Social Topic Distributions

Hakan Akyürek

Mürüvvet Hasanbaşoğlu

July 13, 2020

Overview

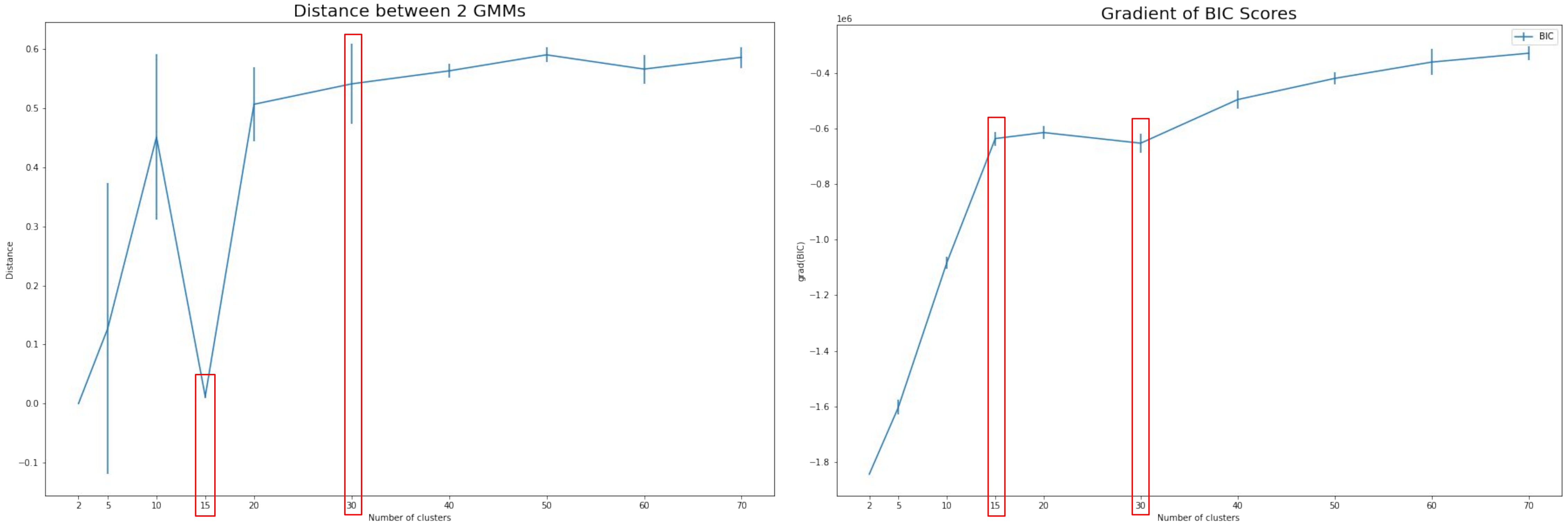
1. **Sentence embeddings**
   - **GMM training**
   - **Topic labeling**

2. **User Clustering**

3. **Experiments**

4. **Next Tasks -** for the next weeks

1. Sentence Embeddings



➔ Train the GMM sentence embeddings models for n_component=15 and n_component=30 using the entire dataset

1. Sentence Embeddings

Topic Labeling:

1. Predict the cluster of all sentence embeddings
2. Compute the Clarity Score

Clarity Scoring: $$\text{score}_a(w) = t_a(w) \log_2 \frac{t_a(w)}{t(w)}$$

```
vectorizer = TfidfVectorizer(norm='l1', dtype=np.float32, stop_words='english')
vectorizer.fit(list_of_sentences)
tw = vectorizer.transform([' '.join(list_of_sentences)]).toarray()[0]
tw_word_dict = dict(zip(vectorizer.get_feature_names(), tw))
```

```
taw_per_cluster = []
for cluster, sentences in clustered_sentences.items():
  taw = vectorizer.transform([' '.join(sentences)]).toarray()[0]
  taw_word_dict = dict(zip(vectorizer.get_feature_names(), taw))
  taw_per_cluster.append(taw_word_dict)
```

1. Sentence Embeddings

Top 10 seed words of GMM n_component=15 model:

