Two metrics were used to analyze inter-annotator agreement: textual agreement and overlap agreement. Textual agreement uses Cohen's kappa score (ref + expl.) to measure agreement over units that either one or both of the annotators have annotated as metaphors. In order to account for multiword, overlap agreement measures the average percentage of overlap over non-identical tags (e.g. "line" vs "line in the sand"). In our first round, we had a kappa score of 0.628, and our average overlap was 48.9% over the labels that were not identical for 20 comments. The second time around, looking at 20 different comments, our average textual agreement went down to 0.492, and our average overlap agreement rose to 64.4%. We then introduced a third annotator. In our third round of agreement testing, we calculated the scores as an average of the three pair-wise scores calculated for each metric based on 20 comments. Our kappa score was 0.593, and our average overlap agreement was 54.4%. We realized that this was because of a disagreement over the word ‘establishment’; once we resolved that, our kappa score became 0.647, which is substantial. Based on these results, the annotators went on to annotate ~ 300 comments each. Then, 10 comments were randomly selected from each of the annotators annotated comments for the other two annotators to annotate as well. This test yielded an almost perfect average kappa score of 0.818 and an average overlap of 66.6%.

March 4 (30 comments)

Random selected

A screenshot of a computer

AI-generated content may be incorrect.

Mean of pairs: