# INLP HW2

# 110550085 房天越

1. How did you do to preprocess your data from the dataset

   I. Data loading
   load the training, validation, and test datasets, which are classified in three JSON file, with label columns extracted from a sample submission CSV file.

   II. Data Augmentation
   Do data augmentation with techniques like synonym replacement, random insertion and random deletion to enhance the training dataset.

   III. Label Conversion
   Converted multi-label annotations into binary matrix format for compatibility with a BERT-based classifier.

   IV. Create a custom PyTorch class named Dataset to handle tokenized text and corresponding labels using the BERT tokenizer.

2. How were the model's hyperparameters chosen? Did you perform hyperparameter tuning? If so, what were the specific steps and results?

   The hyperparameters used in the model are learning rate, training batch size, evaluation batch size, number of epochs, and weight decay. These hyperparameters were chosen based on the best performance.

   No hyperparameter tuning, such as learning rate deduction is performed because the result is good enough, and the number of epochs is small.

3. In your experimental results, which categories of concerns were the most difficult to predict? And which categories were these

concerns most often misclassified as?

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| ineffective | 0.73 | 0.72 | 0.73 | 167 |
| unnecessary | 0.66 | 0.32 | 0.43 | 72 |
| pharma | 0.75 | 0.61 | 0.68 | 127 |
| rushed | 0.70 | 0.73 | 0.72 | 147 |
| side-effect | 0.88 | 0.83 | 0.85 | 379 |
| mandatory | 0.76 | 0.74 | 0.75 | 78 |
| country | 1.00 | 0.05 | 0.10 | 20 |
| ingredients | 0.81 | 0.59 | 0.68 | 44 |
| political | 0.64 | 0.37 | 0.46 | 63 |
| none | 0.78 | 0.40 | 0.53 | 63 |
| conspiracy | 0.92 | 0.22 | 0.36 | 49 |
| religious | 0.00 | 0.00 | 0.00 | 6 |
| micro avg | 0.78 | 0.65 | 0.71 | 1215 |
| macro avg | 0.72 | 0.47 | 0.52 | 1215 |
| weighted avg | 0.78 | 0.65 | 0.69 | 1215 |
| samples avg | 0.72 | 0.67 | 0.68 | 1215 |

This is the result after one epoch, we can take this result for example, we can observe that there are some categories with very low F1-scores, such as “country” and “religious”.

For “country”, we can observe that its precision is very high, but recall is very low. This suggests that the model rarely identifies instance of country correctly and instead classifies them as other more frequent labels like “ineffective” or side-effect, which have high recall values.

For “Religious”, there is no true positives, the model likely misclassifies this category as other classes, such as “none” or “conspiracy”.

4. Building on the previous question, what methods have you tried in your experiment to improve model’s ability to more accurately identify the concerns expressed by users? Please describe both the successful and unsuccessful cases.

For successful cases, data augmentation, such as synonym replacement, random insertion, and deletion techniques are used. Also Fine-tuning BERT, such as utilizing BertForSequenceClassification for multi-label classification is used, in order to leverage its pre-trained knowledge and task-specific fine-tuning.

For unsuccessful cases, reducing learning rate is one of them. Since the best result emerges within very small number of epochs. Reducing learning rate based on the previous F1-score fails because once a worse F1-score is observed, the model has already overfitted and would not have better

performance even with smaller learning rate.