References

[Attardo and Raskin, 1991] Salvatore Attardo and Victor Raskin. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 1991.

[Blinov *et al.*, 2019] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. Large dataset and language model funtuning for humor recognition. In *Proceedings of ACL*, pages 4027–4032, Florence, Italy, jul 2019.

[Cattle and Ma, 2018] Andrew Cattle and Xiaojuan Ma. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of IJCNLP*, pages 1849–1858, Santa Fe, New Mexico, USA, August 2018.

[Chen and Soo, 2018] Peng-Yu Chen and Von-Wun Soo. Humor Recognition Using Deep Learning. In *Proceedings of NAACL*, pages 113–117, New Orleans, Louisiana, jun 2018.

[Clark *et al.*, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceeding of the ACL Workshop BlackboxNLP*, pages 276–286, Florence, Italy, aug 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota, jun 2019.

[Goldwasser and Zhang, 2016] Dan Goldwasser and Xiao Zhang. Understanding satirical articles using common-sense. *TACL*, 4:537–549, 2016.

[González-Ibáñez *et al.*, 2011] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of ACL*, pages 581–586, Portland, Oregon, USA, jun 2011.

[Grave *et al.*, 2018] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of LREC*, 2018.

[Horvitz *et al.*, 2020] Zachary Horvitz, Nam Do, and Michael L. Littman. Context-driven satirical news generation. In *Proceedings of FLP*, pages 40–50, Online, July 2020.

[Hossain *et al.*, 2019] Nabil Hossain, John Krumm, and Michael Gamon. "President vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of ACL*, pages 133–142, Minneapolis, Minnesota, jun 2019.

[Hossain *et al.*, 2020] Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. Stimulating creativity with FunLines: A case study of humor generation in headlines. In *Proceedings of ACL*, pages 256–262, 2020.

[Jain and Wallace, 2019] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of NAACL*, pages 3543–3556, Minneapolis, Minnesota, June 2019.

[Kiddon and Brun, 2011] Chloé Kiddon and Yuriy Brun. That's what she said: Double entendre identification. In *Proceedings of ACL*, pages 89–94, Portland, Oregon, USA, jun 2011.

[Kovaleva *et al.*, 2019] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *EMNLP-IJCNLP*, pages 4365–4374, 2019.

[Mihalcea and Strapparava, 2005] Rada Mihalcea and Carlo Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of EMNLP*, 2005.

[Potash *et al.*, 2017] Peter Potash, Alexey Romanov, and Anna Rumshisky. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of SemEval*, pages 49–57, Vancouver, Canada, aug 2017.

[Purandare and Litman, 2006] Amruta Purandare and Diane Litman. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*. In *Proceedings of EMNLP*, pages 208–215, Sydney, Australia, July 2006.

[Radford *et al.*, 2019] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. `https://openai.com/blog/better-language-models`, 2019. Accessed: 2021-01-20.

[Raskin, 2008] Victor Raskin. *The Primer of Humor Research*. De Gruyter Mouton, Berlin, Boston, 2008.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, Hong Kong, China, nov 2019.

[Reyes *et al.*, 2013] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268, March 2013.

[Stock and Strapparava, 2003] Oliviero Stock and Carlo Strapparava. Getting serious about the development of computational humor. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, page 59–64, 2003.

[Taylor and Mazlack, 2004] Julia M. Taylor and Lawrence J. Mazlack. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1315–1320, 2004.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[Wallace *et al.*, 2015] Byron C. Wallace, Do Kook Choe, and Eugene Charniak. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of ACL*, pages 1035–1044, Beijing, China, jul 2015.

[Weller and Seppi, 2019] Orion Weller and Kevin Seppi. Humor Detection: A Transformer Gets the Last Laugh. In *Proceedings of EMNLP-IJCNLP*, pages 3621–3625, Hong Kong, China, nov 2019.

[West and Horvitz, 2019a] Robert West and Eric Horvitz. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 7265–7272, 2019.

[West and Horvitz, 2019b] Robert West and Eric Horvitz. Unfun.me dataset. `https://github.com/epfl-dlab/unfun`, 2019. Accessed: 2021-01-15.

[Wiegreffe and Pinter, 2019] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP*, pages 11–20, Hong Kong, China, November 2019.

