AI Research Engineer Interview Presentation

<Natthaphat Kitisirisuk>

Presentation Structures

- Take-home Assignment
    - Sentiment Analysis (completed)
    - Image Classification (completed)
    - Named Entity Recognition (not complete)
- Me x Wisesight

Take-home Assignment

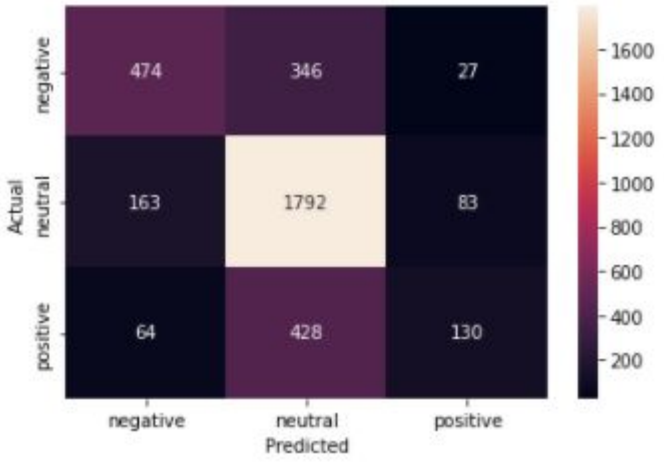
Sentiment Analysis API

Results

- Colab
- Docker

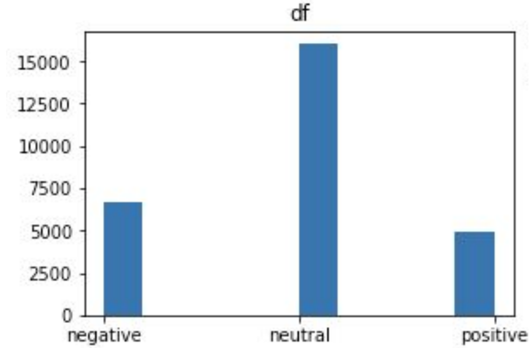
- Logistic Regression is the final model I deployed.
- The model has a high chance to correctly predict negative but not positive and neutral.
- positive texts are likely to be wrong predicted as neutral
- model has a bias toward neutral

Model	Val_Acc
LSTM_1	0.596
LSTM_2	0.61
RNN	0.549
Logistic Regression (ovr)	0.683



Exploring data

	texts	label
	กูไปแก้โครงการที่ รร ตั้งแต่เมื่อวานละ ยังไม่ไ...	neutral
	หิวบาบีก้อน	positive
	จัดปาย เสาร์นี้😌😌	positive
	เซ็ง😏เนื้อติดกะทะ🙊#barbqplaza #wagyubeef	negative
	ปล่อยดอก	neutral