1. food, eat, eating, meat, foods, diet, healthy, vegan, meals, meal
2. milk, cheese, sugar, chocolate, butter, cream, sauce, chicken, oil, tomatoes
3. people, money, need, let, change, going, vote, time, stop, way
4. new, city, park, year, restaurant, university, center, 000, wine, chicago
5. stores, store, buy, grocery, products, market, foods, shop, buying, walmart
6. organic, pesticides, food, foods, conventional, certified, produce, non, farming, buy
7. http, com, www, org, st, https, ave, 10, los, 11
8. people, like, want, time, need, think, going, know, really, make
9. god, science, evidence, logic, scientific, argument, facts, believe, existence, truth
10. know, think, really, read, wrong, question, answer, right, good, post
11. farmers, farm, farming, farms, agriculture, crops, farmer, soil, crop, grow
12. cancer, chemicals, water, disease, health, vaccines, toxic, chemical, fda, bacteria
13. government, tax, corporations, money, people, federal, taxes, capitalism, country, america
14. gmo, monsanto, gmos, genetically, labeling, modified, non, crops, seeds, corn
15. trump, obama, republicans, president, republican, democrats, liberals, hillary, gop, clinton

1. Sentence Embeddings

Top 10 seed words of GMM n_component=30 model:

1. food, eat, eating, foods, healthy, meals, meal, nutrition, cooking, processed
2. cheese, sugar, sauce, butter, milk, chocolate, salad, cream, bread, sweet
3. people, let, time, money, change, world, going, need, life, work
4. michael, john, david, ben, jerry, george, chris, paul, mike, justin
5. stores, store, grocery, foods, buy, products, shop, walmart, shopping, amazon
6. money, tax, pay, taxes, jobs, income, capitalism, wage, economy, wealth
7. organic, food, foods, certified, conventional, buy, organics, produce, non, usda
8. com, http, www, st, org, 10, ave, 30, 11, saturday
9. park, city, restaurant, new, chicago, wine, chef, county, center, york
10. want, care, money, need, make, know, pay, people, kill, control
11. water, drink, milk, coffee, drinking, tea, like, plastic, oil, bottle
12. people, want, like, time, think, going, need, know, really, things
13. stupid, ignorant, sad, ignorance, idiot, dumb, bad, fucking, stupidity, shame
14. cancer, health, disease, vaccines, medical, diseases, vaccine, fda, medicine, antibiotics
15. argument, evidence, logic, facts, truth, true, false, proof, arguments, prove
16. article, read, post, link, thanks, thank, information, reading, comments, comment
17. know, think, really, good, right, like, sure, thing, agree, wrong
18. farmers, farm, farming, farms, agriculture, farmer, crops, agricultural, local, land
19. pesticides, pesticide, herbicides, organic, fertilizers, use, glyphosate, toxic, herbicide, chemicals
20. process, technology, company, research, data, development, percent, information, study, results
21. climate, warming, global, water, earth, carbon, co2, change, planet, fracking
22. plants, plant, bees, garden, soil, seeds, grow, compost, weeds, seed
23. israel, war, jews, hamas, religion, gay, palestinians, people, marriage, rights
24. meat, vegan, cows, animals, chickens, beef, animal, fed, vegetarian, eat
25. science, god, scientific, scientists, universe, evolution, existence, theory, evidence, logic
26. republicans, liberals, republican, liberal, democrats, conservatives, gop, conservative, party, democrat
27. gmo, gmos, genetically, non, labeling, modified, foods, crops, label, corn
28. government, federal, congress, vote, politicians, constitution, state, corporations, democracy, political
29. trump, obama, president, hillary, clinton, bernie, bush, sanders, romney, donald
30. monsanto, fda, seeds, gmo, bayer, dow, seed, roundup, companies, evil

