# When does Multitask Learning work in NLP? An Exploration on Cyberhate in Wikipedia

## Goal

- MTL improves generalization by using domain information in related tasks.

- MTL has shown to be very effective at amortizing data acquisition costs because most problems involve solving related subtasks.

- But, **when** to expect MTL gains in NLP largely remains an open question.

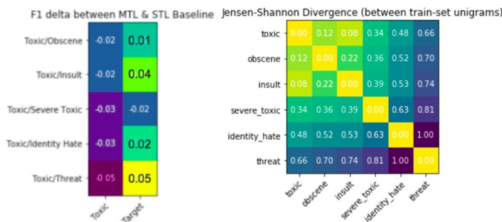- We study this question on an important problem: cyberhate.

## Data & Featurization
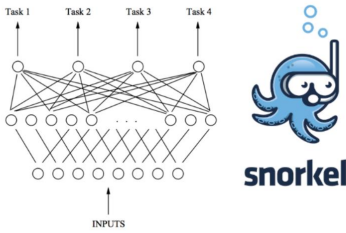


Toxic

Identity Hate
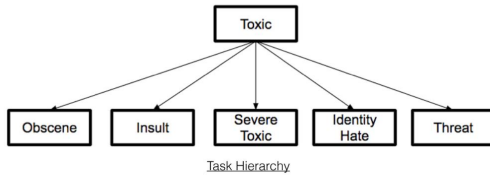


Task Hierarchy

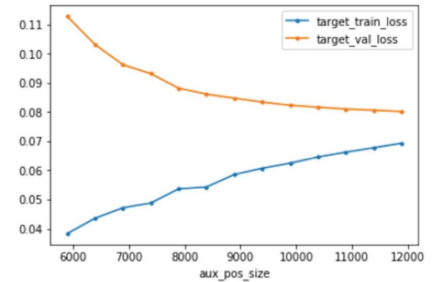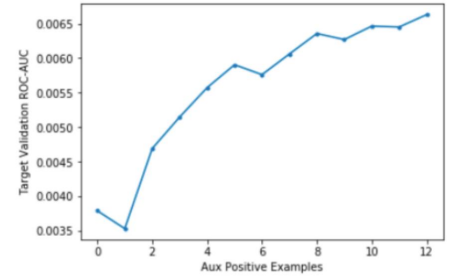| Task | Positive Samples |
|---|---|
| Toxic | 15,294 |
| Obscene | 8,449 |
| Insult | 7,877 |
| Severe Toxic | 1,595 |
| Identity Hate | 1,405 |
| Threat | 478 |
| Total | 153,165 |

- A real-world setting: **class imbalance, noisy labels and class hierarchy**.

- Text cleaned and featured with standard techniques: stop word removal, stemming, tf-idf.

## Model



INPUTS

snorkel

## Results



F1 delta between MTL & STL Baseline

Jensen-Shannon Divergence (between train-set unigrams)

| Task | Reweighting | Oversampling |
|---|---|---|
| Obscene | 0.03 | 0.01 |
| Insult | -0.17 | -0.15 |
| Severe Toxic | 0.01 | -0.04 |
| Identity Hate | 0.02 | -0.01 |
| Threat | 0.08 | 0.01 |

Abraham Starosta
starosta@stanford.edu





## Conclusions

- The more samples for the auxiliary task, the higher the gains from MTL on the main task.

- Class reweighting or positive oversampling don't seem to help MTL with the problem of class imbalance.

- We observe MTL serving as an effective regularizer.

- MTL might help with cyberhate detection.

- Using JSD as a measure of task relatedness in a setting where task dataset sizes have high variance might be deceiving.