NLP HW2

1. **The dataset in this assignment is again, imbalanced. What did you do to deal with this problem? If you left them unchanged, why?**

   The following are the statistical data about our dataset:

	MS	PH	AM	SF	SR	OTHER
Proposition	12%	13%	7%	12%	8%	50%
Precision	48%	64%	63%	78%	73%	67%
Recall	78%	58%	74%	76%	50%	88%

   We can observe that with the decrease in the proposition, the recall results also decrease. Thus, to address this issue, I do the following mechanism to solve these issue:

   - Oversampling : Duplicate samples from the minority class to make its quantity comparable to the majority class. This can be achieved through simple random sampling.

   After doing this mechanism, we get the following result:

	MS	PH	AM	SF	SR	OTHER
Precision	65%	52%	87%	73%	80%	61%
Recall	81%	77%	83%	84%	59%	79%

   Although we get better performance in the validation dataset, we still get not get better results on the test set.

2. **Describe your strategies to solve this problem. Details include input and output format, data preprocessing and model selection should be provided.**

   . The problem formulation is as the following:

   - **Input:** current sentence + n x previous sentence
   - **Output**: the label of the current sentence.
   - **Data preprocessing:**
     We use one-hat vector to encode the label and e use a sliding window approach to capture the existing input.
   - **Model selection:**
     Considering hardware limitations(RAM,memory of CUDA, …), I choose to use a BERT-based model to address this issue instead of using a large language model

3. **Compare your results with different strategies (e.g. different models, different k parameter selection (refer to tips), different**

hyperparameters, etc.) and try to summarize which strategy you think is the best to solve this task, and explain why. Do the results align with your initial thoughts?

As mentioned above, we use bert-based model to solve this problem. The strategy I use to choose model is to choose the model with the best performance on the validation dataset. The following are the experiment result:

	Bert	**Deberta**	**Roberta**	**Xlnet**	**DistillBert**	**Ernie**
Macro-F1	72%	69%	69%	67%	70%	71%