
If your assumptions as a social scientist don't match those of "conventional" NLP...

- Don't give up!
- Modify the code!

Running example from *Brady et al., Science Advances 2021*



SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

How social learning amplifies moral outrage expression in online social networks

William J. Brady[1]*, Killian McLoughlin[1], Tuan N. Doan[2], Molly J. Crockett[1]*

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Moral outrage shapes fundamental aspects of social life and is now widespread in online social networks. Here, we show how social learning processes amplify online moral outrage expressions over time. In two preregistered observational studies on Twitter (7331 users and 12.7 million total tweets) and two preregistered behavioral experiments (*N* = 240), we find that positive social feedback for outrage expressions increases the likelihood of future outrage expressions, consistent with principles of reinforcement learning. In addition, users conform their outrage expressions to the expressive norms of their social networks, suggesting norm learning also guides online outrage expressions. Norm learning overshadows reinforcement learning when normative information is readily observable: in ideologically extreme networks, where outrage expression is more common, users are less sensitive to social feedback when deciding whether to express outrage. Our findings highlight how platform design interacts with human learning mechanisms to affect moral discourse in digital public spaces.

INTRODUCTION

Moral outrage is a powerful emotion with important consequences for society (*1*–*3*): It motivates punishment of moral transgressions (*4*), promotes social cooperation (*5*), and catalyzes collective action for social change (*6*). At the same time, moral outrage has recently been blamed for a host of social ills, including the rise of political polarization (*7*, *8*), the chilling of public speech (*9*), the spreading of disinformation (*10*), and the erosion of democracy (*11*). Some have speculated that social media can exacerbate these problems by amplifying moral outrage (*11*). However, evidence to support such claims remains scarce. Our current understanding of moral outrage is largely based on studies examining its function in small group settings (*2*, *12*), which impose constraints on behavior that are very different from those imposed by online environments (*13*, *14*). There is therefore a pressing need to understand the nature of moral outrage as it unfolds in online social networks.

Foundational research shows that people experience moral outrage when they perceive that a moral norm has been violated (*2*, *15*–*17*), and express outrage when they believe that it will prevent future violations (*5*, *18*) or promote social justice more broadly (*6*). At the same time, however, outrage expressions may be sensitive to factors that have less to do with individual moral convictions, particularly in the context of social media. More specifically, we suggest that online outrage expressions are shaped by two distinct forms of learning. First, people may change their outrage expressions over time through reinforcement learning, altering expressive behaviors in response to positive or negative social feedback (*13*, *19*, *20*). Second, people may adjust their outrage expressions through norm learning, matching their expressions to what they infer is normative among their peers through observation (*21*–*25*). Social media platforms have specific design features that can affect both forms of learning: They deliver highly salient, quantifiable social feedback (in the form of "likes" and "shares"), a central component of reinforcement learning, and they enable users to self-organize into homophilic social networks with their own local norms of expression displayed in newsfeeds (*26*, *27*), which should guide norm learning.

Supporting these hypotheses, recent work demonstrates that social media users post more frequently after receiving positive social feedback (*28*), consistent with a reinforcement learning account. These observations lead to a straightforward prediction that social media users' current moral outrage expressions should be positively predicted by the social feedback (likes and shares) they received when they expressed moral outrage in the past. Furthermore, because moral and emotional expressions like outrage receive especially high levels of social feedback (*29*–*31*), moral outrage expressions may be especially likely to increase over time via social reinforcement learning. Finding evidence for this would contradict the idea that social media platforms provide neutral channels for social expressions and do not alter those expressions.

