Laughing Heads: Can Transformers Detect What Makes a Sentence Funny?

Maxime Peyrard, **Beatriz Borges**, **Kristina Gligorić** and **Robert West**
EPFL
{maxime.peyrard, beatriz.borges, kristina.gligoric, robert.west}@epfl.ch

Abstract

The automatic detection of humor poses a grand challenge for natural language processing. Transformer-based systems have recently achieved remarkable results on this task, but they usually (1) were evaluated in setups where serious vs. humorous texts came from entirely different sources, and (2) focused on benchmarking performance without providing insights into how the models work. We make progress in both respects by training and analyzing transformer-based humor recognition models on a recently introduced dataset consisting of minimal pairs of aligned sentences, one serious, the other humorous. We find that, although our aligned dataset is much harder than previous datasets, transformer-based models recognize the humorous sentence in an aligned pair with high accuracy (78%). In a careful error analysis, we characterize easy vs. hard instances. Finally, by analyzing attention weights, we obtain important insights into the mechanisms by which transformers recognize humor. Most remarkably, we find clear evidence that one single attention head learns to recognize the words that make a test sentence humorous, even without access to this information at training time.

1 Introduction

Humor is a unique feature of human cognition and communication. The computational aspects of humor form a promising research area to advance the semantic understanding, commonsense reasoning, and language abilities of artifical intelligence (AI) systems. In particular, computational tasks such as humor detection and generation offer a rich and challenging testing ground for modern language understanding systems, to the extent that the ability to understand humor is commonly believed to be "AI-complete" [Stock and Strapparava, 2003].

Humorous texts often display complex narrative structures, rely on shared common-sense knowledge, and exploit subtle lexico-grammatical features of language [Raskin, 2008]. Nonetheless, impressive progress in humor detection has recently been made thanks to the advent of transformer architectures [Vaswani *et al.*, 2017] and the pretraining–finetuning paradigm [Devlin *et al.*, 2019]. For instance, Weller and Seppi [2019] finetuned a transformer to classify short texts as funny or serious, obtaining high accuracy on a dataset of Reddit jokes. They, however, neither performed error analysis nor gave insights into what signals are discovered and exploited by their model. In order to guide future research, a better understanding of how current models behave in practice, as well as a sharper outline of their capabilities, appears necessary.

In this work, we propose to leverage aligned pairs of funny and serious sentences collected via Unfun.me, an online game where players make minimal edits to satirical news headlines with the goal of making other players believe the results are serious headlines [West and Horvitz, 2019a]. Previously, West and Horvitz [2019a] released and analyzed 2.8K pairs from Unfun.me in order to better understand how humans create humor. Here, we use an extended dataset with more than 20K new pairs to probe the capacities of modern transformers such as BERT [Devlin *et al.*, 2019]. The minimal modifications between funny and serious headlines allow us to zoom in precisely on where the humor happens and thus perform entirely new kinds of analysis.

West and Horvitz [2019a] found that the difference between funny and serious headlines generated by humans tend to be explained by an opposition along certain dimensions important to the human condition, e.g., *reasonable vs. absurd* or *non-obscene vs. obscene*. We use the annotated, aligned pairs and the identified dimensions of opposition in order to perform fine-grained error analysis of transformer models.

Contributions. We finetune and evaluate standard BERT variants for two different humor detection tasks: (i) the *single-sentence setup*, where models detect whether an input sentence is funny or not, and (ii) the *paired setup* (cf. Sec. 4), where models detect for an input pair of aligned funny and serious sentences which one is funny. Based on the natural pairing of the data, we obtain diverse insights about the data and models:

1. In a careful error analysis, we characterize easy vs. hard instances (Sec. 4.2 and 5.1).
2. Inspection of attention heads reveals that the critical computation takes place in the last transformer layers, confirming that the model picks up on semantic, rather than lexical or syntactic, features (Sec. 5.2).
3. A particularly striking result is the emergence, during finetuning, of a head (the "laughing head") specialized in attending to the funny part of an input sentence. This

head alone detects the humorous part of a sentence five times more accurately than a random baseline (Sec. 5.3).

Code and data,[1] as well as an extended version of the paper (with the appendices referenced here),[2] are available online.

2 Related Work

We discuss two aspects of previous research efforts on computational humor relevant to our work: datasets and approaches.

