

NLP Final Assignment

Fall 2021, CU Boulder

Professor James H. Martin

Students: Sushma Akoju, Waad Alharthi

| | |
|---|----|
| NLP Final Assignment | 1 |
| Introduction | 1 |
| Code and results: | 2 |
| Contribution from Sushma Akoju: | 2 |
| Understanding the Dataset: | 2 |
| Summary of understanding the Dataset: | 3 |
| Goal of the task: | 4 |
| Defining Classification problem: | 5 |
| Details about the dataset: | 5 |
| The analysis for the problem from the dataset: | 5 |
| Findings from Literature review: | 6 |
| RoBERTa classification using HuggingFace Tutorial Approach: | 7 |
| About the approach: | 7 |
| Evaluation: | 7 |
| SpanBERT for Multilabel classification | 8 |
| About this Approach: | 8 |
| Steps: | 8 |
| Evaluation: | 8 |
| Summary of Analysis: | 8 |
| Github repository: | 9 |
| References: | 9 |
| Contribution from Waad Alharthi: | 10 |
| BERT Text classification using HuggingFace Tutorial Approach: | 10 |
| Collaboration and Teamwork: | 11 |

Introduction

We are a team of 2. Each of us explored broadly among most approaches as discussed in homework. We attempt each of following approaches and conducted training and predictions for NER task:

- Using pre-trained BERT to classify binary PCL (HuggingFace)
- Using pre-trained RoBERTa to classify binary PCL (HuggingFace)
- Determining words that contribute towards PCL and adding extra features and adding probability of agreement/disagreement for sampling with 10% shuffling between the two. And use this as a feature. And explore Custom Named Entity extraction for PCL words and

work out the classification approach (just by taking max number of PCL entities in a test sentence). (using SpaCY)

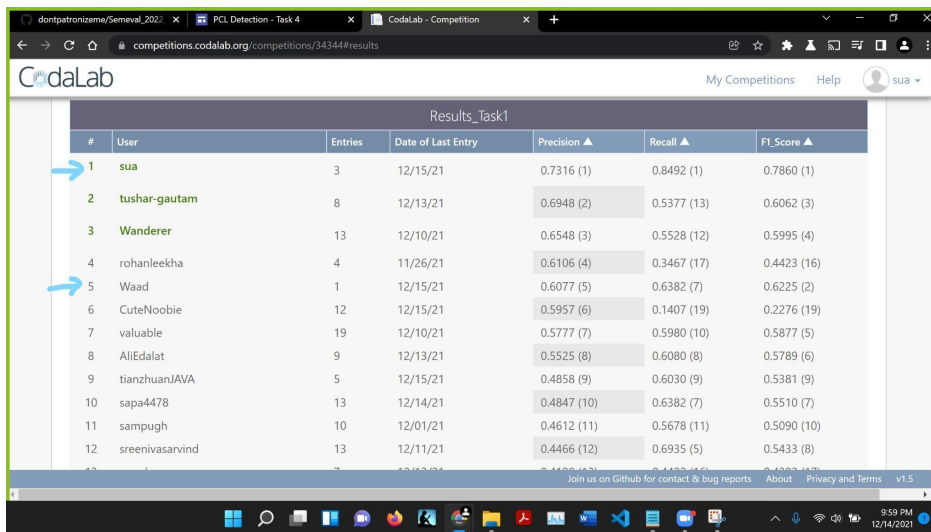
We included prediction results for following in folder task4-hw4/:

- RoBERTa roberta_results.csv (located@ /sushma/task4-1_roberta_sushma.ipynb)
- BERT bert_results.csv (located@ /waad/final_bert_hw4_waad.ipynb)
- Drive folder in case

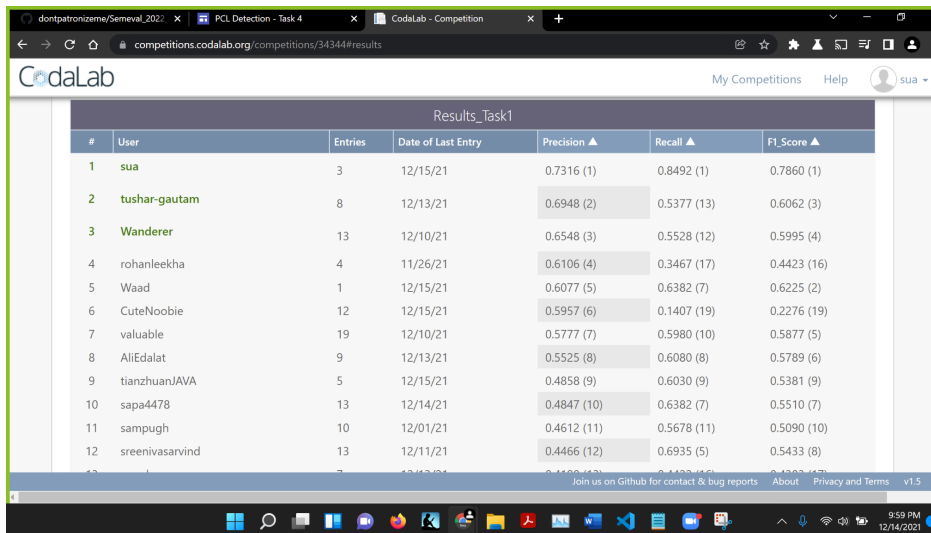
https://docs.google.com/document/d/1Ym5ymzl4wlngul5hDwHBN6l6yX2l5ITGkwaZU_c1gxU/edit?usp=sharing

