This joke is [MASK]: Recognizing Humor and Offense with Prompting

Junze Li[†] **Mengjie Zhao**[‡] **Yubo Xie**[†] **Antonis Maronikolakis**[‡] **Pearl Pu**[†] **Hinrich Schütze**[‡]
[†]EPFL [‡]LMU Munich
junze.li@epfl.ch mzhao@cis.lmu.de

Abstract

Humor is a magnetic component in everyday human interactions and communications. Computationally modeling humor enables NLP systems to entertain and engage with users. We investigate the effectiveness of prompting, a new transfer learning paradigm for NLP, for humor recognition. We show that prompting performs similarly to finetuning when numerous annotations are available, but gives stellar performance in low-resource humor recognition. The relationship between humor and offense is also inspected by applying influence functions to prompting; we show that models could rely on offense to determine humor during transfer. Disclaimer: This paper contains model outputs that are offensive by nature.

1 Introduction

Humor is one of the most attractive phenomena in human communication, providing entertainment and relieving mental stress (Mihalcea & Strapparava, 2005; Lefcourt & Martin, 2012). Humor regulates human communication, as a result, computationally modeling it is expected to improve human-computer interaction experience (Nijholt et al., 2003; Rayz, 2017).

The first attempt to define humor was in ancient Greece, where ideologists and philosophers considered human laughter during comedies as a form of scorn (Morreall, 2020). Superiority theory of humor considers that the laughter is a type of superiority over other peoples' *physical defects or shortcomings*. Incongruity theory focuses on language semantics, claiming that incongruous meanings in the same context lead to a humor effect (Kant, 1790; Schopenhauer, 1883). This type of humor with the form of "setup + punchline" is widely studied in NLP; numerous linguistic resources were created (Mihalcea et al., 2010; Bertero & Fung, 2016; Xie et al., 2020).

Humor recognition (Mihalcea & Strapparava, 2005), striving to identify humorous texts, is the first step to enable NLP models to understand humor. *State-of-the-art models* (Meaney et al., 2021) *rely on transfer learning from large-scale pretrained language models (PLMs)*; finetuning (Devlin et al., 2019) is a common practice. In this work, we further investigate the effectiveness of prompting (Brown et al., 2020; Schick & Schütze, 2020; Liu et al., 2021a) in humor recognition. This is inspired by the fact that recognize humor of the form "setup + punchline" well matches prompting patterns. E.g., we can query a language model to fill in the blank in a pattern "setup + punchline. It is [MASK]", by answering "funny" or "normal". We show that prompting performs comparably to finetuning when numerous annotations are available, but it clearly outperforms finetuning in low-resource transfer scenarios.

Humor is subjective: judgment of what is humorous varies across human personalities, cultural backgrounds, and commonsense knowledge (Chen & Soo, 2018). As in the superiority theory, it

Workshop on Transfer Learning for Natural Language Processing, 2022, New Orleans.

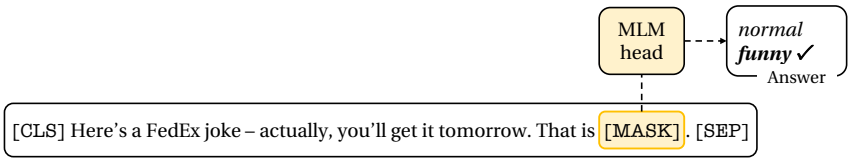

Figure 1: The prompting method for humor recognition.

is likely that some humor recognition datasets include "humorous" content that is offensive, e.g., offensive to women. This is undesirable because a machine-learning based NLP system, e.g., a virtual assistant, should never respond to a user query such as "Tell me a joke!" with text that is offensive (even if some may view the text as humorous). It is thus crucial to identify, mitigate, and reduce offense when modeling humor computationally.

The SemEval-2021 Shared Task shows that the relationship between humor and offense is subtle and challenging to detect (Meaney et al., 2021). In this work, we leverage influence functions (IF; Cook & Weisberg (1982); Koh & Liang (2017)) to identify offense when building humor recognition systems based on transfer learning. For each humorous test example, we locate its influential training examples and then inspect if they were offensive. Since utilizing IF does not incur architectural modifications (Koh & Liang, 2017), we integrate it to both finetuning and prompting. IF identifies humorous and offensive training examples, in both finetuning and prompting. However, the accuracy of identification varies when different resource limitations apply.

