LRG at SemEval-2020 Task 7: Assessing the Ability of BERT and Derivative Models to Perform Short-Edits based Humor Grading

Siddhant Mahurkar
Department of Computer Science & Engineering
Vellore Institute of Technology, Vellore, India
`siddhant.pravin2017@vitstudent.ac.in`

Rajaswa Patil
Department of Electrical & Electronics Engineering
BITS Pilani K. K. Birla Goa Campus, India
`f20170334@goa.bits-pilani.ac.in`

Abstract

In this paper, we assess the ability of BERT and its derivative models (RoBERTa, DistilBERT, and ALBERT) for short-edits based humor grading. We test these models for humor grading and classification tasks on the Humicroedit and the FunLines dataset. We perform extensive experiments with these models to test their language modeling and generalization abilities via zero-shot inference and cross-dataset inference based approaches. Further, we also inspect the role of self-attention layers in humor-grading by performing a qualitative analysis over the self-attention weights from the final layer of the trained BERT model. Our experiments show that all the pre-trained BERT derivative models show significant generalization capabilities for humor-grading related tasks.

1 Introduction

Humor is a communicative ability that produces a sense of laughter by imparting an amusing tone in conversations. If we can teach our models how to generate, detect, and grade humor in language, it would be a big step ahead in Natural Language Processing. Nearly all the existing datasets for humor-related tasks such as "Pun of the Day" (Yang et al., 2015) and "16000 One Liners" (Mihalcea and Strapparava, 2005) focus on humor-detection in a single piece of text as a binary classification task. Previous tasks on humor-grading consist of datasets comprising of originally humorous texts (Potash et al., 2017; Chiruzzo et al., 2019). SemEval-2020 Task-7 (Hossain et al., 2020a) focuses on short-edits based humor-grading. The dataset used for this shared-task was introduced by Hossain et al. (2019). It consists of short-edits applied to a piece of text (news-headlines), which changes it from non-funny to potentially funny. This shared-task consists of two independent sub-tasks based on this dataset. The first sub-task comprises of a regression task, where given the original and the edited news-headline, one must predict the funniness of the edited headline (humor-grading) on a scale of [0,3], where '0' conveys *least funny* and '3' conveys *most funny*. The second sub-task comprises of a classification task, where given two edited versions of the same news-headline, the model must predict the funnier one.

In this work, we propose using BERT (Devlin et al., 2019) and its derivative models (collectively referred to as "BERT models" further throughout the paper) for short-edits based humor-grading. We mainly follow two approaches: In the first one we use pre-trained BERT models with a final regression layer to get a humor-grade for each edited news-headline, and for the second one we use masked-word language-modeling to fine-tune the pre-trained BERT models on the task dataset, which are further trained for the humor-grade regression task. We compare the results of these two approaches for each sub-task. We also test the generalization capabilities of the BERT models for this task by using the FunLines dataset (Hossain et al., 2020b), which is provided as an additional dataset for this shared-task. The code for this work is made publicly available as a GitHub repository.[1]

[1] `https://github.com/rajaswa/bert-humor-semeval-2020`

Original Headline	Edited Headline	Humor Grade
US Navy ship fired **warning** shots at an Iranian boat in the Persian Gulf	US Navy ship fired **tequila** shots at an Iranian boat in the Persian Gulf	2.6
US Navy ship fired **warning** shots at an Iranian boat in the Persian Gulf	US Navy ship fired **texts** shots at an Iranian boat in the Persian Gulf	1.8

Table 1: Example edited news-headlines from Humicroedit dataset

2 Background

Early work in the field of humor detection was based on extracting various corpus-based statistical and linguistic features. Kiddon and Brun (2011) used these features with Support Vector Machine models for humor detection as a binary classification task. Zhang and Liu (2014) worked on humor detection in tweets by extracting nearly fifty humor-related features. These features were used with Gradient Boosting Regression Tree based models. Beukel and Aroyo (2018) showed that adding homophones and homographs as extra features to the existing linguistic features gave a small but significant improvement for humor detection in one-liner jokes. Chen and Soo (2018) implemented a Convolutional Neural Network model with Highway Networks to train an end-to-end neural network for humor detection in English and Chinese languages.

