Results From Each Test Case

Movieshort Properties

	movieshort_baseline	movieshort_one_change	movieshort_two_change	movieshort_three_change	movieshort_gradient
Num Successful examples	280	268	254	131	267
Num errors	20	17	20	19	15
Num positive	127	89	55	20	118
Num negative	153	179	199	111	149
% positive	45.357	33.209	21.654	15.267	44.195
% negative	54.643	66.791	78.346	84.733	55.805

Movieshort_baseline vs movieshort_one_change

- Num overlapping: 260
- Same sentiment: 132
- Different sentiment: 128
- Among the 260 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
0	147	-2.938	-5.0
1	113	-6.486	-5.0

- Average confidence shift overall -3.977272727272727
- Total Swings = 132
- Positive Swings = 15
- Negative Swings = 71
- Zero Swing = 46

Movieshort_one_change vs movieshort_two_change

- Num overlapping: 242
- Same sentiment: 172
- Different sentiment: 70
- Among the 242 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift

0	161	3.266	0.0
1	81	0.484	0.0

Same sentiment excluding the ones that already changed in one_change: 83
Different sentiment excluding the ones that already changed in one_change: 37
overlapping : 120
Total Swings = 83
Positive Swings = 21
Negative Swings = 17
Zero Swing = 45
Average Swing = -0.27710843373493976
Average Swing = -0.27710843373493976
Median Swing = 0.0

Movieshort_baseline vs movieshort_two_change
different = 134
Same = 120
Overlap = 254
Total Swings = 120
Positive Swings = 12
Negative Swings = 43
Zero Swing = 65
Average Swing = -2.65

Movieshort_two_change vs movieshort_three_change

- Num overlapping: 123
- Same sentiment: 110
- Different sentiment: 13
- Among the 123 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
0	100	-0.357	0.0
1	23	-0.714	0.0

Same sentiment excluding the ones that already changed in one and two change: 44
Different sentiment excluding the ones that already changed in one and two change: 10
overlapping : 54

Total Swings = 44
Positive Swings = 6
Negative Swings = 14

Zero Swing = 24
Average Swing = -2.727272727272727
Average Swing = -2.727272727272727

Movieshort_baseline vs movieshort_three_change
Number that are different are 75
Number that are the same are 56
Overlap is 131
Average confidence shift overall is -2.857142857142857
Total Swings = 56
Positive Swings = 5
Negative Swings = 21
Zero Swing = 30

movieshort_baseline vs movieshort_gradient

- Num overlapping: 267
- Same sentiment: 263
- Different sentiment: 4
- Among the 267 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
0	149	0.372	0.0
1	118	0.214	0.0

Total Swings = 263
Positive Swings = 30
Negative Swings = 21
Zero Swing = 212
Average Swing = 0.3041825095057034

Movielong Properties

	movielong_baseline	movielong_one_change	movielong_two_change	movielong_three_change	movielong_gradient
Num Successful examples	298	291	290	145	
Num positive	143	109	81	41	
Num negative	155	182	209	104	

Num errors	2	6	6	1	

Movielong_baseline vs movielong_one_change

-

Sentiment	Num reviews	Average confidence shift	Median confidence shift
0	155		
1	143		

Same = 200
Different = 91
Total Swings = 200
Positive Swings = 18
Negative Swings = 80
Zero Swing = 102
Average Swing = -2.96
Average Swing = -2.96
Median Swing = 0.0

Movielong_one_change vs movielong_two_change

-

Sentiment	Num reviews	Average confidence shift	Median confidence shift
0	121		
1	79		

Same = 138
Different = 59
Total Swings = 138
Positive Swings = 8
Negative Swings = 48
Zero Swing = 82
Average Swing = -3.246376811594203
Average Swing = -3.246376811594203
Median Swing = 0.0

Movielong_two_change vs movielong_three_change

Sentiment	Num reviews	Average confidence shift	Median confidence shift

0	104		
1	41		

Same 103
Different 39
Total Swings = 103
Positive Swings = 11
Negative Swings = 27
Zero Swing = 65
Average Swing = -2.1359223300970873
Average Swing = -2.1359223300970873
Median Swing = 0.0

movielong_baseline vs movielong_gradient

Yelpshort Properties

	yelpshort_baseline	yelpshort_one_change	yelpshort_two_change	yelpshort_three_change	yelpshort_gradient
Num Successful examples	296	294	288	269	292
Num errors	4	1	4	3	4
Num 1	65	49	83	125	67
Num 2	33	97	114	66	34
Num 3	30	93	50	24	20
Num 4	42	42	27	32	42
Num 5	126	13	14	22	129
% 1	21.959	16.667	28.819	46.468	22.945
% 2	11.149	32.993	39.583	24.535	11.644
% 3	10.135	31.633	17.361	8.922	6.849
% 4	14.189	14.286	9.375	11.896	14.384
% 5	42.568	4.422	4.861	8.178	44.178

yelpshort_baseline vs yelpshort_one_change

- Num overlapping: 294
- Same sentiment: 62

- Different sentiment: 232
- Among the 294 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
1	65	-1.552	0.0
2	33	-2.667	0.0
3	28	-3.889	-5.0
4	42	-9.0	-10.0
5	126	-5.556	-5.0

