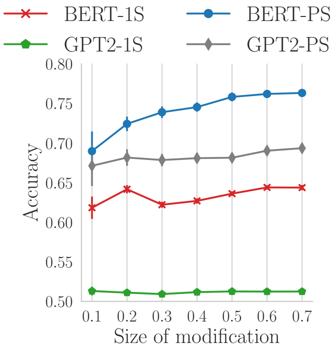


Figure 1: Humor detection accuracy as function of size of modification (lexical difference measured as Jaccard distance between token sets of funny and serious sentences in pairs).

4.2 Accuracy by Humor Type

We report performance per humor type in Table 2, which compares the performance of BERT-1S and BERT-PS against the GPT2 baselines (GPT2-1S predicts funny if the sentence is less likely than a threshold chosen to maximize accuracy on the training data; GPT2-PS predicts the less likely sentence in a pair to be the funny one). In general, performance is higher than on the full test set, probably because these are more standard instances of humor likely to be seen often in the training set. We also observe that different models perform best for different types. For instance, in the paired setup, models perform well when detecting humor in *non-obscene/obscene* pairs, but this type is one of the hardest in the single-sentence setup. Interestingly, the GPT2-PS baseline works better than BERT-PS for the *non-violence/violence* type of humor, which might be explained by the prevalence of violence in general text, sometimes used for funny purposes, and sometimes not, such that the model has difficulty capturing violence as a dimension of humor. Also, there is little improvement from GPT2-PS to BERT-PS for the *reasonable/absurd* type of humor, possibly because this type is mostly marked by sentence surprisal, which is explicitly captured by GPT2. Finally, the best model, BERT-PS, performs significantly worse for the more abstract *good/bad intentions* type than for other types.

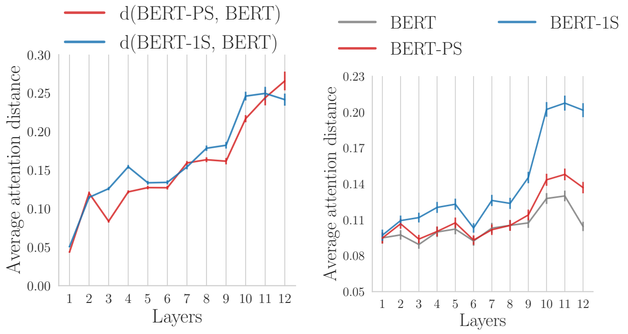
These findings highlight the particular strengths and weaknesses of existing models and can inform future work. We release a simple tool to repeat this analysis, so other researchers can easily benchmark their new models of humor detection.

5 Analysis of Transformer Model Behavior

In this section, we leverage the structure of the Unfun.me data to perform a deeper analysis of BERT-1S and BERT-PS.

5.1 Models Perform Better for Pairs with Large Modifications

We begin by measuring whether finetuned models perform differently for pairs with small vs. large lexical difference, which, for a given pair of sentences, we define as the Jaccard distance between token sets. We select subsets of the test set where all pairs have a distance greater than x and report the accuracy of BERT-1S and BERT-PS, as functions of x, in Fig. 1. For reference, we also report the GPT2-1S and GPT2-PS baselines.



(a) Attention distance to BERT

(b) Attention distance between funny and serious

Figure 2: Average per-layer attention distance: (a) between finetuned models (BERT-1S and BERT-PS) and non-finetuned BERT, for fixed sentences; (b) between funny and serious sentence in pairs, for fixed models.

While BERT-1S and BERT-PS perform significantly better on the subsets with larger modifications, GPT2 baselines do not see significant accuracy changes. The accuracy increase is particularly strong for BERT-PS, probably because pairs with a large lexical difference inherently have more information about their semantic difference. Despite never seeing sentences in pairs, the accuracy of BERT-1S also increases significantly with modification size. A potential explanation might be that funny sentences for which humans could not find a small modification in order to remove the humor may also be easier to recognize as funny.

