

# Tasks

Hee Hwang and Sudarshan Raghavan

April 26, 2021

## 1 Datasets

ACL : Citation Intent Classification  
Hyper: HyperPartisan News Detection  
RCT : Randomized Controlled Trials

### 1.1 Size of Datasets

Task	train	dev	test	Classes
ACL	1688	114	139	6
Hyper	516	64	65	2
RCT-sample	500	30212	30135	5
RCT-20k	180040	30212	30135	5

### 1.2 Size of Augmented Data

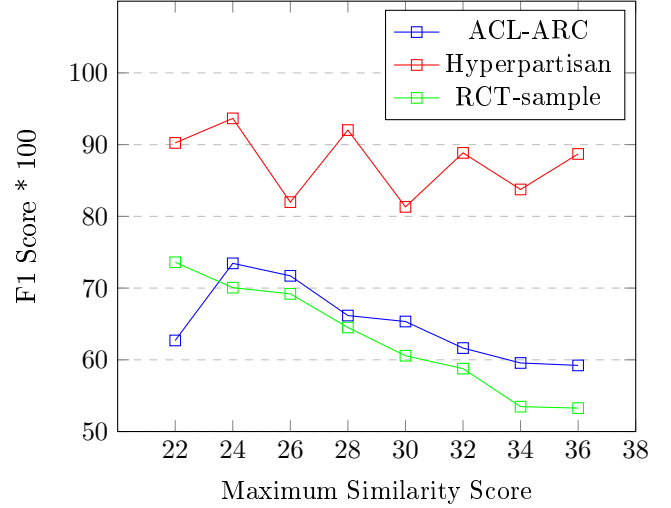
Max Score	ACL	Hyper	RCT
Baseline	1688 (100%)	516 (100%)	500(100%)
24	1746 (103%)	551 (106%)	686(137%)
26	1815 (107%)	567 (109%)	831(166%)
28	1981 (117%)	606 (117%)	1065(213%)
30	2253 (133%)	656 (127%)	1484(296%)
32	2842 (168%)	742 (143%)	2105(421%)
34	3848 (227%)	911 (176%)	2952(590%)
36	5819 (344%)	1127 (218%)	4196(839%)

## 2 Augmentation

### 2.1 Augmentation by Maximum Similarity Score

Max Score	ACL	Hyper	RCT
Baseline	62.70	90.24	73.60
24	<b>73.45</b>	<b>93.66</b>	70.06
26	71.71	81.98	69.19
28	66.17	92.03	64.50
30	65.34	81.32	60.57
32	61.64	88.85	58.77
34	59.56	83.75	53.47
36	59.22	88.70	53.27

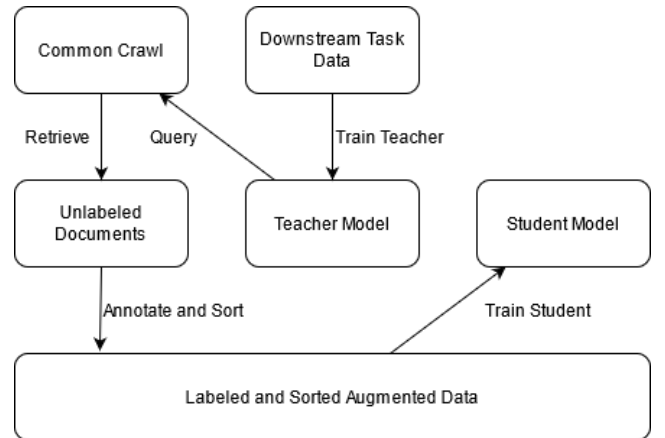
Augmentation by Similarity Score



## 3 Baseline models:

- An off-the-shelf RoBERTa model that has been finetuned to perform classification for each of the downstream tasks

## 4 Augmentation Model



## 5 Algorithm

1. Extract failed test examples from the baseline model
2. Retrieve passages/sentences from Common Crawl
3. Apply augmentation strategy (i)-(iii)
4. Augment all the labelled CC data to the training data
5. Retrain RoBERTa on the augmented training set

## 6 Augmentation Strategies

- Strategy (i)  
Use baseline model (Teacher) to perform unsupervised labelling on retrieved CC data
- Strategy (ii)  
Using a task specific binary classifier, filter out retrieved CC data that is "out-domain"  
Use baseline model (Teacher) to perform unsupervised labelling on the filtered "in-domain" CC data
- Strategy (iii)  
Using a task specific binary classifier, filter out retrieved CC data that is "out-domain"  
Use ground truth labels of failed test examples and assign labels to the filtered "in-domain" CC data