Smash at SemEval-2020 Task 7: Optimizing the Hyperparameters of ERNIE 2.0 for Humor Ranking and Rating

J. A. Meaney **Steven R. Wilson** **Walid Magdy**
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
{jameaney, steven.wilson}@ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

The use of pre-trained language models such as BERT and ULMFiT has become increasingly popular in shared tasks, due to their powerful language modelling capabilities. Our entry to SemEval uses ERNIE 2.0, a language model which is pre-trained on a large number of tasks to enrich the semantic and syntactic information learned. ERNIE's knowledge masking pre-training task is a unique method for learning about named entities, and we hypothesise that it may be of use in a dataset which is built on news headlines and which contains many named entities. We optimize the hyperparameters in a regression and classification model and find that the hyperparameters we selected helped to make bigger gains in the classification model than the regression model.

1 Introduction

Verbal humor uses a variety of linguistic features, such as synonymy, wordplay, and phonological similiarities, as well non-linguistic features like world knowledge, to produce a comic effect. That such a broad set of skills are required to understand humor, has led several researchers to deem that computational humor is an AI-complete problem (Stock and Strapparava, 2006; Binsted et al., 2006). There is a relatively longstanding body of research into humor detection in a limited domain, such as knock-knock jokes (Taylor and Mazlack, 2004), one-liners (Mihalcea and Strapparava, 2006) and humorous news articles from the satirical news publication *The Onion* (Mihalcea and Pulman, 2007). However, the use of shared tasks has attracted more attention and interest in the field since 2017. While previous challenges have focused on collecting Twitter data (Potash et al., 2017; Castro et al., 2018), SemEval 2020 (Hossain et al., 2020) took an original approach and generated the data by collecting news headlines and then asking annotators to edit one word in the headline to make it humorous (Hossain et al., 2019). These headlines emulate those of *The Onion*. The edits shown below indicate the location of the substitution and the word to be inserted.

Table 1: Example of Funlines Headline

Type	Text	Edit	Score
Original	Mitch McConnell thinks tax reform will take longer than Trump claimed		
Edit 1	Mitch McConnell thinks tax reform will take longer than Trump <claimed/>	haircut	2.8
Edit 2	Mitch McConnell thinks tax <reform/>will take longer than Trump claimed	return	1.6

The edited headlines were then rated for humor by subsequent annotators. Sub-task A was to predict the mean funniness score of the edited headline. In sub-task B, the systems saw two edits of the same headline, and predicted which one had achieved the higher mean funniness score.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1049–1054
Barcelona, Spain (Online), December 12, 2020.

2 Previous Work

Excluding work on puns, there have been three humor detection shared tasks in recent years: Semeval 2017 (Potash et al., 2017), HAHA 2018 (Castro et al., 2018) and HAHA 2019 (Chiruzzo et al., 2019). As the tasks and data have varied between them, direct comparison is not possible. However, a comparison of approaches to the tasks shows some interesting trends.

Semeval 2017's entries were evenly divided between feature engineering approaches and deep learning systems, with both achieving competitive results. The highest ranking team in the official results for task A, SVNIT (Mahajan and Zaveri, 2017), used an SVM with incongruity, ambiguity and stylistic features. The second highest-ranking team, Datastories (Baziotis et al., 2017) opted for a Siamese bi-LSTM with attention. Interestingly, a remarkably simple system prevailed in task B: Duluth (Yan and Pedersen, 2017) used the probability assigned to the text by a bigram language model instead of the output of a classifier to make predictions.

Entries to HAHA 2018 were divided along similar lines. The winning system used Naive Bayes and ridge regression models optimized with an evolutionary algorithm (Ortiz-Bejar et al., 2018) with the runner up using a bi-LSTM with attention (Ortega-Bueno et al., 2018).

HAHA 2019 saw a sea change towards the use of transfer learning models, such as BERT (Devlin et al., 2018) and ULMFiT (Howard and Ruder, 2018). These models leverage large amounts of data and transformer attention models to learn contextual relations between words. Adilism (Ismailov, 2019) used multilingual BERT base uncased and extended the language model training without labels, before finetuning their system with the dataset labels. The second place system used an ensemble of a BERT model and ULMFiT, with Naive Bayes and SVM classifiers. The majority of the top entries to this task used BERT in some way, although one noted that it did not improve performance as expected (Ortega-Bueno et al., 2019).