5.2 Global Attention Patterns

To better understand BERT-1S and BERT-PS, we now investigate their attention patterns. The models have 12 layers and 12 attention heads per layer, for a total of 144 heads. For a given sentence s of length $|s|$, each head computes $|s|+2$ self-attention distributions (where the $+2$ comes from the two special "[CLS]" and "[SEP]" tokens).

More finetuning happens in later layers. First, we compare the attention distribution patterns of BERT-1S and BERT-PS after finetuning to those of BERT before finetuning. We write $A_{h_i}^{M}(s)$ for the i-th attention distribution (associated with the i-th input token) of model M for head h on sentence s. The distance between the attention distributions of two models M_a and M_b at head h on sentence s is given by the average Jensen–Shannon divergence

$$D_s^h(M_a, M_b) = \frac{1}{|s|+2} \sum_{i=1}^{|s|+2} JS\left(A_{h_i}^{M_a}(s), A_{h_i}^{M_b}(s)\right). \quad (1)$$

Following the work of Clark *et al.* [2019], we average this quantity over the test set $\mathscr{T}$ to obtain the average attention distance $D^h(M_a, M_b)$ between the two models at head h:

$$D^h(M_a, M_b) = \frac{1}{|\mathscr{T}|} \sum_{s \in \mathscr{T}} D_s^h(M_a, M_b). \quad (2)$$

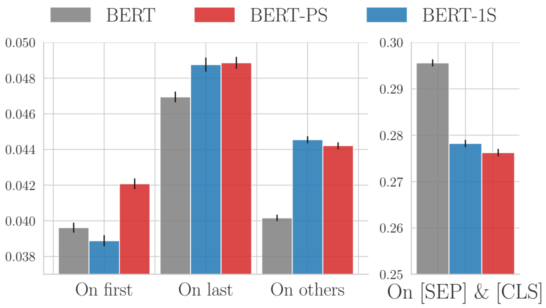


Figure 3: Average total attention paid to special positions and tokens for BERT, BERT-1S, and BERT-PS.

Furthermore, all heads in the same layer can be averaged to produce an average attention distance $D^L(M_a, M_b)$ between the two models at layer L. In Fig. 2(a), we report the attention distance D^L between finetuned models (BERT-1S and BERT-PS) and BERT. For reference, in Appendix A, we show the attention distances for each head. For both BERT-PS and BERT-1S, attention patterns significantly drift away from plain BERT's attention patterns as depth increases. It appears that finetuning modifies BERT more in later layers, in the more semantic stages, while preserving low-level syntactic processing in the earlier layers.

Attention difference between funny and serious increases with depth. Next, we measure the difference in attention between a funny sentence and its serious counterpart, computed as follows, for a model M, sentence pair (s_j, s_k), and head h:

$$D_M^h(s_j, s_k) = \frac{1}{|s_j|+2} \sum_{i=0}^{|s_j|+2} JS\left(A_{h_i}^M(s_j), A_{h_i}^M(s_k)\right). \tag{3}$$

For this analysis, we have to restrict ourselves to cases where $|s_j| = |s_k|$. Again, all heads in the same layer can be averaged to obtain an average attention distance between funny and serious in that layer. Averaging this quantity over all sentence pairs (for BERT, BERT-1S, and BERT-PS) yields Fig. 2(b). For both BERT-PS and BERT-1S, the difference in attention patterns between funny and serious sentences increases with depth, especially in the last three layers. This difference is significantly larger for finetuned models than for BERT and particularly large for BERT-1S. In fact, for BERT-1S, there is a jump at layer 10 observed in both Fig. 2(a) and Fig. 2(b); we shall come back to this observation in Sec. 5.3, where we study the behavior of one particular head in layer 10.