Code and results:

Leaderboard results:



| # | User | Entries | Date of Last Entry | Precision ▲ | Recall ▲ | F1_Score ▲ |
|----|-----------------|---------|--------------------|-------------|-------------|-------------|
| 1 | sua | 3 | 12/15/21 | 0.7316 (1) | 0.8492 (1) | 0.7860 (1) |
| 2 | tushar-gautam | 8 | 12/13/21 | 0.6948 (2) | 0.5377 (13) | 0.6062 (3) |
| 3 | Wanderer | 13 | 12/10/21 | 0.6548 (3) | 0.5528 (12) | 0.5995 (4) |
| 4 | rohanleekha | 4 | 11/26/21 | 0.6106 (4) | 0.3467 (17) | 0.4423 (16) |
| 5 | Waad | 1 | 12/15/21 | 0.6077 (5) | 0.6382 (7) | 0.6225 (2) |
| 6 | CuteNoobie | 12 | 12/15/21 | 0.5957 (6) | 0.1407 (19) | 0.2276 (19) |
| 7 | valuable | 19 | 12/10/21 | 0.5777 (7) | 0.5980 (10) | 0.5877 (5) |
| 8 | AliEdalat | 9 | 12/13/21 | 0.5525 (8) | 0.6080 (8) | 0.5789 (6) |
| 9 | tianzhuanJAVA | 5 | 12/15/21 | 0.4858 (9) | 0.6030 (9) | 0.5381 (9) |
| 10 | sapa4478 | 13 | 12/14/21 | 0.4847 (10) | 0.6382 (7) | 0.5510 (7) |
| 11 | sampugh | 10 | 12/01/21 | 0.4612 (11) | 0.5678 (11) | 0.5090 (10) |
| 12 | sreenivasarvind | 13 | 12/11/21 | 0.4466 (12) | 0.6935 (5) | 0.5433 (8) |



| # | User | Entries | Date of Last Entry | Precision ▲ | Recall ▲ | F1_Score ▲ |
|----|-----------------|---------|--------------------|-------------|-------------|-------------|
| 1 | sua | 3 | 12/15/21 | 0.7316 (1) | 0.8492 (1) | 0.7860 (1) |
| 2 | tushar-gautam | 8 | 12/13/21 | 0.6948 (2) | 0.5377 (13) | 0.6062 (3) |
| 3 | Wanderer | 13 | 12/10/21 | 0.6548 (3) | 0.5528 (12) | 0.5995 (4) |
| 4 | rohanleekha | 4 | 11/26/21 | 0.6106 (4) | 0.3467 (17) | 0.4423 (16) |
| 5 | Waad | 1 | 12/15/21 | 0.6077 (5) | 0.6382 (7) | 0.6225 (2) |
| 6 | CuteNoobie | 12 | 12/15/21 | 0.5957 (6) | 0.1407 (19) | 0.2276 (19) |
| 7 | valuable | 19 | 12/10/21 | 0.5777 (7) | 0.5980 (10) | 0.5877 (5) |
| 8 | AliEdalat | 9 | 12/13/21 | 0.5525 (8) | 0.6080 (8) | 0.5789 (6) |
| 9 | tianzhuanJAVA | 5 | 12/15/21 | 0.4858 (9) | 0.6030 (9) | 0.5381 (9) |
| 10 | sapa4478 | 13 | 12/14/21 | 0.4847 (10) | 0.6382 (7) | 0.5510 (7) |
| 11 | sampugh | 10 | 12/01/21 | 0.4612 (11) | 0.5678 (11) | 0.5090 (10) |
| 12 | sreenivasarvind | 13 | 12/11/21 | 0.4466 (12) | 0.6935 (5) | 0.5433 (8) |

Code:

https://drive.google.com/drive/folders/1uuGDBwTI3U0r7-ECaG1obs_Yi6Q8YcUH?usp=sharing

Results:

1. Sushma's work:

<https://drive.google.com/drive/folders/1g6oECi0EhR9aaZD5PNZvNal08L8rYRdE?usp=sharing>

2. Waad's work:

<https://drive.google.com/drive/folders/1hwTGQ267D7US6XiggBizPVg5wtXvxtyz?usp=sharing>

Contribution from Sushma Akoju:

Understanding the Dataset:

The paper describes about 7 types of relations that are popular in the community.

