Fact checking Application

NLP Final Project 09th March 2023

Application Details

- Create an automated fact check application. Idea based on the findings and dataset from the PUBHEALTH paper.
- This application takes as input a claim text checks it against a news article and predicts if the claim is True/False. This application also generates an explanation text for the claim based on the text of the news article.
- Metrics: For claim classification we will use Precision/Recall/F1 macro scores. For explanation which is a summarized version we will use Rouge metrics for unigram (R1), bi-gram (R2) and longest common subsequence (RL)

Dataset Example

- Claim: *A cat has tested positive for rabies.*
- Evidence text (News article): *She says the cat was likely exposed to a rabid bat, and that's how it became infected. The Department of Health and Welfare says the cat in Owyhee County was behaving aggressively and bit its owner…….* (Truncated)
- Ground Truth Label: *True*
- Ground truth Explanation*: Idaho health officials say a cat has tested positive for rabies for the first time in 27 years*.
- Subject: *News, Health*

Tasks

Task 1 (Text classification):

Input: Claim Text, Evidence Text

Prediction: True/False

Compare prediction label against GT label.

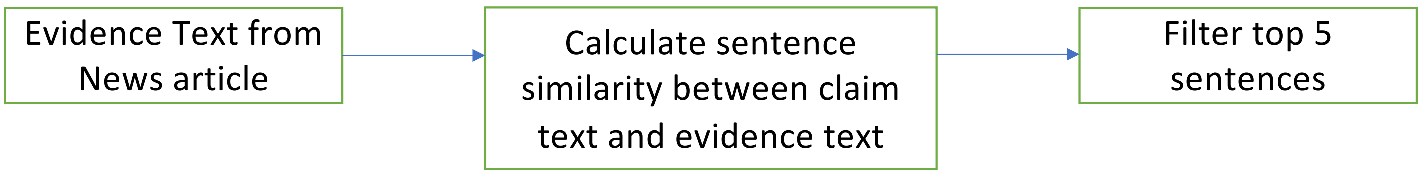
Task 2 (Text Generation):

Input: Evidence Text

Prediction: Generated Text. Compare predicted text against generated text.

Pre-processing

- Pick top 5 similar sentences from evidence text (top_k)



- Similarity based on cosine similarity between claim text and sentences from evidence text using sentence transformer model
- Issues found: The published paper code on GitHub was incorrectly picking most dissimilar sentences instead of similar sentences for top_k. (k=5)

- Second approach - Create pairs of claim text, one sentence from evidence text

Task 1 (Claim Classification)

- Use the entire dataset
- We switched to DistillBert model from Bert Base as it makes experimentation easier
- DistillBert is about 40% smaller compared to Bert Base models. 66M params vs 110M params for bert-base-cased model
- Trained using GPU on M1 Mac using Pytorch
- Fine-tuned for 5 epochs
- We didn't really explore further by changing weights etc.
- 0: True, 1: False, 2: Mixture, 3: Unproven

	precision	recall	f1-score	support
0	0.8021	**0.7513**	0.7759	**599**
1	0.5788	**0.7191**	0.6414	**388**
2	0.3053	**0.2886**	0.2967	**201**
3	0	**0**	0	**45**
Accuracy			**0.6383**	
Macro Avg	0.4216	**0.4397**	0.4285	**1233**
Weighted avg	0.6216	**0.6383**	0.6271	1233

Task 1 (Claim Classification)

- What is we map everything but True to False
- In this case 0: True, 1: False/Unproven/Mixture
- Fine-tuned for 2 epochs
- Model used: DistillBertUncased
- Setup: M1 Mac 1 GPU

	precision	recall	f1-score	support
0	0.7964	**0.803**	0.7997	599
1	0.8124	**0.806**	0.8092	634
accuracy			0.8045	1233
macro	0.8044	**0.8045**	0.8044	1233
weighted	0.8046	**0.8045**	0.8046	1233

