Problem Statement: Finding out the contributing statement from given PDF file.

Algorithm:

* Pre-processing of trial dataset to convert the data in form of CSV file.
* Training of BERT using the data obtained after pre-processing.
* Use the trained model to predict the whether a given statement in the PDF is contributing statement or not. Save the results in CSV file.
* Append all the contributing statements predicted by the model which are predicted in test data using the trained model.
* Generate abstractive summary of predicted contributing statements.
* Compare the abstractive summary and abstract using ROUGE score and save all the results in a file.

Summarization model:

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence models) uses self-supervised objective Gap Sentences Generation (GSG) to train a transformer encoder-decoder model. Pegasus’ pretraining task is intentionally similar to summarization: important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary. By using PEGASUS we can find abstractive summary that is trained on large corpus of text.

Limitation of the Algorithm:

* Finding out the contributing statement depends upon the context of the data. The existing dataset contains examples from NLP domain specifically.
* As the dataset is NLP related, it will give higher accuracy in case of NLP domain research paper whereas accuracy will be low when we try to generalize the model for other domain.

Overcoming the Limitation of the Algorithm:

* Augment the dataset including the scientific papers and their contribution of other domain.
* Fine tuning the BERT model and optimizing the parameter.
* Fine tuning the Google Pegasus model for contribution statement abstractive summary.
* Use positional features within the paragraph while performing pre-processing.