3 System Overview

3.1 Why ERNIE 2.0?

As BERT models are trained on a masked-language model and sentence prediction task, they capture mainly word-level and sentence-level information. By comparison, ERNIE 2.0 (Sun et al., 2020) - henceforth ERNIE - aims to capture more lexical, syntactic and semantic information in corpora, by training on eight different tasks in a continual pre-training framework. Knowledge masking features among these eight tasks, and is implemented by treating a phrase or entity as an entire unit, instead of masking the constituent words. The distinction in how BERT and ERNIE learn is illustrated in how they learn the following sentence: Harry Potter is a series of fantasy novels written by J. K. Rowling

- Learned by BERT: `[mask] Potter is a series [mask] fantasy novels [mask] by J. [mask] Rowling`
- Learned by ERNIE: `Harry Potter is a series of [mask] [mask] written by [mask] [mask] [mask]`

BERT captures co-occurrence information of 'J' with 'K' and 'Rowling', however it does not capture information about the entity J. K. Rowling. By modelling this entity as a single unit, ERNIE claims to be capable of extrapolating the relationship between Harry Potter and J. K. Rowling (Sun et al., 2019). Furthermore, ERNIE is trained on a wide varieties of domains, including encyclopedias and news articles, giving the model a lot of knowledge of named entities. This is of great benefit in the Funlines dataset, which is built on news headlines, and therefore features a large number of named entities, particularly politicians. This may help the model to infer the relationship between Mitch McConnell and Trump in the example from table 1.

3.2 Text Preprocessing

The dataset featured the original headline, with the word which had been replaced in angle brackets, and the substitute word separate. We rendered the edited headlines by placing the word in angle brackets into the sentence. This did not give our model access to the keyword, or to the original headlines.

For ERNIE models, we preprocessed the data as follows: We lowercased the texts and tokenized them into word pieces, this was implemented with a greedy longest-match-first system to tokenize them given the vocabulary. As is conventional for ERNIE, we then added a [CLS] token to the start of each text, and a [SEP] token to the end of each text, with an additional [SEP] replacing the [CLS] in the second text for pairs of texts (e.g. task 2). We also padded sequences to a maximum length of 128.

3.3 Baseline

For task 1, we create two baselines, one which predicted a constant value, and the other which predicted the mean value, using scikit learn (Pedregosa et al., 2011).

Table 2: Baselines for Task 1

System	RMSE
Predict Constant Value	0.7214
Predict Mean Value	0.6968

For task 2, we created three baselines. In the first, we always predicted the same label. The second baseline was a trigram language model built on KenLM (Heafield, 2011), using a dataset containing around 200,000 news headlines from 2012-2018 editions of the Huffington Post[1]. Similarly to the approach taken by the Duluth team (Yan and Pedersen, 2017) in SemEval 2017, we reasoned that the funnier of the two headlines would be the least similar to real news headlines, so we selected the sentence that had a lower log probability according to the model. However, this performed worse than the first baseline.

The third baseline was a trigram model built the headlines labelled as sarcastic from a sarcastic news dataset (Misra and Arora, 2019). These headlines came from *The Onion*, which the competition dataset seeks to emulate. Here we reasoned that the funnier headline would have a higher log probability under this language model. Predicting labels in this way was an improvement over the other two baselines, suggesting that the unique data generation methods in this challenge succeeded in emulating satirical headlines in some way.

Table 3: Baselines for Task 2

System	Accuracy
Predict Constant Value	0.4475
200k Huff Post Headlines	0.4314
28k Onion Headlines	0.4546

3.4 Model Configuration

For the transfer learning models, we used ERNIE base which has 12 layers, a hidden size of 768 and 12 self-attention heads. We used a maximum sequence length of 128, a dropout probability of 0.1 and the Adam optimizer. To finetune for task 1, we built a fully connected layer with mean square error as the loss function. For task 2, after the fully connected layer, we added a softmax layer and used cross entropy as the loss function.

