Now that a model has been trained it can be applied to the corpus. The overall idea is simple. The model take a a string and scores it as positive or negative. The data must be sorted by publication, the candidates must be identified within that publication’s articles, a granular level of analysis must be chosen (sentence, paragraph or article level), sentiment analysis model needs to be applied to that granularized string, and a ratio score of positive to negative sentiment needs to be computed. Due to its simplicity and strength, regular expressions are used to drive most of the identification process

                The regular expression approach is a brute force approach that requires careful condensation of word combinations. Luckily, there are few target candidates and the corpus consist of published work, which should avoid shorthand naming substitutions like “Liz” or colloquial names like “Joey”. In terms of combinations, the regex approach is computationally cheap. Unique first names like “Bernie” or “Kamala” are safe identifiers.  Candidates like Elizabeth Warren and Joe Biden with more commonly used first names, require a combination of their last name and another combination of first and last name I.E “Elizabeth and Elizabeth Warren”.  In the case of Bernie Sanders, the name Sanders procides some potential false captures because of Sarah Huckabee Sanders. Overall, the group used eight search terms to feed into the Regex which python computes almost instantaneously over the dataset. Potential drawbacks of regular expressions are false captures and data loss. As the code automates away much of the validation process, the group will continue to be mindful and monitor the data on a more granular level to make sure it is representative of accurate captures.

                The next decision is what level to run the sentiment model over the corpus. The first option is to run the analysis at the article level.  Using topic modeling or some sort of custom scorer, articles relating to a specific candidate can be identified and scored by publication as a positive or negative article for that candidate. This approach may be explored later, but it wasn’t deployed in the most naïve model.  Another option is to aggregate the corpuses on a paragraph level.  This approach captures bias that spans more than one sentence. Many candidate’s mentions on the paragraph level, may refer to pronouns in place of the candidate names throughout the paragraph.  The drawback with this approach, is many times within a paragraph several candidates will be mentioned.  Simply scoring a paragraph and assigning that value to all candidates mentioned in the paragraph will lead to false positives. A solution to this would be to exclude paragraphs that mention more than one candidate. However, this can lead to sample size issues. The last option, is to explore the data at the sentence level

Modeling on the sentence level, two scoring systems have been developed.  The first captures all candidate mentions and scores those sentences for each candidate. The second approach excludes any sentences wherein multiple candidates are mentioned. One can debate if using sentences with multiple mentions is problematic, but overall the initial results do not seem to be all that affected by segregating the data. The advantage of deploying data segregation on the sentence level over the paragraph level is the amount of data that retained allows for a stronger scoring system.  Paragraph level data leaves us with substantially less initial data to score on(multiple sentences per paragraph), and many more collisions between candidates will occur on the paragraph level. One potential optimizing solution Is to use coreference resolution to rename pronouns with their proper nouns. Later models may explore such tools.

Initial results show some interesting trends. The NYT seems to have by far the largest positive position on a candidate, giving Biden nearly 57% positive coverage. Meanwhile, Sanders is being given only 46% positive coverage and Warren is being given about 50% positive coverage. CNN seems to have a very balanced range for sentiment, with all candidates landing between 45-47% positive coverage. Fox news surprisingly gives Biden slightly positive coverage (51%) and give Bernie and Warren positive coverage of 45% and 46% respectively. The Washington post seems to favor Biden with 49% positive coverage, while Bernie and Warren receive around 44%.

The results are preliminary but encouraging. In the next iteration more focus needs to be given to publications which are showing bias in initial testing. The New York Times stands out as putting their finger on the scale, which is strange considering its’ reputation as the most intellectually balanced paper in America. Additional testing over politically charged language may also be a possible route to pursue. NLTK should allow us to identify adjective usage around topics that are politically charged. I.E topics such as Medicare for All, taxes and spending as well as language such as “electability”, “age”, and “bi-partisan”. The topics will be explored with word clouds and distributions.