Transfer Learning for Humor Detection by Twin Masked Yellow Muppets

Aseem Arora
Indian Institute of Technology Patna, India
aseem_1911mc02@iitp.ac.in

Gaël Dias
University of Caen Normandie, Caen, France
gael.dias@unicaen.fr

Adam Jatowt
University of Innsbruck, Austria
adam.jatowt@uibk.ac.at

Asif Ekbal
Indian Institute of Technology Patna, India
asif.ekbal@iitp.ac.in

Abstract

Humorous texts can be of different forms such as punchlines, puns, or funny stories. Existing humor classification systems have been dealing with such diverse forms by treating them independently. In this paper, we argue that different forms of humor share a common background either in terms of vocabulary or constructs. As a consequence, it is likely that classification performance can be improved by jointly tackling different humor types. Hence, we design a shared-private multitask architecture following a transfer learning paradigm and perform experiments over four gold standard datasets. Empirical results steadily confirm our hypothesis by demonstrating statistically-significant improvements over baselines and accounting for new state-of-the-art figures for two datasets.

1 Introduction

Humor has been studied in fields such as Psychology (Kline, 1907; Wolff et al., 1934) and Linguistics (Bergen and Binsted, 2003; Attardo, 2017). In Natural Language Processing, the tasks of humor classification (Peyrard et al., 2021; Ziser et al., 2020; Meaney, 2020; Weller and Seppi, 2019) and generation (Yamane et al., 2021; Garimella et al., 2020) have recently gained importance although they have been subject of reflection for some time (Mihalcea and Strapparava, 2005; Ritchie, 2009)[1].

Humor can be expressed in different forms (examples in Table 1). In body-punchlines, the humorous effect is brought by the incongruity or the violation of the expectation formed by the body. In Puns, polysemous words or homophones can be used to cause humor. In short stories, the surprising ending emphasizes the humorous connotation.

Most related works on humor classification have treated the different forms of humor independently. Here, we hypothesize that different forms of humor are closely related, both in terms of vocabulary (e.g. taboo content, community-based humor) and constructs (e.g. surprising effect, incongruity, polysemy). So, processing the different forms of humor in shared settings should help improving classification performance over individual settings.

Joke 1	*[Body] What's the difference between a baby and a car? [Punchline] A car isn't burried in my backyard.*
Joke 2	*[Pun] Why was the musician arrested? He got in treble.*
Joke 3	*[News headline] China minister warns seduction of laws by western nations. [One word substituted] China minister warns seduction of kangaroos by western nations.*
Joke 4	*[Story] A linguistics professor was lecturing his class one day. 'In English', he said, 'A double negative forms a positive. In some languages, though, such as Russian, a double negative is still a negative. However, there is no language wherein a double positive can form a negative.' A loud voice from the back of the room piped up, 'Yeah, right'.*

Table 1: Examples of different forms of humor.

For that purpose, we design a shared-private multitask architecture, where a shared representation layer is learned based on two different tasks (masked language modelling and classification). The frozen shared layer is then combined with a fined-tuned private layer to account for each individual type of humor. Empirical results over Reddit (Weller and Seppi, 2019), Humicroedit (Hossain et al., 2019), Shortjokes (Weller and Seppi, 2019) and Puns (Yang et al., 2015) datasets demonstrate that our method steadily improves over baselines and accounts for new state-of-the-art figures for two datasets.

2 Related work

Initial attempts have been proposed by Mihalcea and Strapparava (2005), where humor-specific stylistic features and content-based features are combined to classify short sentences. Purandare and Litman (2006) compute acoustic-prosodic features, such as pitch and energy, in addition to the linguistic features within spoken conversations.

[1] Some efforts have recently tackled multimodal information (Choube and Soleymani, 2020; Hasan et al., 2021).

