**Research Question**: Does adding argument-level textual features help the argumentation identification task in scientific papers?

**Task 2**

**Experiment Design**

1. Conduct a series of experiments to predict argumentative components (and argumentative relations).
2. For each model type, conduct an experiment with and without argumentation-level textual features.
3. **Logistic regression, similar to the paper**
4. **BERT:**
5. **SciBERT:** [From Allen AI for scientific text](https://github.com/allenai/scibert):
6. LLM Prompting: use open source LLM, Llama

**Task 1: Reproduce the feature sets described in** [**Section 4.1**](https://aclanthology.org/2021.eacl-main.171.pdf) **using Python libraries for Dr. Inventor’s Corpus.**

**a) n-grams:** unigram, bigram, and trigram of Internet Argument Corpus (IAC) corpus. The IAC corpus consists of posts from conversations in online forums on a range of controversial political and social topics such as Evolution, Abortion, Gun Control, and Gay Marriage.

Generating n-grams is pretty straightforward in Python but focus on Dr. Inventor's corpus.

**(b) argument lexicons:** Two lists of 20words representing agreement and disagreement as done from Rosenthal et al.' work. I found the annotated dataset is downloadable from Sara Rosenthal's [website](https://www.cs.columbia.edu/~sara/data.php)  Please check it and/or extract lexicons.

**c)Sentiment and opinion lexicon:** Not relevant for scientific paper argumentation mining, Opinion mining could be relevant.

**d) Hedge Features:** [From Dr. Tan's paper](https://arxiv.org/pdf/1602.01103) Could you please check the paper's project page [here](https://chenhaot.com/pages/changemyview.html) to see whether the Data /Code of that paper includes Hedge Feature extraction? If not, we can email Dr. Tan about the hedge feature extraction code. Or you can implement the code by yourself to extract hedge features from Dr. Inventor’s Corpus.

**e) PDTB discourse marker:** Not needed for now

f) ) **Modal verbs:** Can be extracted using Python's Spacy library. See the code snippet [here](https://stackoverflow.com/questions/59713284/use-spacy-models-to-find-modal-verb-for-languages-fr-es-ru)

g) **Pronouns**: Not needed

h) **Textual entailment:** We need to frame this feature as per our project. Research papers have hypotheses and experiments. Look at some articles about textual entailment in scientific articles and let me know.

i) **Lemma overlap**: Not needed

j) **Negation:** This can be detected using Python code.

**Dataset: Dr. Inventor Dataset of Argumentation Mining:**

The original Dr. Inventor Corpus has only four rhetorical label, the argumentation label is absent. I have downloaded the original corpus from this [link](http://taln.upf.edu/drinventor/). The zip file is attached: DRI\_Corpus.zip.

The argumentation tagging on Dr. Inventor's corpus was done by Lauscher et al. I have attached the paper and the corpus: compiled\_corpus\_DrInventor.zip. Not that, this corpus has the 40 articles but *only argumentation tagging*, You can find the details of their argumentation tagging from the attached paper.

**Use the articles from only compiled\_corpus\_DrInventor.zip**