# Predicting Humor in Headlines

Christian Kavouras, James Blackburn, Joseph Nollette, Joshua Valdez

# Task Description

- Implement a regression based model to predict humor in a dataset of headlines
- Dataset consists of headlines taken from the Humicroedit dataset
- https://competitions.codalab.org/competitions/20970#learn_the_details-overview

# Task Description

- Single word edits to headlines for comedic effect
- Labeled scores on a 0-3 scale of “funniness”
- Averaged across 5 human scores

# Approach

- We use a single regression pre-trained BERT model
- BERT then fine-tuned on our headline data
- Evaluated as a single-regression task with RMSE loss
- Note: Due to access to external GPUs, we are able to fine-tune BERT itself

# Approach

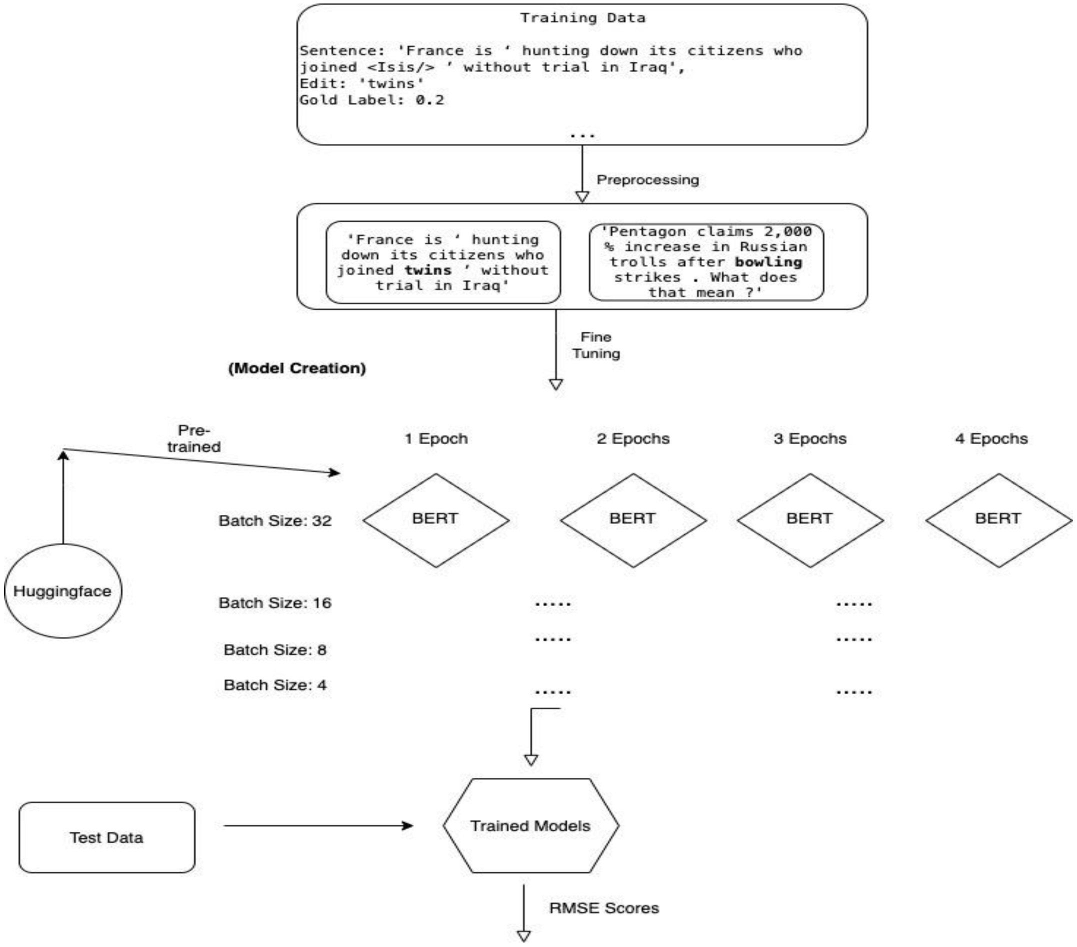
## *Features*

- Inputs to the model are raw sentence embeddings
- Tokenized using vanilla BertTokenizer

## *Hyperparameters*

- Tuned models across varying epochs:
  - 1, 2, 3, 4
- Tuned with various batch-sizes:
  - 32, 16, 8, 4
- Adam Optimizer