- Unbalanced power relations: The help to underprivileged people is often assumed to be expected by the underprivileged. By allowing the right to receive, accepting such help is completely in the hands of the person or group receiving such help.
- Shallow solution: There are some news feeds that common news readers have outgrown. One such example is how a person in power donates a large sum to a minority organization or a NGO or weaker section in the community. Because such news is heavily focussed on the act of giving and much less on how receiver utilizes such donation or how such donation impacts for the long term gain of that weaker section of community seems less focussed. There is research in journalism conducted by Japanese news agencies - their main goal is to follow up sensational stories. All the news stories that are sensationalized and we never bring closure or followup to what happened in the society. We need this kind of information to understand shallow solutions.
- Presupposition: This seems like a common aspect often encountered. Unlike research communities, the general public or people who directly work in news media often go with uncited resources and do not mention the source of information. Identifying such presupposing instances seems to be not an easy task.
- Authority voice: Community-based leadership is often held on a single/small group of people as the leaders of the group. A special case is when one person assumes such a leadership, naturally and the rest of the community naturally accepts it. Along with such responsibilities comes the burden of authority, authenticity and integrity. Such instances bring forth the underlying biases, the way leader/s suggest solutions to community problems.

- Metaphor: The burden of expressing something subtly comes with a larger penalty in journalism. Such expressions have been long held and are in fact a part of social responsibility of the journalist. On the other hand, such subtle expressions also contain hidden PCL. For example, it is not far off from detecting the presence of PCL. Some writers/journalists prefer euphemisms, but it can be helpful to detect underlying problems being presented in a softer way. As suggested by the authors, euphemisms are one such example of hidden PCL. It is also difficult to detect since they are not “explicit” in sentences. They may require advanced techniques, but paraphrasing short phrases and identifying synonymous phrases can help to bring out the true meaning of the subtle expressions in language.
- Compassion: As described by the authors, this presents with a certain poetic quality yet still describes it as needy.
- The poorer, the merrier: this is also commonly used in news coverage, how a vulnerable person or community is asked to come forward to speak while such vulnerable communities do come forward and feel good speaking up. There was one aptly raised question on NPR and 538 politics that suggests “who is listening to vulnerable people?”.

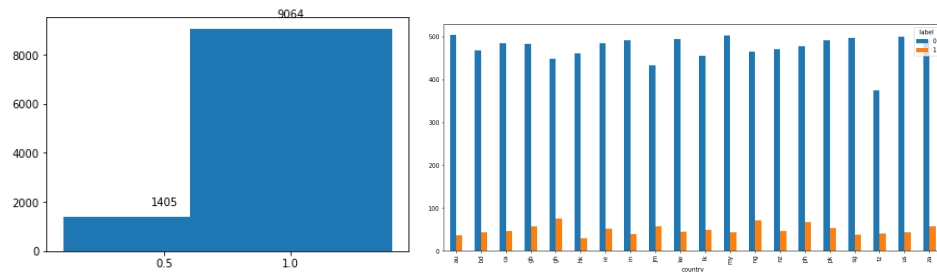
Summary of understanding the Dataset:

In general news coverage over online news feeds or video news coverage, there is certain overcompensation for leading a group or overcompensation for being part of a vulnerable community itself. Such insufficient compensation of social exchange, power dynamics, it seems to have raised the question of PCL. Since the dataset was picked up from news journals, news feeds, this seems to be the case of how journalism, news coverage, audiences and recipients of news contributed to cognitive bias, confirmation bias and other implicit biases that are pointed out by each of 7 categories. Since a lot of recommended articles follow the pattern of suggesting similar articles, this research work and task 4 seems very relevant to addressing some of the core of how the audience, people covered in news, people who cover the news - all together contribute towards underlying patterns such as PCL. This classification task is important to first identify such texts and classify them. I cite this resource for topic-wise stance detection using crowdsourcing, which discussed relevant topics on news feeds. <https://arxiv.org/pdf/2010.03640.pdf>

The dataset consists of 10469 sentences of which 9476 sentences that were labelled as non-patronizing and 993 patronizing/condescending labels for Binary classification task.

| | par_id | art_id | keyword | country | text | orig_label | probability |
|-------|--------|--------|---------|---------|------|------------|-------------|
| label | | | | | | | |
| 0 | 9476 | 9476 | 9476 | 9476 | 9476 | 9476 | 9476 |
| 1 | 993 | 993 | 993 | 993 | 993 | 993 | 993 |

Agreement vs Disagreements Country Wise distribution of PCL:



Multi-labels Analysis:

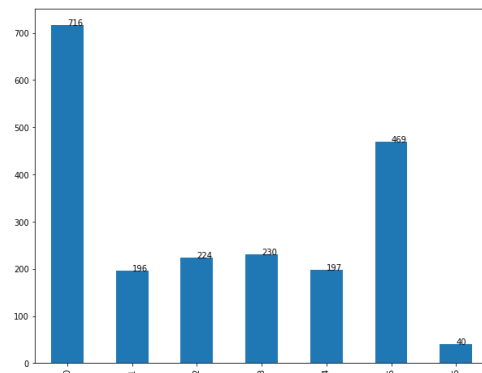
The multi-label analysis indicates that there is an imbalanced distribution of labels.