Boundary tokens are particularly important. In Fig. 3, we measure the total amount of attention paid to the first and last words, averaged over the test set, for BERT, BERT-1S, and BERT-PS. Here, total attention is obtained by summing the attention received by the respective position over all attention distributions in the model. We observe that, while (non-finetuned) BERT already attends more to the last word than to other words (as also observed by Kovaleva *et al.* [2019]), both BERT-1S and BERT-PS attend significantly more than BERT on the last word. Additionally, BERT-PS also attends significantly more to the first word than BERT. Finally, BERT attends a lot to special tokens, whereas both finetuned models redirect part of this attention toward actual words. These results confirm prior work, which has established that the humor in satirical headlines tends to be particularly associated with the first word and last word of the headline [West and Horvitz, 2019a; Hossain *et al.*, 2019].

5.3 Head 10-6: The "Laughing Head"

We found a particularly intriguing attention pattern on head 10-6 of BERT-1S. This head seems to specialize on attending to the "funny token" of a funny sentence, i.e., the token chosen by a human to remove the humor. (We focus here on pairs where exactly one token from the satirical version was modified.) This is clearly shown by the attention maps of Fig. 4, which plot the average attention paid by each head

1. to the modified (i.e., "funny") token in funny sentences (Fig. 4(a));
2. to the non-modified tokens in funny sentences (Fig. 4(b));
3. to the new token that replaces the "funny token" in serious sentences (Fig. 4(c));
4. to the other tokens in serious sentences (Fig. 4(d)).

Fig. 4(a) shows that head 10-6 is clearly special and attends strongly to the modified ("funny") token of funny sentences without activating in other cases. In particular, it does not activate for the same position in serious sentences (Fig. 4(c)). A simple rule that predicts the funny token to be the one to which head 10-6 pays most attention achieves an accuracy of 37%, nearly five times that of predicting a random token (8%).

West and Horvitz [2019a] showed that surface features such as parts-of-speech (POS) tags and position in the sentence were highly associated with whether a token was edited (e.g., final tokens were particularly likely to be edited). If head 10-6 only recognized such surface features, it should—unlike what we observe empirically—regularly activate for serious sentences as well. Investigating further, we explicitly compared the ability of head 10-6 to detect the funny token in funny sentences to three baselines: (1) predicting the final token, (3) predicting the rightmost token with the POS tag overall associated most with edited tokens, and (3) predicting the most surprising token (i.e., with the lowest likelihood according to GPT2). The best baseline (predicting the final token) correctly detects the funny token in 12% of instances, three times less frequently than head 10-6 (37%). Predicting the most frequent POS tag achieves 11%, and predicting the most surprising token, 10%. We further tested whether head 10-6 mostly recognizes surprising words by measuring its activation on the "funny" position when the respective token is replaced by a random token, finding that in this case head 10-6 activates only 60% as much as with the original token, which indicates that the activation of head 10-6 is only partially due to surprisal.

The existence of such a head is particularly striking given that BERT-1S never observes pairs of sentences, but only single sentences with a funny or serious label. One could have imagined that, even if a model developed the ability to detect the funny token, this property could be distributed inside the model. Yet, the jumps in attention distances observed in Fig. 2(a) and 2(b) are mostly explained by this one head.

This supports the hypothesis that BERT-1S learns to detect humor by first identifying a particularly important feature of

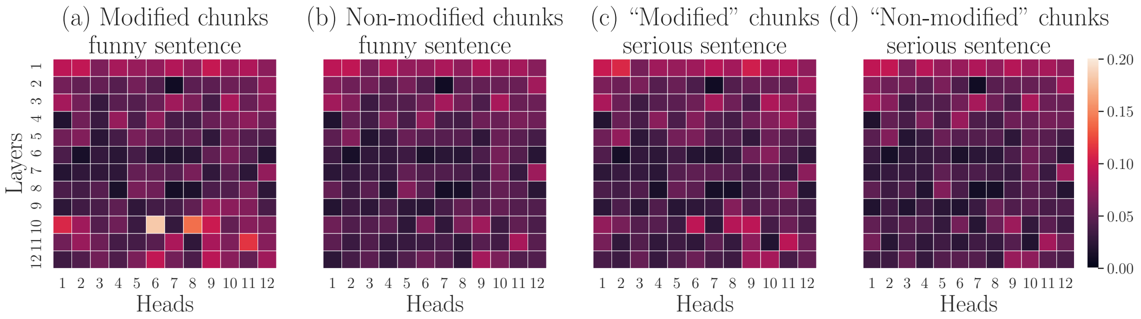


Figure 4: Average attention paid by BERT-1S to (a) the modified (i.e., "funny") token in funny sentences, (b) the non-modified tokens in funny sentences, (c) the new token that replaces the "funny token" in serious sentences, and (d) the other tokens in serious sentences. Lighter colors represent higher average attention.

