Week 2 Sampling, Crowd-sourcing & Reliability

Nak Won Rim

20 tweets related to COVID-19

+

Dodds et al. (2015)'s method

(9 → 5 scale)

10 coders

+

1 “deviant” coder

Please evaluate…

the sentiment of the tweet itself
vs
your feeling toward the tweet

10 coders

+

1 “deviant” coder

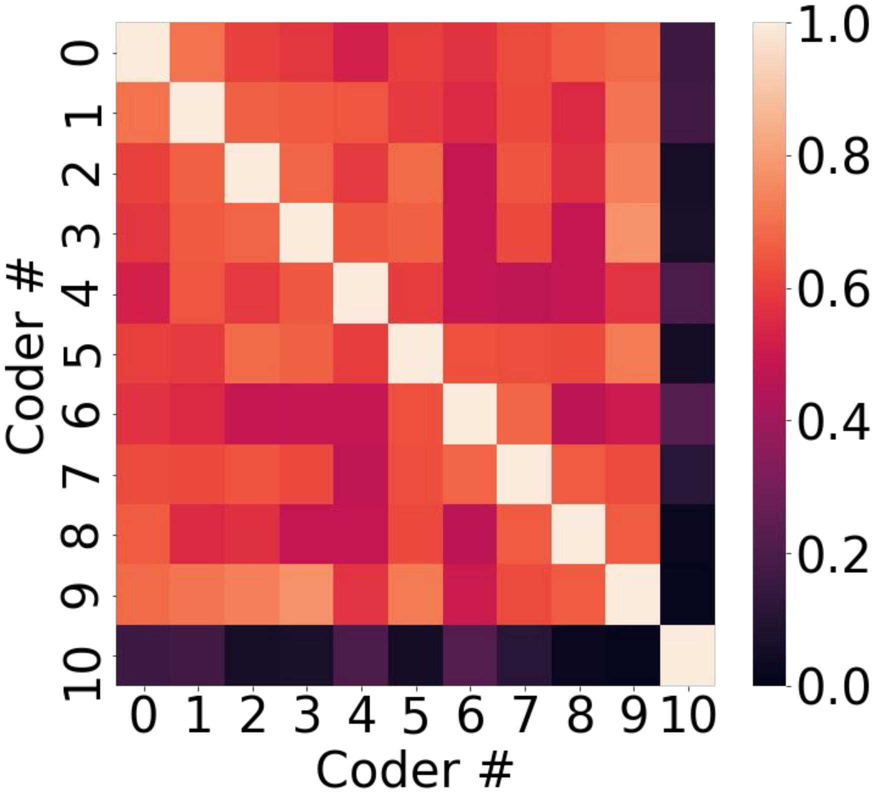
Please evaluate…

the sentiment of the tweet itself

vs