We make the following **contributions**: **(i)** We investigate the effectiveness of prompting, the new transfer learning paradigm in NLP, for humor recognition. Prompting performs similarly to finetuning when enough data is available, but gives stellar performance in low-resource scenarios. **(ii)** We integrate influence functions to finetuning and prompting, and show that the identified influential training examples indeed contain offensive contents. This characterization implies that PLM-based humor recognition systems rely on offensive contents to determine the presence of humor, which is an undesirable behavior and should be avoided in computational humor.

2 Method

We start with introducing finetuning and prompting for humor recognition. Next, we introduce procedures of estimating the influence of individual training examples for a test example.

Finetuning. We follow the standard practice of finetuning BERT as Devlin et al. (2019). We randomly initialize a linear classifier layer and then stack it on top of BERT. All the parameters are then finetuned using the humor recognition dataset. Vector of `[CLS]` is used to represent the sentence.

Prompting is a recent method of utilizing the PLMs (Brown et al., 2020; Schick & Schütze, 2020; Liu et al., 2021a). Unlike finetuning that randomly initializes a new classifier head, prompting reuses the masked-language model (MLM) head of BERT, requiring no extra new parameters. Prompting reformulates a sentence $\mathbf{x}$ using a pattern $f_{prompt}(\cdot)$ which contains the `[MASK]` token. After that, the PLM is asked to fill in the `[MASK]` token by selecting a token from $\mathcal{V}$=$\{y_1, y_2\}$, using the MLM head. In this work, we use $f_{prompt}(\mathbf{x}) = \mathbf{x}$ *That is* `[MASK]` . and $\mathcal{V}$ = {"funny", "normal"}. Figure 1 demonstrates prompting with a concrete example. To compute the model's performance, we use mapping "funny" $\rightarrow$ 1; "normal" $\rightarrow$ 0 such that standard metrics like accuracy can be computed.

Influence Functions. Following Han & Tsvetkov (2020), we leverage influence functions (Cook & Weisberg, 1982; Koh & Liang, 2017) to identify the most influential training examples. For each test example, we compute an influence score of every training example. Denoting an annotated training example as $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, the optimal parameters obtained on the annotated data $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$ as $\hat{\theta} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{z}_i, \theta)$, where θ refers to model parameters, $L(\mathbf{z}, \theta)$ refers to the loss, and n refers to the number of examples. An intuitive way of computing the influence of an example $\mathbf{z}_i$ is to inspect how $\hat{\theta}$ changes, when excluding or including $\mathbf{z}_i$ into the training dataset; i.e., $\mathcal{I}(\mathbf{z}_i)$ =

$\hat{\theta}_{-\mathbf{z}_i} - \hat{\theta}$, where $\hat{\theta}_{-\mathbf{z}_i} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n-1} \sum_{z\neq\mathbf{z}_i} L\left(\mathbf{z}_i, \theta\right)$. But this is computationally expensive because it requires retraining the model for every example $\mathbf{z}_i$. Cook & Weisberg (1982) propose influence functions as an efficient approximation by considering up-weighting a training example $\mathbf{z}_i$ by a small value ϵ_i. Because we are more interested in the change of model loss/predictions, we compute the influence of a training example $\mathbf{z}_i$ *to the prediction of a test example* $\mathbf{z}_{\text{test}}$, following Koh & Liang (2017):

$$\begin{aligned}\mathcal{I}(\mathbf{z}_i, \mathbf{z}_{\text{test}}) &\stackrel{\text{def}}{=} \frac{dL(\mathbf{z}_{\text{test}}, \hat{\theta})}{d\epsilon_i} \\ &= -\nabla_\theta L(\mathbf{z}_{\text{test}}, \hat{\theta})^\top (\frac{1}{n}\sum_{j=1}^{n} \nabla_\theta^2 L(\mathbf{z}_j, \hat{\theta}))^{-1} \nabla_\theta L(\mathbf{z}_i, \hat{\theta}).\end{aligned}$$

For each test example $\mathbf{z}_{\text{test}}$, we compute an influence score of each training example and then conduct z-normalization. We can then inspect whether or not the most influential training examples contain offense and then correct or discard them.

