Before 8th Sept 2020:

* Looked through the data manually. Organised them into the various conferences and located their acceptance statuses.
* Manually went through a number of reviews to try and figure out the most important reasons for papers to get accepted/rejected.
* Converted that into a map of reasons to each section of the paper to try and isolate techniques for each part hat could potentially be used to narrow down on the decisive factors.

8th Sept. 2020:

* Worked on dataloading (focusing on the ACL and arxiv papers for now. Over 10700 of them). This included the titles and the acceptance statuses. Other sections can be added according to our requirements.
* Started working on the EDA of the titles. Worked on length histograms. Not a lot of valuable information from here. Just the distrobutions.
* Worked on the gram analysis. This is pretty useless right now, but after stopwords are removed and stemming is done this should make a lot more sense. Should remove proper nouns too but in the titles of these papers, I highly doubt there will be any proper nouns anyways.

9th Sept 2020:

* Worked on the stopword removal and got it done.
* Now testing for unigram analysis gives it a bit more clarity. Some few points about this are in the documentation.
* Worked on the bigram analysis too. Got some results but nothing concrete to have a good decision boundary. The notes are in the documentation.

10th Sept 2020:

* Added the code to scrape authors and abstracts and the actual intended conference as well. Now we can have probabilities for each conference publishing a given paper.
* Built the grammar checker. Implemented it on the abstract + title. Also extracted the full papers as well and did the grammar checking as well. Plotted the histograms for these two as well. The results and seperability are in the notebooks.
* Extracted the author info: like mail ids and stuff. This gives us info about companies and universities. This could have a big impact. Stilll need to check it.

11th Sept 2020:

* Tried to work on extracting references. This is a VERY tedious process and it took me forever but its got way too many edge cases. Didn’t work for very long and this was pretty frustrating as well. I have an approach in mind after all the scraping is done but scraping is pretty complex.

14th Sept 2020:

* Keeping the references aside for now, scraping is done with the title abstract and the rest of the paper, the conference name and the e-mails.
* Cleaned the mail ids and the conference names as well. All data apart from the references are ready for now. Running it through a very basically TF-IDF classifier (only text to start off with)
* Converted the conferences into one hot encoded vectors. Not using these for classifications though as it is a bit weird with empty samples. Will get back to this after an initial exploration.
* Using TF-IDF to just see what results are like. Using just an acceptance state I get over 70% f1 score. But this does not really mean anything as we do not know which conference we will get accepted into during testing!
* Using the one hot encoding system with conference names gives a couple of issues due to the none hot encoding problem. Working on fixing it right now but it is a bit slow.
* Fixed th encoding issue and worked on a preliminary TF-IDF and OnevsRest Classifier. The data is massive so this works well as it trains quickest. Around 63% f1 on a small subset of the train data. Got the probability predictions to work as well. This does take time on the full dataset though.
* Read up on the readability easy index and picked the couple of important ones.

15th Sept 2020:

* Started working on the grammar checker and quantified it in the database.
* Went through all the readability indexes one by one and checked them all out to evaluate them against one another. This took a lot of time but I think I found the right indexes to use after all that work.
* Then implemented the grammar checker and text pipelines. Struggled with doing this as it was textual and input. Had to put it forward to the next day.

16th Sept 2020:

* Continued working on the composite pipeline. Tried a lot of combinations and composite structures to combine the texual and numeric inputs and finally picked the one with the best possible performance. This took a lot of time and was a pretty slow and frustrating process but we do have a pipeline now with the accuracy of about 65-70% with class predictions. This cannot be looked at as a metric though as the probabilities make so much more sense and there is an edge case that needs to be discussed as well.

17th Sept 2020:

* Hyperparamter tuning and checking with aspects like class weights and stuff.
* Working on the references as well. Trying to quantify that in some way, still.