Humor detection datasets. Dataset creation approaches usually extract texts considered humorous online, e.g., on Twitter [Potash *et al.*, 2017; Zhang and Liu, 2014] or Reddit [Weller and Seppi, 2019], and, in parallel, collect serious texts from other sources to induce balanced datasets. For instance, Mihalcea and Strapparava [2005] collected 16K one-liner punchlines and matched them with 16K news headlines. Similarly, Yang *et al.* [2015] matched 2.4K puns with 2.4K headlines. Finally, the "Stierlitz" dataset [Blinov *et al.*, 2019] consists of 60K Russian jokes paired with news headlines. Others have directly collected pairs of funny and serious sentences. West and Horvitz [2019a] collected pairs via Unfun.me, an online game where players propose small edits to turn satirical news headlines into serious-looking ones. The Humicroedit [Hossain *et al.*, 2019] and FunLines [Hossain *et al.*, 2020] datasets were obtained via the reverse approach, where humans edited serious headlines into funny ones. The researchers have used their datasets to study humor perception and creation. Here, we use an updated version of the data collected by West and Horvitz [2019a] to study automatic humor detection methods.

Humor detection approaches. Researchers have applied humor detection techniques to several kinds of humor, e.g., irony [Wallace *et al.*, 2015; Reyes *et al.*, 2013], sarcasm [González-Ibáñez *et al.*, 2011], or satire [Goldwasser and Zhang, 2016]. Various baselines have been proposed, such as statistical *n*-gram analysis [Taylor and Mazlack, 2004] or classifiers based on human-crafted features [Purandare and Litman, 2006; Kiddon and Brun, 2011]. Following the evolution of the NLP field, pretrained word vectors lead to further improvements [Yang *et al.*, 2015; Cattle and Ma, 2018]. Finally, deep learning approaches have become ubiquitous [Chen and Soo, 2018; Blinov *et al.*, 2019]. In particular, transformers [Vaswani *et al.*, 2017] such as BERT [Devlin *et al.*, 2019] are now used to recognize funny sentences [Weller and Seppi, 2019] and generate funny texts [Horvitz *et al.*, 2020]Many researchers have aimed to understand how humans generate and perceive humor [Raskin, 2008]. Some of the resulting theories have been empirically verified. For instance, an analysis of the Unfun.me dataset [West and Horvitz, 2019a] produced evidence in favor of the *General Theory of Verbal Humor* [Attardo and Raskin, 1991]. The original and modified ("unfunned") headlines are generally opposed to each other along certain dimensions important to the human condition (e.g., *reasonable vs. absurd*, *high vs. low stature*, or *non-obscene vs. obscene*). On the contrary, despite impressive recent progress on automatic humor detection, the trained models have not been analyzed in depth. For instance, Weller and Seppi [2019] and Blinov *et al.* [2019] focus on reporting the performance of their transformer models. Yet, to further advance the field of computational humor and NLP in general, it is important to understand the capabilities of these models.

[1] https://github.com/epfl-dlab/laughing-head

[2] https://arxiv.org/abs/2105.09142

3 Data

To investigate humor detection using transformer models, we employ data particularly well-suited for a fine-grained understanding in controlled settings: pairs of funny and serious sentences with a small lexical difference collected via the Unfun.me game and previously released by West and Horvitz [2019a]. Unfun.me is an online game that incentivizes players to make minimal edits to satirical news headlines with the goal of making other players believe the results are serious news headlines.

We use an extended dataset of 23,113 pairs [West and Horvitz, 2019b],which we randomly split into 18,832 training pairs, 2,414 validation pairs, and 1,867 testing pairs. Additionally, as part of the game, a subset of pairs has been annotated with quality ratings measuring how well the unfunning process worked, i.e., whether the unfunned sentence was perceived as serious by other humans. These annotations come from other players who evaluated the quality of the unfunned sentences (for details, see West and Horvitz [2019a]). From these annotations, we form a restrictive *high-quality test set* of instances that received the maximum score according to all annotators, consisting of 754 pairs. We later refer to this test set as "HQ".

Finally, a subset of the test set (254 pairs) comes with manual annotations from two trained annotators capturing the opposition that leads to humor in the pair (cf. Sec. 1). In this work, we use the following seven types of opposition (listed with examples; satirical versions in bold; multiple oppositions may apply to the same pair; "∅" refers to empty string):

1. normal/**abnormal**: *Bush picks* {***laser***, *rural*} *background for presidential portrait*
2. possible/**impossible**: *City opens new art* {***jail***, *museum*}
3. non-violence/**violence**: *Russian officials promise low* {***death***, *highway*} *toll for Olympics*
4. good/**bad** intentions: *BP ready to resume oil* {***spilling***, *drilling*}
5. reasonable/**absurd** response: *general motors reports record sales of new* {***disposable***, ∅} *car*
6. high/**low** stature: *Hollywood mourns passing of* {***16th or 17th Lassie***, *Robin Williams*}
7. non-obscene/**obscene**: *Tiger Woods announces return to* {***sex***, *golf*}

The advantage of such data for our analysis are threefold: (1) a naturally paired setup (2) with minimal modifications between funny and serious sentences and (3) additional annotations allowing us to investigate the performance of machine learning systems in fine-grained ways.

