EndTimes at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with BERT and Ensembles

Chandan Kumar Pandey and **Chirag Singh** and **Karan Mangla**
Samsung R&D Bangalore
{chandan.p, c.singh, karan.mangla}@samsung.com

Abstract

This paper describes Humor-BERT, a set of BERT (Devlin et al., 2019) Large based models that we used to solve the SemEval-2021 Task 7: Detecting and Rating Humor and Offense (Meaney et al., 2021). It presents pre and post processing techniques, variable threshold learning, meta learning and Ensemble approach to solve various sub-tasks that were part of the challenge. We also present a comparative analysis of various models we tried. Our method was ranked 4^{th} in Humor Controversy Detection, 8^{th} in Humor Detection, 19^{th} in Average Offense Score prediction and 40^{th} in Average Humor Score prediction globally. F1 score obtained for Humor classification was 0.9655 and for Controversy detection it was 0.6261. Our user name on the leader board is ThisIstheEnd and team name is EndTimes.

1 Introduction

The purpose of this paper is to present different approaches that we tried towards various sub-tasks in SemEval-2021 Task 7 Detecting and Rating Humor and Offense task (Meaney et al., 2021). It consists of following sub-tasks:

- Task 1a: Classifying text based on whether they are humorous or not.
- Task 1b: Predicting Humor rating score
- Task 1c: If the text is classed as humorous, predict if the humor rating would be considered controversial, i.e. the variance of the rating between annotators is higher than the median
- Task 2a: predict how generally offensive a text is for users. This score was calculated regardless of whether the text is classed as humorous or offensive overall.

We participated in all of the above sub-tasks. Task 1a and Task 1c are classification problems with F1 score as metric whereas, Task 1b and Task 2a are regression problems with RMS error values as scoring criteria. All the above tasks shared the same training, development and test set (Meaney et al., 2021). Training set consisted of 8000 text sentences, containing labels for all the subtasks. Size of development and test set was 1000 text sentences each.

Humor, is an interesting linguistic challenge. It's an abstract concept and depends a lot on audience and their interpretation of the jargon used to generate humor. Notion of humor changes from culture to culture, age group, gender and social status of target audience. It is a subjective and personal phenomenon. What's humorous to one person could be non humorous or even offensive to the other. In this task, labels and ratings were collected from a balanced set of age groups from 18-70 and variety of genders, political stances and social status (Meaney et al., 2021). Humor detection and rating tasks have been the defined before but the task of detecting Controversy (Meaney et al., 2021) is new and interesting. It captures the variance in humor rating due to variance in age, gender and other demographic features of the user.

This task is also unique in the way that it combines humor and offense rating tasks for the same dataset. What is humorous to one user could be offensive to other depending on their demographic characteristics.

2 Related Works

Humor and offense related tasks have been pursued before as well. Various NLP techniques from classical N-gram techniques (Taylor and Mazlack, 2004) to transfer learning over pre-trained language models like BERT (Devlin et al., 2019) have been tried

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1215–1220
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

for humor related tasks. (Yang et al., 2015) extract humor anchors and use a k-NN based classifier to detect humor. (Chen and Soo, 2018) used CNN and highway network whereas, (Weller and Seppi, 2019) used BERT fine tuning to classify humor in text. Apart from text, multimodal data for humor was curated by (Hasan et al., 2019). In other works on multimodal data (Yang et al., 2019) tag video comments data automatically and use audio data alone to predict humor. They used Random Forest and CNN over MFCC features. (Chen and Lee, 2017) take transcripts from TED talk and detect audience laughter using CNN.

3 System overview

In development phase, we split the data as training:validation :: 7200:800, but for evaluation phase we take entire 8000 texts for training and 1000 size development set. Later on, we also tried bagging, which will be described later.

3.1 Models

Humor is an abstract linguistic concept, hence a model which can understand the nuances of language should be great for detecting and rating humor, offense and controversy in text. Naturally we tried one of the best language model, BERT and modified it according to the task at hand. We describe below models for each sub tasks separately.

3.1.1 Task 1a: Humor Detection

Various models that we tried for this sub-task are following:

1. BERT-Large fine tune [**BERT-L**]
2. Fully connected layer over BERT-Large [**BERT-FFN**]
3. BERT-Large textual entailment [**BERT-ENT**]
4. CNN over BERT embedding [**BERT-CNN**]
5. BERT-Large based Ensemble [**BERT-ENS**]

For BERT-Large fine tune model(**BERT-L**), we used uncased BERT Large model and fine tuned over train set. We selected the best performing model over development data. For this we selected COLA as task type as it is GLUE task which models standard binary classification. It simply uses softmax classifier over CLS token from last layer of BERT Large model.
In the second approach (i.e. for **BERT-FFN**), we tried a 128 size fully connected layer over CLS token from last layer and then two class softmax classifier to get humor probability scores. We also used dropout after CLS and fully connected layer. We also tried textual entailment kind of classification model over BERT Large (**BERT-ENT**), for this task. Each of the text data is modified to indicate whether humor is implied from any given text. For example training example:

