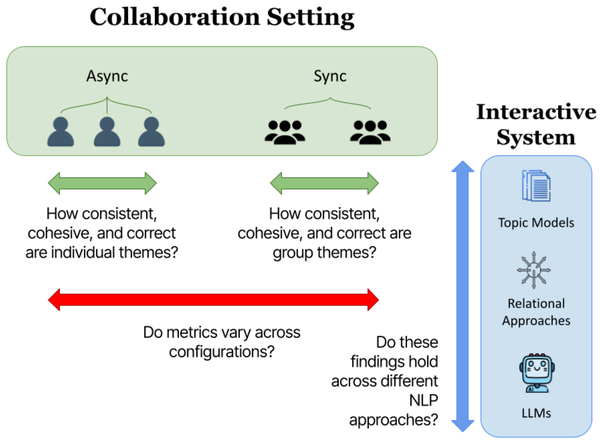
Effects of Collaboration on the Performance of Interactive Theme Discovery Systems

Alvin Chen, Dananjay Srinivas, Alexandra Barry, Maksim Seniw, Maria Pacheco

Boulder Language And Social Technologies

Motivation

NLP-assisted interactive systems enable qualitative researchers to work with extremely large datasets, but no unified framework exists for evaluating tool use across real-world settings. We created a three-pronged framework and evaluated three types of interactive system on two common collaboration settings.



Experiment Procedure

Collaboration Settings

Asynchronous collaboration: 3 coders per system look at the same subset independently with no communication.

Synchronous collaboration: Coders work together in real time without prior review. We recruit 2 groups of 3 coders for each system. No inter-group communication.

Interactive Systems

Topic Model: Fang et al. 2023's interface that allows merging, splitting, and editing topics using topic words.

Relational Approaches: Pacheco et al. 2023's system that learns user-generated labels based on good/bad examples and rules based induction.

LLM-Based System: Chew et al. 2023's framework for iteratively improving prompts to help LLMs label documents based on human preferences.

Coding Protocol

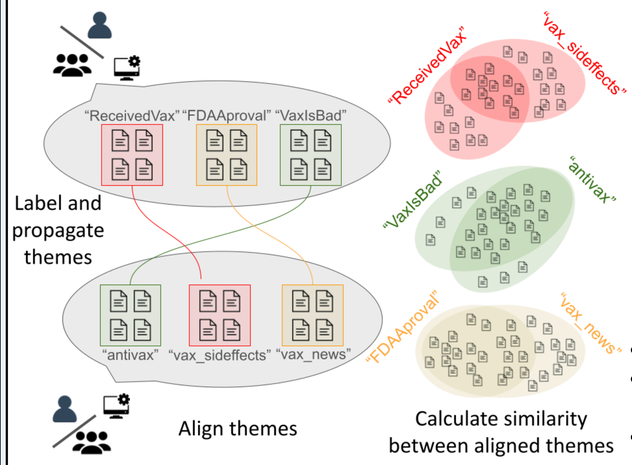
Dataset: 85k tweets about Covid-19 vaccine.

Procedure:

- Using the same starting clusters, participants look through each cluster to identify themes.
- For topic model, users use editing operations iteratively to get an ideal model.
- For the other two, users give an example subset which is used to generalize to the whole dataset.

Results and Evaluation

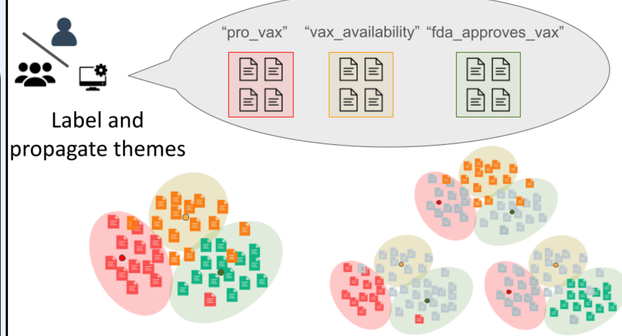
Consistency: Do different annotators/groups find the same themes?