3 Datasets

We leverage the human-annotated HaHackathon Dataset (**HHD**) of SemEval 2021 Task 7 (Meaney et al., 2021). An HHD example consists of a text and two annotations: A binary humor label and an offense score ranging from 1 to 5. A humorous and not offensive example is "Why do birds fly south in the Winter? Because its too far to walk! " while a humorous but offensive example is "Getting a girlfriend is a lot like getting a car. The more money you have, the more options you have.".

HHD contains humor texts covering eight categories including Sexism, Body, Origin etc. In this work, we limit our scope to gender-related humor, and extract related examples with surface form matching using the keywords in category "Sexism". We call our constructed dataset Gender Humor Dataset (**GHD**). We enforce two constraints when creating GHD: **(i)** For gender-related texts and the ones not related to gender, we keep the same amount of humorous and non-humorous texts. This makes sure that a model cannot make predictions through simple heuristics. **(ii)** We use 80% examples for training and 20% examples for testing. Appendix Figure 3 plots the offense score distribution of GHD training examples. We observe a long-tail distribution, where the maximum offense score is 4.7 and the average offense score is 0.64.

4 Experiment

We conduct a series of experiments to answer the following research questions. **RQ1**: How well does prompting perform in humor recognition, compared to finetuning when the amount of labeled data varies? **RQ2**: Do the influential training examples correspond to humorous test examples contain offensive contents? **RQ3**: What is the difference between the identified offensive training examples in finetuning and prompting?

In our experiments, we use `BERT-base-uncased` (Devlin et al., 2019) as our PLM. We implement our models using PyTorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2020). Detailed experiment configurations are in Appendix §A.2.

4.1 Comparing Finetuning and Prompting Humor Recognition Methods (RQ1)

We firstly compare the performance of finetuning and prompting of recognizing humor with HHD. We conduct experiments using all of the training examples ("Full data") as well as randomly sampled few-shot data. For few-shot experiments, we repeat each experiment five times and report mean and standard deviation. Figure 2 left shows the results. It can be observed that prompting has an overall better performance than finetuning in the few-shot scenarios; the performance difference between these two methods decreases as more examples become available. Lastly, prompting and finetuning perform similarly when all HHD training examples are used.

	Prompting		Finetuning	
	F1-score	**Acc.**	**F1-score**	**Acc.**
16-shot	**74.20** (10.65)	**67.86** (5.57)	67.32 (15.03)	58.38 (4.07)
32-shot	**79.68** (1.15)	**70.90** (0.96)	75.66 (0.73)	62.28 (1.37)
64-shot	**80.36** (0.52)	73.21 (0.46)	80.29 (2.25)	**74.64** (3.66)
128-shot	**83.26** (0.83)	**78.08** (0.84)	81.07 (2.88)	74.48 (4.32)
Full data	92.85	91.00	**93.00**	**91.30**

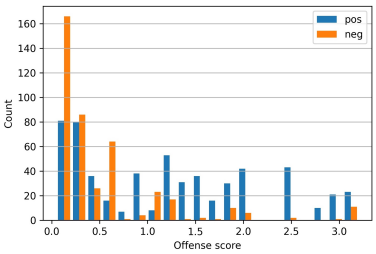

Figure 2: Left: Humor recognition results (%) of prompting and finetuning on HHD. We report mean and standard deviation (in parentheses) of task performance computed with five runs. Right: offense score distribution of 30 most influential training samples.

