Characterizing the Language of Online Communities and Its Relation to Community Reception

UNIVERSITY OF WASHINGTON
ELECTRICAL ENGINEERING

Trang Tran and Mari Ostendorf
{ttmt001,ostendor}@uw.edu

Transformation, Interpretation & Analysis of Language

Background & Motivation

- Online communities
  - Platforms for online interactions: comments sections, discussion forums
  - Often organized around a common topic
  - Prior work has shown that users tend to write in the style of the community
- Our focus: role of ***style*** vs. ***topic*** in
  - Community identity
  - Community endorsement
  - Language of people involved in multiple communities
- Data: Reddit
  - Thousands of communities
  - Quantifiable reader endorsements available

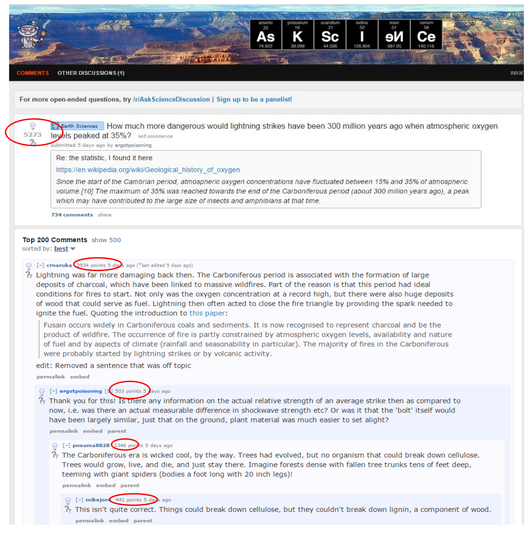
Modeling Style vs. Topic

- Style: hybrid word-POS tags trigram LMs, motivated by work in genre detection
  - word_only: 156K words
  - tag_only: 38 POS tags
  - hybrid_15K: 15K most common words + 38 POS tags
  - hybrid_530: 500 most common words (balanced) + 30 next most common words from each subreddit
    - ➔ vocab = 854 words + 38 POS tags
- Topic: LDA-{100, 200}
- Caveat: can't completely separate style/topic

Subreddit Statistics

subreddit	# posts	# comments	% k ≤ 0
askmen	4.5 K	1.1 M	10.6
askscience	0.9 K	0.3 M	9.1
askwomen	3.6 K	0.8 M	7.5
atheism	3.1 K	1.0 M	15.2
changemyview	2.3 K	0.5 M	16.7
fitness	2.4 K	0.9 M	8.6
politics	4.9 K	2.2 M	20.8
worldnews	9.9 K	6.0 M	23.6
merged_others	28.0 K	14.2 M	13.2

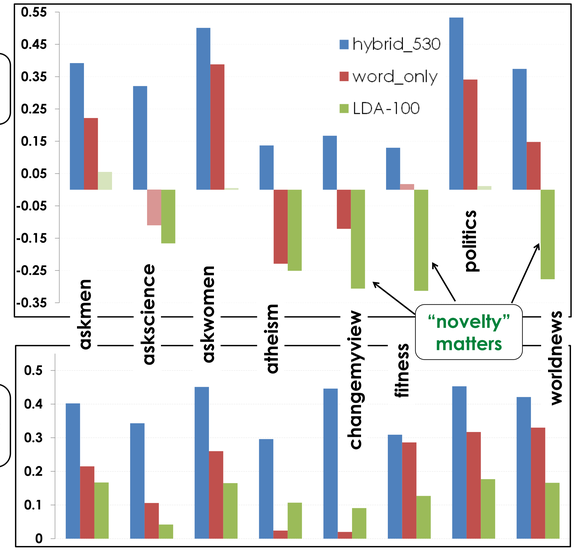
merged_others: constructed from 9 other subreddits: books, chicago, nyc, seattle, nfl, science, running, explainlikeimfive, todayilearned (76K – 5M comments)



Experiment Setup & Scoring

- Document d_i: group of comments by {thread, author}
- Scoring $s_{i,j}$: similarity between d_i and subreddit j
  - Style: $p(d_i \mid LM_j)$
  - Topic: $\frac{1}{3}(sim_{i,j,[1]} + sim_{i,j,[2]} + sim_{i,j,[3]})$
- Community classification: $\hat{j} = arg\max_j s_{i,j}$
- Correlation analysis:
  - Normalize scores: $s_{i,j} - s_{merged\_others}$
  - Spearman rank correlation with: {thread's karma, author's k-index}

Correlation with Thread's Karma



Correlation with Author's k-index

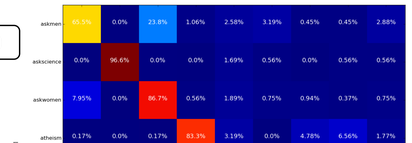
Average Classification Accuracy

Model	Threads	Authors
tag_only	27.6%	18.8%
hybrid_530	**86.5%**	**51.0%**
hybrid_15k	69.4%	46.6%
word_only	68.9%	46.8%
LDA-100	71.5%	27.5%
LDA-200	69.6%	27.7%

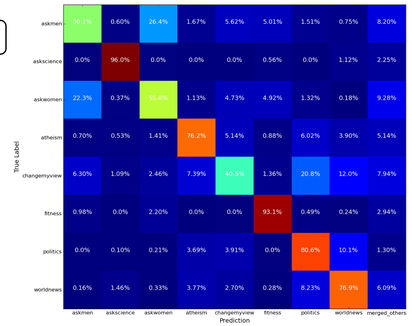
Fewer words

Thread Classification Confusion Matrices

hybrid-530



LDA-100



Questions

- How accurately can we identify a community using style vs. topic?
- What is the best computational model of style?
  - Decide based on performance in classifying community that a discussion comes from
- Do people/discussions that conform more to "community norms" tend to be more highly endorsed?
  - Assess using correlation with author k-index and discussion thread karma
  - For style match vs. topic match
- What are the participation patterns of successful multi-community participants?
  - Is a successful user successful in multiple communities?

Analysis & Findings

- Community identity:
  - Best results for style: POS + general words + community jargon
  - word n-gram ≈ topic models ➔ style is more indicative
  - Users' comments: more topically diverse and therefore harder to classify
- Community endorsement:
  - More positively correlated with style than topic
  - Model with best classification accuracy = model most positively correlated with endorsement
- Users tend to "specialize": among active users (100+ comments)
  - Max k-index ≤ 5: median 6 subreddits
  - Max k-index ≥ 100: median 3 subreddits; only 4 out of 42 have another 50+ k-index