More recent approaches for humor related tasks are focused around transformer-based architectures (Ismailov, 2019; Mao and Liu, 2019; Weller and Seppi, 2019). In this work, we experiment with BERT and its derivative transformer models. BERT (Devlin et al., 2019) applies a bidirectional training of transformers with Masked Language Modelling and Next Sentence Prediction tasks. It is pre-trained on the Wikipedia (2,500 million words) and the Book Corpus (800 million words) datasets. RoBERTa (Liu et al., 2019) is based on BERT but has been trained for a longer time with more amount of data (10 times that of BERT). Unlike BERT, RoBERTa is not trained for the Next Sentence Prediction task. DistilBERT (Sanh et al., 2019) is a distilled version of BERT, with a 40% size reduction over BERT while retaining 97% of its language understanding capacity. ALBERT (Lan et al., 2019) is a lite BERT with an 89% parameter reduction over the BERT-base model. It uses a self-supervised loss that focuses on inter-sentence coherence. BERT can be used with a classification head in humor detection tasks. A significant improvement over previous baselines proves that self-attention layers of BERT achieve success by extracting crucial humor-related features (Weller and Seppi, 2019). Ismailov (2019) used a language-modeling based fine-tuning approach with BERT to get better results for humor-grading related tasks. These properties of BERT can be tested for generalization across the other BERT derivative models for short-edits based humor grading.

In this work, we test the ability of the BERT models to perform humor-grading tasks on the Humicroedit dataset (official task dataset) and FunLines dataset (additional dataset). Both of these are English short-edits based humor-grading datasets. The Humicroedit dataset consists of 15,095 edited headlines, and the FunLines dataset consists of 8,248 edited headlines. Both the datasets share the same format. Table 1 shows two different edits of the same news-headline with varying grades of humor.

3 Methodology

Since this is a short-edits based humor-grading dataset, the original text context is also crucial while grading the humor. To incorporate the original text context, we use a two-sentence input based approach with the BERT models, as shown in Figure 1. We concatenate the original and the edited news-headlines (separated and padded with special tokens of the respective BERT model) and feed it to the models in a tokenized format. Further, we inspect the effectiveness of masked language model fine-tuning based pre-training of the BERT models. Inspired from Ismailov (2019), we initialize the BERT models with their respective fine-tuned language models weights. The implementation details of the models are described in Section 4.
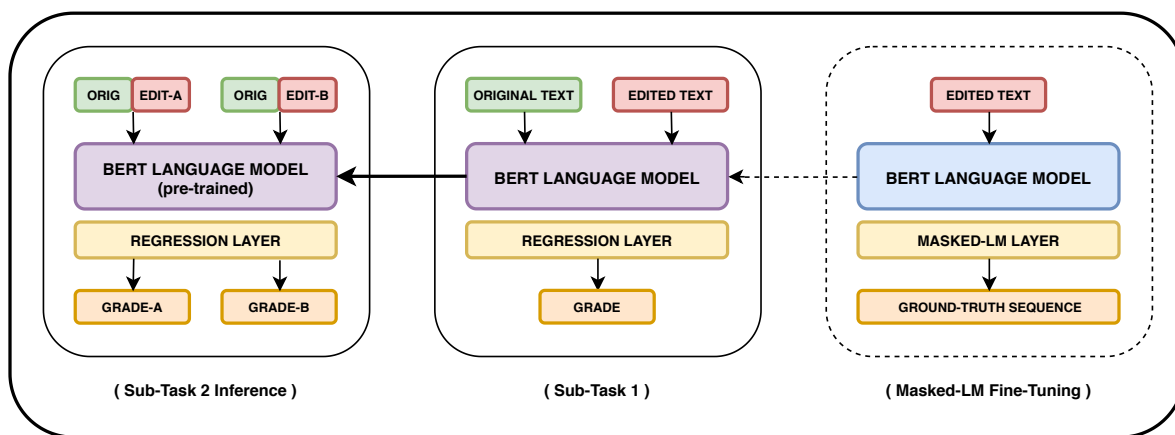
Figure 1: System Architecture