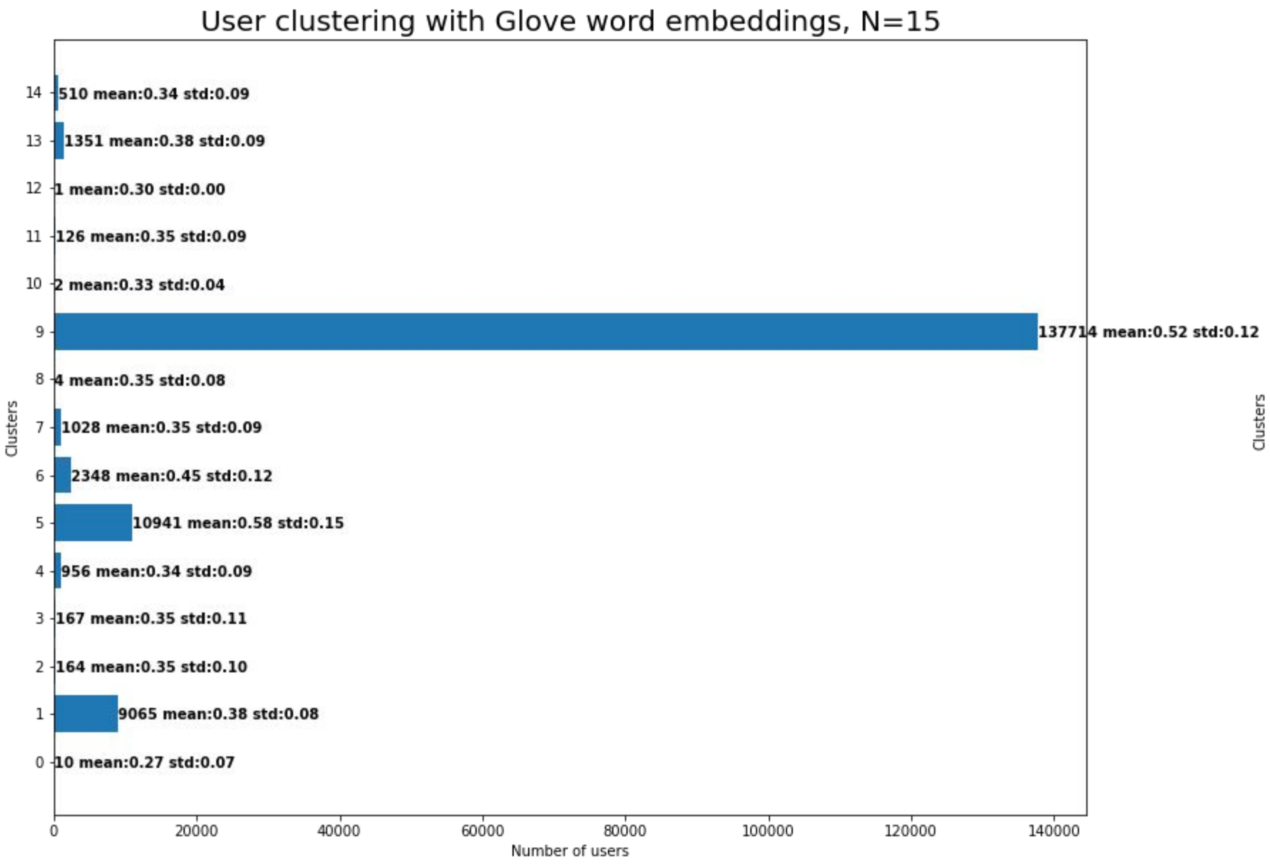
2. User Clustering

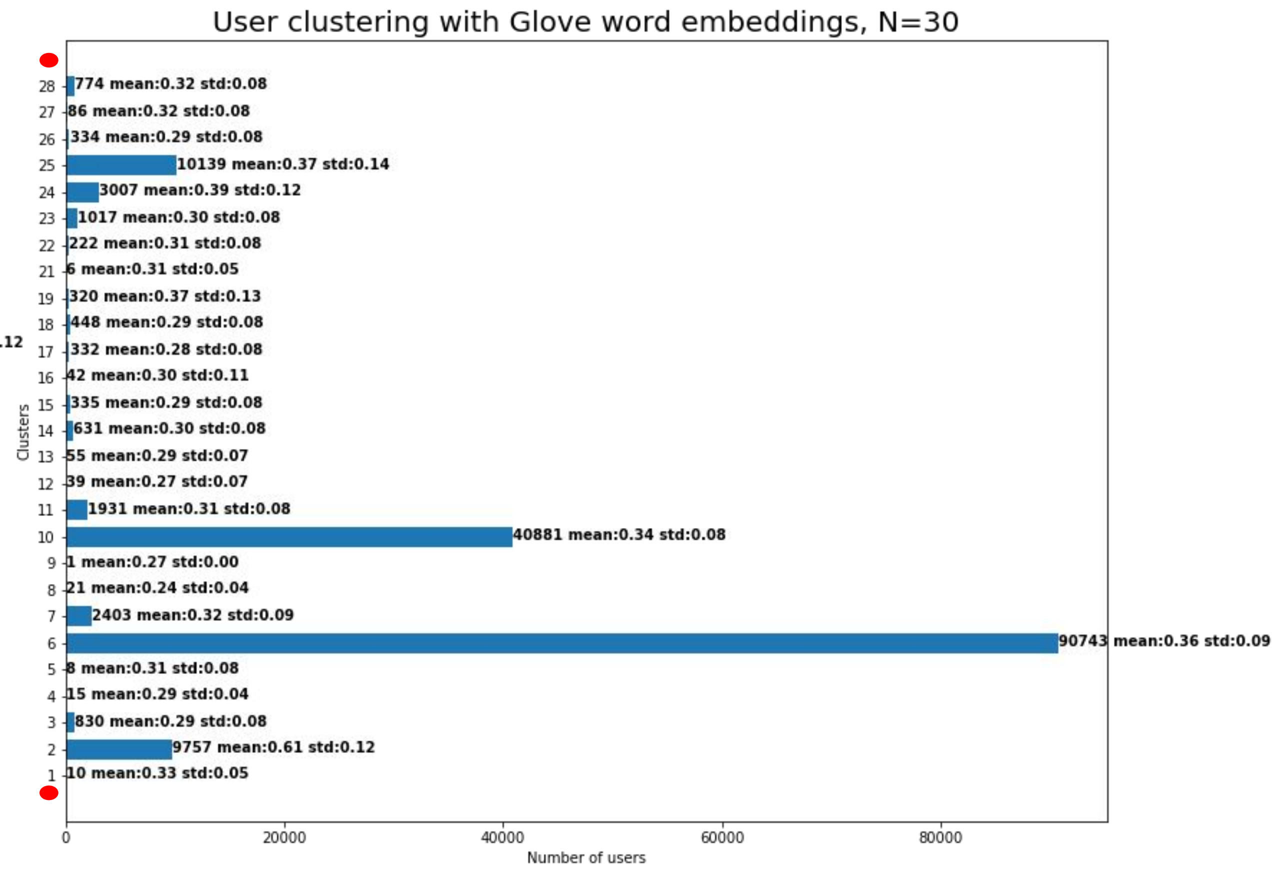
1. **Define the user dataset:**

```
{
    dataset_name: {
        author_id+’$$’+author_name: concatenated user comments
    }
}
```