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 1–7
November 20–23, 2022. ©2022 Association for Computational Linguistics

Zhang and Liu (2014) tackle humor recognition in tweets based on phonetic, morpho-syntactic, lexico-semantic, pragmatic and affective features. Bertero and Fung (2016) combine hierarchical continuous representations with high-level features (e.g. structural features, antonyms, sentiment) to predict humor of body-punchlines in TV-sitcoms dialogues. Chen and Soo (2018) propose a Convolutional Neural Network (CNN)-based architecture combined with highway networks (Zilly et al., 2017). Weller and Seppi (2019) propose a new task, which consists in recognizing whether a joke is funny or not. For that purpose, they build the Reddit dataset and design a straightforward BERT architecture, which competes with human perception. Further experiments on Puns and Shortjokes, show that contextualized embeddings are strong representations for humour recognition, also upgrading (Chen and Soo, 2018) results. Wang et al. (2020) design a multilingual model based on a pre-trained (Chinese, Russian, Spanish) BERT, that is fine-tuned on inter-sentence relationship and sentence discrepancy prediction for body-punchlines. Similar works are proposed by (Ziser et al., 2020) to recognize humorous questions in product Q&A systems, and (Xie et al., 2021), who formalize uncertainty and surprise for body-punchlines in English.

3 Shared-Private Multitask Architecture

In order to take advantage of the different humor types, we propose a shared-private multitask architecture (Liu et al., 2017). The model depicted in Figure 1 consists of a **frozen shared BERT** (Devlin et al., 2019) layer, which is pre-trained on two different tasks to account for different humor types, and a **private BERT** layer, which is fine-tuned on each dataset independently.

3.1 MLM Pre-trained BERT (+MLM)

Although it is known that BERT representations are able to account for the humorous language (Weller and Seppi, 2019), we propose to fine-tune them by Masked Language Modeling (MLM) (Devlin et al., 2019) over a large dataset that embodies a wide spectrum of different forms of humor (here, ShortJokes). The objective is to improve the original language model and utilize it as the common representation resource for all the classification tasks.

3.2 BERT Shared Layer (+Class)

In order to account for a generalized (aka. shared) representation of humorous utterances, we propose to fine-tune the MLM pre-trained BERT (§3.1) based on a classification task stating whether some text is humorous or not, by taking different humor type samples as input. To account for the widest spectrum of humor forms, a specific dataset is built from Reddit, Humicroedit, Shortjokes and Puns, which is balanced to avoid the predominance of a given humor type (details in §4). Formally, each input sentence is fed to the shared BERT layer and the embedding for the $[CLS]$ token, $h_{CLS} \in \mathrm{R}^d$, is used as sentence embedding. This latter representation is then fed to a classification layer, comprised of a fully connected layer followed by softmax function. Training is performed using cross-entropy.

3.3 Shared-private Model

The shared-private architecture combines a BERT shared layer (§3.2) and a private BERT layer (§3.1), and is trained for the task of humor classification for each dataset independently. The private layer is fine-tuned for the specific task at hand, while the shared BERT is kept frozen to preserve the already learned information of different humor types. As such, classification is decided based on the general information about humor and the specific codes of a given humor type. Formally, each input sentence is fed to both shared and private BERT layers to obtain the corresponding sentence embeddings, i.e. $h^s_{CLS} \in R^d$ and $h^p_{CLS} \in R^d$. The concatenation of these representations $[h^s_{CLS}, h^p_{CLS}]$ is then input to a classification layer, comprised of a fully connected layer followed by softmax function. Training is performed using cross-entropy.

