Toward Effective Automated Content Analysis Via Crowdsourcing

Jiele Wu,[b,n] Chau-Wai Wong,[n]
Xinyan Zhao,[u] Xianpeng Liu[n]

[b] Beijing Institute of Technology, China
[n] North Carolina State University, USA
[u] University of North Carolina at Chapel Hill, USA

ICME 2021

Complex Tasks are Difficult for Crowdsourcing

- Coding/annotation by crowdsourcing was shown to be effective when measuring relatively ***objective features***.

- However, latent ***subjective features*** are difficult for crowdsourcing:
  - Lack of validated tools to measure complex subjective semantic features, e.g., emotion, frame, moral reasoning.
  - Online workers' response quality tend to deteriorate as they work longer.

- <u>A Core Question: How to balance quality and efficiency in crowdsourcing coding/annotation of difficult tasks?</u>

Proposed Solution: Quality-Aware Annotation System

- Proposed quality-aware semantic annotation system:
  - **Qualifying**: Select MTurk workers who are capable of complex coding.
  - **Monitor** MTurk workers' performance and provide **feedback** over time.

- Tested the system through a task of labeling emotions of tweets related to the Flint water crisis.
  - 11 emotions: anger, disappoint, sorrow, fear, and worry, satisfied, hope, sympathy, grateful, surprise and sarcasm.[1,2]
  - We had each tweet labeled 5 times for 9,287 tweets, resulting in a total of 42,980 labels.[3]

[1] R. S. Lazarus, *Emotion and Adaption*. Oxford University Press, 1991.

[1] Y. Jin et al., "Toward a publics-driven, emotion-based conceptualization in crisis communication: Unearthing dominant emotions in multi-staged testing of the integrated crisis mapping (ICM) model," 2012.

[3] K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov, "Crowd-sourced text analysis: Reproducible and agile production of political data," 2016.

Qualifying Process

Real-Time Performance Monitoring

Qualification → **Coding/ Annotation** → **Feedback**

Qualification

1) **Training session**: Background, instructions, 5 training questions.

2) **Test session**: One is qualified if the score over 15 tweets passes a baseline.

Coding/ Annotation

1) A worker codes 20 randomly selected tweets (5 have ground-truth labels).

2) Ground-truth data (N = 100) labeled by human experts.

Feedback

1) **Quality score**: Percentage of correctly answered questions out of 5 embedded ones.

2) Must maintain cumulative quality score > 60% to work on subsequent tasks.