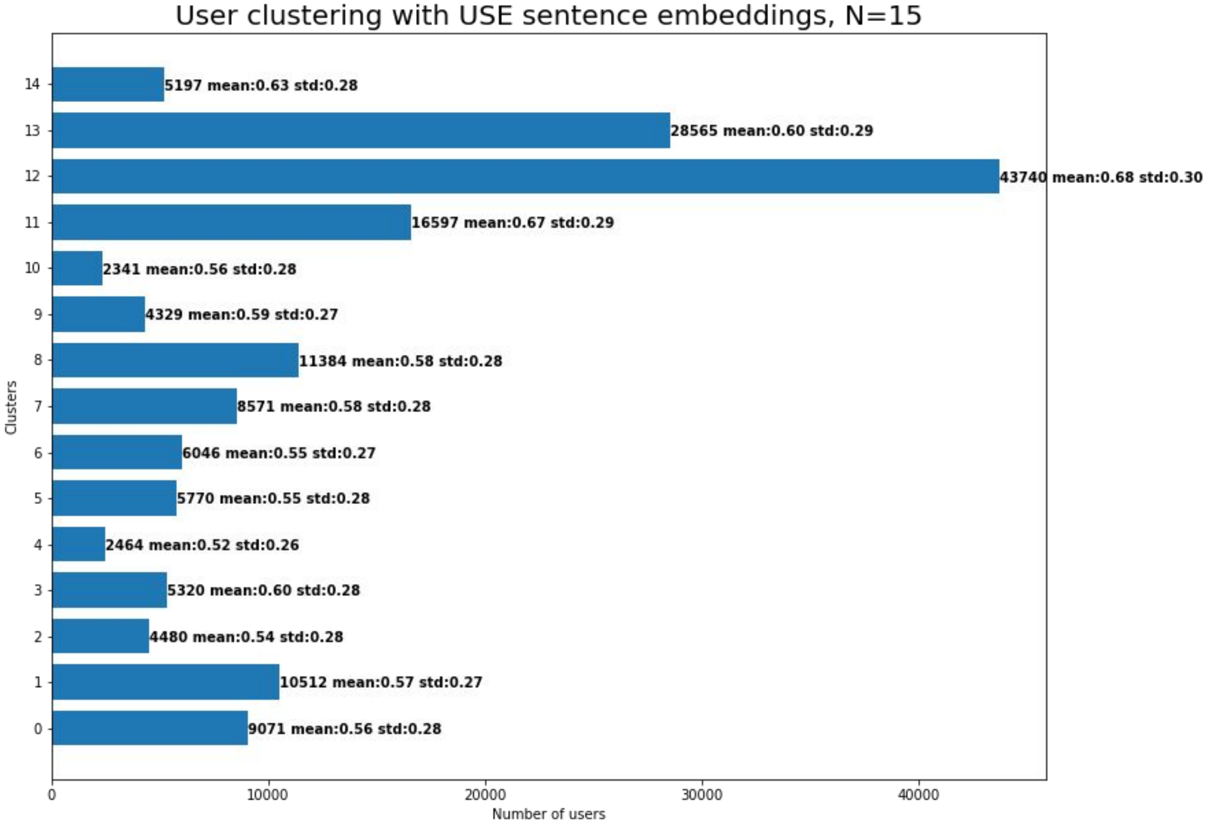
2. **Compute probabilities of Users**
    a. Get probability distribution of Users per cluster
    b. Max probability cluster is the cluster that the User belongs to