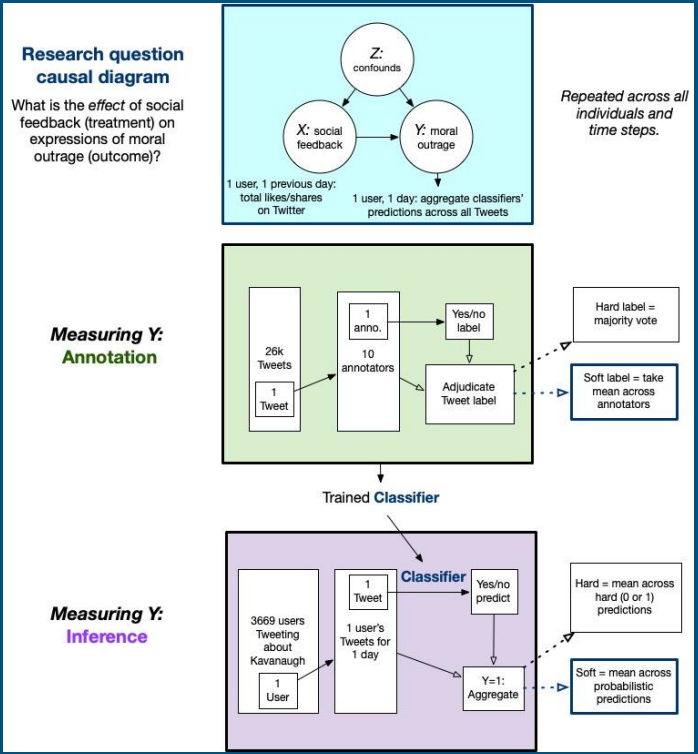
However, reinforcement learning alone is unlikely to fully explain the dynamics of online moral outrage expression. Social media users interact with others in large social networks, each with its own norms of expression (*27*). Every time a user logs onto a platform, their newsfeed immediately provides a snapshot of the communication norms currently present in their network (*26*). This information is likely to guide norm learning, where users adjust their behavior by following what others do, rather than responding to reinforcement (*21*–*23*, *32*–*35*). Crucially, reinforcement learning and norm learning processes might interact with one another. When individuals can directly observe which actions are most valuable, they rely less on reinforcement learning (*22*, *36*). Thus, moral outrage expressions might be guided more by norm learning than reinforcement learning when normative information is readily observable in a network.

We tested our hypotheses across two preregistered observational studies of Twitter users and two preregistered behavioral experiments in a simulated Twitter environment. Collectively, this work demonstrates that social media users' moral outrage expressions are sensitive to both direct social feedback and network-level norms of expression. These findings illustrate how the interaction of human psychology and digital platform design can affect moral behavior in the digital age (*26*, *34*, *37*, *38*).

[1]Department of Psychology, Yale University, New Haven, CT, USA. [2]Department of Statistics and Data Science, Yale University, New Haven, CT, USA.
*Corresponding author. Email: william.brady@yale.edu (W.J.B.); mj.crockett@yale.edu (M.J.C.)

Downloaded from http://www.science.org on November 18, 2021

Pipeline: Brady et al., Study 1



Why this paper?

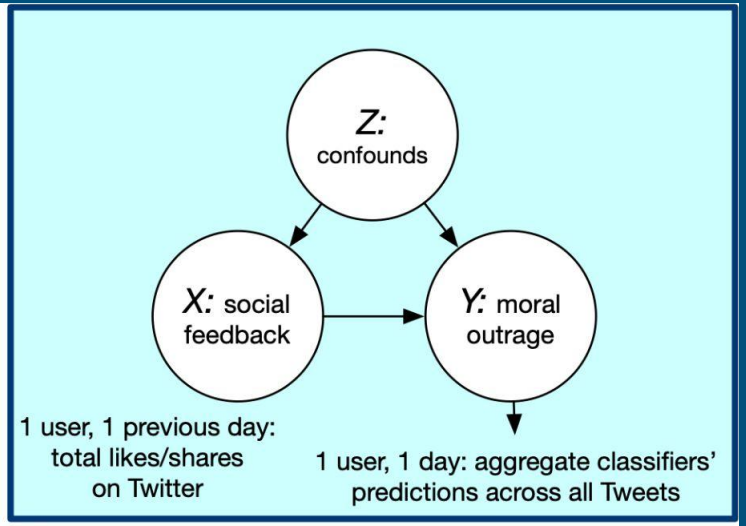
1. Social science **expertise** matters
2. Thorough data and research design
3. This tutorial provides **opportunities** for future research and is not a critique of this particular study.

Research question causal diagram

What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?

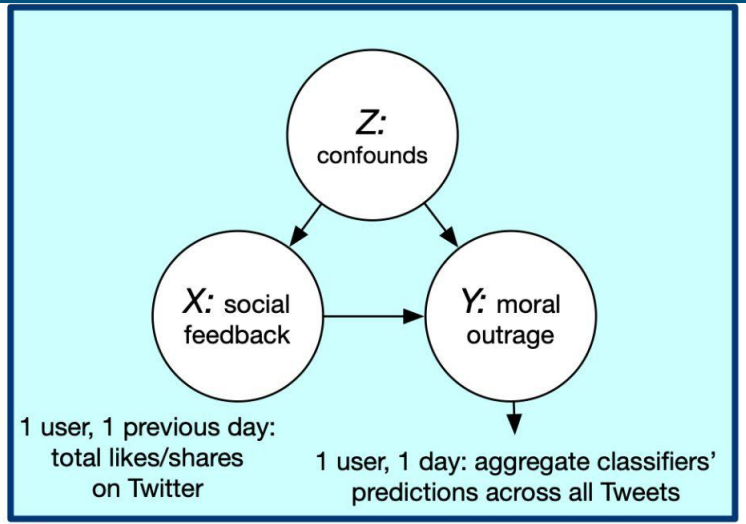
Research question causal diagram

What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?