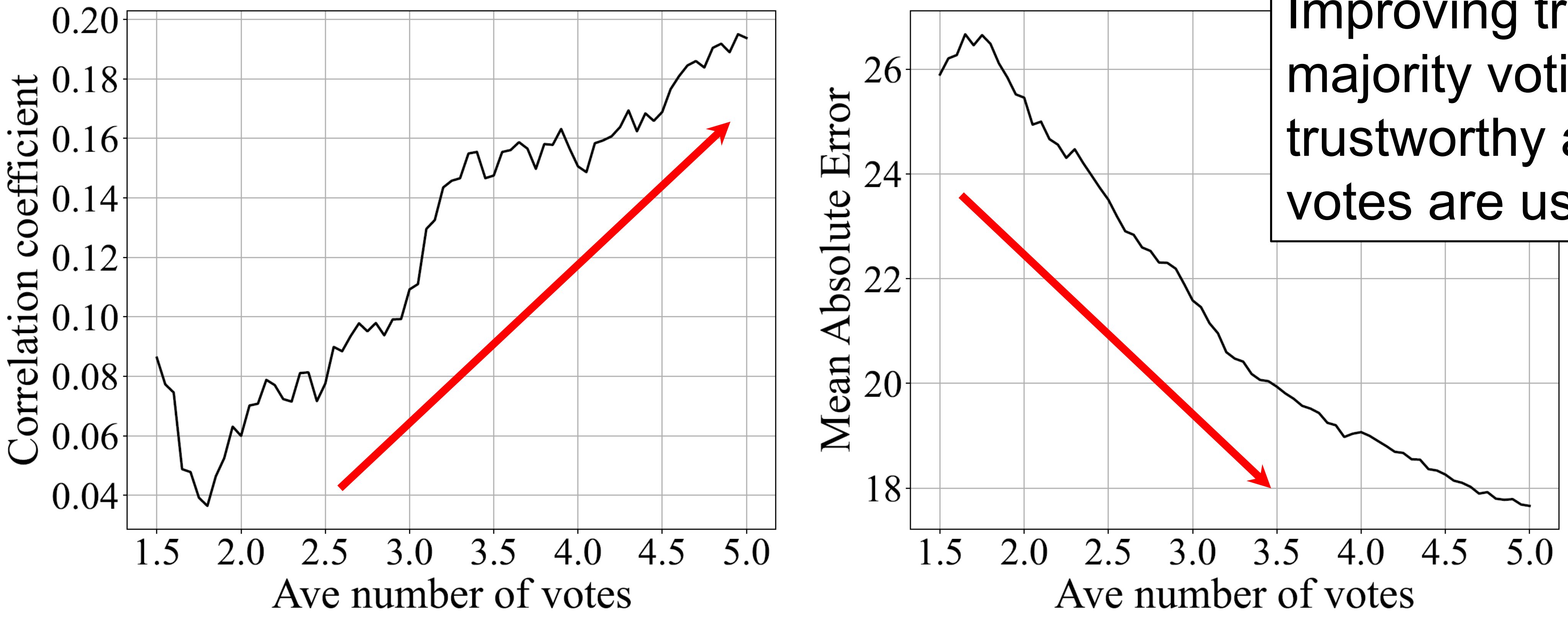
RESULTS

1. Quality Control is a Must for Complex Coding Tasks

- The qualifying process can identify eligible workers:
  - 150 out of 1,030 MTurk workers were interested in & capable of doing complex coding task.

- The real-time performance monitoring is effective in removing weak workers:
  - 11% workers could not maintain cumulative quality scores above the minimally qualifying score, 60%.
  - They were disqualified from subsequent tasks.

2. Majority Voting is Consistent with Experts Labeling



Improving trend → majority voting is more trustworthy as more votes are used.

Majority-voting quality scores evaluated on 20 tweets to be labeled.

3. Majority Voting Results Are Learnable

- We characterized the *learnability* using the generalization capability of a *powerful learning system*, e.g., a fine-tuned deep neural network.
- We show that majority-voting based labels can be learned, achieving a classification accuracy around 70%–80%.

Weight scheme	Valence		Resiliency		Attribution	
	Balanced acc		Balanced acc		Balanced acc	
	Ave	Gain	Ave	Gain	Ave	Gain
Equal weight	70.3	-	78.4	-	72.8	-
Design 1	68.8	−1.5	79.5	1.1	72.5	−0.3
Design 2	70.3	0	79.9	1.5	**75.6**	**2.8**
Design 3	**70.9**	**0.6**	**81.0**	**2.6**	73.2	0.4

Weighted voting can improve labels' quality

Discussions & Recommendations

- Challenges for labeling multiple-emotion tweets:
  - Intuitive emotions (anger) tend to mask the less intuitive ones (sarcasm).
  - Workers tend to just report one primary label rather than all emotions.
  - Solutions: i) Adapt a multiple-label task into a single-label task. ii) Craft a quality metric to encourage the discovery of secondary labels.

- Workers may unintentionally label own emotions instead of tweets’ emotions.
  - Solution: In addition to the initial training, constantly remind workers of the coding/annotation rule.

- Coding accuracy of tweets vary from 10% to 100%.
  - Solution: Select easier questions for lower performing workers.

Conclusion

- We have proposed a crowdsourcing system that can harvest a large number of high-quality labels for complex coding tasks.
- We have shown that labels aggregated based on majority voting are accurate, consistent, and learnable.

Welcome to our poster!