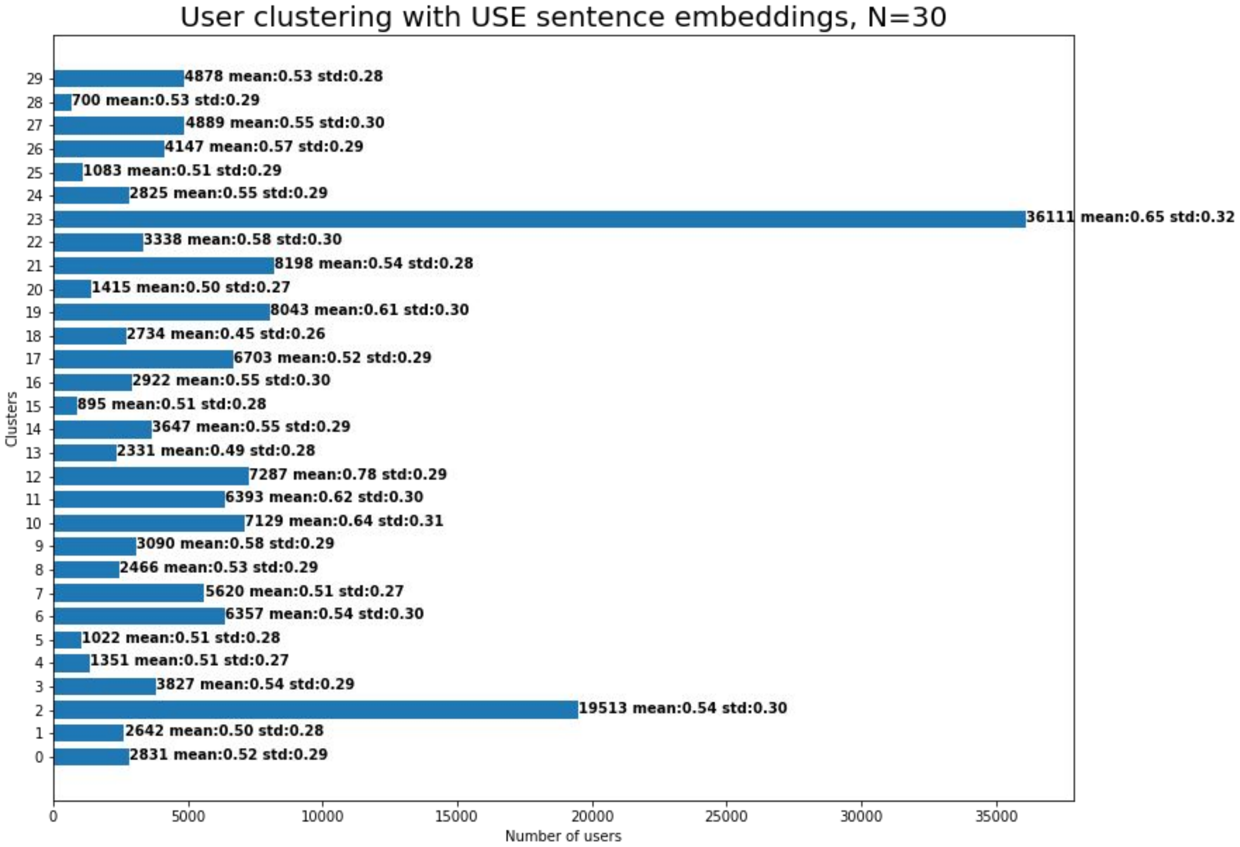
2. User Clustering





2. User Clustering





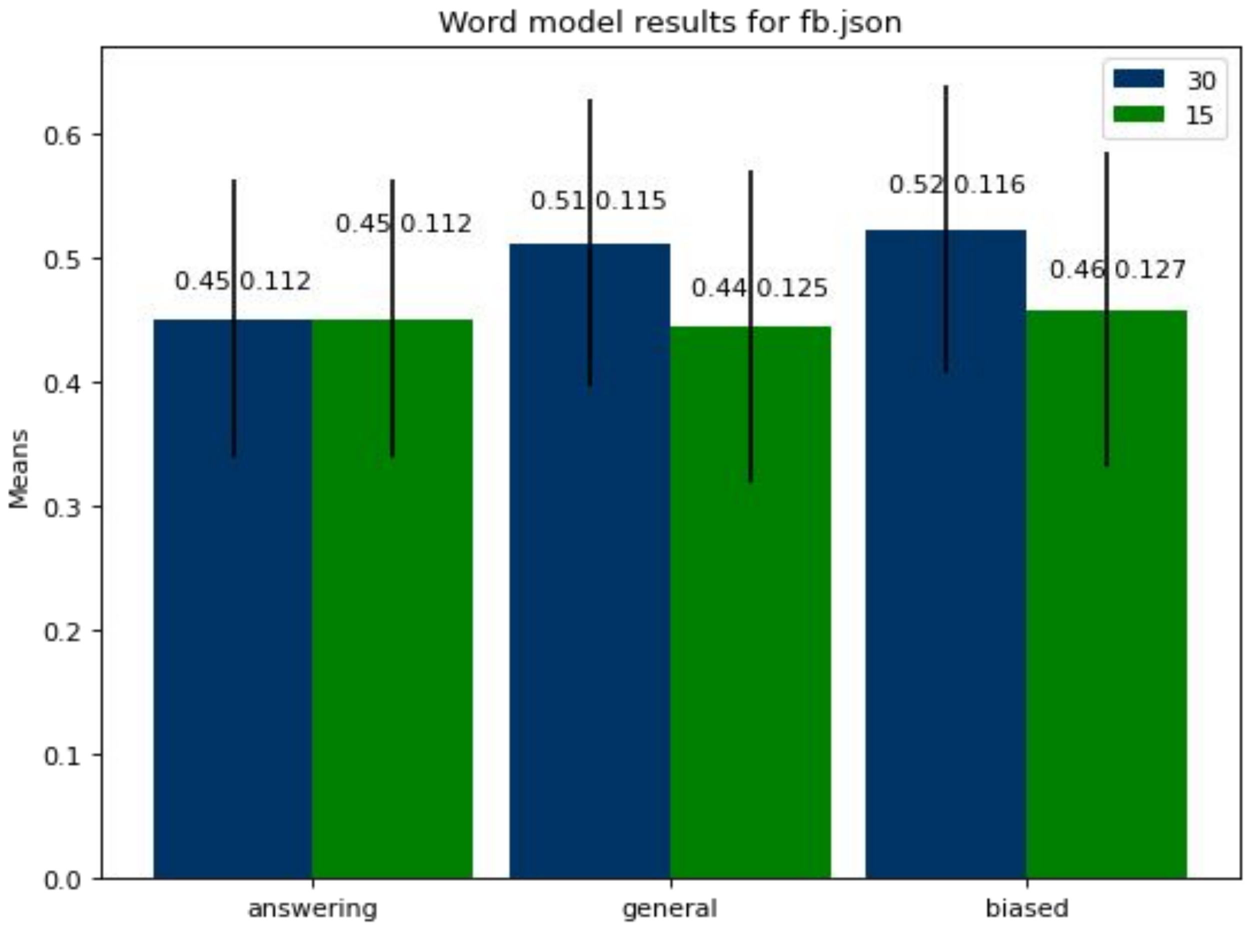
3. Experiments

- We migrated all experiments to user basis from comment basis.
  - Comments for each user is concatenated
- 3 different datasets; Facebook, Quora, Washington post
  - User counts mostly aren't enough in the others
- 2 models, more incoming next week: GMM's with word embeddings, 15 and 30 clusters
- 5 random user sets; general, biased, unbiased, forum, news
- Predefined articles, selected randomly
  - 150 articles per dataset

Data sets			# of articles	# of articles with comment	# of relative comments
Biased		Facebook	5013	4705	298996
		Food Babe	15	15	3944
		Food Revolution	78	60	2966
		Organic Authority	66	0	0
		Organic Consumers	64	0	0
Unbiased	Forum	Cafe Mom	86	85	1962
		Disqus	36	36	6150
		Quora	567	523	4196
		Reddit	81	78	2371
		US Message Board	0	0	0
	Newssites	Chicago Tribune	2283	78	281
		Huffington Post	880	0	0
		LA Times	1522	77	374
		NY Post	106	0	0
		NY Times	438	137	16128
		USA Today	95	22	259
		Washington Post	1563	943	84669