Research question causal diagram

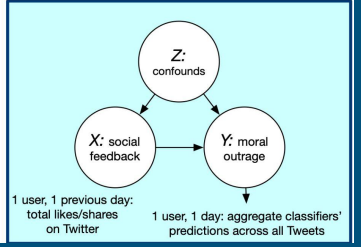
What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?



Repeated across all individuals and time steps.

Research question causal diagram

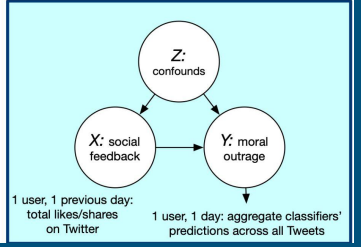
What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?



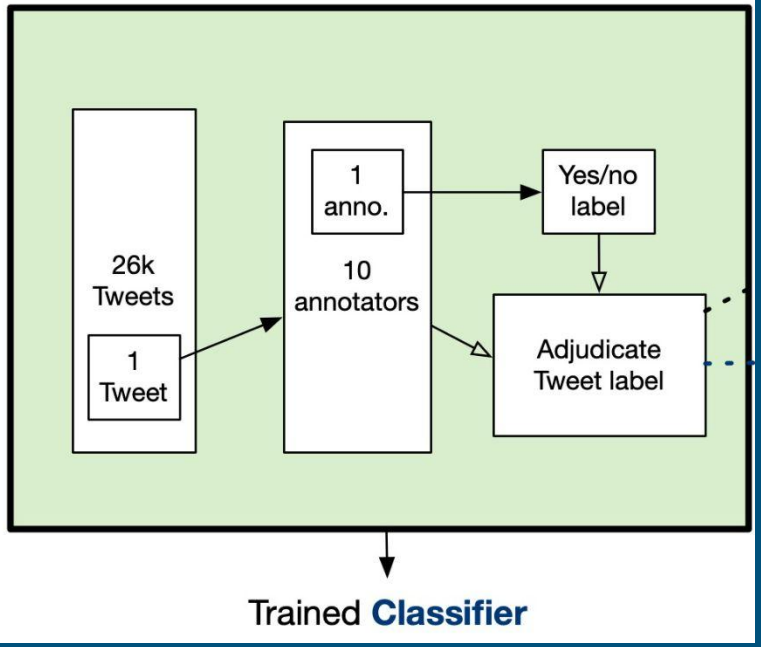
Measuring Y: Annotation

Research question causal diagram

What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?