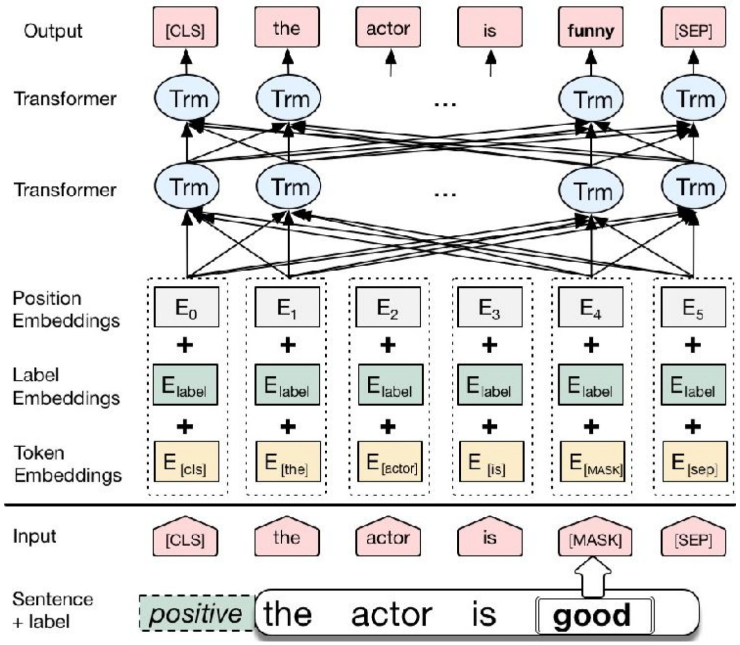
# System Design



# System Design

(Schematic of BERT architecture:)

In our case, labels are decimal values in [0, 3]



# Results

results for subtask1:

baseline: 0.57471

BERT single regressor:

task A RMSE results:

| epochs | batch=32 | batch=16 | batch=8 | batch=4 |
|---|---|---|---|---|
| 1epochs | 0.54199 | 0.54312 | 0.53396 | 0.53438 |
| 2epochs | 0.53324 | 0.54381 | 0.5383 | 0.54047 |
| 3epochs | 0.54728 | 0.54737 | 0.56014 | 0.55393 |
| 4epochs | 0.55964 | 0.55747 | 0.56742 | 0.56569 |

# Discussion

- For our primary task, our BERT model performs impressively well
  - (Butts up against the baseline)

- Can expand from a single regression
- Can further fine tune a downstream, feature based model on our input

# System Improvement

- Explored possibilities for improvement of BERT
  - Conditional Random Field?
  - Feed-Forward Neural Network?
- Used output embeddings from BERT as input to FFNN
- Adjusted five hyperparameters along three values each to measure improvements

# System Improvement

Hyperparameters

| | | | |
|---|---|---|---|
| Epochs | 2 | 4 | 8 |
| Batch Size | 4 | 8 | 16 |
| Learning Rate | 2E-6 | 2E-5 | 2E-4 |
| Dropout Rate | 0.2 | 0.3 | 0.4 |
| Hidden Layers | 1 | 2 | 3 |

# System Improvement

Best 3 results

| RMSE | Epochs | Batch Size | Learning Rt | Dropout Rt | Hid. Layers |
|---|---|---|---|---|---|
| 0.56156 | 8 | 4 | 2E-4 | 0.4 | 2 |
| 0.5634 | 8 | 4 | 2E-4 | 0.4 | 1 |
| 0.56541 | 4 | 8 | 2E-4 | 0.2 | 2 |

# Adaptation Task

Can this system predict the humor content of Reddit posts?

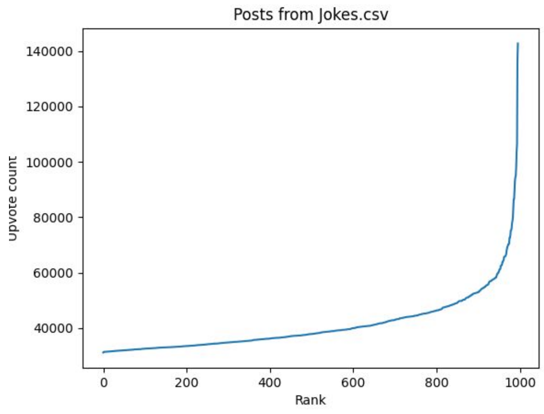
| Non-Humorous | Humorous |
|---|---|
| r/news | r/Jokes |
| r/worldnews | r/dadjokes |
| r/askscience | r/TheOnion |
| r/movies | r/oneliners |
| r/politics | r/fifthworldproblems |
| r/wallstreetbets | r/wheredidthesodago |