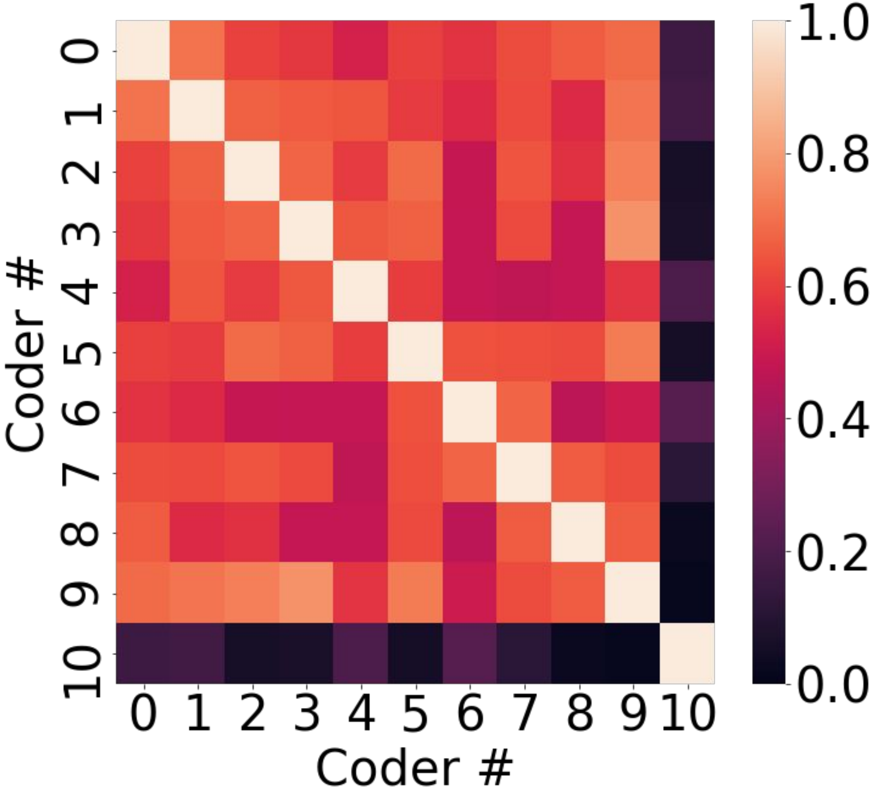
your feeling toward the tweet

Cohen's weighted κ



—

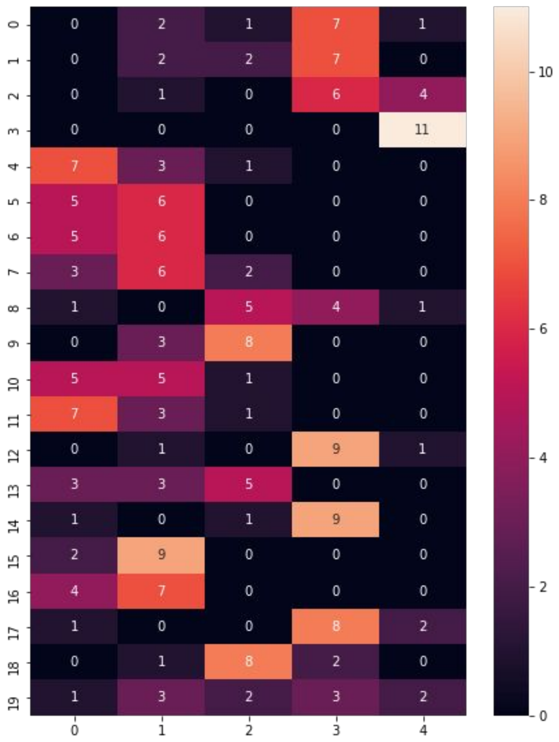
Cohen's weighted κ

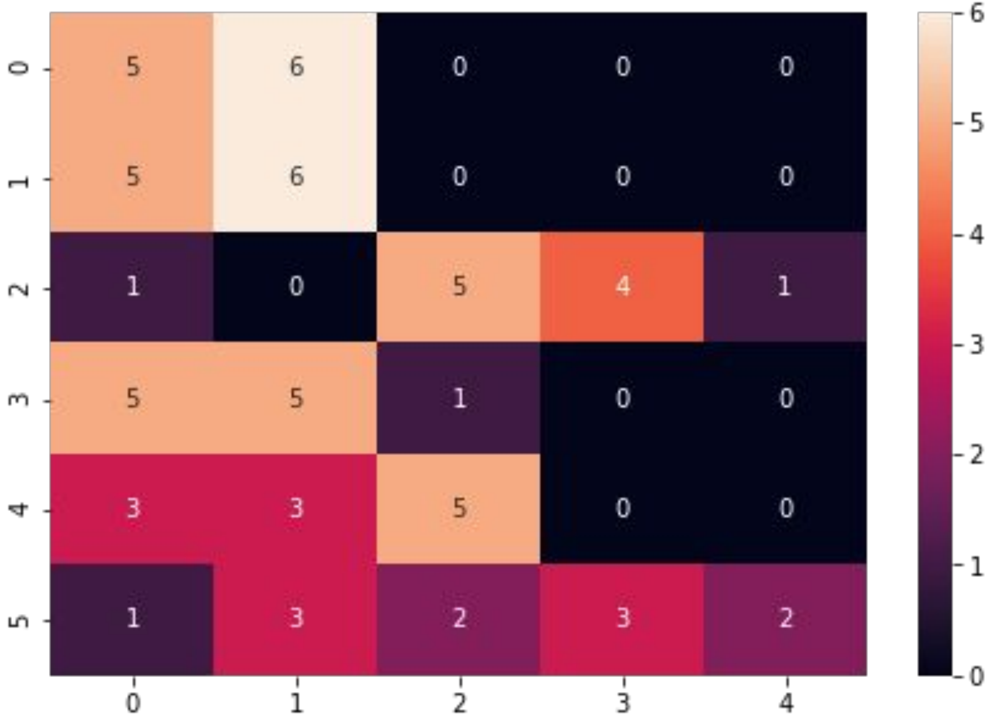


	With Deviant Coder	Without Deviant Coder
Cohen's weighted κ (averaged)	.52	.61
Krippendorff's α	.66	.77

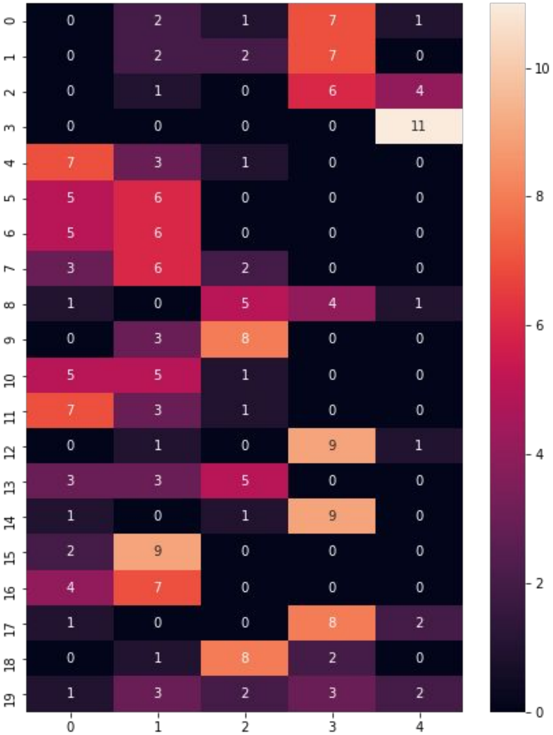
→ One “troll” can decrease the score quite a lot!