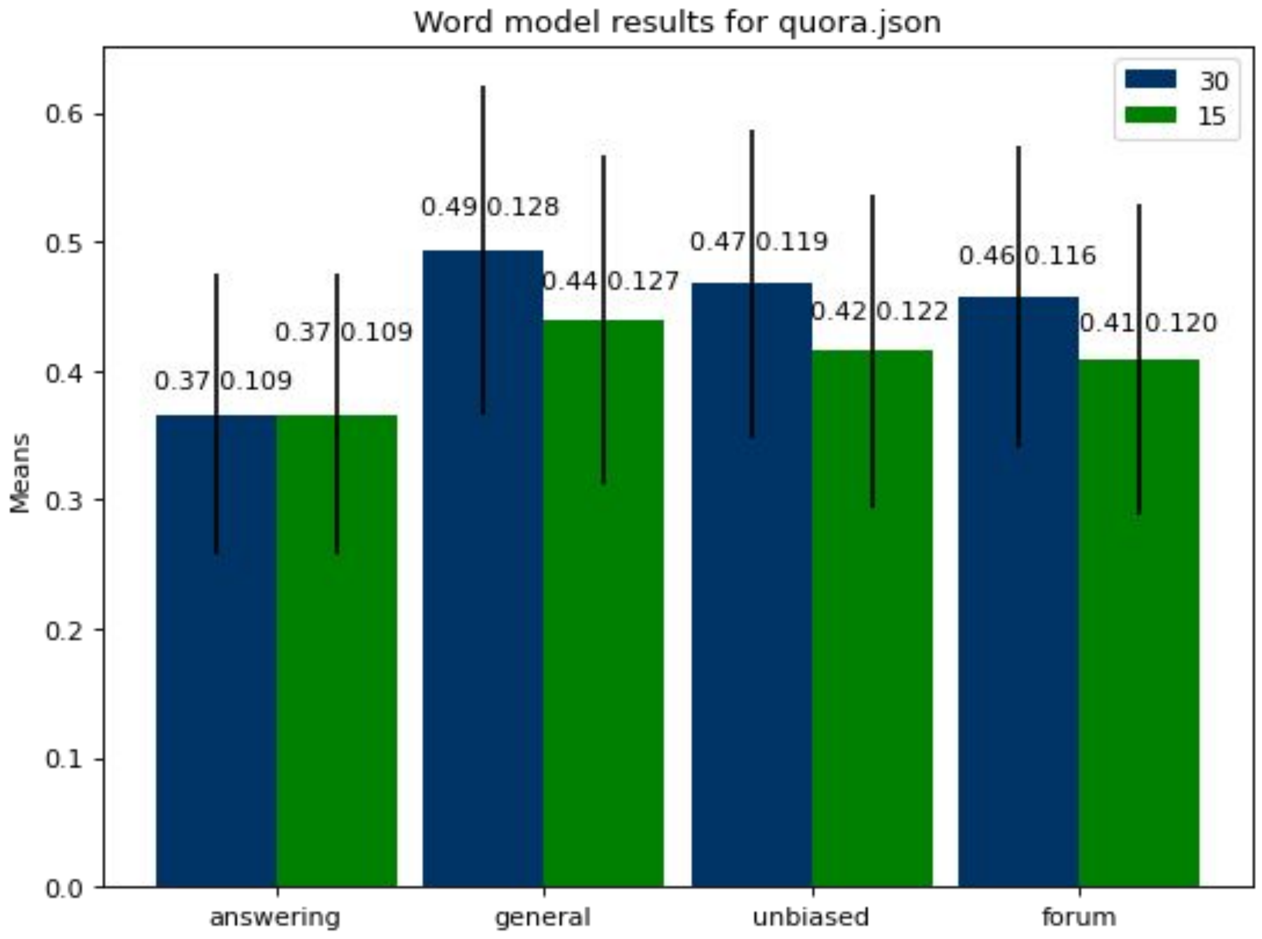
3. Experiments

- Facebook
  - Got the worst results.
  - Got even worse scores in biased subset.
  - While mean is highest, std in average is lowest.



