Final Presentation – ML Course

TWEET SENTIMENT EXTRATION
INTRODUCTION

- Anabel Ovide
- Markel Azpeitia
- Alejandro Villoria
- Ike Ofodile

GITHUB REPO

https://github.com/alejandrovgonzalez/mlProject
An NLP PROBLEM

- Predict word/sentence that expresses tweet sentiment

<table>
<thead>
<tr>
<th>textID</th>
<th>text</th>
<th>selected_text</th>
<th>sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>I’d have responded, if I were going</td>
<td>I’d have responded, if I were going</td>
<td>neutral</td>
</tr>
<tr>
<td>1</td>
<td>Sooo SAD I will miss you here in San Diego!!!</td>
<td>Sooo SAD</td>
<td>negative</td>
</tr>
<tr>
<td>2</td>
<td>my boss is bullying me...</td>
<td>bullying me</td>
<td>negative</td>
</tr>
<tr>
<td>3</td>
<td>what interview! leave me alone</td>
<td>leave me alone</td>
<td>negative</td>
</tr>
<tr>
<td>4</td>
<td>Sons of ****, why couldn`t they put them on t...</td>
<td>Sons of ****,</td>
<td>negative</td>
</tr>
</tbody>
</table>
MODEL: RoBERTa

- Robustly Optimized BERT Pretraining Approach
- But, what is BERT?
- BERT is a Deep learning model for natural language processing tasks.
DATA AUGMENTATION

- Back translation
- Random insert
- Random deletion
- Synonym replacement
DATA AUGMENTATION

- Back translation
- Random insert
- Random deletion
- Synonym replacement

CAREFUL!
POST PROCESSING

In texts with multiple consecutive whitespaces, the prediction is moved to a wrong location:

Text and predicted text:

“is _ back_ home _ now _ _______ gonna _ miss _ everyone”

If we move the prediction by the number of extra whitespaces:

“is _ back_ home _ now _ gonna _ miss _ everyone”  Correct prediction
POST PROCESSING

In texts with multiple consecutive whitespaces, the prediction is moved to a wrong location:

Text and predicted text:

“is back home now gonna miss everyone”

Score improved!!

If we move the prediction by the number of extra whitespaces:

“is back home now gonna miss everyone”  Correct prediction
Creating tokens

- Library tokenizer
- Save elements text into tokens and ids
- roBERTa reads this tokens
- Problem? For “!!!”, “???”,”...” one single token.
- Less combinations to train.
- Solution? Split tokens!
Creating tokens

- Library tokenizer
- Save elements as phrases
- roBERTa reads the phrases

- Problem: !!!, ???, ... one token
- Less combinations to train.
- Solution? Split tokens!

Doesn’t work... :(
Underestimate with Jaccard

- Usually notebooks overestimate
- If underestimate higher Jaccard score

Text: My dog is beautiful

Selected Text: g is beautiful

Train with...

is beautiful
Underestimate with Jaccard

- Usually notebooks overestimate
- If underestimate higher Jaccard score

My dog is beautiful

g is beautiful

Train with...

It works!!

is beautiful
DECAYING LEARNING RATE

- Reduce model’s learning rate for each epoch.
- Found learning rate values by plotting loss in validation and trial and error.
- Helps the model understand more complex patterns.
- Prevents overfitting.
DECAYING LEARNING RATE

- Reduce model’s learning rate for each epoch.
- Found by plotting loss over training epoch and trial and error.
- Helps the model understand more complex patterns.
- Prevents overfitting.

Keep trying!
LABEL SMOOTHING

- Regularization Technique
- Used when the loss function is cross entropy
- Softmax activation function used on the model
- Tensorflow - `CategoricalCrossentropy`

Private Score: 0.71147 → 0.71383
Public Score: 0.70469 → 0.70987

After Label smoothing
LABEL SMOOTHING

- Regularization Technique
- Used when the loss function is cross entropy
- Softmax activation function used on the model
- Tensorflow - CategoricalCrossentropy

Private Score: 0.71147 → 0.71383
Public Score: 0.70469 → 0.70987

Cool!! After Label smoothing
RESULTS

Final score: 0.71

- Natural Language Processing
- Transformers (BERT, RoBERTa, etc)
- Better understanding of ML concepts