Total Swings = 62
Positive Swings = 0
Negative Swings = 27
Zero Swing = 35
Average Swing = -3.467741935483871

Correct direction moved: 116
Wrong direction moved: 97

yelpshort_one_change vs yelpshort_two_change

- Num overlapping: 287
- Same sentiment: 108
- Different sentiment: 179
- Among the 287 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
1	47	-0.588	0.0
2	96	0.5	0.0
3	92	-1.667	0.0
4	39	0.0	0.0
5	13	0.0	0.0

yelpshort_baseline vs yelpshort_two_change

Right direction: 203
Wrong direction: 28
Overlap: 288

yelpshort_two_change vs yelpshort_three_change

- Num overlapping: 266
- Same sentiment: 127
- Different sentiment: 139
- Among the 266 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
1	77	1.724	0.0
2	111	0.25	0.0
3	47	-1.786	0.0
4	20	2.917	0.0
5	11	0.0	0.0

yelpshort_baseline vs yelpshort_three_change

Right direction: 217
Wrong direction: 8
Overlap: 269

Yelpshort_baseline vs yelpshort_gradient

- Num overlapping: 292
- Same sentiment: 266
- Different sentiment: 26
- Among the 292 reviews that overlap:

Sentiment	Num reviews	Average confidence shift	Median confidence shift
1	65	0.386	0.0
2	33	0.333	0.0
3	29	0.0	0.0
4	41	0.441	0.0
5	124	0.325	0.0

Right direction (this is a positive change this time): 17
Wrong direction: 5
This is really out of the reviews that are not equal to 1 or 5 for those we can look at confidence
Out of 103 so 16.5% changed

Yelplong Properties

	yelplong_baseline	yelplong_one_change	yelplong_two_change	yelplong_three_change	yelplong_gradient
Num Successful examples	282	287	284	293	
Num errors	18	13	16	7	
Num 1	56	46	57	89	
Num 2	55	80	119	114	
Num 3	28	91	68	55	
Num 4	53	50	30	27	
Num 5	92	20	10	8	
% 1	6.4%	4.52%	20.07%	30.38%	
% 2	19.86%	16.03%	41.9%	38.9%	
% 3	19.5%	27.87%	23.94%	18.77%	
% 4	9.92%	31.7%	10.56%	9.21%	
% 5	32.6%	6.97%	3.5%	2.73%	

yelplong_baseline vs yelplong_one_change

-

Sentiment	Num reviews
1	46
2	80
3	91
4	50
5	50

Same 104
Different 169
Overlapping 273
correct_direction': 51,
'wrong_direction': 98
Total Swings = 104

Positive Swings = 2
Negative Swings = 38
Zero Swing = 64
Average Swing = -2.8846153846153846
Average Swing = -2.8846153846153846
Median Swing = 0.0

yelplong_one_change vs yelplong_two_change

-

Sentiment	Num reviews
1	32
2	44
3	28
4	12
5	5

Same 62
Different 50
Num overlapping examples 112
Total Swings = 62
Positive Swings = 5
Negative Swings = 26
Zero Swing = 31
Average Swing = -2.8225806451612905
Average Swing = -2.8225806451612905
Median Swing = 0.0

yelplong_baseline vs yelplong_two_change

Same 58
Different 211
Overlapping 269
'correct_direction': 127
'wrong_direction': 55

yelplong_two_change vs yelplong_three_change

-

Sentiment	Num reviews
1	19

2	32
3	21
4	7
5	0

Same 39
Different 31
Num overlapping examples 70
Total Swings = 39
Positive Swings = 7
Negative Swings = 8
Zero Swing = 24
Average Swing = 0.07692307692307693
Average Swing = 0.07692307692307693
Median Swing = 0.0

yelplong_baseline vs yelplong_three_change
Same 42
Different 235
Overlapping 277
Correct direction 178
Wrong direction 18

yelplong_baseline vs yelplong_gradient

Final Analysis Results

- **Facts**
    - Many more of the positive sentiment examples give errors in the output than negative sentiments
    - Average confidence shift given the same sentiment was very very low -3.98% in movie short one change which can mean two things either it is a bad indicator or the reviews are not being faithful
    - Movie 2 change confidence -0.277
    - Movie 3 change confidence -2.72
    - For movie short one change half of them changed sentiment
    - 77 produce errors by then end of movie and 175 changed sentiment
        - 78% of the ones that didn't error change their sentiment after three changes
        - This does not mean that every suggestion of the model is faithful but in three guesses it was able to give the right options
    - Yelp went from 116 to 203 to 217 correctly moved
- **Opinions Sort of plus assumptions**

- The model can cohesively put together what a human would expect the reasoning to sound like however when tasked with actually finding phrases that were important to it the model could not actually execute the task properly
    - Although by 3 phrases most of them changed sentiment the details are not so specific
- In yelp if the original sentiment is 2 there is no correct direction for the decision to move so we did not include it
    - Otherwise correct direction is the opposite side it was originally on
- Restaurants we should expect a shift earlier on to show that the model is actually faithful

Average Percent Decrease in Confidence Level
(after one change on movie inputs)

One Change Short	4.00%
Two Change Short	0.30%
Three Change Short	2.70%
One Change Long	2.96%
Two Change Long	3.20%
Three Change Long	2.10%