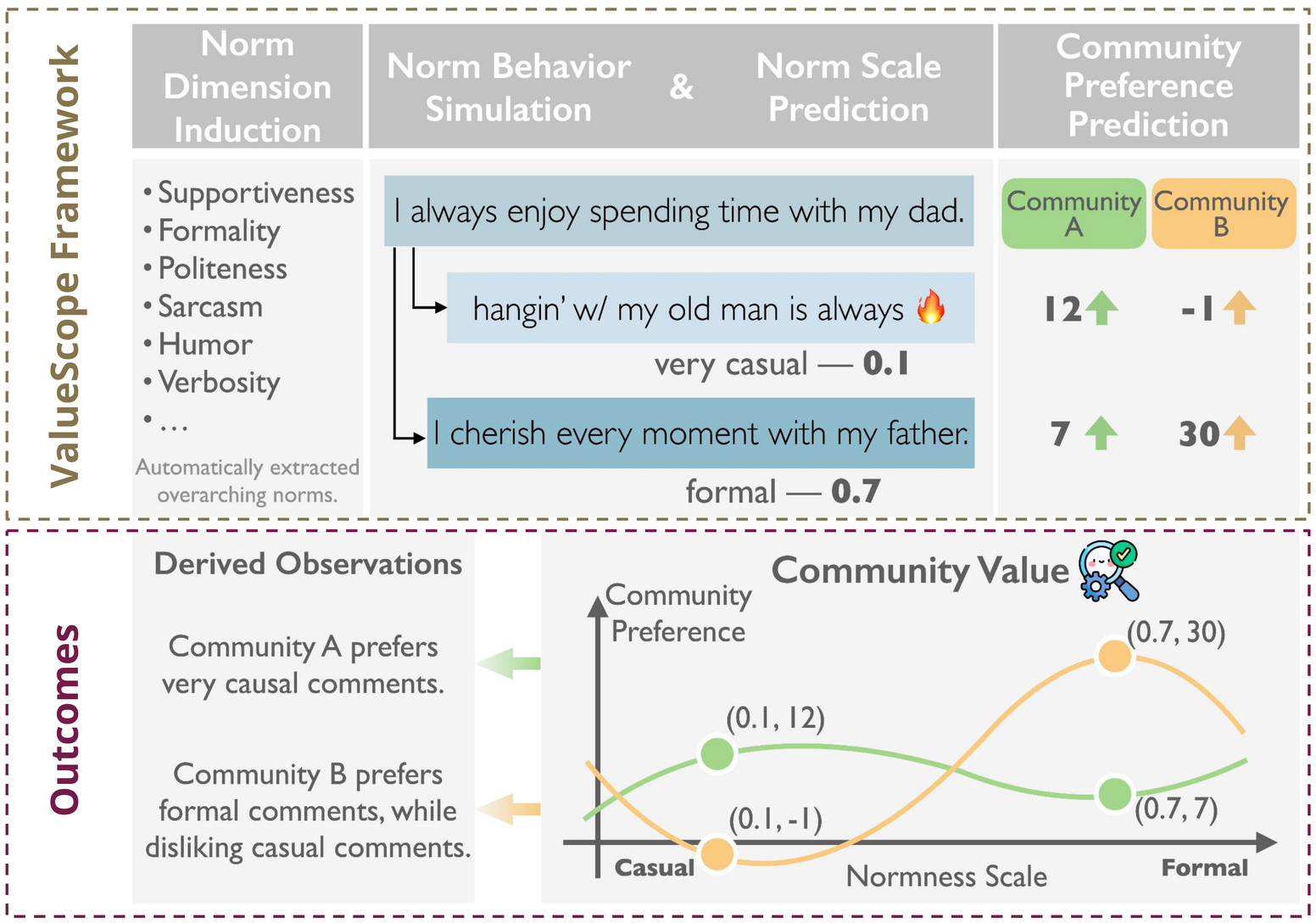
ValueScope: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions

Chan Young Park*, Shuyue Stella Li*, Hayoung Jung*, Svitlana Volkova, Tanushree Mitra, David Jurgens, Yulia Tsvetkov

{chanpark, stelli, hjung10}@cs.washington.edu

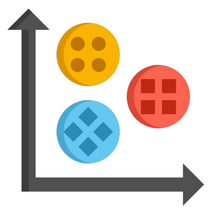
We measure implicit norms&values in community by:

1. Introducing **ValueScope**, a framework to **quantitatively decompose** social values into **social norms** and **community preferences**.
2. Studying **dynamics in social interactions** grounded in social science theory.
3. Providing quantitative evidences that **even similar communities exhibit diverse norms**, advancing practical applications of social norm studies.