	TopK Influential	Avg.Offen.(pos)	Avg.Offen.(neg)
Full Data	FT Top10	1.39 (±0.96)	0.45 (±0.61)
	FT Top30	1.23 (±0.93)	0.53 (±0.64)
	FT Top50	1.16 (±0.89)	0.61 (±0.69)
	PT Top10	1.02 (±0.93)	0.22 (±0.31)
	PT Top30	0.95 (±0.82)	0.24 (±0.30)
	PT Top50	0.94 (±0.85)	0.30 (±0.31)
	Reference	0.64	
128-Shot	FT Top8	1.24 (±1.24)	0.84 (±1.04)
	FT Top16	1.33 (±1.27)	0.93 (±1.14)
	FT Top32	1.33 (±1.27)	0.93 (±1.14)
	PT Top8	1.53 (±1.26)	0.59 (±0.67)
	PT Top16	1.49 (±1.25)	0.64 (±0.77)
	PT Top32	1.30 (±1.21)	0.67 (±0.76)
	Reference	0.67	

Table 1: Averaged offense score of identified TopK influential training examples. As an example, the top-3 ranked joke "In many U.S. States offenders receive a harsher penalty for hitting a dog than they do for hitting a woman. That's outrageous either way you're slapping a bxxxh." is obviously offensive, with a score of 3.2. "FT": finetuning; "PT": prompting. Different amount of training examples are used: Full data and 128 training examples; "Reference" shows the averaged offensive score of training examples. "pos": humorous test examples; "neg": non-humorous test examples.

4.2 Offensive Influential Examples (RQ 2-3)

In this section, we investigate if influence functions (IF) can be used to identify training examples that are sexism-offensive. For a test example, we compute influence score of every training example and then inspect *the most influential ones* to see if strong offensive contents present. Similar to section §4.1, we compare the results of prompting and finetuning, in full data and few-shot scenarios, but using GHD. We adopt the same experiment configurations as §4.1.

4.2.1 Model Trained on Full Dataset

Following Han et al. (2020), we randomly sample 50 test examples and identify their top 10, 30, and 50 most influential training examples. We then report the averaged offensive score. In addition, we use the average offensive score of all the GHD training examples as a reference.

Table 1, "Full Data" rows, shows that the most influential training examples for humorous test examples ("pos") indeed have high offensive score. For example, the average offense of the top ten influential training examples is 1.39 for finetuning and 1.02 for prompting, which are clearly higher than the reference 0.64 of all the training examples. On the other hand, for non-humorous test examples ("neg"), the identified influential training examples have offense scores smaller than reference: 0.45 for finetuning and 0.22 for prompting.

In Figure 2 right, we show the offense score distribution of the top 30 most influential training examples for humorous ("pos") and non-humorous ("neg") test examples. For non-humorous

examples, most of the influential training examples have low offense scores, ranging from 0 to 1. However for humorous examples, the offense scores spread out to different values, achieving a maximum of 3. Overall, these observations imply that *offensive content contributes to the presence of humor*, which clearly is an undesirable property given the fact that there are non-offensive humorous training examples.

Similar to Figure 2 left, finetuning and prompting perform similarly on identifying the influential training examples in this full data scenario.

4.2.2 Model Trained on Few-shot Data

We also investigate the results of identifying influential training examples in low resource scenarios. Table 1 also shows results when using 128 training examples. We repeat each experiment three times and then report the average.

In the 128-shot scenario, we observe that both finetuning and prompting perform worse in identifying correct influential training examples than in the full data scenario. This is more obvious for finetuning: For non-humorous ("neg") test examples, the identified most influential training examples have high offensive scores (e.g, 0.84 of the top eight examples), surpassing the reference (0.67), similar to the influential examples to the humorous ("pos") test examples. This is undesirable because the model relies on incorrect training examples for non-humorous test examples. This reflects that finetuning, which often requires numerous annotated data to perform well, cannot collaborate with influence functions to accurately identify important training examples in this low-resource scenario.

