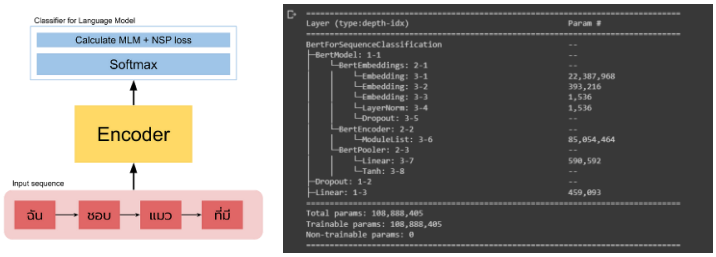
Khunanon Rattanagoses
6332003721

# KAGGLE JobTopGun job title classification

## Proprocessing

- Normalize the text with normalize from pythainlp.
- Tokenize the sentence by using word_tokenize by pythainlp in which the dictionary is customed for specific vocab like 3d, 2d, and เวชสำอาง etc.
- Strip and make to it lower case on the sentence.
- Remove web syntax.
- Change informal word to suitable word, ex "ผจก." : "ผู้จัดการ", 'จนท.' : "เจ้าหน้าที่", etc.
- Remove the word which indicates position including province, district, landmark position, and zone.
- Remove the date and time range e.g.จันทร์ ถึง อาทิตย์, month, year, etc.
- Remove the word which showing work details. For example, “full-time”, “part-time”, “จบใหม่”, “grade”.
- Remove all numbers except the ‘2D’ and ‘3D’ words.
- Remove punctuation characters.
- Correct the incorrect words or undefine words by using correct() from pythainlp for the Thai language and autocorrect from Speller for the English language, the incorrect word is chosen by the word_effect_map which the word is counted as a frequency on data and unique rows.
- Delete stop word both Thai and English and job applications.

## Modeling




Credit: https://medium.com/@chameleontk/ทำความเข้าใจ-bert-098589715545

- I decided to finetune BERT model architecture with a pre-trained Geotrend/bert-base-en-th-cased model which supports both Thai and English language. Because The BERT model proposes the self-attention mechanism which focuses on the important token in the sequence and bidirectional for built context-based representations, I thought it can learn some of the significate tokens in each category.
- First, I split the data into train 80%, validation 10% and test 10% for finding the efficient text preprocessing. After I got a suitable score, I used all the entire dataset to train with 25 epochs, a batch size 32 and maximum sequence length of 32.

## Results

- The result from above process is achieved a score of 0.68196 on the public score and 0.67685 on the private score.