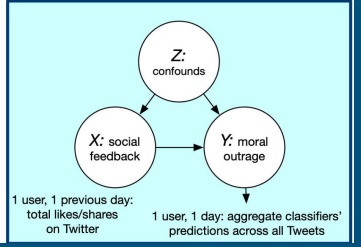


Measuring Y: Annotation

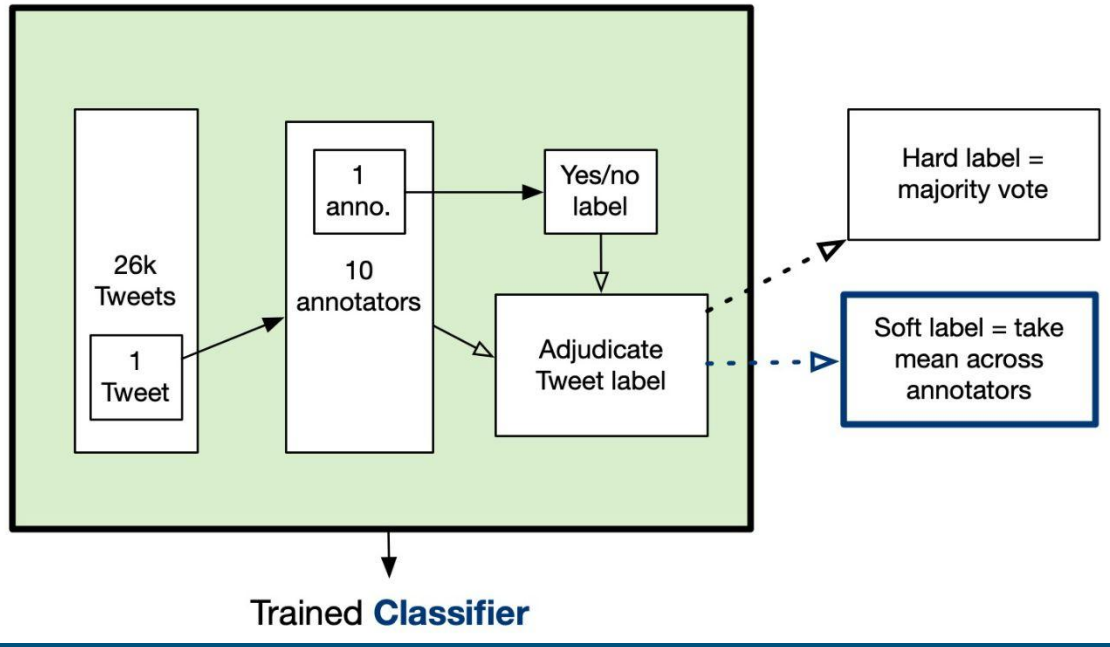


Research question causal diagram

What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?

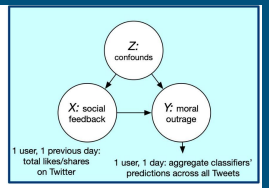


Measuring Y: Annotation

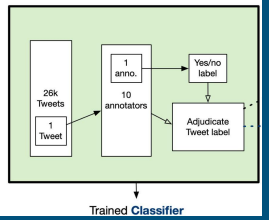


Research question causal diagram

What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?



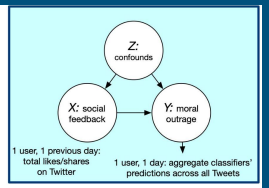
Measuring Y: Annotation



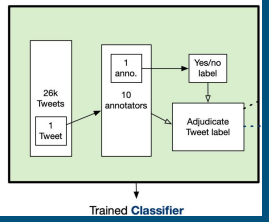
Measuring Y: Inference

Research question causal diagram

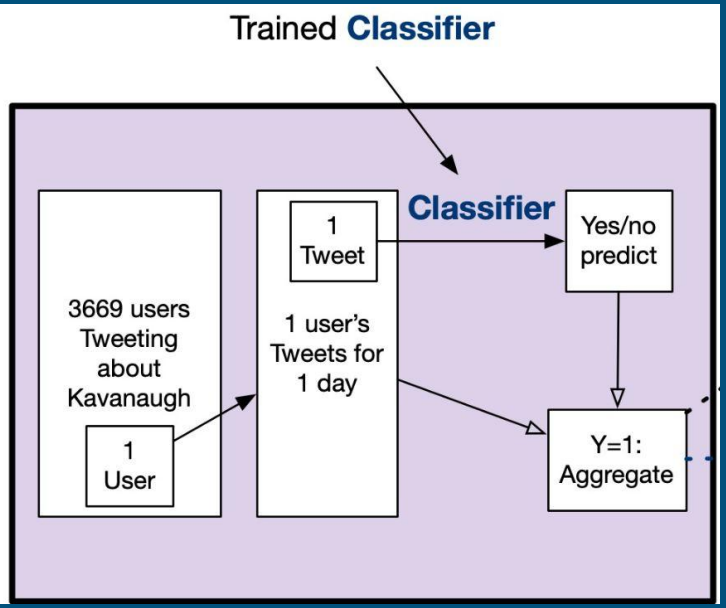
What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?