4 Models

We train standard transformer models to detect humor in the two setups described below.

Single-sentence setup. We first ignore the pairing between funny and serious sentences. The sentences are only associated with a binary label indicating whether they are funny or not. An encoder E maps a sentence s_i into a feature space, and a classifier converts the sentence representation $E(s_i)$ into a binary prediction. We train the model with the binary cross-entropy loss by finetuning the full pretrained model with back-propagation.

Paired setup. Next, we exploit the natural pairing of our data. Here, each funny–serious pair is first randomly ordered and then associated with a binary label indicating whether the first sentence is the funny one. This setup is naturally modeled via Siamese networks. For a given encoder E and an input pair (s_i, s_j), a classifier is trained on top of the concatenation of the feature representation of the two sentences, $[E(s_i), E(s_j)]$. Siamese networks have been successfully used recently in tasks based on the comparison of two sentences [Reimers and Gurevych, 2019].

Encoders used. The main focus of our analysis is on the BERT model, thus we employ BERT-base, an encoder with 12 layers and 12 heads per layer. Additionally, we report the performances of two BERT variants: distilBERT, a simplified and smaller alternative to BERT, and RoBERTa, a variant of BERT pretrained with an improved training setup. Each of these can be either used in the single-sentence setup (denoted "1S") or in the paired setup (denoted "PS"). For each setup, the BERT variants can be either fully finetuned or kept with weights frozen and only the classifier layer being finetuned. For reference, we also report the performance of baseline encoders without pretraining: BOW represents sentences by the average of their fastText word vectors [Grave *et al.*, 2018]; LSTM is a vanilla LSTM architecture trained on top of pretrained fastText vectors; and TRANSFORMER is a vanilla transformer with the same architecture as BERT, but trained from scratch without pretraining. Language models such as GPT2 [Radford *et al.*, 2019] are also important baselines, as it is often believed that humorous sentences are more surprising (i.e., have a lower probability according to the language model). In the paired setup, a natural baseline thus predicts the sentence with the lower GPT2 likelihood to be funny. In the single-sentence setup, a simple approach consists of predicting funny whenever the sentence probability is below some threshold previously identified via search on the training set. Later, we also use sentence probability scores provided by GPT2 in the analysis (cf. Sec. 5).

Error bars and significance level. Throughout the rest of the paper, error bars in plots represent bootstrapped 99% confidence intervals, and when we say that a change or difference is "significant", we technically mean that $p < 0.01$ in a paired t-test for comparing means.

4.1 Accuracy on Unfun.me Dataset

In Table 1, we report the accuracy of each encoder described above in both the single-sentence and the paired setup for both the full test set (*Full*) and the subset of the test set whose pairs were annotated as high-quality (*HQ*).

The results indicate that humor detection is a challenging task, especially in the single-sentence setup: simple baselines such as BOW, LSTM, and TRANSFORMER barely improve upon random prediction. Only finetuned BERT architectures are significantly better than random, with the exception of RoBERTa, which achieves above-chance accuracy (57.5%) even without finetuning. In comparison, systems perform much better in the paired setup, where the classifier can take its decision based on the information from both sentences. The best model, distilBERT with finetuning, achieves an accuracy of 77.7%. Both BERT and RoBERTa achieve similar accuracy, with no significant difference. Overall, performance on the full test set and the HQ subset is similar, with a tendency towards slightly lower performance on the high-quality instances. However, the differences between the high-quality set and the full set are not significant, indicating that the full test set performance is a good indicator of performance on high-quality instances. Given the popularity of BERT and the similar performance of BERT, distilBERT, and RoBERTa, we focus on BERT encoders in the rest of the paper: BERT with no finetuning (simply called BERT), BERT finetuned on the single-sentence setup (BERT-1S), and BERT finetuned on the paired setup (BERT-PS).

	1S		PS	
	Full	HQ	Full	HQ
No pretraining				
BOW	.511	.509	.515	.513
LSTM	.512	.511	.606	.598
TRANSFORMER	**.522**	**.526**	**.611**	**.607**
Pretraining, no finetuning				
GPT2	.526	.522	**.704**	**.682**
BERT	.536	.531	.689	.675
distilBERT	.534	.529	.685	.669
RoBERTa	**.575**	**.568**	.684	.675
Pretraining and finetuning				
BERT	.645	.641	.766	.737
distilBERT	**.651**	**.647**	**.777**	**.758**
RoBERTa	.649	.640	.755	.751

Table 1: Accuracy of various standard models for both setups: single-sentence (1S) and paired (PS). Datasets are balanced, so random baselines have accuracy 0.5. Best values per block in bold.