Notations:

'Unbalanced_power_relations': 0, 'Shallow_solution': 1, 'Presupposition': 2, 'Authority_voice': 3, 'Metaphors': 4, 'Compassion': 5, 'The_poorer_the_merrier': 6

```
pd.DataFrame(pcl_label_counts.sum(axis=0)).T
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|-----|-----|-----|-----|-----|-----|----|
| 0 | 716 | 196 | 224 | 230 | 197 | 469 | 40 |



Goal of the task:

To train an NLP model to classify the PCL/No PCL task

Defining Classification problem:

Classification of each sentence to PCL or non-PCL sentence by using a Language model.

Details about the dataset:

- Agreement between PCL and No-PCL labels:
Add agreement scores i.e. for all labels marked as odd 1 or 3 indicate disagreement i.e. only 50% agreement among 2 annotators. Only one annotator marked the label as PCL true. All others with even numbered labels are the ones who have 100% agreement. *Reference: README.txt provided with dataset.*

As per data analysis, there is 13.42% disagreement between the two annotators and 86.579% agreement between the two annotators.

So the confidence level is expected to be higher for 86.58% labels.

| | par_id | art_id | keyword | country | text | label | orig_label |
|-------------|--------|--------|---------|---------|------|-------|------------|
| probability | | | | | | | |
| 0.5 | 1405 | 1405 | 1405 | 1405 | 1405 | 1405 | 1405 |
| 1.0 | 9064 | 9064 | 9064 | 9064 | 9064 | 9064 | 9064 |

The keyword wise counts for each PCL/NON-PCL labels are as follows:

| keyword | disabled | | homeless | | hopeless | | immigrant | | in-need | | migrant | | poor-families | | refugee | | vulnerable | | women | |
|---------|----------|----|----------|-----|----------|-----|-----------|----|---------|-----|---------|----|---------------|-----|---------|----|------------|----|-------|----|
| label | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| country | 947 | 81 | 899 | 178 | 881 | 124 | 1031 | 30 | 906 | 176 | 1053 | 36 | 759 | 150 | 982 | 86 | 1000 | 80 | 1018 | 52 |

This suggests that keyword-wise labels indicate true PCL labels. To rephrase it, presence of keywords are correlated with labels based on only true presence of PCL or not.

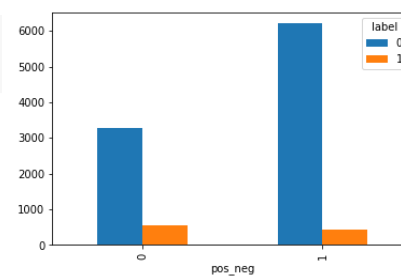
For example, the keyword homeless appears in 899+178 sentences, but the sentence with "homeless" keyword marked as Not PCL 899 times and sentence with "homeless" keyword marked as PCL 178 times. This also suggests that by examining one of the sentences with "homeless", the sentence does have contextual information for one of the phrases.

The analysis for the problem from the dataset:

- Additionally, I did sentiment analysis of all sentences in the dataset using Distilbert pretrained model. The label wise groups over sentiments is as follows:

```
sentimentwise_grp['keyword']
```

| pos_neg | label | |
|---------|-------|------|
| 0 | 0 | 3273 |
| | 1 | 554 |
| 1 | 0 | 6203 |
| | 1 | 439 |



This additional information suggests that out of 9476 non_PCL sentences, 3273 are negative sentiment sentences and remaining 6203 are positive sentiment. Similarly, out of 993 sentences, 554 sentences have negative sentiment and the remaining 439 have positive sentiment.

This suggests that positive/negative sentiment does not correlate directly with PCL/Non-PCL labels. In fact there is a high probability most positive sentiment sentences are 65.4% non-PCL labels and 44.2% of positive sentiment sentences can be PCL sentences even though the number of PCL sentences are only 993. So we can conclude the sentiment analysis has less correlation w.r.t any one label. Due to such a small number of PCL sentences annotated, this becomes a hard problem as described in the research paper.

- I did extract POS tags for each of sentences