Selecting Communities & Norms to Investigate

1. Cluster the top 5k subreddits by topic
2. Obtain top 5 norms per topic group
3. Compile norm–frequency ranking
4. Find the most interesting norms from the list



Sarcastic–Genuine | **Supportive–Toxic** | **Humorous–Serious** | **Polite–Rude** | **Formal–Casual** | **Concise–Verbose**

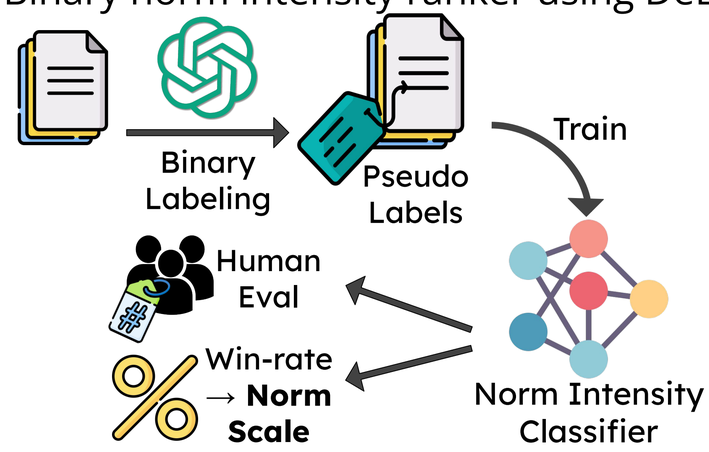
r/askmen, r/askwomen, r/asktransgender

r/stocks, r/wallstreetbets, r/pennystock

r/democrats, r/republican, r/libertarian

Quantifying Normness

Binary norm intensity ranker using DeBERTa



Quantifying Preference

DioloGPT: GPT-2 trained on 147M comments

We fine-tune DioloGPT for each subreddit

$$\mathcal{L} = -\sum_{i \in \text{batch}} \log \frac{e^{h(c_i, r_i^+)}}{e^{h(c_i, r_i^+)} + e^{h(c_i, r_i^-)}}$$

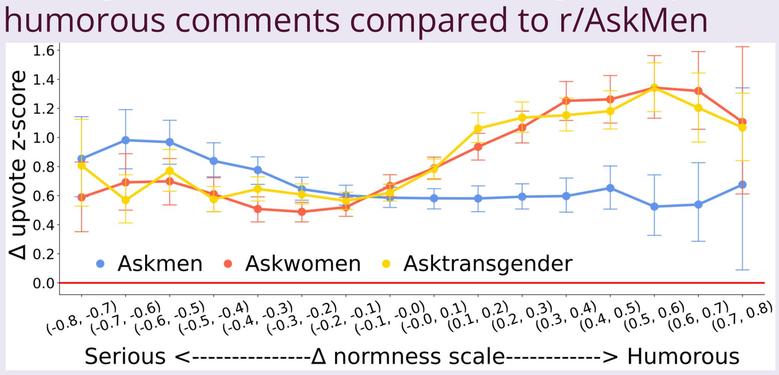
Evaluation: binary acc. on comment pairs

Downstream Applications

1. Community moderation.
2. Recommender systems on norm preference instead of topic.
3. RLC(ommunity)F: preference generator for training community specific LLMs.

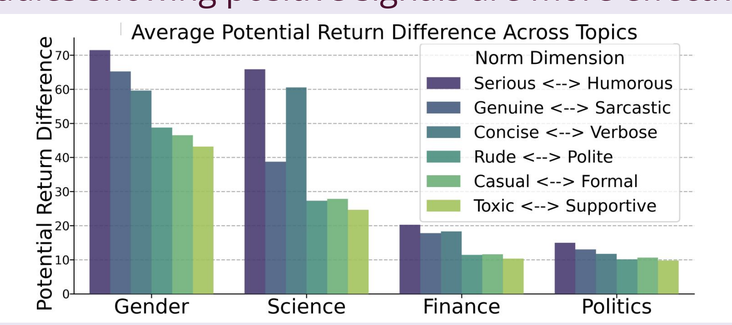
Return Potential Model Plot

Finding: r/AskWomen & r/AskTransgender prefers humorous comments compared to r/AskMen



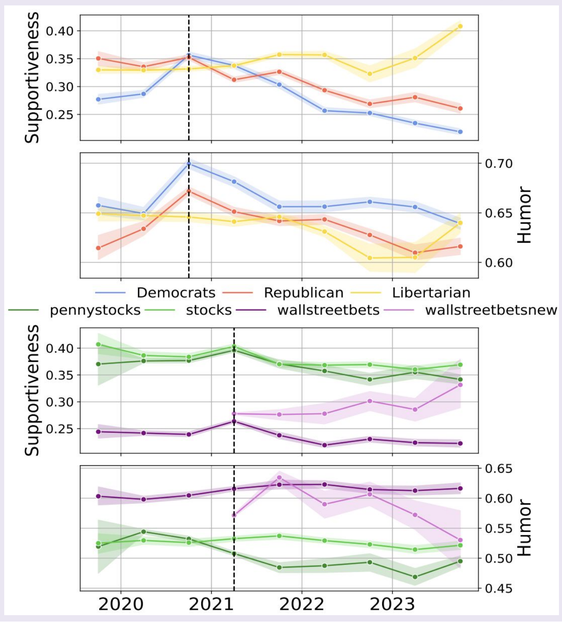
Rewards or Punishment?

Finding: Online communities tend to prioritize using rewards to regulate behaviors, aligning with studies showing positive signals are more effective.



Impacts of External Events

Findings: Norms of online communities may be influenced by external events, such as the 2020 U.S. Elections and the creation of new spinoffs.



Predicting Changes in Norms

Finding: Norms with high preference magnitude but low consensus is likely to witness upcoming change.

R^2	NI-only	NI+CR
Politeness	0.17	0.23
Supportiveness	0.04	0.10
Sarcasm	0.13	0.14
Humor	0.27	0.28
Formality	0.17	0.18
Verbosity	0.09	0.09

Norm Intensity (NI): magnitude of (dis)approval.

Crystallization (CR): level of consensus.

→ Both NI and CR are significant indicators of temporal shifts in NI