4 Datasets

Literature datasets. *Puns* (Yang et al., 2015) contains humorous quotes in the form of puns. In particular, negative instances have been extracted to minimize domain differences, i.e. by ensuring similar word dictionary and text length. We use the splits provided by Weller and Seppi (2019) for this dataset. *Reddit* (Weller and Seppi, 2019) contains body-punchline type jokes collected from *reddit.com* along with the number of upvotes on each joke. Punchlines are then labeled as humorous or non-humorous based on a cut-off value for upvotes. *Humicroedit* (Hossain et al., 2019)

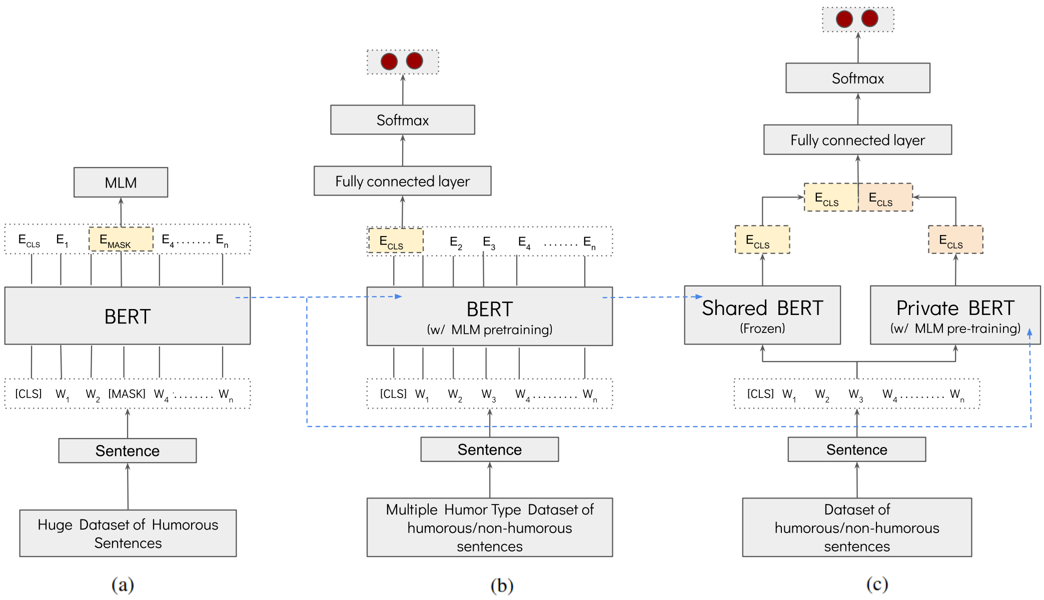

Figure 1: Overall architecture: (a) Masked language modeling; (b) Shared layer; (c) Shared-private model. Dashed arrows indicate from which model the weights of the BERT modules are initialized.

Puns						Reddit						Humicroedit						Shortjokes						Shared			
Train		Validation		Test		Train		Validation		Test		Train		Validation		Test		Train		Validation		Test		Train		Validation	
Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
1,809	1,810	152	149	155	147	9,719	9,719	304	304	304	304	9,652	9,652	2,419	2,419	3,024	3,024	171,831	171,031	10,849	10,720	10,889	10,680	31,723	31,638	4,752	4,795

Table 2: Training, validation and test splits by number of positive and negative instances for five datasets.

consists of news headlines with corresponding edits, where one word is substituted to cause incongruity. Here, the original news headlines are taken as non-humorous, while the edited headlines are taken as humorous. *ShortJokes*, first found on Kaggle[2] and then replicated by Weller and Seppi (2019), gathers puns, body-punchlines and short text jokes, ranging from 10 to 200 characters. Details of the datasets are given in Table 2.

Shared dataset. A dataset of humorous and non-humorous samples is specifically built to train the shared BERT layer (§3.2). We include all training samples from Puns, Reddit, and Humicroedit, while for Shortjokes, only 21,000 training samples are included to guarantee balance of different types of humors. Similarly, the validation set contains a total of 9,547 samples built from all validation samples of Puns, Reddit, and Humicroedit, while for ShortJokes, only 3,800 validation samples are included. This dataset is only used for pre-training and as such does not include a test split.