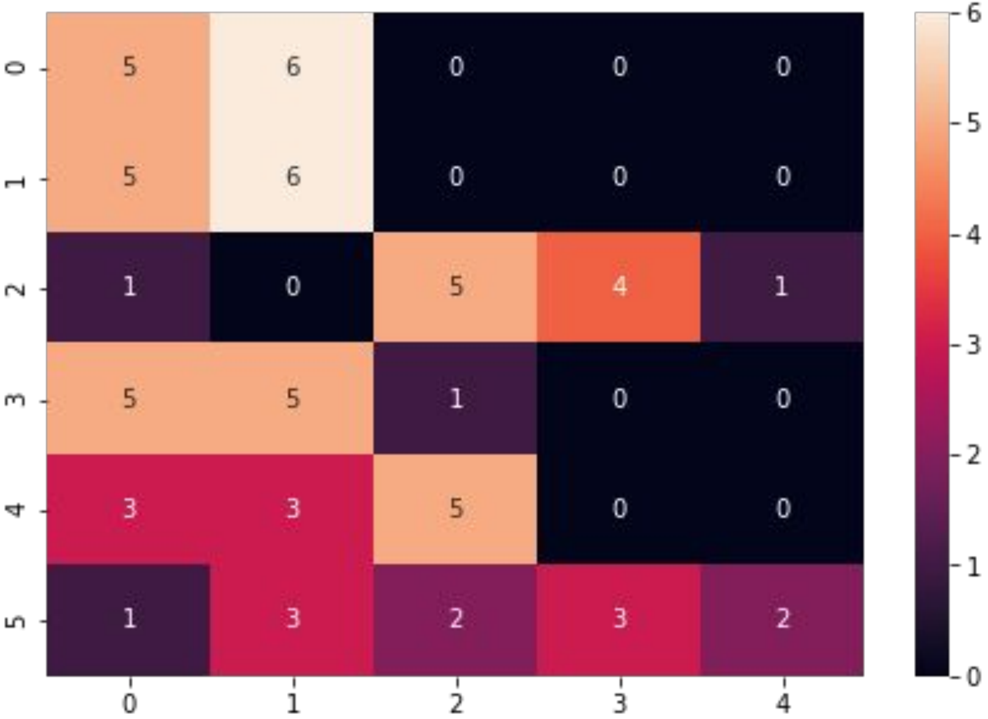
Problem in vote majority





Problem in vote majority



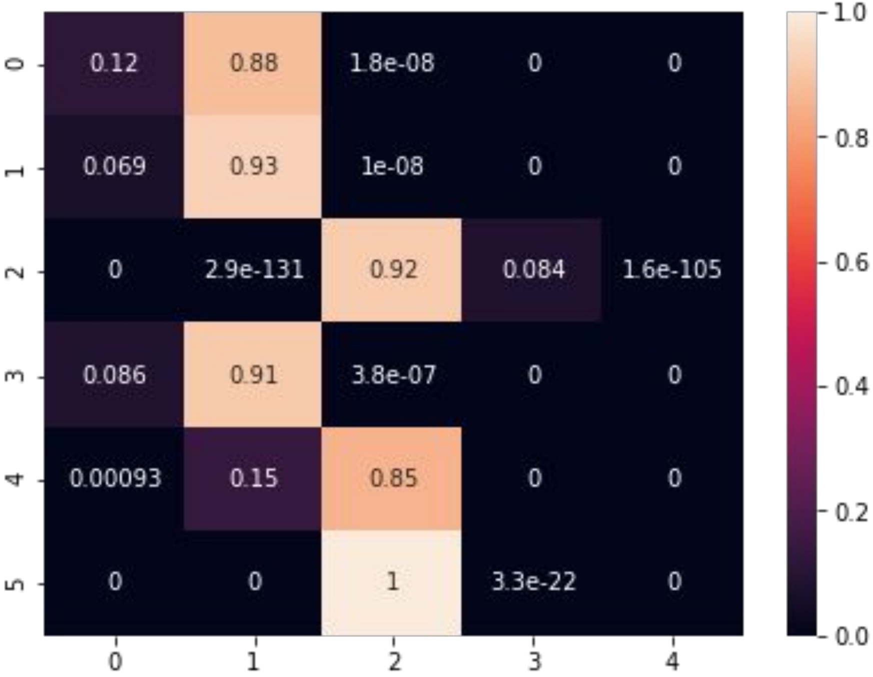


How can we account for variability in coder accuracy and break ties?