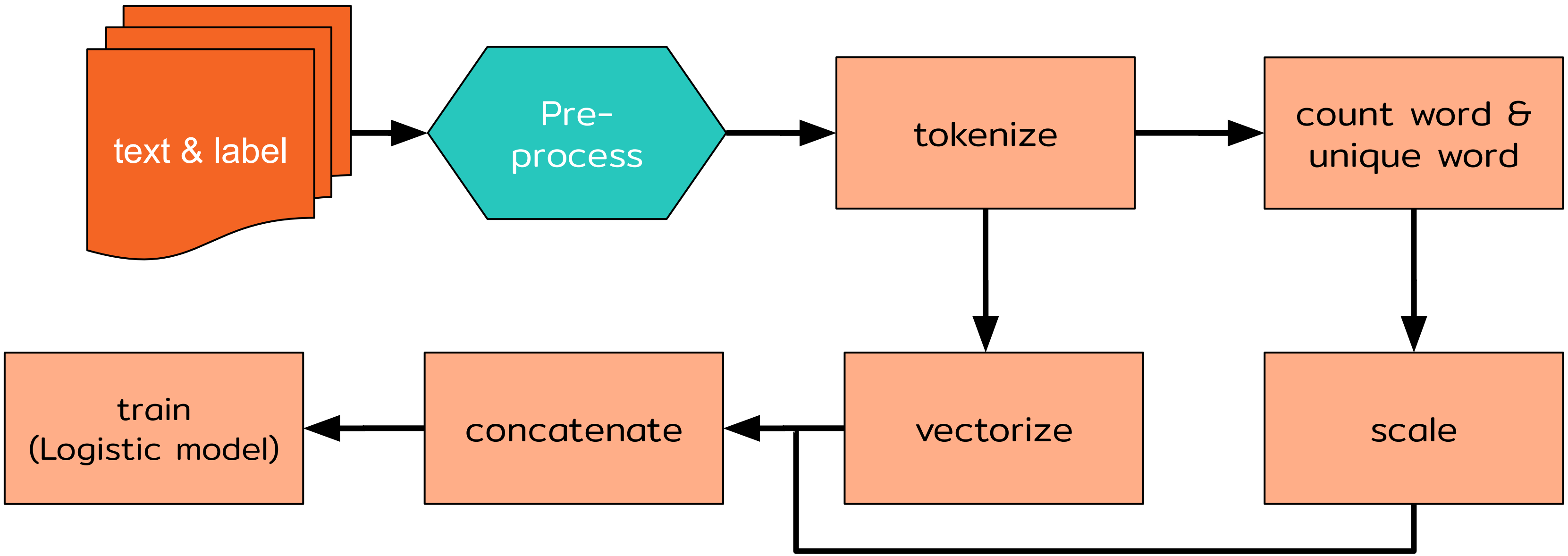
- This data contain text which is the mixture of Thai, English, emoji, etc.
- Informal language
- range from 1-508 words
- 3 classes total : 27504 texts
  - positive : 4877
  - neutral : 15984
  - negative : 6643

Pre-processing data

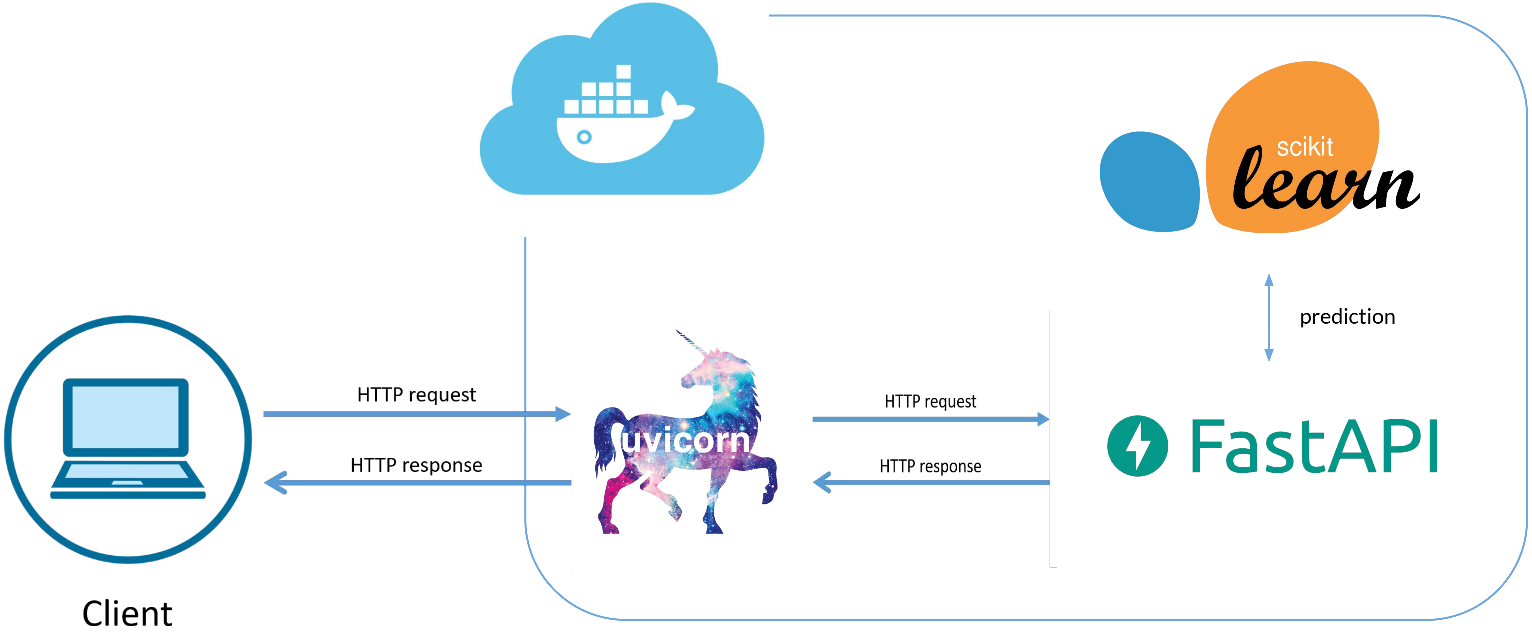
- tokenize text
- vectorize text
- count word and unique word
- scale wc and uwc
- concat vectorized text with scaled wc,uwc