I'm the Michael Jordan of lazy sports analogies.# 1
becomes
I'm the Michael Jordan of lazy sports analogies. ### It is humorous. # 1

Where label 1 represents the presence of humor in original sentence whereas, in case of textual entailment instance it represents whether *"It is humorous"* is implied by the text sentence. # is just a separator for illustration here. We also used entailment model to augment the dataset by adding various paraphrases of the sentence *"It is humorous"*. For example, the original text above can be augmented in the following way:

I'm the Michael Jordan of lazy sports analogies. ### It is humorous. # 1
I'm the Michael Jordan of lazy sports analogies. ### It is funny. # 1
I'm the Michael Jordan of lazy sports analogies. ### It is not humorous. # 0

In another approach (i.e. **BERT-CNN**), we tried freezing the BERT parameters and used it only to get the sentence embedding from the last layer. We took word embeddings for each word in the sentence from the last layer of BERT and then applied a one layer 1-D CNN over those embeddings, then a fully connected layer and softmax layer to detect humor.

BERT-Large based Ensemble model (i.e. **BERT-ENS**), is our best model which gave F-1 score of 0.9655. First thing we tried is Bagging of BERT-large models. The base model used in bagging was the model from approach 2. We split the train data in 10 random datasets (i.e. bags) of size 7200(train) and 800(validation). Train 10 base models over each of those datasets and select the best performing model on validation set of size 800. Now combine these 10 models using soft/hard voting ensembling approach. We predict the labels using best models from each of the bag and then take a majority vote among the 10 selected models. We tried variants of this approach where each of the 10 models had different hyper-parameter values, for example different values of dropout and maximum sentence length varies between 64 to 128. In another approach, we sum the softmax score of each epoch of all the 10 bags and then use argmax to predict the labels for each bag. After that, we take a majority vote among 10 such models created (This was our best performing model). The development data of size 1000 was used for hyper-parameter tuning and selection of the final ensemble model. We used **Binary cross entropy** as loss function.

3.1.2 Task 1b: Humor Rating

Taking hint from Task 1a, we tried fewer models that we felt would perform better for this sub-task. Following models were tried

1. Fully connected layer over BERT-Large [**BERT-FFN**]
2. BERT-Large based Ensemble [**BERT-ENS**]

The architecture of models 1 is same as the corresponding models in Task 1a (3.1.1), except for the last layer and loss function. Here instead of two class softmax layer we used a linear regression layer to predict the humor rating and used **Mean squared Error** and **Root Mean squared Error** as metric to be minimized. For model in approach 3, we use the same setting as described in Task 1a (3.1.1), except that instead of voting method we used averaging of humor rating predicted by all the base 10 models. As in Task 1a, models from each bag are selected based on which model has least RMS error i.e. the best model among all epochs. In other approach, We took the average of rating predictions from all the epochs that were trained for 10 bags (This was our best model). We also tried bagging models with different hyper-parameter values as in Task 1a. The best hyper-parameter values are described in 4.3.

3.1.3 Task 1c: Humor Controversy

We tried models similar to the Task 1a (3.1.1) as described below.

1. BERT-Large fine tune [**BERT-L**]
2. Fully connected layer over BERT-Large [**BERT-FFN**]
3. BERT-Large based Ensemble [**BERT-ENS**]

The architecture of models 1 and 2 are same as the corresponding models in Task 1a (3.1.1), except for additional meta learning of softmax thresholds to predict whether the humor is controversial or not. There is class imbalance here and so models were less confident in detecting controversy in humor. Also, it's a challenging task, in the sense that detecting controversy is not so obvious. In order to balance the odds we tried various softmax score threshold values to predict controversy. Instead of taking argmax we tried different values of softmax probability score for Controversy class. For, standard binary classification a threshold of 0.5 is used to predict a particular class. We tried different lower values of threshold in favour of Controversy class, since model were not very confident in detecting it. Threshold value 0.1 worked best for approach 1 and 2.
In 3^{rd} approach we use bagging ensemble setting similar to Task 1a (3.1.1) with the difference that we used softmax score averaging of models rather than a voting, and thereafter we used threshold tuning. Threshold value 0.15 worked best for our method. All the threshold values were obtained by fine tuning over development dataset of size 1000.

3.1.4 Task 2a: Average Offensiveness Score

This is a regression task, hence we tried models similar to that of Task 1b (3.1.2).

1. Fully connected layer over BERT-Large [**BERT-FFN**]
2. BERT-Large based Ensemble [**BERT-ENS**]

The architecture, loss function and all other settings are same as that of corresponding models in Task 1b (3.1.2).

4 Experiments

In this section, we describe the dataset and various experiments performed. For classification tasks we used **Binary Cross-Entropy** loss, whereas for regression tasks we used **Mean Squared Error** as metric to be minimized.