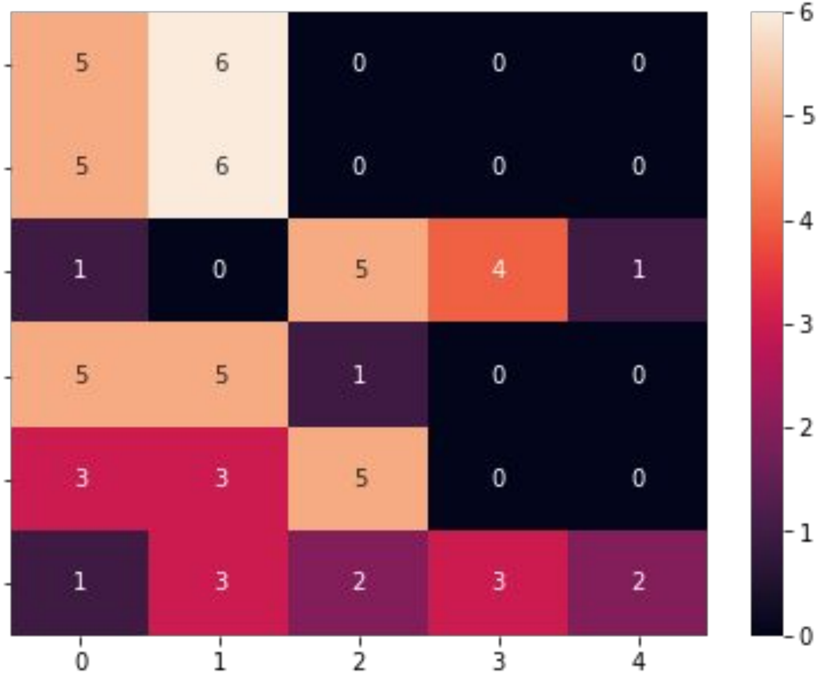
→ Give more weights to better coders!

—

pyanno Model B (Dawid & Skene)