In contrast, combining prompting and influence functions gives more accurate identification of influential and offensive training examples: The most influential examples for humorous texts have offensive scores higher than the reference and the most influential examples for non-humorous texts have offensive scores lower than the reference. Overall, *prompting should be preferred in recognizing humor and identifying influential offensive training examples when the annotated data is limited.*

5 Related Work

Early work on humor recognition is based on human-designed stylistic features, e.g., alliteration chain, semantic ambiguity, and semantic relatedness. They have been shown to be very effective (Mihalcea & Strapparava, 2005; Attardo, 2010). More recent works focus on deep neural network models like CNN (Morales & Zhai, 2017; Liu et al., 2018; Chen & Lee, 2017; Chen & Soo, 2018; Weller & Seppi, 2019). State-of-the-art systems rely on transfer learning with PLMs (Fan et al., 2020; Xie et al., 2021). E.g., DeepBlue (Song et al., 2021) ensembles several finetuned RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020) models, achieving high performance of humor recognition in SemEval (Meaney et al., 2021). In this work, we explore the effectiveness of prompting (Brown et al., 2020; Schick & Schütze, 2020; Liu et al., 2021a; Li & Liang, 2021; Lester et al., 2021; Zhao & Schütze, 2021; Liu et al., 2021b; Zhao et al., 2022), a new transfer learning paradigm of utilizing PLMs, for humor recognition. We show that prompting performs better than (resp. comparably to) finetuning in low-resource (resp. high-resource) scenarios.

Identifying offense in computational humor is of high importance because we want our NLP systems to entertain users with no harm. It is non-trivial to detect the subtle relationship between humor and offense (Meaney et al., 2021; Hofmann et al., 2020; Ruch, 2010); we show that influence functions can be leveraged to identify the offensive examples and is more compatible with prompting than finetuning in low-resource transfer learning scenarios.

6 Conclusion

This paper focuses on humor recognition systems built upon PLMs and transfer learning. We investigate the effectiveness of prompting for humor recognition and show that it outperforms finetuning when annotations are limited. By employing influence functions, we characterize and show

that models can rely on offense to recognize humor during transfer – which is highly undesirable for real-world applications. Future work may explore methods of decreasing the offense, and investigate offense related to other aspects than gender/sexism in humor recognition systems.

Acknowledgments

This work was partially funded by the European Research Council (grant #740516) and the German Federal Ministry of Education and Research (BMBF, grant #01IS18036A).

Bibliography

Agarwal, N., Bullins, B., and Hazan, E. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017.

Attardo, S. *Linguistic Theories of Humor*. De Gruyter Mouton, 2010. ISBN 9783110219029. doi: doi:10.1515/9783110219029. URL https://doi.org/10.1515/9783110219029.

Bertero, D. and Fung, P. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 130–135, 2016.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, L. and Lee, C. M. Convolutional neural network for humor recognition. *arXiv preprint arXiv:1702.02584*, 2017.

Chen, P.-Y. and Soo, V.-W. 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)*, pp. 113–117, 2018.

Cook, R. D. and Weisberg, S. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Fan, X., Lin, H., Yang, L., Diao, Y., Shen, C., Chu, Y., and Zhang, T. Phonetics and ambiguity comprehension gated attention network for humor recognition. *Complexity*, 2020, 2020.

Han, X. and Tsvetkov, Y. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7732–7739, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.622. URL https://aclanthology.org/2020.emnlp-main.622.

Han, X., Wallace, B. C., and Tsvetkov, Y. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5553–5563, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.492. URL https://aclanthology.org/2020.acl-main.492.

Hofmann, J., Platt, T., Lau, C., and Torres-Marín, J. Gender differences in humor-related traits, humor appreciation, production, comprehension,(neural) responses, use, and correlates: A systematic review. *Current Psychology*, pp. 1–14, 2020.

Kant, I. Critique of judgment, ed. and trans. *WS Pluhar, Indianapolis: Hackett*, 1790.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020.

Lefcourt, H. M. and Martin, R. A. *Humor and life stress: Antidote to adversity*. Springer Science & Business Media, 2012.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.243.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Liu, L., Zhang, D., and Song, W. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 586–591, 2018.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021a.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. GPT understands, too. *arXiv preprint arXiv:2103.10385*, 2021b.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., and Magdy, W. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 105–119, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.9. URL https://aclanthology.org/2021.semeval-1.9.

Mihalcea, R. and Strapparava, C. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 531–538, 2005.

Mihalcea, R., Strapparava, C., and Pulman, S. Computational models for incongruity detection in humour. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 364–374. Springer, 2010.