We follow a masked word prediction based language modeling approach to fine-tune the BERT language models on the entire dataset. Since the datasets do not come from an open-domain source and are limited to news headlines, all the words in the text are masked for prediction. We use only the edited humorous texts for language modeling pre-training. All the layers of the language models are trained with a maximum sequence length of 256 tokens for masked word prediction. The trained weights from these fine-tuned language models are used to initialize the model weights for sub-task 1.

For sub-task 1, humor grading is considered a regression task over the humor grades of the edited news-headlines. We experiment with two types of weight-initializations for this task: weights from the language model fine-tuned on the dataset and the original pre-trained BERT model weights. We fine-tune the BERT models for the regression task by appending a fully-connected layer. This layer takes in the pooled output embedding from the BERT models and returns a humor grade value. The models are trained with the mean-squared error between the predicted and ground-truth humor grades as an objective function. The model weights are optimized for this objective function by Adam optimizer. For sub-task 2, we directly use the sub-task 1 models for a zero-shot inference over the pairs of edited news-headlines, as shown in Figure 1. After getting the humor grades for both the edits of a news-headline, we label the one with a higher grade value as the more humorous one.

4 Experimental Setup

For all our experiments, we use the PyTorch[2] implementations of the smallest base variants of the pre-trained BERT models by Hugging Face[3] (Wolf et al., 2019). The data distribution across the Training-Development-Test splits used for our experiments is shown in Table 2. We validate our models with the mean-squared-error loss for the development set of sub-task 1. The official evaluation metric for sub-task 1 (humor grade regression) is the root-mean-squared-error (RMSE) loss. Whereas, the official evaluation metric for sub-task 2 (humor grade based multi-class classification) is the categorical accuracy. Since we directly use the sub-task 1 models for a zero-shot inference over the sub-task 2 data samples, there is no training involved for sub-task 2. Hence, we treat the entire sub-task 2 dataset (i.e., training + development + test) as the test set for our experiments. We also report the accuracy results on the official test set for the Humicroedit dataset. We create our own data splits for sub-task 1 on the FunLines dataset, as shown in Table 2. Humicroedit and FunLines datasets share a similar format, but their backgrounds differ with respect to the annotations. To check the generalization of BERT models in terms of short-edits based humor detection, we perform cross-dataset inferences. For this, we use models trained on the Humicroedit dataset for an inference over the FunLines dataset and those trained on the FunLines dataset for an inference over the Humicroedit dataset. Further, we use bert-viz tool by Vig (2019) to get a visualization over the self-attention weights of the BERT models.

---

[2] https://pytorch.org/
[3] https://huggingface.co/

Dataset	Sub-Task	Training	Development	Test
Humicroedit	1	9652	2419	3024
	2	9381	2355	2960
FunLines	1	5274	1322	1652
	2	-	-	1958

Table 2: Data distribution across the Training-Development-Test splits (number of samples)