```
pcl_df.groupby(['tag', 'label']).count()[17:].T
```

| tag | CC | CD | DT | EX | FW | IN | JJ | JJR | JJS | LS | MD | NN | NNP | NNPS | NNS | | | | | | | | | | | | |
|---------|------|------|-----|-------|------|-----|----|-----|-----|-------|------|-------|------|------|-----|-----|-----|---|------|-----|-------|------|-----|----|---|---|-------|
| label | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | | | | | | | | | | | | |
| sent_id | 1872 | 7934 | 490 | 39020 | 4685 | 720 | 82 | 179 | 24 | 53577 | 6070 | 40486 | 4271 | 1419 | 140 | 910 | 106 | 5 | 4254 | 580 | 85382 | 9518 | 521 | 58 | 7 | 2 | 35639 |

- The POS tag and sentiment labels are less correlated to multi label classification tasks as well.
- However from all of the above analysis, it is apparent that without context, the PCL is difficult to be trained.
- The dataset created with POS tags is located at :
<https://drive.google.com/file/d/1q3Qu5hsgwyqK3xqIRU9LJdTk1JFV6Hh2/view?usp=sharing>

RoBERTa classification using HuggingFace Tutorial Approach:

About the approach:

This is the same approach described in the paper which gave accurate results. I finetune the Roberta-base model with configuration provided in the paper and evaluate the results.

Defining input parameters, probabilities:

The input parameters as per RoBERTa base model configuration used in the paper, 10 epochs with batch size of 32 with a random seed set at 1. I used the default learning rate from HuggingFace i.e. 2e-07.

[2300/2300 9:52:12, Epoch 10/10]

| Step | Training Loss | Validation Loss | Accuracy | Precision | Recall | F1 |
|------|---------------|-----------------|----------|-----------|----------|----------|
| 500 | 0.206800 | 0.255699 | 0.906310 | 0.526786 | 0.567308 | 0.546296 |
| 1000 | 0.089800 | 0.461223 | 0.912046 | 0.550847 | 0.625000 | 0.585586 |
| 1500 | 0.026700 | 0.564355 | 0.915870 | 0.580000 | 0.557692 | 0.568627 |
| 2000 | 0.008500 | 0.578284 | 0.924474 | 0.637363 | 0.557692 | 0.594872 |

Evaluation:

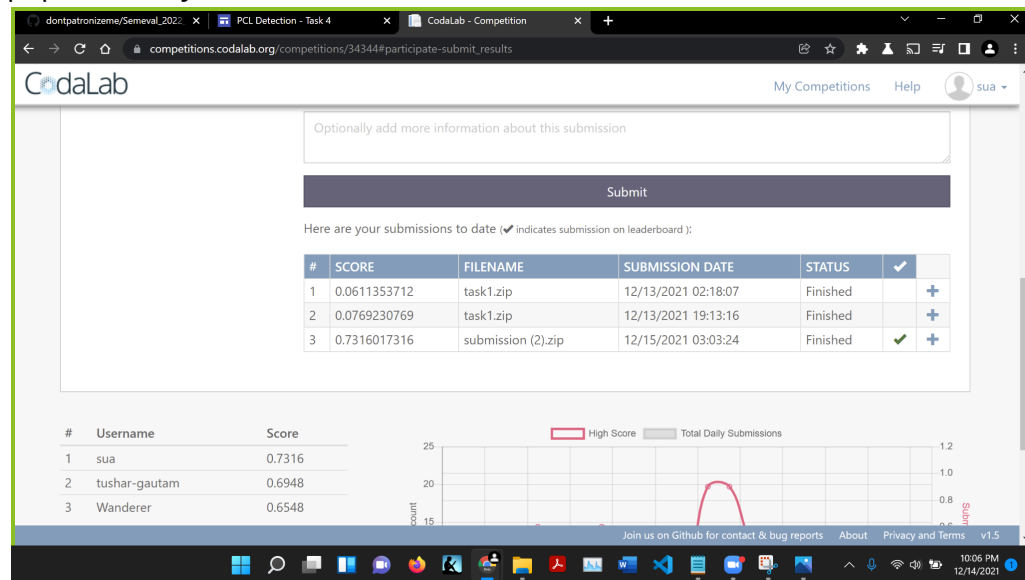
1. Code: [Code Binary label roberta](#)
2. Results: [Binary label Roberta output](#)
3. Results:

From evaluation script provided by the organizers:

```
!cat scores.txt

task1_precision:0.7316017316017316
task1_recall:0.8492462311557789
task1_f1:0.786046511627907
```

F1-score from leaderboard: 0.73 and I was ranked first. Of Course we would expect this since so far Roberta Model with configuration suggested in the paper already does better.



4. Total duration of training: 28 hours over Google Colab, excluding the number of times the browser with Colab session crashed. Training a batch size of 32 requires TPU. Training a batch size of 32 over GPU has not worked on Colab as browser crashes.

5. Additionally we want to be able to “explain” why Roberta does well for binary label classification. Additionally due to computational resource limitations, I could not repeat this experiment for multi-label classification tasks since 32 batch over TPU suffered Colab crashes.

KeyBERT for Detecting and Extracting the keywords and phrases.

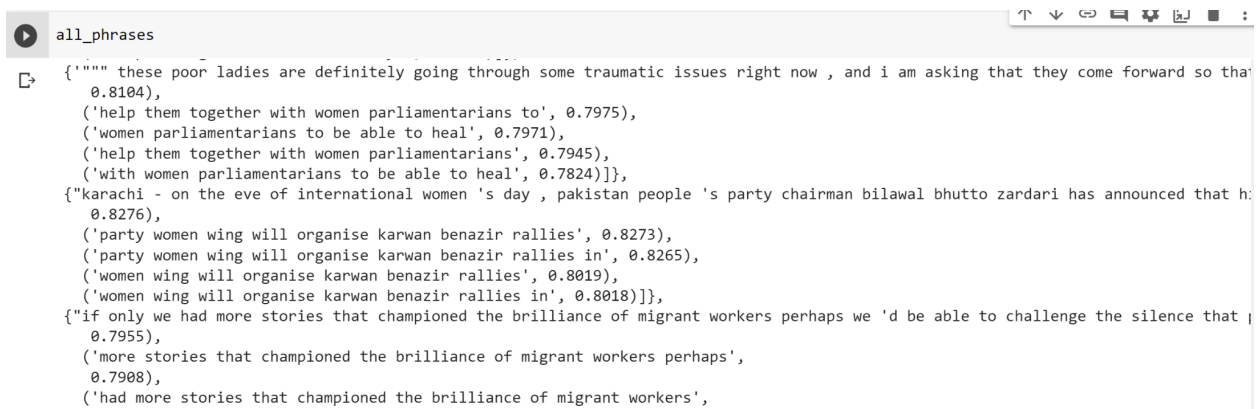
About this Approach:

In an attempt to explore the span of texts in a given sentence that consists of keywords, I found KeyBERT. KeyBERT is a keyword extraction technique that uses BERT embeddings to extract keywords. The keywords of interest are 'hopeless', 'migrant', 'immigrant', 'disabled', 'refugee', 'in-need', 'homeless', 'vulnerable', 'women', 'poor-families'. KeyBERT extracts n-grams which use CountVectorizer to get candidate n-grams which is an input to BERT and calculates distance between candidate n-grams and document and which is then compared for relevancy to the sentence. It minimizes distance to document and maximizes distance to candidate.

Evaluation:

1. Code: [KeyBERT for Phrase extraction as a feature](#)
2. Results: Did not complete classification task but completed analyzing the results of keywords and phrases:

The phrases do include keywords and we can search for phrases that have max scores and select the phrases with keywords which can serve as a feature.