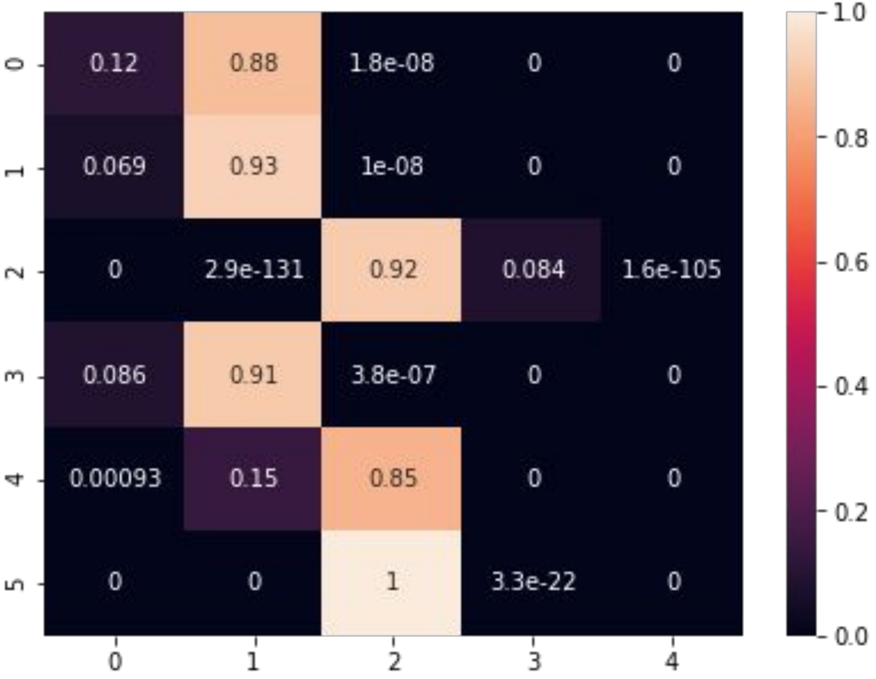


Inferred Labels



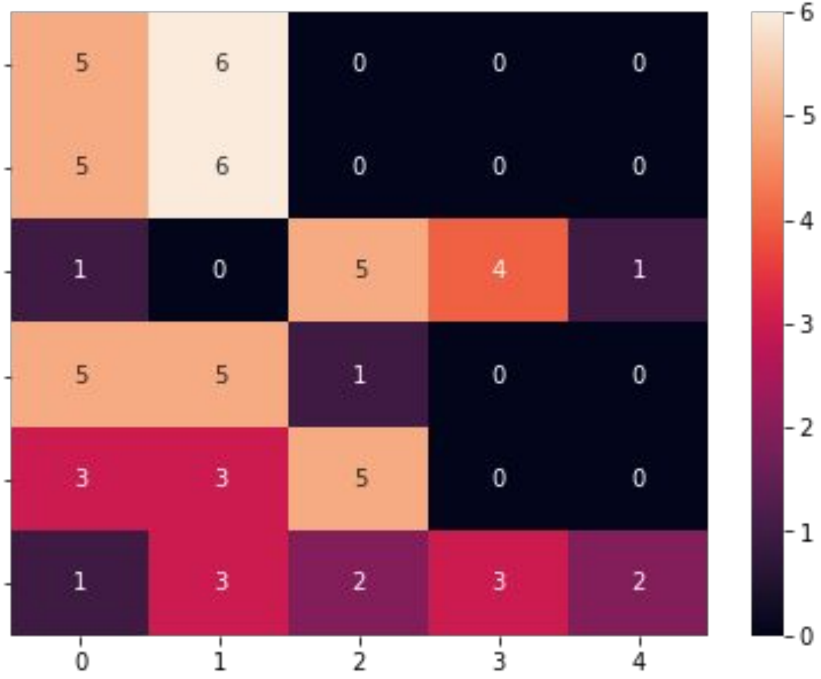
Votes

pyanno Model B (Dawid & Skene)



Inferred Labels

	0	1	2	3	4
0	0.12	0.88	1.8e-08	0	0
1	0.069	0.93	1e-08	0	0
2	0	2.9e-131	0.92	0.084	1.6e-105
3	0.086	0.91	3.8e-07	0	0
4	0.00093	0.15	0.85	0	0
5	0	0	1	3.3e-22	0

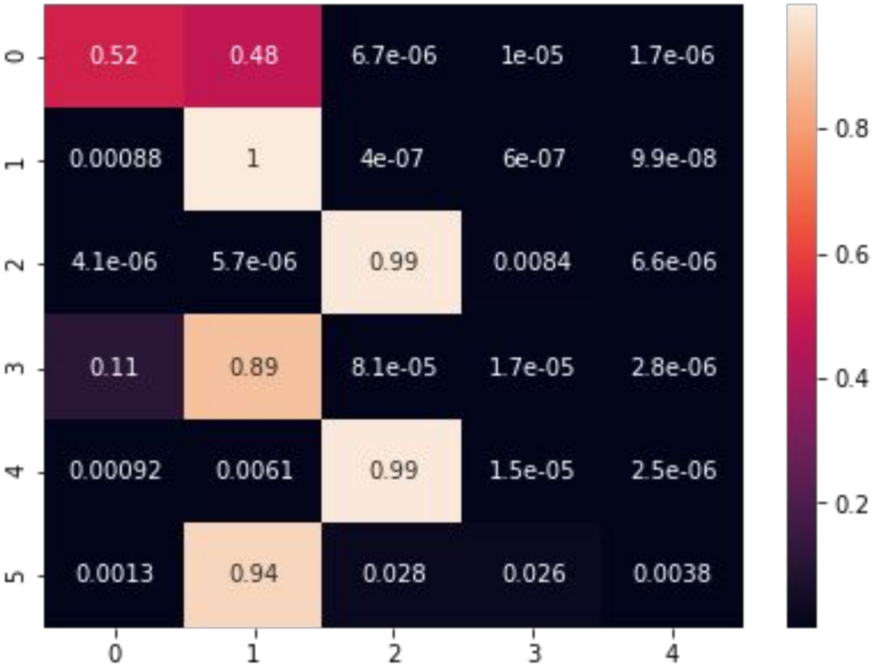


Votes

	0	1	2	3	4
	5	6	0	0	0
	5	6	0	0	0
	1	0	5	4	1
	5	5	1	0	0
	3	3	5	0	0
	1	3	2	3	2

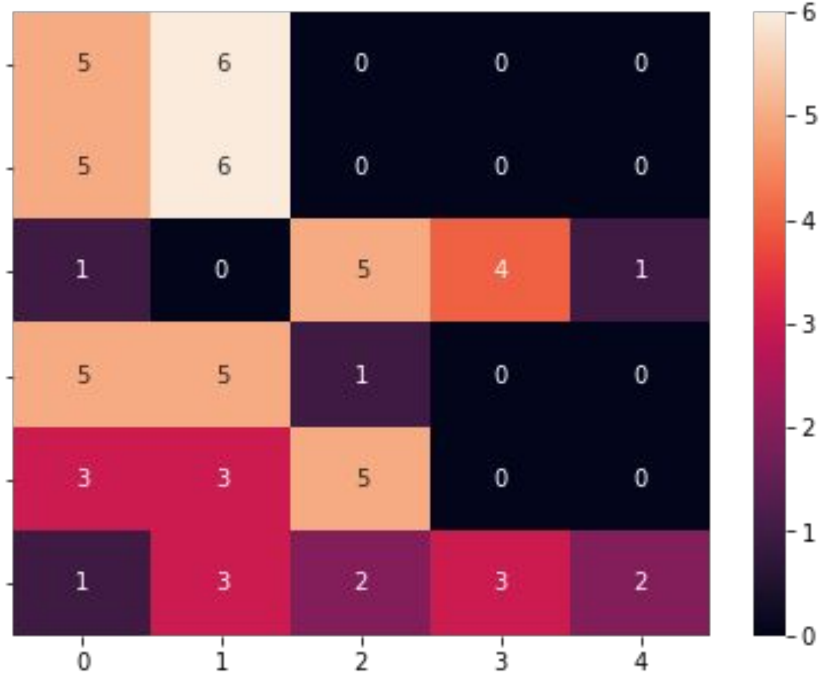
Model Bθ (Rzhetsky et al.)