Task 1 (Claim Classification)

- Use claim, sentence pairs
- In this case 0: True, 1: False/Unproven/Mixture
- Fine-tuned for 1 epochs
- Model used: DistillBertUncased
- Setup: M1 Mac 1 GPU,
- Training time – 12+ hours
- Could not try further due to resource constraints

	precision	recall	f1-score	support
0	0.6838	**0.798**	0.7365	599
1	0.7734	**0.6514**	0.7072	634
accuracy			0.7226	1233
macro	0.7286	**0.7247**	0.7219	1233
weighted	0.7299	**0.7226**	0.7214	1233

Task 1 (Claim Classification)

- The best True/False [False/Unproven/Mixture] classification model has recall and f1-score of around 80% using top_k approach. But let's look at some failure cases

- Example 1 (Fails to detect unrelated text):

claim	The new supplement InteliGEN can boost brain function
top_k	the aircraft will also be used to evacuate injured, elderly and young people. authorities urged a mass exodus from several towns on the southeast coast, an area popular with tourists during the summer holiday season, warning that extreme heat forecast for the weekend will further stoke the fires. temperatures are forecast to soar above 40 degrees celsius (104 degrees fahrenheit) along the south coast on saturday, bringing the prospect of renewed firefronts to add to the around 200 current blazes. "the priority today is fighting fires and evacuating, getting people to safety," prime minister scott morrison told reporters in sydney. "it is going to be a very dangerous day.
gt_label	FALSE
pred_label	TRUE

Task 1 (Claim Classification)

- Example 2 (In this case the explanation expected the claim to be more rigorous)
- GT Explanation: "Not only does this story neglect to provide any caveats regarding research abstracts presented at conferences, it omits the number of subjects in the study. One of the most important pieces of context for this study was that it only had 12 subjects. That needed to be in the article. In general, the term "artifical pancreas" builds unrealistic hope for this technology for patients,"

claim	Artificial Pancreas Continues to Show Promise
evidence_text	One the one hand, the article tells us that the technology is emerging, the algorithm is still being developed, and the whole approach is still being researched in future studies. The end of the article explains the artificial pancreas technology, implying that the computer linkage between monitor and pump is the novel part that is still under development. — is provided. What's missing is at least some emphasis, ideally early, that this research is quite experimental at this point, with some discussion of the steps between this small study and the technology potentially going to market. It could've been clearer about how the "closed loop" technology is supposed to improve on the available devices.
gt_label	FALSE
pred_label	TRUE

Task 2 (Explanation Generation)

- Pre-processing – We will use the <claim_text, top_k text> as inputs and try to generate explanation text
- This is abstractive summarization as the explanation was written by human annotators

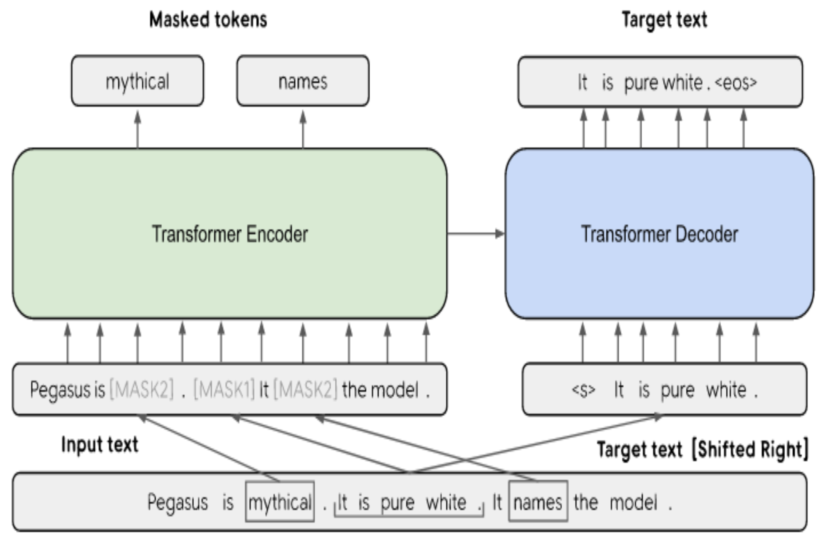
Task 2 (Explanation Generation)