<table>
<tr><th>texts</th><th>label</th><th>processed</th><th>wc</th><th>uwc</th></tr>
<tr><td>กูไปแก้โครงการที่ รร ตั้งแต่เมื่อวานละ ยังไม่ไ...</td><td>neutral</td><td>กู|ไป|แก้|โครงการ|ที่|รร|ตั้งแต่|เมื่อวาน|ละ|ย...</td><td>17</td><td>16</td></tr>
<tr><td>หิวบาบีก้อน</td><td>positive</td><td>หิว|บาบี|ก้อน</td><td>3</td><td>3</td></tr>
<tr><td>จัดปาย เสาร์นี้😌😌</td><td>positive</td><td>จัด|ปา|ย|เสาร์|นี้|xxwrep|😌</td><td>7</td><td>7</td></tr>
<tr><td>เซ็ง😏เนื้อติดกะทะ🙈#barbqplaza #wagyubeef</td><td>negative</td><td>เซ็ง|😏|เนื้อ|ติด|กะทะ|🙈|#|barbqplaza|#|wagyubeef</td><td>10</td><td>9</td></tr>
<tr><td>ปล่อยดอก</td><td>neutral</td><td>ปล่อย|ดอก</td><td>2</td><td>2</td></tr>
</table>

Data processing flow



Sentiment API



What you I have learned

- I found labeling really quite a bit matter. I believe some label should be different. But language is sentiment, so different person might have different feeling.
- FastAPI, I've never work with FastAPI before. It was a great experience
- I've tried to use deep learning (LSTM), but my models' performance were quite low and overfitting. I think it might take too much time tuning and training, so I tried to use supervised learning. Surprisingly, it has higher accuracy than deep learning models.
- I wasted a lot of time training deep learning model, so I might need to be caution.
- Might have to try the method that take less time and easier to implement first, before trying more complicate method.

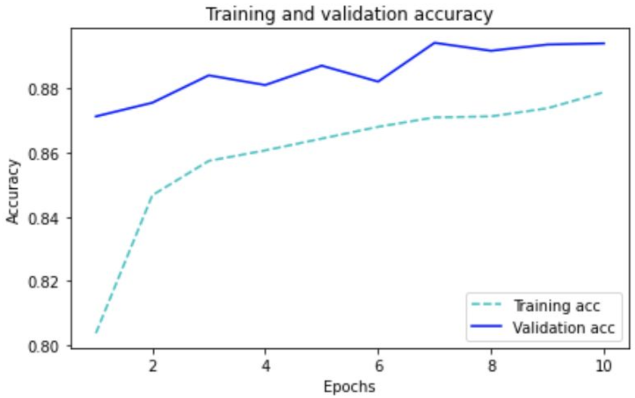
Image Classification

Result

- [Colab]
- [Docker]
  - /api/classify
  - (document)/docs

- I use deep learning to train dataset
- use pre-trained model VGG16
- 3 classes classification (cats, dogs, others)
- Trained for 10 epoch, the train accuracy is 0.8788 and validation accuracy is 0.894
- From graph, the model seem underfitting, but the actual train acc and val acc is quite close.