5 Experimental setups

All models have been implemented using PyTorch (Paszke et al., 2019) and Hugginface (Wolf et al., 2019) libraries. All models are based on BERT base[3]. The embedding size d for h_{CLS} is 768. For training BERT with the MLM objective, each word is masked with a probability of 0.15, and we use a batch size of 6 and a learning rate of 2×10^{-5}. For training on the humor classification task, for both the shared BERT and shared-private architecture, we use a batch size of 16 and a learning rate of 2×10^{-5}. We use the Adam optimizer with a default weight decay of 0.01. For each dataset, the model is trained for 4 epochs. The best model is saved based on the development set accuracy results. Code and datasets are available at `https://github.com/aseemarora1995/humor-detection`.

6 Results Analysis

Experimental results are illustrated in Table 3. We report mean accuracies and F1 scores over 5 runs, along with standard deviation values. Our proposed model *BERT Shared&Private (+MLM +Class)*

[2]https://www.kaggle.com/abhinavmoudgil95/short-jokes

[3]https://huggingface.co/bert-base-uncased

	Puns		Reddit		Humicroedit		Shortjokes	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
BERT	90.71 ± 1.07	90.70 ± 1.07	70.43 ± 2.00	69.43 ± 2.64	80.18 ± $\underline{0.23}$	80.10 ± $\underline{0.23}$	98.55 ± 0.08	98.55 ± 0.08
BERT (+MLM)	90.88 ± 0.48	90.88 ± 0.47	70.96 ± 1.76	70.13 ± 2.22	80.62 ± 0.40	80.62 ± 0.40	98.58 ± 0.05	98.58 ± 0.05
BERT Shared (-MLM +Class)	88.08 ± 1.12	88.06 ± 1.13	66.15 ± 0.65	65.47 ± 0.73	78.84 ± 0.65	78.79 ± 0.71	95.48 ± 0.46	95.48 ± 0.46
BERT Shared (+MLM +Class)	88.94 ± 0.95	88.93 ± 0.95	66.37 ± 0.65	65.71 ± 0.81	79.32 ± 0.60	79.30 ± 0.58	95.88 ± 0.38	95.88 ± 0.38
BERT Shared&Private (-MLM -Class)	91.19 ± 0.55	91.19 ± 0.55	68.95 ± 2.53	67.26 ± 3.60	80.61 ± 0.47	80.55 ± 0.48	98.62 ± 0.06	98.62 ± 0.06
BERT Shared&Private (-MLM +Class)	91.13 ± 1.51	91.12 ± 1.51	68.75 ± 2.17	67.45 ± 2.92	80.17 ± 0.33	80.10 ± 0.36	98.57 ± 0.06	98.57 ± 0.06
BERT Shared&Private (+MLM -Class)	91.72 ± 0.95	91.71 ± 0.94	69.41 ± 1.29	68.34 ± 1.57	80.49 ± 0.76	80.41 ± 0.87	98.56 ± 0.05	98.56 ± 0.05
BERT Shared&Private (+MLM +Class)	**93.25**† ± $\underline{0.37}$	**93.25**† ± $\underline{0.37}$	**73.55**† ± $\underline{0.41}$	**73.40**† ± $\underline{0.39}$	**81.36**† ± 0.31	**81.35**† ± 0.30	**98.77**† ± $\underline{0.03}$	**98.77**† ± $\underline{0.03}$

Table 3: Accuracy and F1 scores averaged over 5 runs together with standard deviation values (±) for four datasets. † means statistical difference with BERT base in terms of t-test (two-tailed p-value < 0.05). **Bold** values mean maximum Accuracy and F1 score, and underline stands for the smallest values of standard deviation.