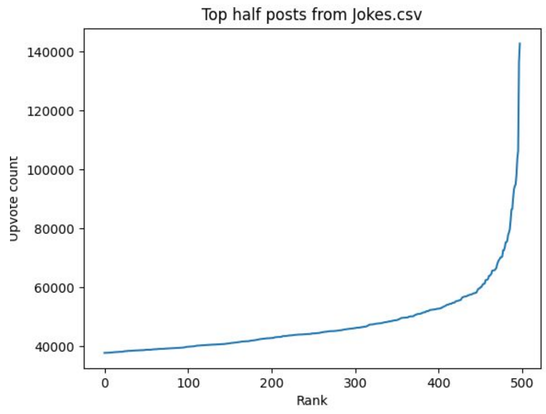
# Adaptation Task

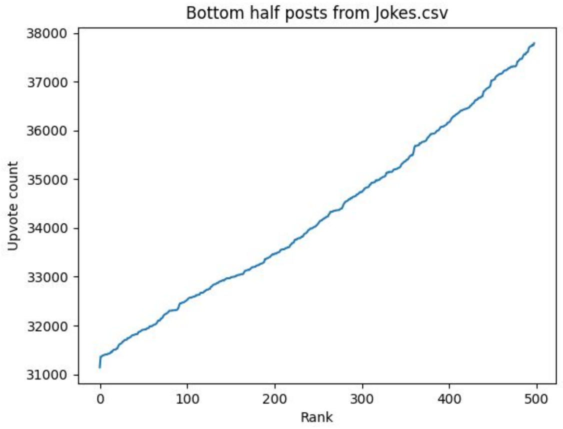
- Obtain top 1,000 posts of all time from each subreddit
- Normalize upvote scores to the range 0.0 to 3.0
- Calculate Spearman correlation of predicted humor and upvote score for each subreddit

- Prediction: stronger correlation in humorous subreddits than in non-humorous ones

# Upvote Distribution

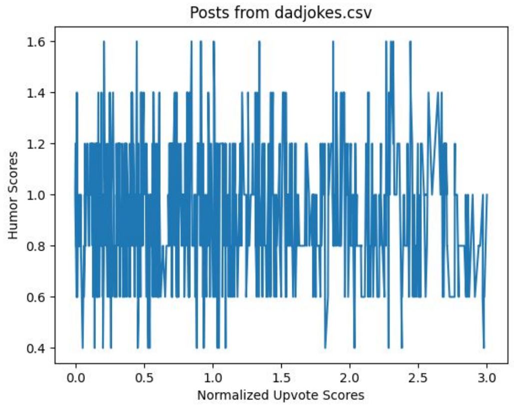


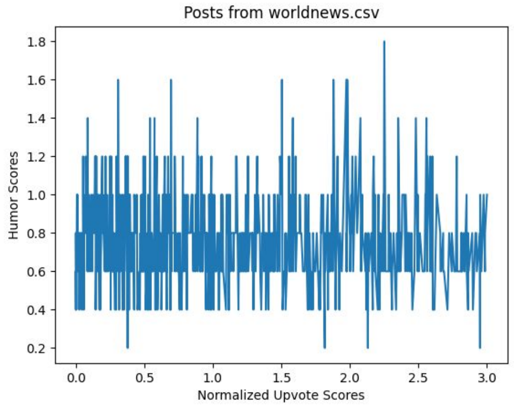



# Adaptation Results

Spearman correlation: 0.001897




Spearman correlation: -0.034297



# Future Improvements

Gold Annotations:

- Gather human-valued annotations for reddit data

Pseudo-Labeling:

- Predict humor scores for additional reddit data (noisy labels)
- Feed noisy-labeled reddit data back into system for supplemental training

# Related Reading

BERT paper: (Devlin et al., 2018) https://arxiv.org/pdf/1810.04805.pdf

Aspect-based sentiment analysis - ideas from a similar task domain (Xu, et al., 2019) https://arxiv.org/pdf/1904.02232.pdf

Winning Model for same task (Spanish Dataset) http://ceur-ws.org/Vol-2421/HAHA_paper_3.pdf