```
all_phrases
{
  "these poor ladies are definitely going through some traumatic issues right now , and i am asking that they come forward so that",
    0.8104),
  ('help them together with women parliamentarians to', 0.7975),
  ('women parliamentarians to be able to heal', 0.7971),
  ('help them together with women parliamentarians', 0.7945),
  ('with women parliamentarians to be able to heal', 0.7824)],
  {"karachi - on the eve of international women 's day , pakistan people 's party chairman bilawal bhutto zardari has announced that h":
    0.8276),
  ('party women wing will organise karwan benazir rallies', 0.8273),
  ('party women wing will organise karwan benazir rallies in', 0.8265),
  ('women wing will organise karwan benazir rallies', 0.8019),
  ('women wing will organise karwan benazir rallies in', 0.8018)],
  {"if only we had more stories that championed the brilliance of migrant workers perhaps we 'd be able to challenge the silence that":
    0.7955),
  ('more stories that championed the brilliance of migrant workers perhaps',
    0.7908),
  ('had more stories that championed the brilliance of migrant workers',
```

The phrases do include keywords and we can search for phrases that have max scores and select the phrases with keywords which can serve as a feature. We can use scores for each of keywords towards a given sentence as a feature for the classification task. By creating a column for each keyword and its corresponding score if it exists otherwise score it as zero.