4.1 Dataset

We used the dataset provided by the SemEval-2021 Task 7 organizers (Meaney et al., 2021). Dataset consisted of 8000 English text sentences for training and 1000 text sentences for development and evaluation each. We trained our model on the training dataset only, no external dataset other than that was used. Dataset split for each model has already been described in section 3.

4.2 Pre-processing

We removed various special characters which do not contain any useful information. We also expanded emoji symbols to their meaningful definition using existing preprocessing package spaCy since they are really important for tasks such as humor detection and humor rating prediction. Afterwards, texts were lower cased and tokenized using Sentencepiece (Kudo and Richardson, 2018) tokenizer.

4.3 Hyperparameters

We tried various values for the hyper-parameters. The one that worked best for us is described in table 1.

Hyper-parameter	Value
Batch Size	32
Learning rate	2e-5
Maximum Sentence length	64, 128
CLS layer Dropout	0.3
FFN layer Dropout	0.5
Number of Bags	10
Number of Epochs	25
FFN Activation	*tanh*
FFN hidden size	128

Table 1: Hyper-parameter values

Changing Maximum sentence length didn't make much difference in results. Activation function $tanh$ worked better than $relu$ and $gelu$. CLS layer dropout is dropout after last layer of BERT. FFN layer dropout is applied after Feed forward network layer of hidden size 128. Dropout was required because the model was too complex for 8000 size dataset. No of Bags is the no of bags/no of base models we trained during Bagging ensemble approach. All the models were developed using Tensorflow (Abadi et al., 2015) library. Training was done on NVIDIA Tesla P-40 GPUs.

5 Results and Analysis

We describe the results for each of the sub-tasks and analyse the results. Baseline for classification task was Naive Bayes model with bag of words features, and for the regression task, it was Support Vector Regression.

Model	F1 Score
Baseline	0.884
BERT-L	0.926
BERT-FFN	0.935
BERT-ENT	0.912
BERT-CNN	0.937
BERT-ENS	**0.966**

Table 2: Humor Detection Results

Model	RMS Error
Baseline	0.861
BERT-FFN	0.667
BERT-ENS	**0.654**

Table 3: Average Humor Score Results

Model	F1 Score
Baseline	0.462
BERT-L	0.567
BERT-FFN	0.586
BERT-ENS	**0.626**

Table 4: Humor Controversy Results

Model	RMS Error
Baseline	0.642
BERT-FFN	0.522
BERT-ENS	**0.469**

Table 5: Average Offensiveness Score Results

We see in table 2 that applying a Feed Forward network with dropout layer improves the result. This is expected since a FFN layer after BERT results in non-linear combination of BERT features. Dropout is required because model is more complex now. BERT-CNN also performs reasonably well because of efficacy of CNN in learning and combining $N - Gram$ features. Not to mention, here we have frozen the BERT layer so no of parameters is less. Relatively poor performance of BERT-ENT can be attributed to the fact that this task is not suitable for Textual Entailment classification. **BERT-ENS** outperforms other models by huge margin. Here we see the effect of bagging as a way to reduce error due to variance. Different base models learn different features and peculiarities. BERT-Large is a complex and strong classifier, so understandably, it has high variance. Combining multiple BERT-Large models using voting mechanism provided quite a significant jump in F1 score.

For Average Humor score results in table 3 we again see that a single BERT model with FFN does perform well but bagging reduces the error caused by high variance. We see same pattern for Average Offensiveness score. In fact the effect of bagging is even more prominent there.
Meta-learning in form of variable softmax threshold worked really great for Humor Controversy detection. Humor Controversy is even more abstract than Humor and hence difficult to detect. There is class imbalance as well, since, very few examples are actually Controversial. So, we had to lower the softmax probability threshold values for Humor Controversy class. Individual BERT model is a weak learner in this case that's why combining them results in a huge jump in F1 score. On comparing our results with the baseline models we see that even a single BERT-Large model outperforms the baseline by large margin. This solidifies our notion that a great language model like BERT will always be a top performer in NLP tasks. And, combining the BERT with ensemble methods has potential to outperform other competing models.

6 Conclusion and Future Work

Humor is a abstract linguistic construct. That is why a strong language model like BERT-Large performs really well on these tasks. Our models were ranked under 10 in 2 tasks and under 20 in 1 task, this shows that BERT based models outperform other models. Especially, if many BERTs are combined using a good ensemble technique. For future work ensemble methods like stacking and blending could be tried. Also, different models with different hyper parameter values could be combined in a more effective way to get better results.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Lei Chen and Chong Min Lee. 2017. Predicting audience's laughter during presentations using convolutional neural network. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 86–90, Copenhagen, Denmark. Association for Computational Linguistics.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and* *the 11th International Joint Conference on Natural Language Processing*.

Julia M Taylor and Lawrence J Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Zixiaofan Yang, Bingyan Hu, and Julia Hirschberg. 2019. Predicting humor by learning from time-aligned comments. In *INTERSPEECH*, pages 496–500.