5 Discussion

For the final submission, we use BERT with masked word language model fine-tuning for both the sub-tasks. On the final test-set leaderboard, our system ranks 14th for sub-task 1, with a RMSE loss of 0.53318 and 13th for sub-task 2, with an accuracy of 0.62177. Further, in the post-evaluation experiments, we analyze the performance of all the BERT models on both the Humicroedit and the FunLines dataset. Table 3 shows the performance metrics of the BERT models both with and without the masked word language model fine-tuning along with the cross-dataset inferences. Overall, the BERT models perform better without masked word language model fine-tuning. This shows the capability of the BERT models to perform humor-grading tasks without the need for any language model fine-tuning on the humor dataset. All the BERT models perform significantly well for the cross-dataset inferences, showing their generalization capabilities across datasets from different backgrounds. Smaller BERT models (ALBERT: 11M parameters and DistilBERT: 65M parameters) do not show a significant drop in performance for sub-task 1 as compared to their bigger counterparts (BERT: 110M parameters and RoBERTa: 125M parameters). But we see a relatively lower accuracy on sub-task 2 for these smaller models. For sub-task 1, BERT slightly outperforms all the other models on both the datasets. Whereas for sub-task 2, RoBERTa gives a better performance than the rest of the models, with a significant gain on the FunLines dataset. Overall, all of the BERT models perform better on the Humicroedit dataset, which might be attributed to its relatively larger size as compared to the FunLines dataset.

BERT DERIVATIVE MODEL	TRAINING DATASET	MASKED-LM FINE TUNING	HUMICROEDIT			FUNLINES	
			TASK-1 (TEST SET) (RMSE)	TASK-2 (TEST SET) (ACC)	TASK-2 (ALL DATA) (ACC)	TASK-1 (TEST SET) (RMSE)	TASK-2 (ALL DATA) (ACC)
BERT (base-uncased)	Humicroedit	Yes	0.533[*]	0.622[*]	0.622	0.643	0.605
		No	**0.53**	0.619	**0.643**	0.601	0.616
	FunLines	Yes	0.671	0.601	0.598	0.527	0.6
		No	0.631	0.614	0.604	**0.521**	0.677
RoBERTa (base)	Humicroedit	Yes	0.533	**0.628**	0.641	0.591	0.628
		No	0.541	0.596	0.643	0.591	0.596
	FunLines	Yes	0.668	0.507	0.499	0.56	0.493
		No	0.696	0.587	0.581	0.541	**0.718**
ALBERT (base-v2)	Humicroedit	Yes	0.582	0.537	0.552	0.669	0.516
		No	0.577	0.55	0.556	0.658	0.548
	FunLines	Yes	0.673	0.521	0.524	0.572	0.553
		No	0.704	0.517	0.525	0.568	0.541
DistilBERT (base-uncased)	Humicroedit	Yes	0.572	0.541	0.536	0.647	0.541
		No	0.562	0.577	0.574	0.648	0.572
	FunLines	Yes	0.678	0.543	0.541	0.557	0.536
		No	0.67	0.545	0.549	0.551	0.563

Table 3: Post Evaluation Experiments (*final submission)

BERT models use the [CLS] token as the pooled output representation for the entire input sequence. The [CLS] token is ultimately used for the humor-grading tasks in our experiments. In order to interpret the BERT models' decisions, we analyze the multi-head self-attention weights of the [CLS] token from the final layer of the trained BERT model, as shown in Figure 2. Taking a closer look at the attention-visualization over these attention weights reveals that the model learns to assign more attention to certain parts of the text, which play a role in imparting funniness to the text as a whole. In general, some attention heads (2nd and 12th head) learn to assign more attention to the edited words as compared to other words in the text. The attention from the 2nd and 12th head is represented by orange-colored connections in Figure 2.