Model	Val_Acc
CNN_1	0.847
CNN_2	0.84
RestNet50	0.625
VGG16	0.894



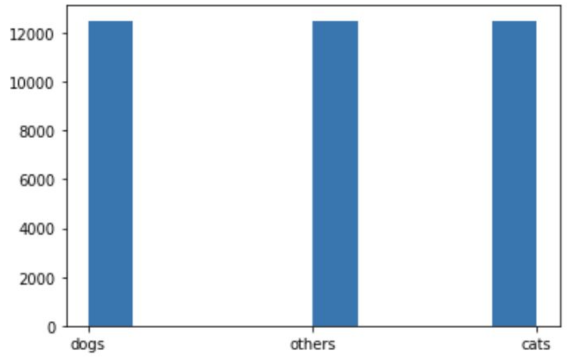
Exploring data

- I found the given dataset didn't come with label, and contain images of food, places etc.
- I decide to use open dataset for training.

- 3 classes : total 37500 images
  - cats : 12500
  - dogs : 12500
  - others : 12500

Preprocessing

- create image data generator for augmentation
  - rotate
  - rescale
  - resize
  - horizontal flip
  - height shift



What you have learned

- From the training history, the accuracy tend to rise further, so I might have to train for more than 10 epoch.
- I want to try make validation accuracy has the value close to train accuracy more than it is right now.

Named Entity Recognition

Preprocessing

- Data contain 28055 texts (from sentiment dataset)
- split word, tag pos and name entities.
- Colab

words	pos	ner
[ตะ, เดือน, ได]	[NN, VV, NN]	[O, O, O]
[คุณ, ลูกค้า, สามารถ, เข้าไป, อ่าน, รายละเอียด...	[NN, NN, AX, VV, VV, NN, NN, AV, VV, PS, PU, N...	[B-PERSON, I-PERSON, O, O, O, O, O, O, O, O, O...
[25, นะ]	[NU, PA]	[B-ORGANIZATION, I-ORGANIZATION]
[มา, ทาน, ที่, สาขา, เซ, นท, รัล, รัตนาธิเบศร์...	[AV, VV, PS, NN, NN, NN, NN, NN, PU, AX, NN, N...	[O, O, O, B-LOCATION, I-LOCATION, I-LOCATION, ...
[ใช้, นาวา, รา, , มา, , 4, , ปี, , ไม, วิ่...	[VV, NN, NN, PU, AV, PU, NU, PU, CL, PU, NN, V...	[O, O, O, O, O, O, B-TIME, I-TIME, I-TIME, O, ...
[รถยนต์, โตโยต้า, ไฮลักซ์, รุ่น, ที่, หาย, ไป,...	[NN, NN, NN, NN, CC, VV, AV, VV, PU, NN, PU, N...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O]
[สงสาร, นาง, นะ, , เจอ, บุหรี่, ไฟฟ้า, ก็, ยี...	[VV, NN, PA, PU, VV, NN, NN, CC, VV, AV, VV, P...	[O, B-PERSON, I-PERSON, I-PERSON, I-PERSON, I-...
[กว่า, จะ, ร็อก, เท่า, วันนี้, , ใน, ทุกวัน, ...	[CC, AX, VV, VV, NN, PU, PS, NN, NN, PR, CC, V...	[O, O, O, O, B-DATE, O, O, O, O, O, O, O, O, O...
[อยุธยา, ไม่, เห็น, มี]	[NN, NG, VV, VV]	[B-LOCATION, O, O, O]
[ใคร, จะ, ใส่, ชุด, ไทย, ไป, กิน, , ค่า, ชุด,...	[PR, AX, VV, NN, NN, AV, VV, PU, NN, NN, VV, A...	[O, O, O, O, B-LOCATION, O, O, O, O, O, O, O, ...

What you have learned

- I’ve never done NER before, so this is my first time learning about it, especially tagging.
- I spent too much time studying about tagging and tried several libraries, so I might need to work more on time management.
- I ended up with PyThaiNLP which I found very effective for Thai.

Me x Wisesight

I want to do this

- While I was taking a shower, I came up with an idea to classify text to identify 'สามกีบ', 'สลิ่ม', 'IO' and 'neutral'. This might help filter out some text and message to display in some occasion.

Finished