	Modified token	Other tokens
Funny sentences	.240	.134
Serious sentences	.062	.095

Table 3: Perturbation analysis: Percentage of decisions changed when masking modified vs. other tokens. The difference is significant for funny ($p < 10^{-6}$), but not for serious ($p > 0.01$), sentences.

the data: what is the smallest change that distinguishes funny from serious? We further confirm this insight with a perturbation experiment: for each sentence in the test set, we iteratively mask each word and observe how the classification of BERT-1S changes. Intuitively, we expect that a system recognizing the funny token of a sentence would often switch its decision from funny to serious when the funny token is masked. In Table 3, we report the percentage of decision changes resulting from masking the modified token vs. masking other tokens, for both funny and serious sentences. Masking the modified (funny) token in funny sentences changes the decision from funny to serious 24% of the time, compared to only 13% when other tokens are masked. Furthermore, which token is masked does not make a significant difference (with respect to decision changes) in serious sentences. This confirms that the model has, to some extent, learned to recognize not only if a sentence contains humor, but also where the humor is located.

The effect of head 10-6—which we call the "laughing head"—is large and robust. It remains present and strong even when changing random seeds. Furthermore, we show in Appendix B that a laughing head also emerges in distilBERT but, interestingly, not in RoBERTa.

6 Discussion and Conclusion

We started by evaluating transformer-based architectures on the task of humor detection and then leveraged the unique paired structure of the data to obtain novel insights into how transformers deal with the task, finding that finetuned BERT models tend to perform better in cases with larger lexical differences between the funny and serious sentences in the pair. We also observed varying accuracy across humor types, with models being particularly strong at identifying humor when the funny and serious sentences are opposed along the *non-obscene/obscene* dimension, but struggling more with the *good/bad intentions* and *non-violence/violence* dimensions. An analysis of attention patterns revealed that finetuning mostly modifies the last transformer layers and that models attend to funny and serious sentences differently. This difference grows with layer depth significantly more than for non-finetuned BERT. This indicates that the critical computation takes place in the last transformer layers and that the model picks up on semantic, rather than lexical or syntactic, features. We also found that finetuned models redirect part of the attention dedicated to special tokens ("[CLS]" and "[SEP]") by non-finetuned BERT toward actual words, and particularly towards the last word, in line with the micro-punchline structure of typical satirical headlines [West and Horvitz, 2019a].

Our most striking finding pertains to the emergence of a "laughing head" that specializes on attending strongly to the funny parts of funny sentences. This head alone predicts which words "contain the humor" with an accuracy nearly three times as high as the best baseline.

Our core analyses rely on an investigation of attention heads. Jain and Wallace [2019] warned that attention patterns do not directly imply explanations of model decisions. However, following the subsequent recommendations of Wiegreffe and Pinter [2019], we always carefully compared the attention of finetuned models against frozen-weight versions (BERT without finetuning), allowing us to discover significant and meaningful qualitative changes happening during finetuning that enable the model to go from random to significantly-above-random accuracy. Our results thus shed lights on the inner workings of humor detection models.

Overall, this work shows that, although humor detection remains a challenging task, existing models can already capture highly nontrivial features of what makes satirical headlines funny. Moreover, our characterization of easy vs. hard instances can guide future research efforts to further help computational models recognize humor.

Acknowledgments

Thanks to Zachary Horvitz and Eric Horvitz for insightful feedback. With support from Swiss National Science Foundation (grant 200021_185043), European Union (TAILOR, grant 952215), and gifts from Google, Facebook, Microsoft