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 attempt 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.(Hossain et al., 2019)

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 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 (Devlin et al., 2018) 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.
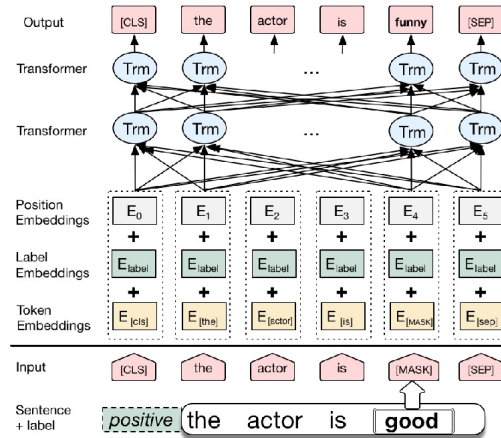
To complete the pretraining phase, sixteen distinct models were generated from embeddings of the labeled training instances, by varying the number of training epochs from one to four, and employing batch sizes of 4, 8, 16, and 32 instances. A root mean square error is employed as the loss function.

4 Results

Below is a table of the RMSE values for the sixteen combinations of epochs and batch size.

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.

5 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.

6 Conclusion

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.