Inferred Labels



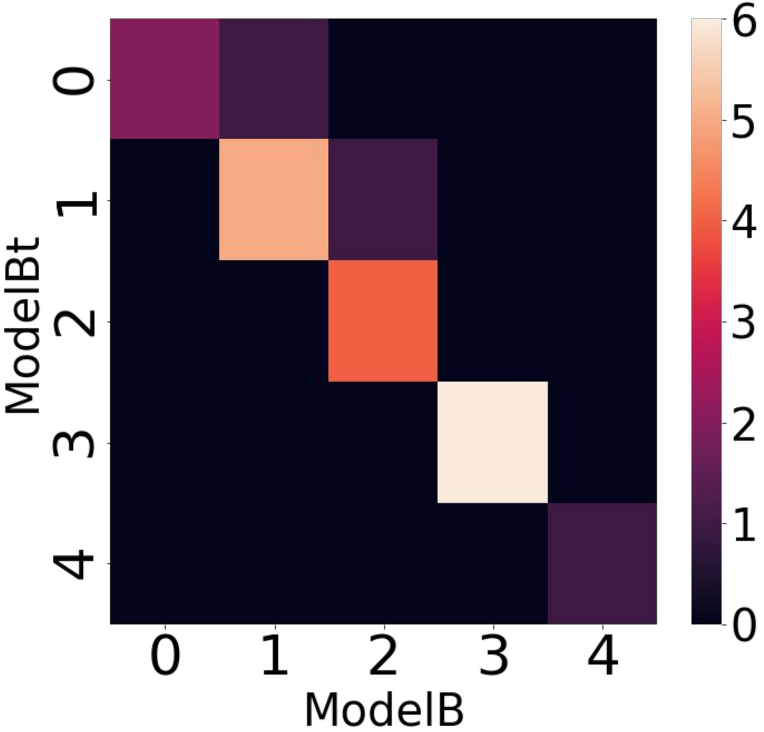
	0	1	2	3	4
0	0.52	0.48	6.7e-06	1e-05	1.7e-06
1	0.00088	1	4e-07	6e-07	9.9e-08
2	4.1e-06	5.7e-06	0.99	0.0084	6.6e-06
3	0.11	0.89	8.1e-05	1.7e-05	2.8e-06
4	0.00092	0.0061	0.99	1.5e-05	2.5e-06
5	0.0013	0.94	0.028	0.026	0.0038

