AI / NLP developer task

The task is to investigate the data set of sentences from the abstract and introduction of annotated research articles.

The dataset is publicly available on UCI ML Repository:
https://archive.ics.uci.edu/ml/datasets/Sentence+Classification
and can be downloaded here:
https://archive.ics.uci.edu/ml/machine-learning-databases/00311/

Description of the dataset is provided within the downloadable file.

There are several questions that we would like to be answered about the dataset:

1) Automated labelling of unlabelled sentences.

2) Keyword insights: are there some words or phrases that can indicate that the sentence belongs to a particular class?

3) Investigation of the usual text flow in the abstract and introduction. Are certain types of sentences more likely to be at the beginning / end of the abstract or introduction? Is there a time correlation between types, e.g. certain sentence types more often follow after others?

4) Any other insights that can be extracted from this dataset.

The task is open ended, i.e. we do not require to answer all questions, concentrating deeply on one or a subset of them is totally possible. If anything in the description is not specified you can either ask for clarification or are welcome to make an assumption. As long as you can argue for your design choice and assumptions that is fine.

Besides the actual code of the solution, it is also important that you keep notes for a presentation on how you approached the task and what problems you encountered. During the interview we would like to get a walkthrough on how you solved the problem, a presentation with some notes and pictures on your thought process and architecture / data model are very helpful there. We may also iterate a bit on your solution and challenge certain assumptions to see how your solution holds up to changes and what you would need to change to make it work under the new assumptions.