```

{ 'rome : italy 's far-right interior minister matteo salvini on friday launched a blistering attack on prosecutors in sicily after tl
  0.1801),
  ('immigrant', 0.155),
  ('migrant', 0.1539),
  ('vulnerable', 0.1459),
  ('homeless', 0.0847)]},
{ 'in an act of defiance against hungarian authorities , which had suspended trains to western europe , between 1,200 and 2,000 refug
  0.3842),
  ('migrant', 0.3407),
  ('homeless', 0.2155),
  ('immigrant', 0.1889),
  ('women', 0.1439)]},
{ 'the problem is most cocoa is produced by poor families who can not afford fertilisers and pesticides , the experts noted .' : [['po
  0.4391),
  ('homeless', 0.1368),
  ('hopeless', 0.1211),
  ('immigrant', 0.1155),
  ('migrant', 0.1142)]},

```

SpanBERT for Multilabel classification

About this Approach:

Upon reading a couple of online papers, towards each of the categories (class labels) in multi-label classification task for this PCL task, I found <https://arxiv.org/pdf/2109.04666.pdf>. This paper is heavily designed towards the task of span of text detection by using SpanBERT by making use of AutoPhrase for phrase mining. However we do not attempt to phrase mining. The pipeline suggested in this paper is very relevant towards mining phrases surrounding keywords in this PCL multi-label classification. Since the PCL task and the authors of PCL paper do depend on the span of texts and the number of span of text that contribute towards each of 7 labels, I would like to explore SpanBERT for multi-label classification.

Findings from Literature review:

I want to explain this with an logical reasoning example:

"the demographics of pakistan and india are very similar . poverty is a widespread issue . according to the fao , 40 percent of children in Pakistan are malnourished and underweight due to lack of access to adequate food . and this is not because there is n't enough ; pakistan is the 8th largest food producing country , however , 50 percent of the population is food insecure .

with the massive income inequality that persists , rha is a brilliant movement . we collect leftover or extra food from restaurants and distribute it to the homeless and hungry in the locality ."

The part of the sentence, *"the demographics of pakistan and india are very similar . poverty is a widespread issue . according to the fao , 40 percent of children in Pakistan are malnourished and underweight due to lack of access to adequate food . and this is not because there is n't enough ; pakistan is the 8th largest food producing country , however , 50 percent of the population is food insecure ."* is well founded in evidence that supports the claim that there is malnourishment among 40% of children in Pakistan, but the subsequent sentence suggests the reason is not about lack of enough food being produced since Pakistan is 8th largest food producing country in the world.

But soon after this sentence, there is “jump” in context and no “evidence” to next claim i.e. *with the massive income inequality that persists*

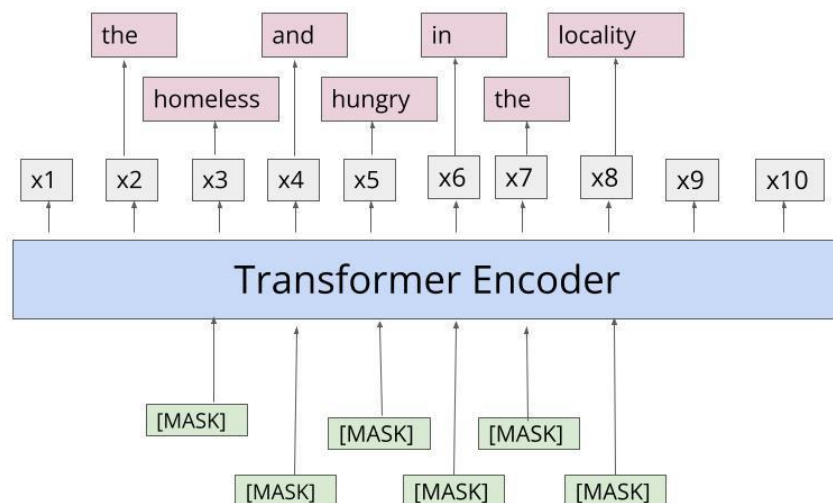
There is no grounding or evidence that supports the claim about income inequality. There is also not enough evidence to suggest that the Rha movement is helping or has improved these numbers or any excerpts from people that received help. This is a clear case of presence of PCL.

Suppose if each claim in this paragraph requires some evidence, then the final inference that Rha movement is helping is less logically supported by the immediate claim that entails it. The immediate claim being massive income inequality which does not have any evidence, is a claim with weak or rather no evidence. There is no correlation established in this paragraph that malnourishment is the cause of income inequality. Such subtle gaps between sentence-evidence pairs for some claims, suggests that this is logically weak causal inference. Such lack of evidence also can be recognized from sudden change in context i.e from income inequality to Rha movement and its apparent success. There is also no evidence of improvement in success from the *rha is a brilliant movement . we collect leftover or extra food from restaurants and distribute it to the homeless and hungry in the locality* . This also seems like an excerpt from the people who were organizing the movement. The fact that this movement was brilliant - it is not clear if this has been appreciated by the writer or the organizer.

I believe that a change in type of task can help to identify this pattern. Suppose we used SpanBERT to recognize the span of texts and by SpanBERT

In the given sentence, the span of words ***‘the homeless and hungry in the locality’*** is masked. The Span Boundary Objective (SBO) is defined by the x3 and x8 highlighted in pink. This is used to predict each token in the masked span.

Keyword from the task4 is ***“homeless”***



$$L(\text{homeless}) = L_{MLM}(\text{homeless}) - L_{SBO}(\text{homeless}) \quad \text{--- equation 1}$$

The loss is summation of losses:

$$L_{MLM}(\text{homeless}) = -\log P(\text{homeless} | x_3) \quad \text{--- equation 2}$$

$$L_{SBO}(\text{homeless}) = -\log P(\text{homeless} | x_2, x_8, P_2) \quad \text{--- equation 3}$$

x_2 - the start of the span boundary

x_8 - the end of the span boundary

P_2 - the position of x_3 (homeless) from the starting point (x_2)

We utilize this information to find how good the model is in predicting the word homeless.

Let the masked span of tokens be: x_s, \dots, x_e

$$\text{SBO function: } Y_i = f(x_{s-1}, x_{e+1}, P_{i-s+1}) \quad \text{--- equation 4}$$

$P_1, P_2 \dots$ are relative positions w.r.t x_{s-1}

SBO function is a two-layered feed forward network with GeLU activation.

$$\text{First hidden representation: } h_0 = [x_{s-1}; x_{e+1}; P_{i-s+1}] \quad \text{---}$$

equation 5

$$\text{Second hidden representation: } h_1 = \text{LayerNorm}(\text{GeLU}(W_1 h_0)) \quad \text{---}$$

equation 6

$$\text{Second hidden representation: } y_i = \text{LayerNorm}(\text{GeLU}(W_2 h_1)) \quad \text{---}$$

equation 7

Second hidden representation:

$$L(x_i) = L_{MLM}(x_i) + L_{SBO}(x_i) = -\log P(x_i | X_i) - \log P(x_i | y_i) \quad \text{--- equation 8}$$

Summary:

With spanBERT, the idea is instead of masking random tokens, we select a span of tokens to mask and expect that it predicts masked tokens.

Steps:

Evaluation:

6. Code: [Code SpanBERT Task4 Multilabel classification](#)
7. Results: [Multilabel classification Output](#)

```
!cat scores.txt

task2_unb:0.3726114649681529
task2_sha:0.2298850574712644
task2_pre:0.2807017543859649
task2_aut:0.15730337078651682
task2_met:0.26666666666666666
task2_com:0.3585858585858586
task2_the:0.0
task2_avg:0.2379648818377749
```

SpanBERT for Single label classification

Evaluation:

1. Code: [SpanBERT Task4 binary label classification](#)
2. Results: [Binary label classification Output](#)