	1S		PS	
Type	GPT2	BERT	GPT2	BERT
normal/abnormal	.493	.630	.815	.849
possible/impossible	.518	.665	.790	.821
non-violence/violence	**.553**	.657	.842	.816
good/bad intentions	.532	.606	.702	.723
reasonable/absurd	.537	**.704**	.889	.907
high/low stature	.510	.647	.872	.892
non-obscene/obscene	.516	.582	**.918**	**.989**

Table 2: Accuracy per humor type as annotated in the Unfun.me dataset, for both the single-sentence (1S) and paired (PS) setup, and for finetuned BERT encoders and GPT2-based baselines.

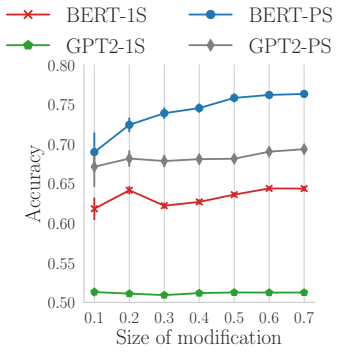
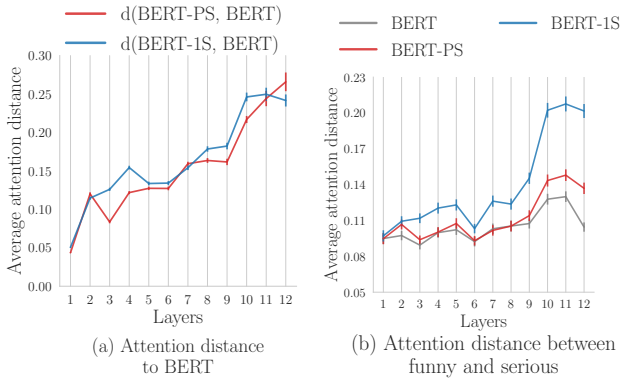


Figure 1: Humor detection accuracy as function of size of modification (lexical difference measured as Jaccard distance between token sets of funny and serious sentences in pairs).



(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.

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.

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 $A^M_{h_i}(s)$ 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

$$D^h_s(M_a,M_b) = \frac{1}{|s|+2}\sum_{i=1}^{|s|+2} JS\left(A^{M_a}_{h_i}(s), A^{M_b}_{h_i}(s)\right). \quad (1)$$

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:

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

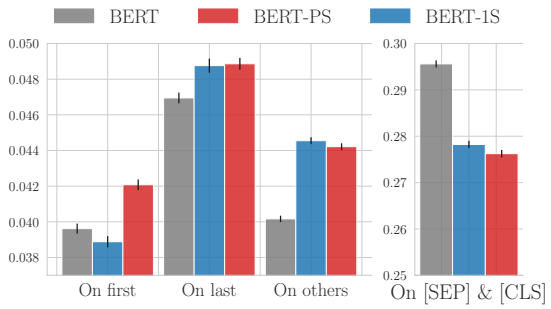


Figure 3: Average total attention paid to special positions and tokens for BERT, BERT-1S, and BERT-PS.

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:

$$D^h_M(s_j, s_k) = \frac{1}{|s_j|+2} \sum_{i=0}^{|s_j|+2} JS\left(A^M_{h_i}(s_j), A^M_{h_i}(s_k)\right). \quad (3)$$

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

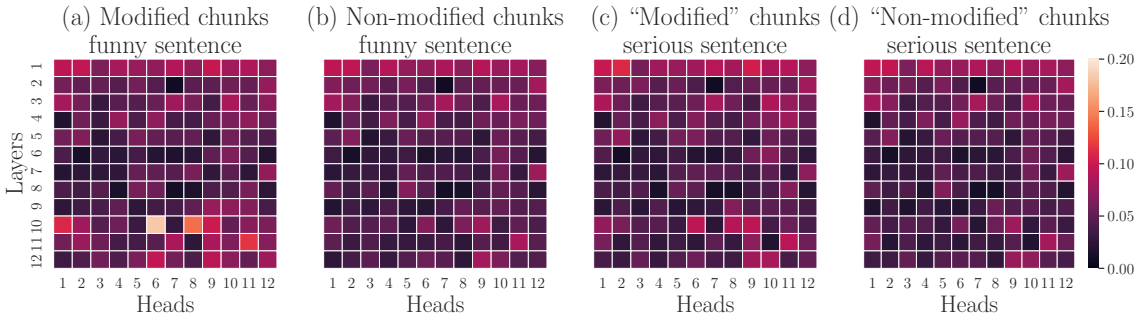


Figure 4: Average attention paid by BERT-1S to (a) the modified (i.e., "funny") token in funny sentences, (b) the non-modified tokens in funny sentences, (c) the new token that replaces the "funny token" in serious sentences, and (d) the other tokens in serious sentences. Lighter colors represent higher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

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.