Votes

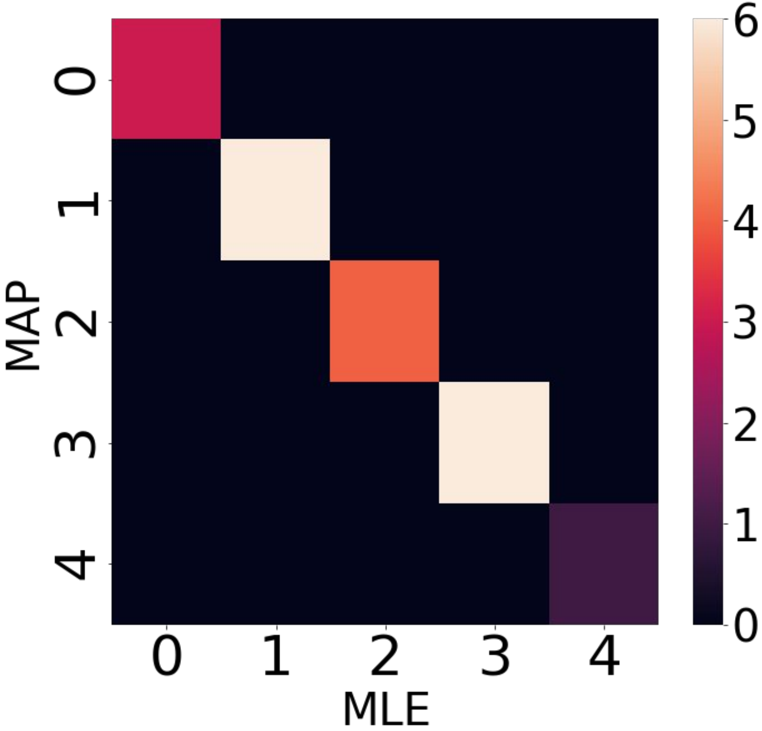


	0	1	2	3	4
	5	6	0	0	0
	5	6	0	0	0
	1	0	5	4	1
	5	5	1	0	0
	3	3	5	0	0
	1	3	2	3	2

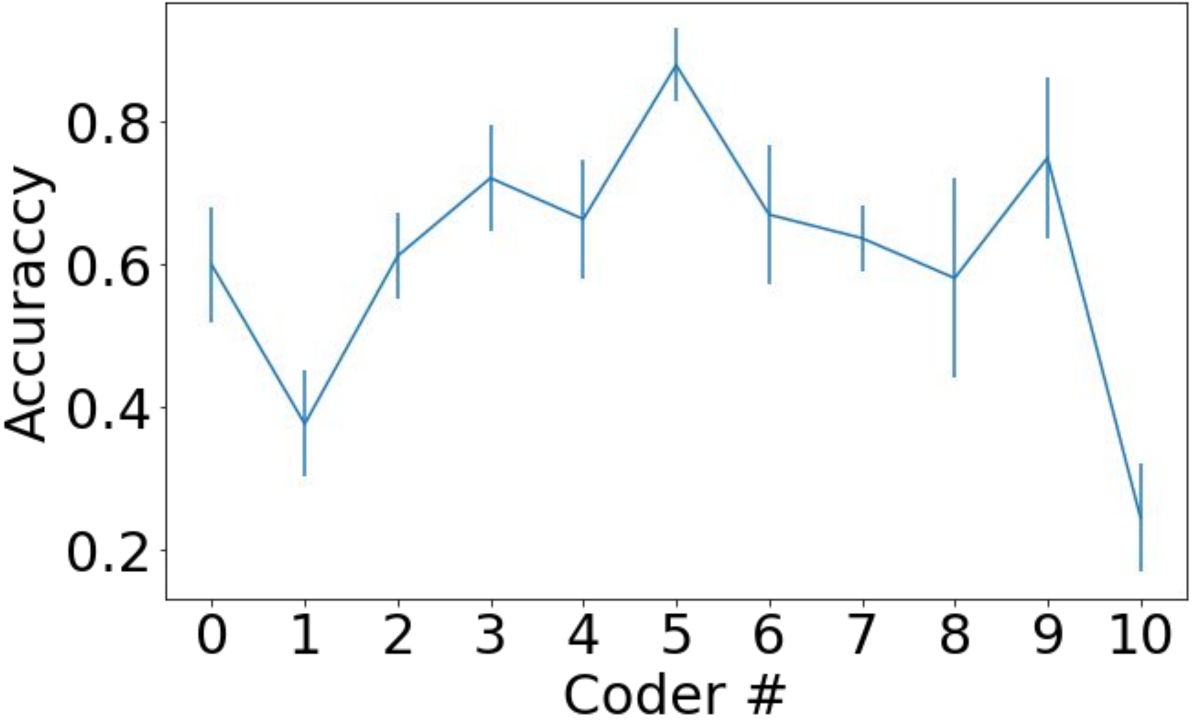
Model B vs Model B0



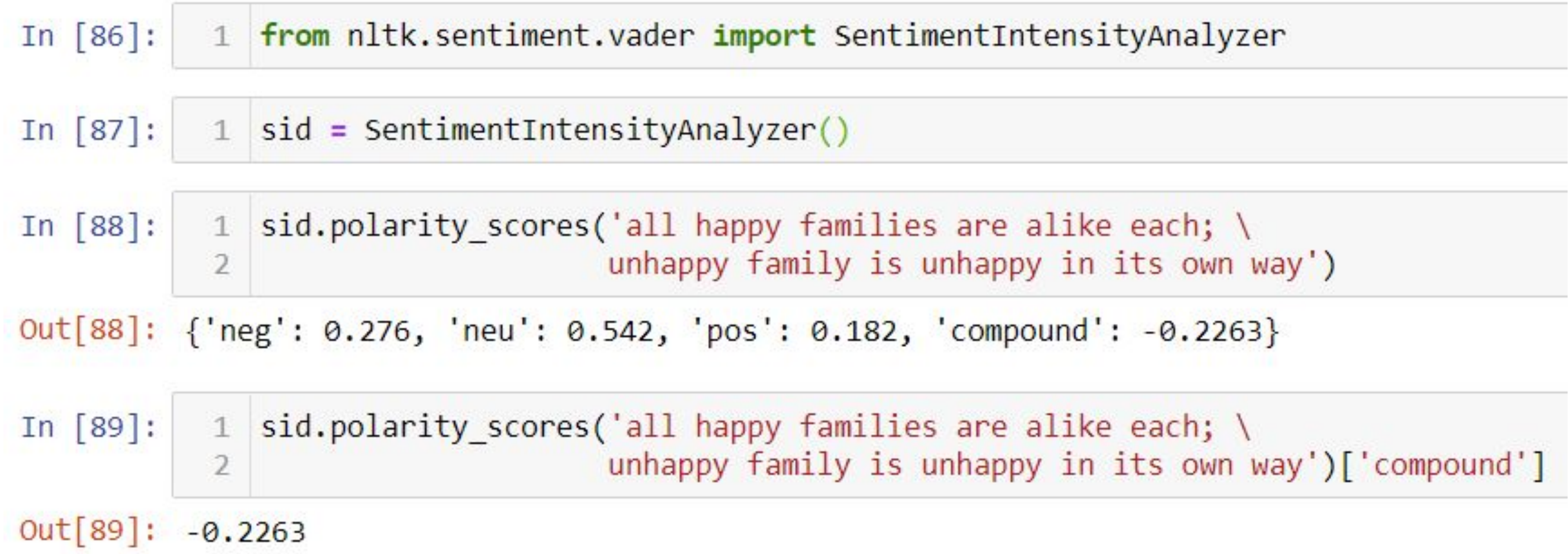
MAP vs MLE (Model Bθ)



