Goals:

1. Are the results that we get from Mturk correct/meaningful ?

2. What are the issues that are stopping us from getting that correct result or why are the workers giving us that kind of result? So, we can formulate our task in a way so we can avoid those problems.

We annotated 16 articles by 5 workers.

Analysis of Goal 1:

* 15/16 articles have a majority label by the workers (mturk-label) **(0.93)**
* 8/16 articles mturk-label matches with their source-label **(0.50)**  
  So, articles do not follow the political affiliation of the outlet.
* 12/16 articles mturk-label matches with our manual annotation, one was not rated by us **(0.75)**

Now, I manually checked the 4 non-matched articles. For 3 out of 4, I have a disagreement with Prahlad. In one, we agreed, but a third annotator disagrees, and it is a long article.

668c47ca-f35a-4075-b4b5-30cfc870a7e5:

Prahlad: Right (Used inference about authors thought on trump & assumed author familiarity with trump’s tone)

Salvi: Left (The article seemed like a sarcastic piece on Trump, and in ending, it is also shown the author fears for future trump presidency)

4036ab43-daf0-4209-83b0-969969cdbdf2:

Prahlad: Left (Inferred that the author is supporting Obama using prior knowledge of Obama administrations against intention towards Cuban embargo)

Salvi: Center (The only thing that seemed remotely biased to me was slight criticism of Obama bypassing Congress. Otherwise it all seemed facts)

7e9742ef-5cf4-4790-bba1-c633f5be2156:

Prahlad: Right (He thinks author is trying to prove if a fallout happens, that is not trump’s fault)

Salvi: Center (Things like one-sided citation of Fox news, but a fallout would not be his fault, suggests a right bias, but admitting trump has many faults equals that out)

**6e2ec9f5-eb67-4422-84d8-3d68b1a50220:**

Prahlad: Left (Thought it was proving FBI trying to sabotage, but that was only till 2nd paragraph)

Salvi: Left (After reading full article, seems it is more right-biased than left)

Nooha: Right (Full article explains Hilary’s Email leak as an issue in a greater way than FBI trying to sabotage election with it)

So -

- For the non-matched articles, the bias is ambiguous or subtle and subjective judgement would differ for those articles. We need to rely on the majority vote.

- Workers do not infer decisions in the article. **Workers cannot understand inferred bias.**

**-** Workers can identify political affiliation of an article if it is not too subtle, **given we choose workers with good political knowledge**

- Articles with long snippets have high chance of wrong decisions

**Give the non-matched articles to the workers who have correctly identified the articles so far.**

Analysis of Goal 2:

* 15/16 articles have a majority label by the workers (mturk-label) (0.93)  
  So there is a high chance that there will be a majority-label. But there is always going to **one or two** which are different from the majority agreement.   
  1:1:3 - 7/16  
  1:3 - 4/16  
  All agree – 4/16
* Issues of these workers with different judgements (judgements other than the majority)
  1. Did not spend much time in the task (In short, **garbage annotators**)
  2. No valid reason/ No valid markers – annotations might be correct, but since none of them are a must-to-do they chose to avoid that
  3. **Workers are not reading full article/ all the snippets/ reading just the title**
  4. **Have wrong political perspective about an Entity**, i.e., praising Trump is a thing left biased articles do, or condemning Hilary is a thing left biased articles would do.
  5. **Having wrong notion of Bias/No Bias**, i.e., praising one side, but not bashing another side is not a bias according to them.
* **Recommendations for pre-screening workers (with a smaller reward):**
  1. **We must fix our definition of bias or no bias. Then we must check if the workers can identify these.** For example, the definition of **No Bias**   
     - talking about both sides in a good/bad way, but with a balance.   
     - Just representing facts
  2. **Check if workers can identify when an article talking positively/negatively about an entity and political party.**
  3. **Check if workers can identify the political affiliation of that entity.** Then if they can identify if they can reach to correct bias label from that affiliation (included in definition/ example).
  4. **Check if workers can identify if an article is preferring a certain side.**
  5. **Check if they find No preference, do they explain that in the reasoning.** [Include in definition/example/ as an option in reasoning]
  6. **Ask worker about their political affiliation**, we can do a correlation check with their labels post-task
* **Other design choices along with the []s:**
  1. If we are recording reasoning, make it a must-to-do (it was optional previously)
  2. Markers have no effect eventually, I have tried having a better graphical tool for that, but still that is not perfect.
  3. Instead of having one article per hit, include multiple but with greater reward, so we can attain the services of a good worker throughout.
  4. Have a wider rating range (Left, Lean Left, Center, Lean Right, Right)