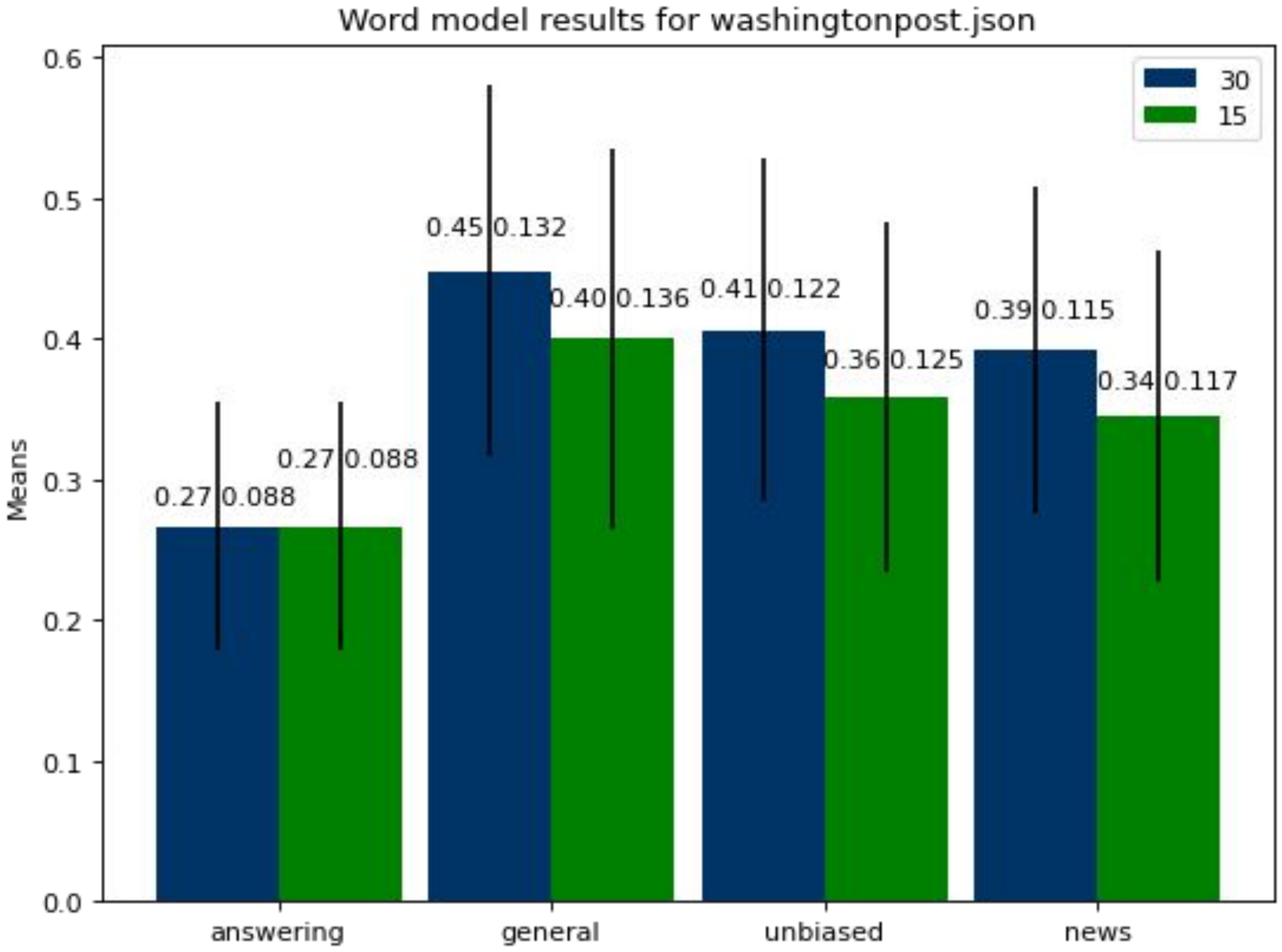
3. Experiments

- Quora
  - Got better results in unbiased subset and even better results in forum subset.
  - Users tend to stick to the article topic more.
  - Got the closest result to answering users.



3. Experiments

- Washington post
  - Got the best results.
  - Std is higher in model with 15 clusters.



3. Experiments

- Generally GMM models with 15 clusters return less mean but higher standard deviation.
- Sentence experiments take a long time, because of embedding calculation. However, early results show that sentence embeddings have larger distances.
- The expectation is actually fulfilled for users when word embeddings are used.
- Standard deviations tend to become smaller as the random user set becomes more specific.

4. Next Tasks

- Tidy up the code in the notebook
- Prepare for the final presentation
- Start working on the report

Questions

References

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
- Sridhar, Vivek Kumar Rangarajan. "Unsupervised topic modeling for short texts using distributed representations of words." *Proceedings of the 1st workshop on vector space modeling for natural language processing*. 2015.
- Fine tune GloVe embeddings using Mittens
- roamanalytics/mittens: A fast implementation of GloVe, with optional retrofitting
- https://medium.com/analytics-vidhya/basics-of-using-pre-trained-glove-vectors-in-python-d38905f356db
- https://scikit-learn.org/stable/modules/mixture.html
- https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4
- https://stackoverflow.com/questions/26079881/kl-divergence-of-two-gmms
- https://medium.com/@sourcedexter/how-to-find-the-similarity-between-two-probability-distributions-using-python-a7546e90a08d
- Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Sung, Y. H. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Angelidis, Stefanos, and Mirella Lapata. "Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised." *arXiv preprint arXiv:1808.08858* (2018).