Coder Accuracy



Inferred label vs Vader

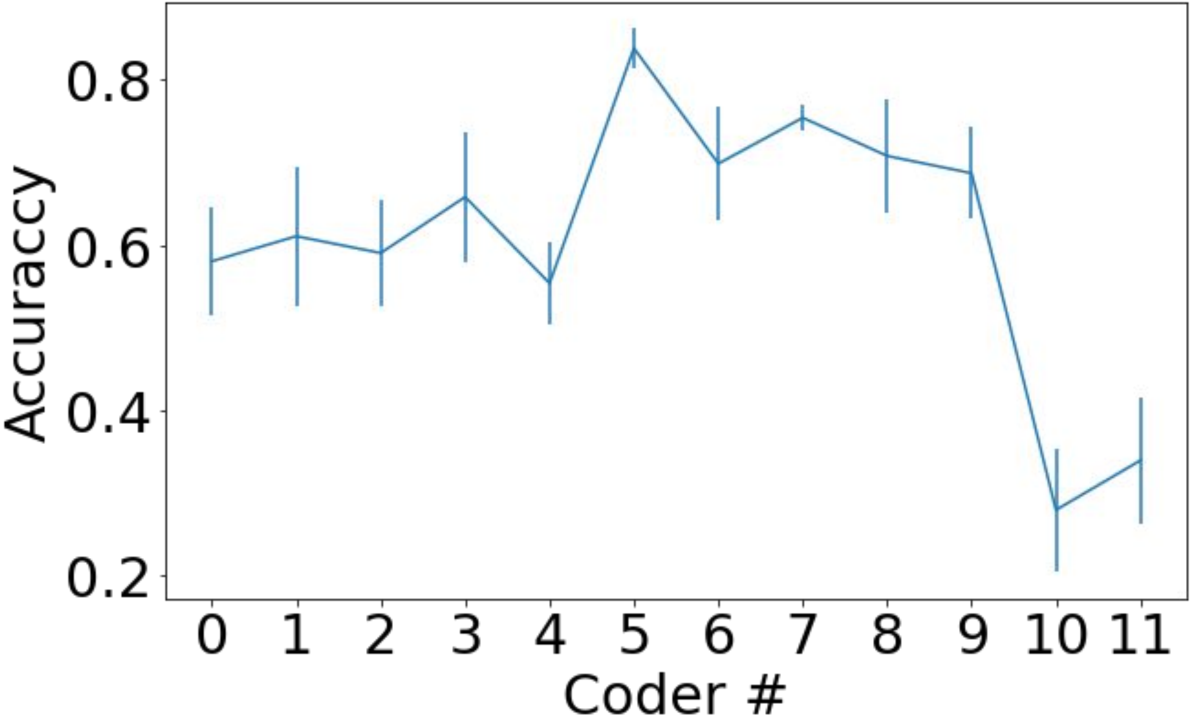


```
In [86]: from nltk.sentiment.vader import SentimentIntensityAnalyzer

In [87]: sid = SentimentIntensityAnalyzer()

In [88]: sid.polarity_scores('all happy families are alike each; \
                             unhappy family is unhappy in its own way')

Out[88]: {'neg': 0.276, 'neu': 0.542, 'pos': 0.182, 'compound': -0.2263}

In [89]: sid.polarity_scores('all happy families are alike each; \
                             unhappy family is unhappy in its own way')['compound']

Out[89]: -0.2263
```

Inferred label vs Vader

Cohen's weighted κ	.35
Pearson's ϱ	.61
Spearman's ϱ	.63

Inferred label vs Vader



Conclusion

- Humans seems quite reliable even in sentence-level sentimentality annotations
- Algorithms does not seem to conform with human annotations (at least vader)

- Give instructions well