Measuring Y:
Annotation

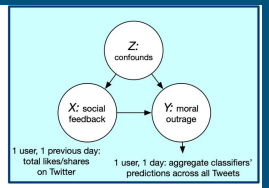


Measuring Y:
Inference

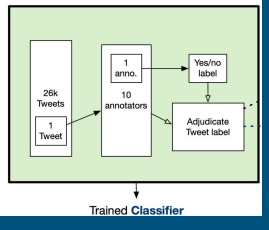


Research question causal diagram

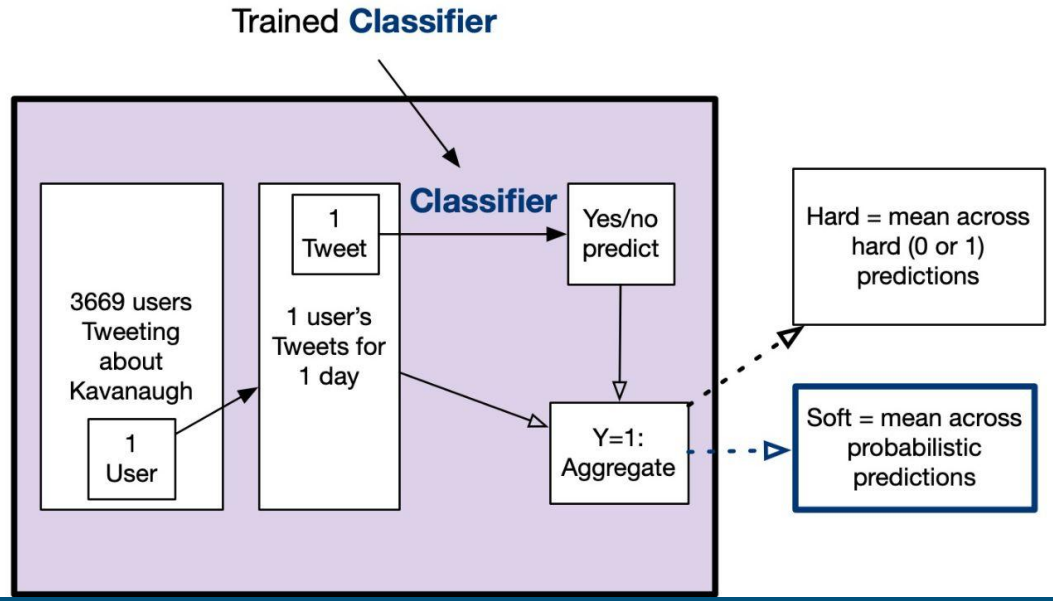
What is the *effect* of social feedback (treatment) on expressions of moral outrage (outcome)?



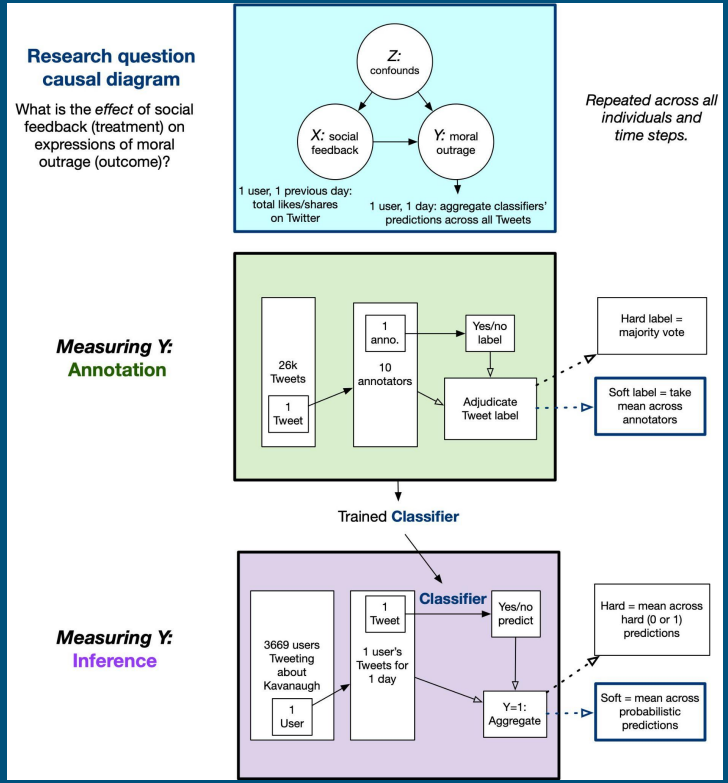
Measuring Y:
Annotation



Measuring Y:
Inference



Pipeline: Brady et al., Study 1



High level tutorial goals

1. Understand a **real-world social science research study** that has an “aggregated supervised classification pipeline”
2. **Soften (make probabilistic)** aspects of this pipeline
3. Be able to make slight **modifications** to “off-the-shelf” codebases (huggingface, sklearn)

Specific goals

	101 (prerequisites)	201 (this tutorial)
(1) Annotation	• Measure annotator agreement • Hard label = majority vote across annotators	• Assume disagreement is a useful signal • Soft label = take mean across annotations
(2) Classification: training & evaluation	• Hard classifier = Train with the hard labels	• Soft classifier = Train with Cross Entropy Loss on soft labels • Evaluate calibration error
(3) Aggregating predictions at inference time	• Mean across hard predictions at inference time (“classify and count (CC)”)	• Mean across soft predictions at inference time (“probabilistic classify and count (PCC)”)

Pointers to 301+ extensions

Code