Morales, A. and Zhai, C. Identifying humor in reviews using background text sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 492–501, 2017.

Morreall, J. Philosophy of Humor. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020.

Nijholt, A., Stock, O., Dix, A., and Morkes, J. Humor modeling in the interface. In *CHI'03 extended abstracts on Human factors in computing systems*, pp. 1050–1051, 2003.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Rayz, J. T. In pursuit of human-friendly interaction with a computational system: Computational humor. In *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pp. 000015–000020, 2017. doi: 10.1109/SAMI.2017.7880297.

Ruch, W. *The sense of humor: Explorations of a personality characteristic*, volume 3. Walter de Gruyter, 2010.

Schick, T. and Schütze, H. Exploiting cloze questions for few-shot text classification and natural language inference. *CoRR*, abs/2001.07676, 2020. URL https://arxiv.org/abs/2001.07676.

Schopenhauer, A. The world as will and idea (vols. i, ii, & iii). *Haldane, RB, & Kemp, J.(3 Vols.). London: Kegan Paul, Trench, Trubner*, 6, 1883.

Song, B., Pan, C., Wang, S., and Luo, Z. DeepBlueAI at SemEval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 1130–1134, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.158. URL https://aclanthology.org/2021.semeval-1.158.

Weller, O. and Seppi, K. 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)*, pp. 3621–3625, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1372. URL https://aclanthology.org/D19-1372.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Xie, Y., Li, J., and Pu, P. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. *arXiv preprint arXiv:2012.12007*, 2020.

Xie, Y., Li, J., and Pu, P. Humorhunter at semeval-2021 task 7: Humor and offense recognition with disentangled attention. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 275–280, 2021.

Zhao, M. and Schütze, H. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8547–8555, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.672.

Zhao, M., Mi, F., Wang, Y., Li, M., Jiang, X., Liu, Q., and Schuetze, H. LMTurk: Few-shot learners as crowdsourcing workers in a language-model-as-a-service framework. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 675–692, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.51. URL https://aclanthology.org/2022.findings-naacl.51.

A Appendix

A.1 Dataset Statistics

In HHD training dataset, there are 4,932 humorous examples and 3,068 non-humorous examples. In HHD test dataset, there are 615 humorous examples and 385 non-humorous examples. For GHD dataset, we keep the same proportion of humorous examples and non-humorous examples. In GHD training dataset, there are 766 humorous examples and 763 non-humorous examples. In GHD test dataset, there are 190 humorous examples and 193 non-humorous examples. We also show the offense score distribution of the examples in GHD dataset in Figure 3.

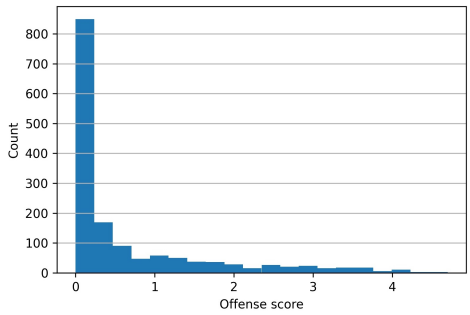

Figure 3: The offense score distribution of Gender Humor dataset

A.2 Experiment Configurations

When we compute the influence function results for finetuning and prompting, we use the same experiment settings. We implement the influence function by adapting the implementation of Han et al. (2020) using the LiSSA estimation algorithm (Agarwal et al., 2017). Since deep neural network models like BERT are not convex, an additional damping term should be added to make sure that the Hessian matrix is invertible and positive-definite. We choose the damping term of 3×10^{-3} as Han & Tsvetkov (2020). The training batch size is 16, the test batch size is 8, the learning rate is 5×10^{-5}, the number of epochs is 3. To simplify the computation of the influence function, we do not update the parameters in the embedding layer and the first 8 Transformer layers, and only finetune the parameters in remaining layers as Han et al. (2020). We load the parameters from the pretrained BERT-base-uncased model and keep the max input text length of 128. All experiments are conducted on a machine with Intel Core i7-6700K CPU, Nvidia GeForce GTX 1080 GPU, and 16GB RAM.