```
!cat scores.txt

task1_precision:0.3283208020050125
task1_recall:0.6582914572864321
task1_f1:0.43812709030100333
```

Summary of Analysis:

1. I learnt about Roberta, SpanBERT and KeyBERT.
2. I learnt about the works and paper and research works in journalisms and detecting euphemisms and metaphors.
3. The PCL task is very helpful. By using additional ranking techniques over Phrase detection and phrase extraction and fine tuning model over extracted phrase extraction features, the model can be improved.

Drive folder:

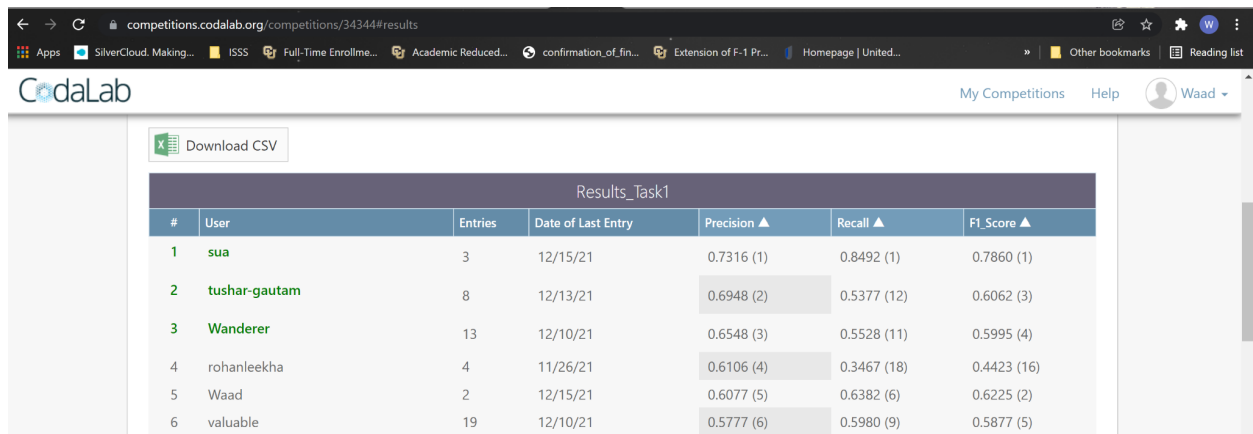
[Sushma's contributions](#)

References:

1. SpanBERT for Euphemism detection and extraction for drug corpus:
<https://arxiv.org/pdf/2109.04666.pdf>
2. RoBERTa Text classification:
<https://towardsdatascience.com/fine-tuning-pretrained-nlp-models-with-huggingfaces-trainer-6326a4456e7b>
3. Hugging Face Tutorial for RoBERTa:
https://huggingface.co/docs/transformers/master/custom_datasets#sequence-classification-with-imdb-reviews

Contribution from Waad Alharthi:

According to the article [1], RoBERTa should be the best model for this given NLP classification task. Since the authors already gave a baseline for RoBERTa I chose to work on “xlnet-large-cased” and “bert-base-cased” [2]. I spent time trying different tokenizing methods that didn’t improve the result at all. Then I struggled with the output printing format. Eventually I copied the authors’ RoBERTa-base baseline collab notebook and simply changed the model name and epochs number to 6. While training for both tasks took longer than anticipated I will add the results as soon as possible. This result shows the score of the code I implemented for subtask1 following the tutorial [3] using “bert-base-cased”



| Results_Task1 | | | | | | |
|---------------|---------------|---------|--------------------|-------------|-------------|-------------|
| # | User | Entries | Date of Last Entry | Precision ▲ | Recall ▲ | F1_Score ▲ |
| 1 | sua | 3 | 12/15/21 | 0.7316 (1) | 0.8492 (1) | 0.7860 (1) |
| 2 | tushar-gautam | 8 | 12/13/21 | 0.6948 (2) | 0.5377 (12) | 0.6062 (3) |
| 3 | Wanderer | 13 | 12/10/21 | 0.6548 (3) | 0.5528 (11) | 0.5995 (4) |
| 4 | rohanleekha | 4 | 11/26/21 | 0.6106 (4) | 0.3467 (18) | 0.4423 (16) |
| 5 | Waad | 2 | 12/15/21 | 0.6077 (5) | 0.6382 (6) | 0.6225 (2) |
| 6 | valuable | 19 | 12/10/21 | 0.5777 (6) | 0.5980 (9) | 0.5877 (5) |

References:

1. <https://towardsdatascience.com/battle-of-the-transformers-electra-bert-roberta-or-xlnet-40607e97aba3>
2. https://huggingface.co/transformers/v2.0.0/pretrained_models.html
3. <https://towardsdatascience.com/fine-tuning-pretrained-nlp-models-with-huggingfaces-trainer-6326a4456e7b>

Collaboration and Teamwork:

We followed this approach to work and collaborate together:

1. We each discussed our thoughts and analysis.

2. We each discussed and shared ideas, approaches we each thought of.
3. We shared feedback and helped each other to explore each of our ideas.
4. We learned from each other and were able to bring nice coordination to each of our thoughts.
5. In the end, we each worked on exploring each of our own ideas and implementing them.
6. We encouraged each other to discuss.