	Puns		Reddit		Humicroedit		Shortjokes	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
BERT Shared (-MLM +Class)	88.08 ± 1.12	88.06 ± 1.13	66.15 ± 0.65	65.47 ± 0.73	78.84 ± 0.65	78.79 ± 0.71	95.48 ± 0.46	95.48 ± 0.46
BERT Shared (-MLM +Class Complete)	85.16 ± 1.22	85.07 ± 1.30	64.57 ± 2.31	63.97 ± 2.41	78.76 ± 0.69	78.70 ± 0.73	98.47 ± 0.05	98.47 ± 0.05
BERT Shared (+MLM +Class)	88.94 ± 0.95	88.93 ± 0.95	66.37 ± 0.65	65.71 ± 0.81	79.32 ± 0.60	79.30 ± 0.58	95.88 ± 0.38	95.88 ± 0.38
BERT Shared (+MLM +Class Complete)	84.24 ± 3.26	84.05 ± 3.41	64.31 ± 2.49	63.04 ± 3.48	78.71 ± 0.63	78.67 ± 0.63	98.48 ± 0.07	98.48 ± 0.07
BERT Shared&Private (+MLM +Class)	93.25 ± 0.37	93.25 ± 0.37	73.55 ± 0.41	73.40 ± 0.39	81.36 ± 0.31	81.35 ± 0.30	98.77 ± 0.03	98.77 ± 0.03
BERT Shared&Private (+MLM +Class Complete)	92.52 ± 0.56	92.51 ± 0.56	71.48 ± 2.13	70.59 ± 3.00	80.38 ± 0.57	80.34 ± 0.59	98.60 ± 0.01	98.60 ± 0.01

Table 4: Accuracy and F1 score averaged over 5 runs together with standard deviation values for four datasets. Complete is appended when the BERT Shared is trained on the complete dataset containg all instances of Puns, Reddit, ShortJokes and Humicroedit.

achieves best mean accuracies and F1 scores for all datasets over all BERT-like variations. This architecture also achieves new state-of-the-art performances for two datasets, as revealed in Table 5. Moreover, our methodology shows the least variations in results as evidenced by minimum standard deviation values for three out of four datasets, thus indicating it is the most robust model.

In Table 3, we present different variations of our model to better assess the contribution of each of its parts. In particular, *BERT (+MLM)*, which pre-trains BERT with the MLM objective and fine-tunes it for each dataset, shows steady improvements in performance and robustness over BERT base models. The BERT Shared variants, which are pre-trained for classification over the shared dataset (§4), evidence transfer results as they are not fine-tuned for each datasets, but instead are kept frozen without private layer. Results show that fine-tuning is necessary. Besides, the introduction of the MLM objective clearly boosts results in all settings. The Shared-private architectures all contain a shared and a private layer, that can be initialized in different ways. In our experiments, we tested all combinations, where both shared and private layers are initialized with the exact same configuration. Results clearly show that the combination of the MLM objective and the classification pre-training ensures superior performance and robustness.

As explained in the §3.2, the shared BERT is pre-trained for humor classification using a balanced shared dataset, To explain the importance of using a balanced dataset, we perform experiments by pre-training the shared BERT on a complete training sets combined from all the four datasets, without taking care of balance between humor types. Results are shown in the Table 4. The *BERT Shared (-MLM +Class)* and *BERT Shared (+MLM +Class)* achieve significantly better results for Puns, Reddit, and Humicroedit datasets as compared to *BERT Shared (-MLM +Class Complete)* and *BERT Shared (-MLM +Class Complete)*, respectively. While for the ShortJokes dataset, the opposite is true. This is because the complete shared dataset contains almost 15 times more samples of ShortJokes as compared to those in the balanced version. This makes the shared BERT biased towards the ShortJokes dataset and the performance for the remaining datasets is affected.