System	Setting	Jaccard	Centroid	Group Avg.
Relational	Sync	**0.36 (0.19)**	**0.98 (0.01)***	**0.52 (0.07)***
	Async	0.30 (0.22)	0.94 (0.07)*	0.44 (0.10)*
LLM-Based	Sync	0.14 (0.08)	0.98 (0.03)	0.44 (0.03)
	Async	**0.17 (0.11)**	0.98 (0.02)	**0.45 (0.03)**
Topic Model	Sync	**0.56 (0.23)**	**0.98 (0.05)****	**0.52 (0.10)**
	Async	0.30 (0.17)	0.96 (0.05)**	0.51 (0.09)

- Sync collaboration finds more consisten themes
- Centroid and Group Average similarity tests produce more significant results
- LLM-based experiments fail all tests of significance

Cohesiveness & Distinctiveness: Does the same annotator/group find unique themes that are different from other themes they found?



System	Setting	All Documents		Top 25% Closest	
		Intra-Theme	Inter-Theme	Intra-Theme	Inter-Theme
Relational	Sync	**0.51 (0.08)***	0.42 (0.05)*	**0.70 (0.09)***	0.52 (0.07)*
	Async	0.45 (0.10)*	**0.34 (0.11)***	0.64 (0.09)*	**0.46 (0.13)***
LLM-Based	Sync	**0.44 (0.06)**	0.40 (0.04)	0.63 (0.07)	0.55 (0.05)
	Async	0.43 (0.05)	**0.39 (0.04)**	0.63 (0.05)	**0.54 (0.05)**
Topic Model	Sync	0.52 (0.10)	0.40 (0.04)	0.56 (0.11)	0.39 (0.05)
	Async	0.52 (0.10)	0.40 (0.04)	0.56 (0.11)	0.39 (0.05)

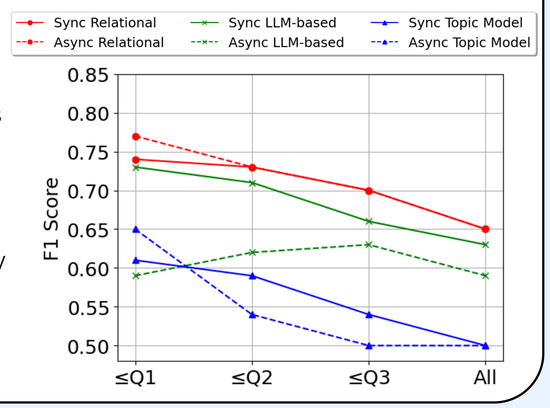
- Sync themes more cohesive but not always more distinctive
- Top 25% closest documents more distinctly show the same results

Correctness: How accurately does the interactive system assign annotations?

- Randomly select 200 documents from each experiment
- Documents evenly distributed by themes and theme similarity (50 per quartile)
- Two annotators check for correctness with a third for tie-breaks
- Overall Krippendorff's Alpha is 0.632

Findings

- Relational system has the highest correctness F1; consistent between sync and async experiments. Other systems vary, likely due to less induction biases.
- Models vary in robustness as documents become less like human-labeled examples.



User Study

- Participants found working in teams to ease the coding process.
- Asynchronous coders were more likely to notice inadequacies of the tool they used.
- Annotators noted the lack of control and agency that the topic modeling system.
- Annotators felt that starting topics often negatively impacted their annotations.

Design Recommendations

Consider collaboration dynamics when designing tools

- Many types of real world collaboration
- Different tools are better for different settings
- Tools need to be evaluated on different settings to provide useful performance metrics

User agency impacts experience and outcome

- Topic model based tools provide less agency
- Perceived agency impacts user experience
- Responsiveness to user input impacts model generalization

LLM-based systems have room for innovation

- LLM inference is expensive and unreliable
- New methods for conditioning LLMs are needed
- LLMs are useful for qualitative coding outside of automating annotation

Conclusions

RQ1: Does collaboration setting measurably affect the quality of resulting code-books?
Yes, across each dimension we evaluate on.

RQ2: Do these findings hold across different NLP approaches?
Maybe, all systems show differences in at least one dimension. Future work should consider other dimensions.

Based on measurable differences, researchers should consider experimental settings and user agency when designing interactive systems.