[Yang *et al.*, 2015] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor Recognition and Humor Anchor Extraction. In *Proceedings of EMNLP*, pages 2367–2376, Lisbon, Portugal, sep 2015.

[Zhang and Liu, 2014] Renxian Zhang and Naishi Liu. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, page 889–898, New York, NY, USA, 2014.

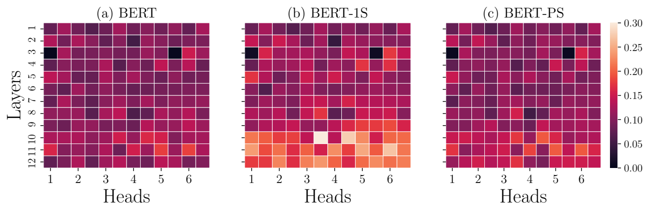

Figure 5: Average, per head, attention distance between funny and serious sentence of each encoder: (a) BERT, (b) BERT-1S, and (c) BERT-PS

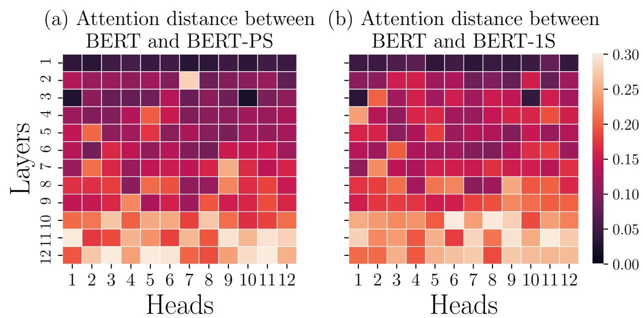

Figure 6: Average, per head, attention distance between finetuned models (BERT-1S in (a) and BERT-PS in (b)) and BERT

A More details on attention patterns

In the main paper, Fig. 2(a) and Fig. 2(b) report attention distances averaged for all heads in a same layer to focus on the effect of depth. For reference, we report in Fig. 6 the same computation as the one resulting from Fig. 2(a) but without averaging heads. Similarly, Fig. 5 reproduces Fig. 2(b) without averaging heads. With these plots, we can also confirm the conclusions from the main, namely that the finetuned models BERT-1S and BERT-PS differ from the non-finetuned BERT more in the last layers and the difference between funny and serious is increasing with depth (significantly more in BERT-1S and BERT-PS than BERT).

Finally, in the main paper, we identified the special head of BERT-1S by carefully comparing its attention patterns on modified and non-modified chunks for both funny and serious sentences. To confirm that his behavior is really special to BERT-1S, we report in Fig. 7 the same analysis for BERT-PS and BERT. We indeed observe that this head 10-6 of BERT-1S is special, since, when compared on the same scale, no other head in other models fires that much.

B Details about the "laughing head"

We here provide additional experiments related to the laughing head phenomena observed in the main paper Sec. 5.3.

B.1 The laughing head is where the finetuning happened

It is particularly intriguing to look at the attention distance per head h between funny and serious sentences for BERT-1S, as shown in Fig. 6 (b). In this figure, the rows are the layers and cells are heads, whose color indicates how large is the average attention difference between funny and serious sentences. We observe that head 10-6 (6-th head of layer 10) is particularly different for funny and for serious sentences.

B.2 The laughing head in other models

We repeat the experiments described in Sec. 5.3 with the distilBERT and RoBERTa architectures also in the single sentence setup. The attention maps on modified/non-modified for funny/serious sentences is reported in Fig. 8. We see that, to some extent, the head 5-6 in distilBERT also exhibits the same pattern as the head 10-6 of BERT-1S. However, no such head emerges in RoBERTa.

B.3 Example of activation of the laughing head

For randomly sampled funny sentences from the test set where the laughing head correctly activated on the modified token, we report the total attention paid to each token in the sentence in Fig. 9.

Figure 7: The columns represent attention paid to: (a) modified chunk on funny sentence, (b) non-modified chunk on funny sentences, (c) modified chunk on serious sentences, and (d) non-modified chunk on serious sentences. The first row is BERT, the second row is BERT-1S (same as Fig. 4 in the paper), and the last row is BERT-PS. Lighter color represents higher average attention.