In Table 5, we present results from the literature, for the all datasets used in our experiments. Our methodology clearly competes with the current state-of-the-art strategies, as it achieves new standards for Reddit and ShortJokes datasets. Nevertheless, Fan et al. (2020) achieve slightly higher performance over Puns. Note that they use other splits than (Weller and Seppi, 2019) and as such results are not directly comparable to all other configurations. But the most important is that they make use of WordNet (Miller, 1995) turning their model resource-dependent. Similarly, Xie et al. (2021) report better results for Humicroedit. How-

	Puns		Reddit		Humicroedit		Shortjokes	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
BERT Large (avg/max)	$91.46 \pm 1.20/92.72$	$91.45 \pm 1.20/92.71$	$68.67 \pm 1.27/69.67$	$67.51 \pm 1.57/68.73$	$82.22 \pm 0.53/82.97$	$82.20 \pm 0.53/82.96$	$98.69 \pm 0.06/98.76$	$98.69 \pm 0.06/98.76$
Weller and Seppi (2019)	93.00	93.10	72.40	-	-	-	98.60	98.60
Fan et al. (2020)	(**93.88**)	(**93.93**)	-	-	-	-	-	-
Xie et al. (2021)	-	-	-	-	(**83.65**)	(**83.63**)	-	-
BERT Shared&Private (avg/max)	$93.25^{\dagger}/93.71$	$93.25^{\dagger}/93.71$	$\mathbf{73.55^{\dagger}/73.85}$	$\mathbf{73.40^{\dagger}/73.69}$	81.36/81.81	81.35/81.80	$\mathbf{98.77^{\dagger}/98.78}$	$\mathbf{98.77^{\dagger}/98.78}$

Table 5: SOTA Accuracy and F1 scores. Results for BERT Large have been computed over 5 runs. † means statistical difference with BERT Large in terms of t-test (two-tailed p-value < 0.05). Results in "()" are discussed in §6 as they are not directly comparable. "-" means the lack of results reported in the literature.

ever, they apply cleaning over the original dataset, and only keep 3,341 examples in total, i.e., 9 times less the size of our dataset. As such, results cannot directly be compared to ours. Moreover, they propose a methodology specific to body-punchlines, which can not be transposed to other forms of humor. Weller and Seppi (2019) use the BERT Large model (unlike BERT base in our case). As they do not report mean results and standard deviation values for all datasets, we replicated their experiments, reported as *BERT Large*. Our strategy evidences gains over BERT Large for three out of four datasets, failing to improve only on Humicroedit. However, it is worth noticing that our model is two-third the size of BERT Large with about 220M parameters as compared to 340M parameters for BERT Large. Moreover, our strategy is less sensitive to variations due to its multitask architecture.

7 Error Analysis

In Table 6, we provide some qualitative results. In particular, our model correctly predicts examples 1, 2, and 3 as humorous, while BERT fails to predict the humorous connotation. These examples clearly specify a certain type of vocabulary, which is common to most forms of jokes. For instance, *dick* is a sexual expletive, *sick* could imply weirdness or creepiness, and *billionaires* is directly linked to money, a classic topic for jokes. As all these topics commonly occur in humor, we can hypothesize that the shared representations correctly capture the semantics of this specific vocabulary.

But some humor contents still remain unsolved by both models. For example, humorous quotes 4, 5, 6, and 7 are odd classified by both models. Example 4 uses the polysemous word *bank* to provoke the funny connotation, but such phenomenon is difficult to be handled by contextualized representations, as the humorous trick is based on the fact that two different representations coexist and form incongruity. Example 5 is understandable only with additional common sense knowledge about *paranoia*, which is unlikely to be dealt with by current

No.	Dataset	Joke	BERT	Ours
1	Reddit	*my boss hates it when i shorten his name to* ***dick*** *mostly because his name is steve*	✗	✓
2	ShortJokes	*when you go to the hospital and there is music playing these are some* ***sick*** *beats*	✗	✓
3	ShortJokes	*no amazon i do not want to sort stuff by price high to low. who are the* ***billionaires*** *who would even make that an option*	✗	✓
4	Puns	*if you have to pay to go to the river we'd better stop at the* ***bank***	✗	✗
5	Reddit	*i went to the library and asked the librarian if she knew where books on* ***paranoia*** *were. she said ¨they're right behind you.*	✗	✗
6	ShortJokes	*politicians are the only people in the world who create problems and then campaign against them*	✗	✗
7	Humicroedit	*[original non-joke]* ***official*** *who works closely with jared kushner, ivanka trump to leave white house.*	✓	✓
		[correct prediction] ***monkey*** *who works closely with jared kushner, ivanka trump to leave white house.*	✓	✓
		[incorrect prediction] ***assassin*** *who works closely with jared kushner, ivanka trump to leave white house.*	✗	✗

