# Working Title

James Blackburn (jimbb82@uw.edu)
Joshua Valdez (jdv2@uw.edu)
Joseph Nollette (nollejos@uw.edu)
Christian Kavouras (cdkavour@uw.edu)

## Abstract

We present a system for predicting the humor content of English-language news article headlines; specifically, the boost in perceived humor resulting from the replacement of one word or phrase in the original headline with a new word, as rated by human judges.

## 1 Introduction

SemEval-2020 Task 7 seeks to expand upon the task of predicting the humor content of chunks of text, and attempts to measure the change in human-judged humor values when individual words are replaced. We will build a system that – given original chunks and edited chunks – will predict the mean humor score of an edited headline.

## 2 Task Description

The primary task of this project is to implement a regression model to predict the degree of humor of brief news headlines, using data from the Humicroedit data set. This data set contains the text of news headlines in which one word has been edited to change a serious headline into a humorous one. All training instances contain the full text of the headline, the word that was replaced, the new word that was put in its place, and a decimal score between 0 (not funny) and 3 (very funny), obtained by taking the average score given by five human judges. The description of this task and its dataset can be found here.[**?**]

The adaptation task we plan to complete is to apply a similar model, trained on both the original and edited headlines referenced above, to tweets. The goal is to predict the humor content of tweets according to the same scale and, by employing a time-series clustering algorithm, explore the correlation of the frequency of humorous posts on social media to certain periods of time.

## 3 Approach

This project uses a single-regression, pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. The model is fine-tuned on our headline data and evaluated as a single-regression task with RMSE loss. The availability of external GPUs allows us to fine-tune BERT itself. The model uses raw-sentence embeddings as input which are then tokenized using BertTokenzer.

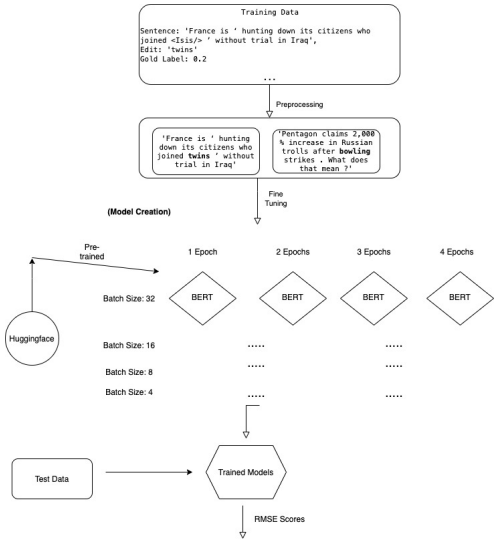


Figure 1: System Architecture for our primary task prediction

## 4 System Overview

We have chosen to interpret the primary task as a regression problem, using a bidirectional encoder representation from transformers (BERT) model to predict the humor content of edited headlines.

This model [**?**] consists essentially of a series of transformer encoders which takes embeddings of labeled headlines as its input and generates a predictied humor score for each instance. The task will be completed by generating pretrained models and then proceeding to a fine-tuning phase to improve performance.

To complete the pretraining phase, sixteen distinct models were generated from embeddings of the labeled training instances, by varying the number of training epochs from one to four, and employing batch sizes of 4, 8, 16, and 32 instances. A root mean square error is employed as the loss function.

# 5 Results

Below is a table of the RMSE values for the sixteen combinations of epochs and batch size. Our Baseline value was formulated using the RMSE across the label values of our training and test data.

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 |

As shown, the best result we have obtained from this initial run is from 2 epochs and a batch size of 32.

# 6 Discussion

From this initial phase, we aim to explore methods for fine-tuning BERT, as well as the potential of expanding our approach from single regression to multiple regression. In a similar task using a multi-lingual data set, the highest performing model makes use of a Gradient Boosted Layer on top of BERT [**?**]. This could be a place of further exploration for improving performance.

For our primary task, our BERT model performs impressively well, only butting up against the baseline.

# 7 System Improvement

We have explored possibilities for the improvement of BERT, such as the use of a Conditional Random Field and/or a Feed-Forward Neural Network, used output from the BERT model as input for the FFNN, and adjusted five parameters along three values each to measure the improvements.

| 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 |

# 8 Adaptation Task

In the making of this program, we obtained the top 1,000 posts from twelve subreddits, six humorous and the other six non-humorous, normalized the upvote scores to the range from 0.0 to 3.0, and calculated the Spearman correlation of predicted humor and upvote score for each subreddit. The prediction we made was that there would be a stronger correlation in humorous subredits than in non-humorous ones - case in point, the Spearman correlation for r/dadjokes was 0.001897, while that of r/worldnews was -0.034297.

# 9 Future Improvements

While our system is potent by itself, it is not perfect. The system could potentially benefit from human-valued annotations for reddit data, predictions for humor scores for additional reddit data (noisy labels), and a subsystem for feeding noisy-labeled reddit data back into the system for supplemental training.

# 10 Conclusion

With all our RMSE values being lower than the baseline value, our implementation model should be accurate enough as it is.

# 11 Works Cited


- 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