- Using entire available dataset.
- Performance on test dataset (n=1233)
- Fine-tuned a t5-small model for 3 epochs
- Tried different models and the PEAGUSUS model without fine-tuning
- PEGASUS clearly works better then t5-small

Model	R1	R2	RL
t5-small (fine-tuned)	0.1918	0.0574	0.1532
PEAGUSUS (no fine tuning)	0.2317	0.07284	0.1737

PEGASUS (briefly)

- Pre-training with Extracted Gap-sentences for Ab-stractive SUmmarization Sequence-to-sequence models
- Use Masked Language modelling (MLM) like BERT and GSG (Gap sentence generation)
- Training data – C4 and HugeNews
- Uses self-supervised objective GSG as there is lack of ground truth data for abstractive summarization
- Two versions PEGASUS BASE (266M params), PEGASUS Large (568M params). For comparison t5-small is 60M params.



Task 2 (Explanation Generation)

- Results after fine-tuning PEAGUSUS
- Fine-tuning for 1 epoch improves ROUGE scores
- Tried two combinations Input length: 256, Output length: 128
- Tried two combinations Input length: 256, Output length: 64
- It seems restricting the output length generating shorter sentences the accuracy is better

Model Name	R1	R2	RL
t5-small (fine-tuned)	0.1918	0.0574	0.1532
PEAGUSUS (no fine tuning)	0.2317	0.07284	0.1737
PEAGUSUS (fine tuned) 256/128	0.2938	0.1085	0.2151
PEAGUSUS (fine tuned) 256/64	0.3084	0.1194	0.2319

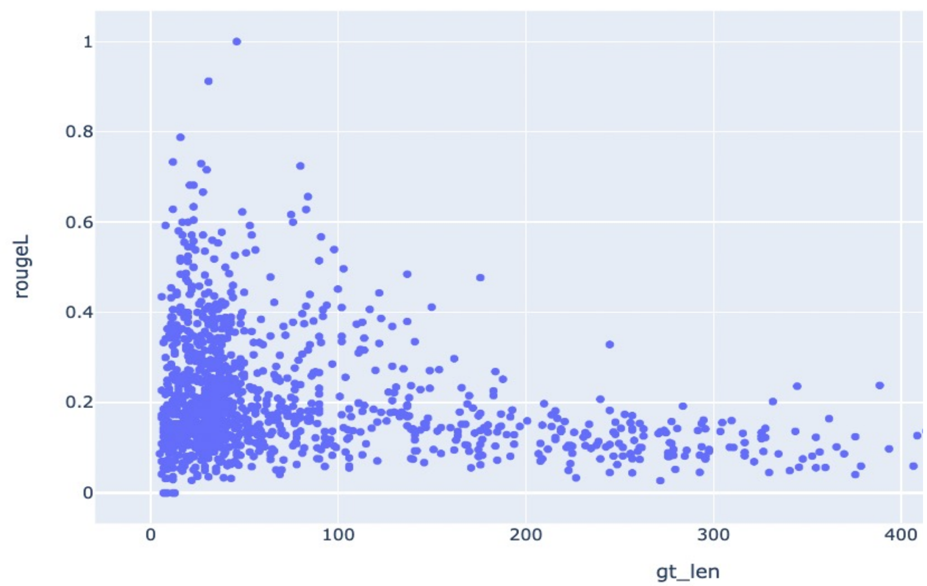
Task 2 (Explanation Generation)

The length of ground truth explanation vs RougeL score for predicted explanations.