Table 6: Error analysis between BERT and our method, and some examples still unsolved.

language models. Example 6 requires some form of reasoning to understand the humorous connotation, which is also unlikely to be solved by language models. Finally, example 7 clearly evidences the limitations of current language models. While the slight variation using the word *monkey* is correctly understood by both BERT and our strategy, the more subtle word replacement with *assassin* is incorrectly handled. Indeed, while the word *monkey* is usually associated to humorous content, this is not so true for *assassin*.

8 Conclusion

Humor is an important part of human communication. In this paper, we hypothesize that different forms of humor share a common background, and as a consequence, additional usage of one form can help in better understanding other forms in humor classification. So, we propose a shared-private multitask architecture that achieves new state-of-the-art performances for two out of four datasets, and evidences strong robustness. This latter issue is crucial for humorous text generation (Jin et al., 2020). Nevertheless, we observe that current models still have limited capacity to understand such complicated forms of humor where polysemy, external knowledge, context, and reasoning are important.

References

Salvatore Attardo. 2017. Humor in language. In *Oxford Research Encyclopedia of Linguistics*.

Benjamin Bergen and Kim Binsted. 2003. The cognitive linguistics of scalar humor. *Language, culture, and mind*, pages 79–92.

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 130–135.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *2018 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 113–117.

Akshat Choube and Mohammad Soleymani. 2020. *Punchline Detection Using Context-Aware Hierarchical Multimodal Fusion*, page 675–679.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Xiaochao Fan, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, and Yanbo Zou. 2020. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111.

Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. Judge me by my size (noun), do you? YodaLib: A demographic-aware humor generation framework. In *28th International Conference on Computational Linguistics (COLING)*, pages 2814–2825.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 12972–12980.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut hair": Dataset and analysis of creative text editing for humorous headlines. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 133–142.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5082–5093.

Linus W Kline. 1907. The psychology of humor. *The American Journal of Psychology*, pages 421–441.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–10.

J. A. Meaney. 2020. Crossing the line: Where do demographic variables fit into humor detection? In *58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (ACL)*, pages 176–181.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 531–538.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Maxime Peyrard, Beatriz Borges, Kristina Gligoric, and Robert West. 2021. Laughing heads: Can transformers detect what makes a sentence funny? In *30th International Joint Conference on Artificial (IJCAI)*, pages 3899–3905.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*. In *2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 208–215.

Graeme Ritchie. 2009. Can computers create humor? *AI Magazine*, 30(3):71–81.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *22nd Annual Conference of the European Association for Machine Translation (EAMT)*, pages 53–59.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers:

State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Harold A Wolff, Carl E Smith, and Henry A Murray. 1934. The psychology of humor. *The Journal of Abnormal and Social Psychology*, 28(4):341.

Yubo Xie, Junze Li, and Pearl Pu. 2021. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 33–39.

Hiroaki Yamane, Yusuke Mori, and Tatsuya Harada. 2021. Humor meets morality: Joke generation based on moral judgement. *Information Processing & Management (IPM)*, 58(3).

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2367–2376.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *23rd ACM International Conference on Information and Knowledge Management (CIKM)*, page 889–898. Association for Computing Machinery.

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. 2017. Recurrent highway networks. In *34th International Conference on Machine Learning (ICML)*, volume 70, pages 4189–4198.

Yftah Ziser, Elad Kravi, and David Carmel. 2020. *Humor Detection in Product Question Answering Systems*, page 519–528.