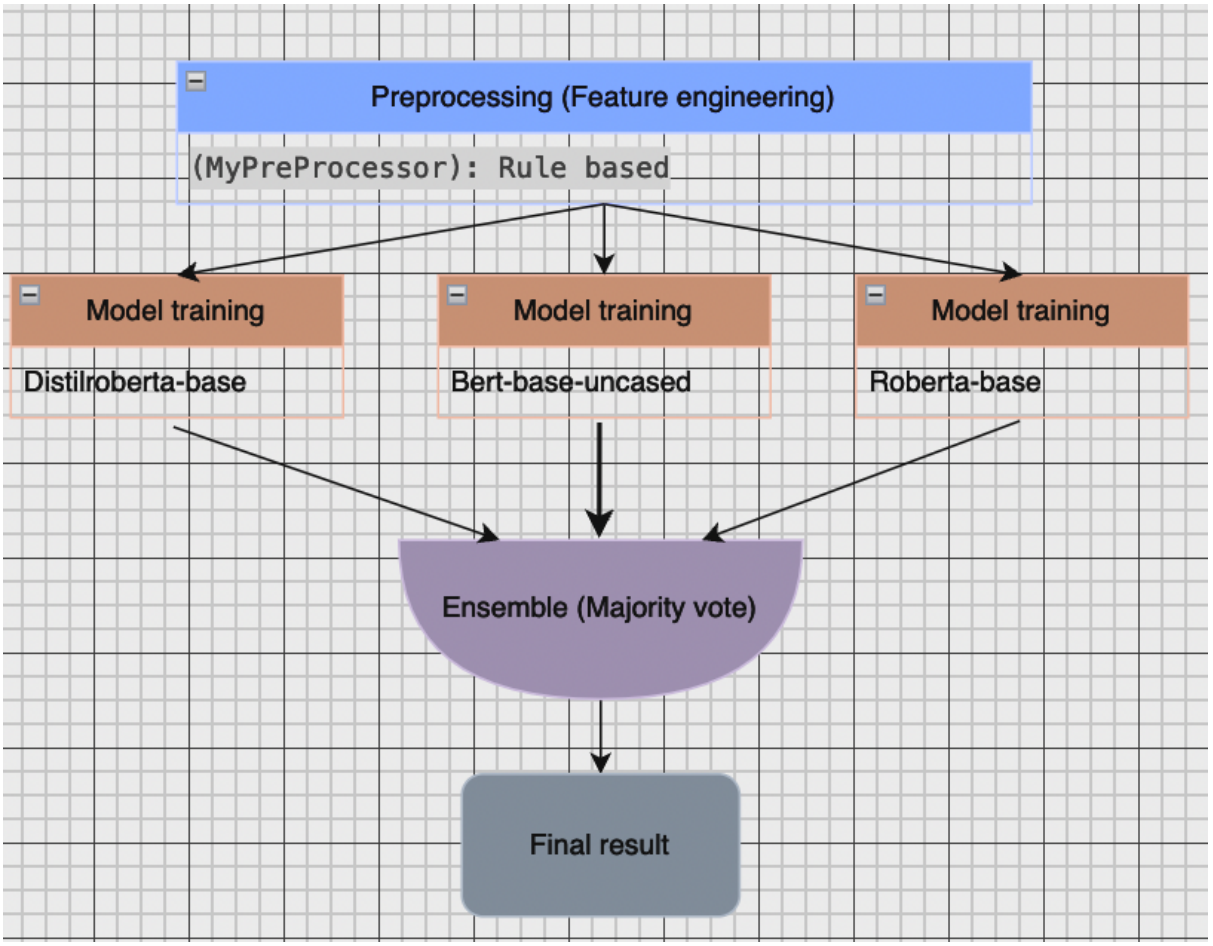
Report

Name: 莊昱陽
Student ID: 111061643
GitHub ID: yuyangdanny
Kaggle name: ABC

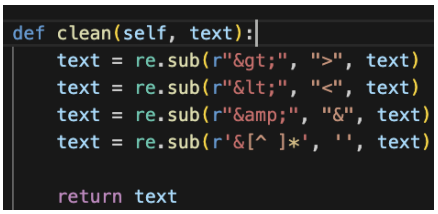
Framework



After read the data, I will do preprocessing (feature engineering) in ‘text’ column and the rule in each row of text is in the picture below:

```
def __call__(self, text):
    # rm html
    html = re.compile(r'<.*?>')
    text = html.sub(r'',text)

    # rm URL
    url = re.compile(r'https?://\S+|www\.\S+')
    text = url.sub(r'',text)

    # rm informal space
    text = self.rm_space(text)

    # clean
    text = self.clean(text)

    # rm metion
    text = re.sub(r'@\S+', '', text)

    # Split connective
    text = re.sub(r'([a-z])([A-Z])', r'\1 \2', text)

    # lower case transformation
    text = text.lower()

    # contractions
    text = self.contractions(text)

    return text
```

1. Remove html
2. Normalize space
3. clean markdown syntax



```
def clean(self, text):
    text = re.sub(r"&gt;", ">", text)
    text = re.sub(r"&lt;", "<", text)
    text = re.sub(r"&amp;", "&", text)
    text = re.sub(r'&[^ ]*', '', text)

    return text
```

4. Remove mention, for example:
   @user -> ‘’
5. Split connective word for example:
   WordSize -> Word size
6. Change into lower case
7. Contractions
   The detail is in code

Model training

I train bert-base-uncased in 3 epochs, roberta, distill roberta both in 2 epochs, and also use a warm-up scheduler in training step.

Ensemble

Finally, I generate a prediction file by ensemble bert-base-uncased, roberta, distill roberta by majority vote.

Experiment

Feature engineer	model	Score (private / public)
Rule based	bert-base-uncased	0.54489 0.56021
Rule based	roberta	0.54738 0.56162
Rule based	distill roberta	0.53641 0.55294
Rule from ekphrasis	bert-base-uncased	0.48748 0.50114
Rule based + Majority vote	**bert-base-uncased roberta distill roberta**	0.55247 0.56895
Rule from ekphrasis + Majority vote	bert-base-uncased roberta distill roberta	0.55031 0.56719