Looking through and organizing the Dataset (just by going through the dataset manually, no code so far to check):

* ACL: All accepted papers are in the acl\_accepted.txt file. The names tally with the paper titles from the dataset.
* Arxiv : For all three types: Ai, CL and LG, the acceptance status is in the review.
* CoNLL : not able to track or find any trace of accepted or not yet. But there are only 22 entries here.
* ICLR folder: accepted status in the review itself.
* NIPS : Need to download the dataset to check maybe? But the papers online say that this is only collection of acceted papers. So this might not be very effective.

Format of the papers and notes for each section:

* Abstract:

1. Full project run through
2. A summary of the quality of language and explanation level? Still not sure. Need to read more.
3. Innovations, new concepts, results

Mostly text – Can maybe use BERT to summarize or directly relate? Mostly short text of about 10 half-lines max

* Intro/Methods:

1. Language/Understandability check

Mostly text as well, maybe use GPT or BERT to summarize or quantify as well?

* Results:

1. Tabular data needs to be dealt with.
2. Combination of information from text and tables.
3. Might contain crucial information about performance or results.

Still thinking about what model to use here.

* Discussion:

1. English check too maybe? Understandability would be crucial here.
2. Innovations/New techniques/effectiveness will be discussed here as well. So really important.

Still thinking about the model.

* Additional info: Authors/Guides/Universities will also have an impact on the results. Need to tally this though. Need to read through a few papers before I can guage all this.

Reasons why papers get negative reviews/rejections and some important points:

* Parts of the paper being too abstract/ vague maybe? Not enough good examples and not explained well. Primary flaws and missing information in the approach. Incomplete and unjustified claims.
* Just because the abstract talks about things triumphantly, does not mean that the paper is accepted. For example, the authors have used terms like “drastically outperforms” and “only available technique” but on reading the reviews they do not actually seem to be true. So the key point is Do not just trust information in the papers. Should there be way in which we can check if these claims are actually true? How are we ever going to evaluate the performance against existing methods and older papers!!
* Huge sentences and grammatical errors never help. So many reviewers have given feedback on grammatical errors as well!
* Contributions just not being substantial enough to be published at that level. No idea on how I plan to check this!

Meeting points:

* Available data :

Papers and all their contents. Reviews are available too but I do not think we can use them directly as we will not have access to them while making the predictions.

We also know if the paper is actually accepted or not, and we have access to some rating on the basis of factors like compactness and simplicity and so on.

* What we need to predict :

Probability of the paper getting published in each level of conference maybe? Or maybe the probability of any particular conference?

Some points:

* The abstract (and the whole paper too for that) always says good stuff, even if it isn’t always true. Always try to check everything before using claims, if possible, especially with performance of techniques.
* Need to find a way to validate results. Not always quantitative in nature. Even if they are well tabulated, how do we check! Only solution right now is to maintain a database maybe?
* The introduction, method and discussion sections need to be tested for how well concepts are explained as papers can be rejected for this too.
* Start with title. And then move on to the next few sections.Intro and discussion imp.

Still left to do :

* Check lengths and validity of various sections with acceptance.
* References : same paper and valid.
* Journal names scrape

Started coding here:

Unigram analysis:

* It is mostly difficult to make good classifications based on just the title words.
* There are still some words that can tip the probabilities though. For example the word “classification” on the title favours a paper rejection rather than an acceptance. Similarly for words like “recognition” in the title as well. So this definitely does happen. But finding a concrete decision boundary…not yet. Also, the fundamental distribution affects this too. So it would be really hard to guage. Impossible to make anything concrete from this.
* Did the bigram analysis too. With all this, bottom line is still that We cannot make a proper decision based on the title text on its own.
* Grammar check has been done on the title + abstract. I have done the check on the full papers as well. Histograms of the accepted and rejected relationships with the grammatical errors are also plotted. It can naturally be seen that higher errors lead to a higher chance of rejection. Needs to be discussed further.
* Worked on the composite pipeline for numeric and textual input for the model. Grammar check, ARI and title+abstract for now. EDGE CASES NEED TO BE DISCUSSED.