Figure 2-1 and Figure 2-3 represent the attention visualization of some of the best predictions from the Humicroedit and FunLines dataset, respectively. In these cases, we observe that the [CLS] token firmly attends the edited words by the 2nd and 12th attention head. On the other hand, Figure 2-2 and Figure 2-4 represent some of the worst predictions from the Humicroedit and FunLines dataset, respectively. Here, the [CLS] token divides its attention relatively more uniformly across the entire input sequence, with relatively lesser attention from the 2nd and 12th attention head. The same trend is observed across most of the samples from both of the datasets for all the BERT models. This reveals the importance of self-attention in the decision making of the BERT models for short-edits based humor grading.
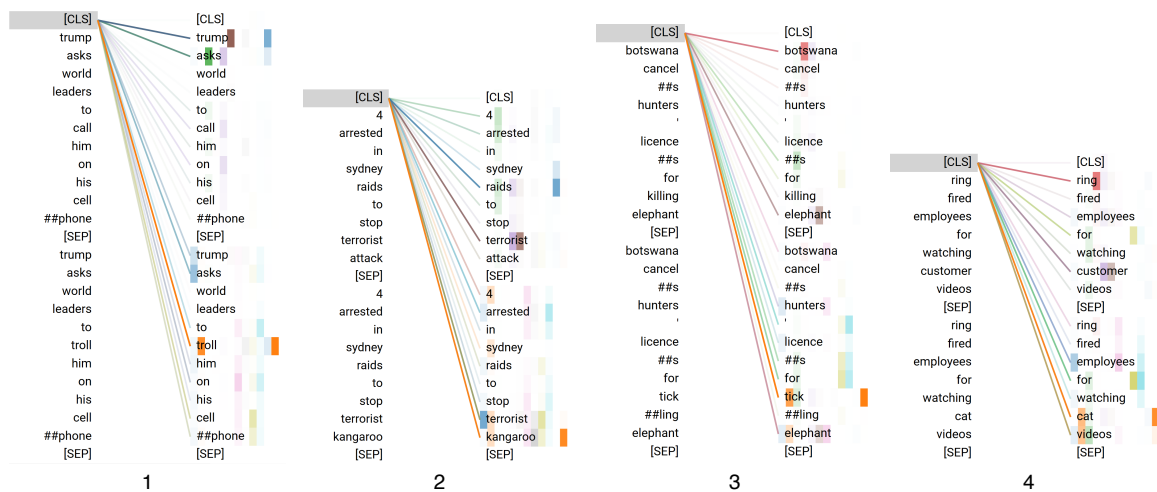

Figure 2: Attention Visualisation over self-attention weights of the BERT model

6 Conclusion

In this work, we tested the ability of BERT and its derivative models for short-edits based humor-grading tasks. Our experiments showed that the BERT models perform better with their original pre-trained weights, and there is no need for language model fine-tuning on the humorous texts. The tests also revealed that the BERT models show good generalization capabilities for humor-grading tasks by performing significantly well with zero-shot inference across the sub-tasks and cross-dataset inferences. The qualitative analysis over the self-attention weights of the BERT model revealed the attentive bias shown by some of the attention heads towards the edited parts of the text. It also showed the importance of self-attention in the decision making ability of the BERT models.

We plan to extend our work by testing autoregressive and generative transformer models for humor-grading related tasks. One can also experiment with different language modeling techniques and different input formats for the short-edits based humor-grading tasks. To further test the abilities of the BERT models for humor-grading, we plan to test our approach on other humor-related datasets such as the "Pun of the Day" and the" 16000 One Liners" datasets.

References

Sven Beukel and Lora Aroyo. 2018. Homonym detection for humor recognition in short text. 10.

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.

Luis Chiruzzo, Santiago Castro, Mathías Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In *IberLEF@SEPLN*.

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, June. Association for Computational Linguistics.

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, June. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. Semeval-2020 Task 7: Assessing humor in edited news headlines. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020b. Stimulating creativity with funlines: A case study of humor generation in headlines. In *Proceedings of ACL 2020, System Demonstrations*, Seattle, Washington, July. Association for Computational Linguistics.

Adilzhan Ismailov. 2019. Humor analysis based on human annotation challenge at iberlef 2019: First-place solution. In *IberLEF@SEPLN*.

Chloé Kiddon and Yuriy Brun. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 89–94, Portland, Oregon, USA, June. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jihang Mao and Wanli Liu. 2019. A bert-based approach for automatic humor detection and scoring.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada, August. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

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, November. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898.