4 Experiments

We experimented with optimizing three hyperparameters: learning rate (1e-06, 0.0001 or 0.001), batch size (16, 32 or 64) and number of epochs (3, 4 or 5). For the sake of brevity, we report only the three

[1] https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection

highest and lowest results for each task. The results reported are the mean of 5 runs, with standard deviation in parentheses.

Table 4: Highest and Lowest Performing Parameters for Task 1

Learning Rate	**Batch Size**	**Epochs**	**RMSE (SD)**
0.0001	*16*	3	*0.5806 (0.011)*
0.0001	64	3	0.5817 (0.005)
0.0001	32	3	0.5829 (0.003)
0.001	16	3	0.5966 (0)
0.0001	16	5	0.6008 (0.010)
1e-06	*64*	*3*	*0.6009 (0.001)*

We noticed remarkably little variation in the task 1 results, regardless of the hyperparameter tweaking. Given that the same learning rate is observed in both high and low-scoring systems, and that there is no observable pattern in terms of batch size, this suggests that another hyperparameter, or variable may help to achieve better results.

By contrast, in task 2, we saw much more variation, with a jump of almost 11% from the lowest to the highest-scoring configuration. A small learning rate of 0.0001, along with a relatively large batch size of 64 featured in all three top results, and the number of epochs was decisive, bringing a 5% increase over at the optimal number - 4. We observed that the lowest learning rate also achieved the lowest scores. However, with too small a learning rate, the network appears not to converge, and varying the other hyperparameters does not impact this.

Table 5: Highest and Lowest Performing Parameters for Task 2

Learning Rate	**Batch Size**	**Epochs**	**Mean Accuracy (SD)**
0.0001	*64*	4	*0.59408 (0.017)*
0.0001	64	5	0.54644 (0.051)
0.0001	64	3	0.5150 (0.053)
1e-06	32	3	0.4911 (0.006)
1e-06	64	5	0.4880 (0.005)
1e-06	*64*	*3*	*0.4846 (0.002)*

5 Conclusion

While transfer learning models have achieved very impressive results on a variety of NLP tasks, the performance on this humor task was not as high as anticipated. Perhaps in a multi-task learning setup, we may have seen better performance. Nonetheless, our work demonstrates the importance of optimizing the hyperparamters of the finetuning layers, which achieved improvements on both tasks, but specifically the classification task.

Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

K Binsted, J Hendler, B van den Bergen, S Coulson, Antinus Nijholt, O Stock, C Strapparava, G Ritchie, R Manurung, H Pain, et al. 2006. Computational humor. *IEEE intelligent systems*, 21(suppl 2/2):59–69.

Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018. Overview of the haha task: Humor analysis based on human annotation at ibereval 2018. In *IberEval@ SEPLN*, pages 187–194.

Luis Chiruzzo, S Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. 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.

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

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Adilzhan Ismailov. 2019. Humor analysis based on human annotation challenge at iberlef 2019: First-place solution. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)*.

Rutal Mahajan and Mukesh Zaveri. 2017. Svnit@ semeval 2017 task-6: Learning a sense of humor using supervised approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 411–415.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 337–347. Springer.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Rishabh Misra and Prahal Arora. 2019. Sarcasm detection using hybrid neural network. *arXiv preprint arXiv:1908.07414*.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. Uo upv: Deep linguistic humor detection in spanish social media. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*.

Reynier Ortega-Bueno, Paolo Rosso, and José E Medina Pagola. 2019. Uo upv2 at haha 2019: Bigru neural network informed with linguistic features for humor recognition. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)*.

José Ortiz-Bejar, Vladimir Salgado, Mario Graff, Daniela Moctezuma, Sabino Miranda-Jiménez, and Eric S Tellez. 2018. Ingeotec at ibereval 2018 task haha: μtc and evomsa to detect and score humor in texts. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Oliviero Stock and Carlo Strapparava. 2006. Laughing with hahacronym, a computational humor system. In *21st conference of American Association for Artificial Intelligence (AAAI-06)*, pages 1675–1678. AAAI.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975.

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.

Xinru Yan and Ted Pedersen. 2017. Duluth at semeval-2017 task 6: Language models in humor detection. *arXiv preprint arXiv:1704.08390*.