There are about 200 samples where the len(top_k) (evidence_text) < len(explanation_text)

These explanations are unusually long either very close to the original evidence_text or more than 200 words.

Ignoring these improves the Rouge scores by ~2%



Task 2 (Explanation Generation)

- Fine tuned model (PEGASUS 256/64)
- Tried 3 decoding strategies
- The context is set to the evidence_text
- p=0.95 (Only nucleus sampling). Results shown here.
- Top_k=10 and p=0.95
- Top_k=50 and p=0.95
- Rouge scores dropped about 1-4%

Model	R1	R2	RL
Fine tuned model	0.3084	0.1194	0.2319
Fine tuned model with nucleus sampling	0.2683	0.1005	0.2023

Task 2 (Explanation Generation)

- Is this a good explanation generation model ?
- The Rouge scores were not informative so we did manual evaluation on ~n=29 instances

Manual Evaluation Setup

Best explanation Rating

1: Irrelevant

2: Incorrect

3: Partially incorrect partially correct

4: Partially correct rest irrelevant

5: Correct

Best explanation Type

All: All predictions are same

1: top-p is correct

2: top-p, top_k=10

3: Default decoding strategy using beam search

Manual Evaluation Examples

Correct	**Rating 5**
GT	An experimental Roche drug designed to treat the “negative symptoms” of schizophrenia failed to meet its main goal in two late-stage studies, dealing a blow to the Swiss drugmaker’s research hopes in the risky area of brain science.
PEGASUS top-p	Swiss drugmaker Roche said on Thursday that a new drug to treat negative symptoms of schizophrenia failed to meet its main goal in a late-stage study.

Irrelevant	**Rating 1**
GT	What's undetermined: As of this writing, Five Guys has not explicitly corroborated the police version of events.
PEGASUS top-p	Police officers in Daphne, Alabama, claimed they were denied service at a Five Guys restaurant on the evening of July 7.

Manual Evaluation Examples

Partially Correct/Incorrect	Rating 3
GT	On any given day in the United States, 18 percent of men and 11 percent of women drink more alcohol than federal guidelines recommend, according to a study that also found that 8 percent of men and 3 percent of women are full-fledged “heavy drinkers.”
PEGASUS top-p	More than half of adults in the United States drink more than recommended levels of alcohol, according to a new study.

Manual Evaluation Results

All: All predictions are similar quality

1: top-p is correct

2: top-p, top_k=10

3: Default decoding strategy using beam search

Annotator 1

	Best explanation Rating	
	count	mean
Best explanation		
1	4	4.250000
2	5	3.800000
3	1	5.000000
All	19	3.210526

Annotator 2

	Best explanation Rating	
	count	mean
Best explanation		
1	6	5.000000
2	15	4.466667
3	4	4.000000
All	3	1.666667

Conclusion & Demo

Things that we were able to finish:

- Trained claim classification model and explanation generation model on entire dataset
- Claim classification: Used top_k vs <claim, evidence sentence> pairs. Achieved an F1-score of 80% on a 2-class classifier.
- Fine tuned t5-small and PEAGUSUS model. Used PEAGUSUS model with and without decoding.
- Manual evaluation on generated explanations which seems to indicate there is subjectiveness in explanations which Rouge score don't quite capture.

Things that we weren't able to look into deeper:

- Classification: We didn't try further on 4 class classification to improve accuracy.
- Classification: We didn't fine tune for longer time on claim sentence text pair.
- Classification: Looking at failure cases for unrelated text
- Generation: Use the input text length as 512 instead of 256. Train on bigger hardware.

Application Demo

- Fact checking model requires a claim and source (evidence text), so the model accepts those as inputs.
- Built basic HTML/CSS/Javascript + Flask backend
- If source is not provided, it pulls the first Google result
- Packaged the app and model into a Docker image
- Model outputs the label (True/False), a confidence score, and an explanation. Graphic is generated with Plotly.