Meta-tuning Language Models to Answer Prompts Better

Ruiqi Zhong **Kristy Lee*** **Zheng Zhang*** **Dan Klein**
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu

Abstract

Large pretrained language models like GPT-3 have acquired a surprising ability to perform zero-shot classification (ZSC). For example, to classify review sentiments, we can "prompt" the language model with the review and the question "*Is the review positive?*" as the context, and ask it to predict whether the next word is "*Yes*" or "*No*". However, these models are not specialized for answering these prompts. To address this weakness, we propose meta-tuning, which trains the model to specialize in answering prompts but still generalize to unseen tasks. To create the training data, we aggregated 43 existing datasets, annotated 441 label descriptions in total, and unified them into the above question answering (QA) format. After meta-tuning, our model outperforms a same-sized QA model for most labels on unseen tasks, and we forecast that the performance would improve for even larger models. Therefore, measuring ZSC performance on non-specialized language models might underestimate their true capability, and community-wide efforts on aggregating datasets and unifying their formats can help build models that understand prompts better.

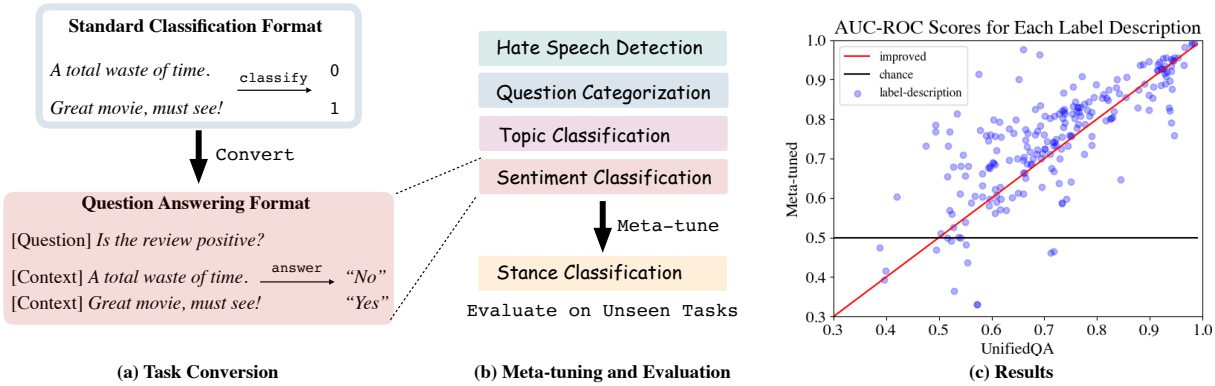
1 Introduction

The task of zero-shot classification (ZSC) is to classify textual inputs using label information or description alone, without seeing any training example of that label (Yin et al., 2019). Large language models - whose only training objective is to predict the next word given the context - have acquired a surprising ability to perform ZSC (Radford et al., 2019; Brown et al., 2020). For example, to classify whether the sentence "*This movie is amazing!*" is positive, we can *prompt* the language model with the context "*Review: This movie is amazing! Positive Review?* ___ ", and check whether the next word is more likely to be "*Yes*" or "*No*" (Zhao et al., 2021). To better convert ZSC into a language modelling (LM) task that an LM model is likely to answer correctly, a lot of recent works focus on finding better prompts (Shin et al., 2020; Schick and Schütze, 2020b,a; Gao et al., 2020).

Nevertheless, large language models' ability to answer prompts is a by-product, but not the main specialization, of the LM training objective. For example, to classify review sentiment, we can ask the model to answer "*Yes/No*" to the question "*Is the review positive*?" (Figure 1 (a)). We find that UnifiedQA (Khashabi et al., 2020), a model trained to answer general questions and initialized with T5 (770 parameters) (Raffel et al., 2019), can achieve 0-shot accuracy 92% on SST-2 (Socher et al., 2013), while GPT3 (175 B parameters) achieves 80% accuracy (Zhao et al., 2021). Specialization in QA makes a 200x smaller model answer prompts better.

Our work takes a step further and directly trains the model to specialize in ZSC. This requires us to 1) unify different classification tasks into the same format, and 2) gather a dataset of classification datasets and label descriptions for training (Section 2). Since our method **learn**s 0-shot **learn**ing by fine-**tuning** on a dataset of datasets, we name our approach **meta-tuning**. Our method lies in between the two extremes of 1) prompting *general*-purpose language models out of the box to perform ZSC, and 2) fine-tuning them on *specific* tasks: we make the model specialized in ZSC, but still able to generalize to new unseen tasks.

We focus on binary classification tasks and unify them into a "*Yes*"/"*No*" QA format (Clark et al., 2019), where the input is provided as the context and the label information is provided in the question (Figure 1 (a)). Using this format, we gathered a diverse set of classification datasets from 43 different sources listed on Kaggle, SemEval, HuggingFace, and other papers. These tasks range from hate speech detection, question



(a) Task Conversion **(b) Meta-tuning and Evaluation** **(c) Results**

Figure 1: **(a)** We convert the datapoint format from classification to question answering. We manually annotated label descriptions (questions) ourselves. **(b)** We finetune the UnifiedQA (Khashabi et al., 2020) model (with 770 M parameters) on a diverse set of tasks, and evaluate its 0-shot classification (ZSC) performance on an unseen task. **(c)** For each label description (question) we evaluate the AUC-ROC score for the "Yes" answer, and each dot represents a label description. The x-value is the ZSC performance of UnifiedQA; the y-value is the performance of our Meta-tuned model. In most cases, the y-value improves over the x-value (above the red line) and is better than random guesses (above the black line) by a robust margin.

categorization, sentiment classification to stance classification, etc, and the genre ranges from textbooks, social media, to academic papers, etc. In total, these datasets contain 204 unique labels, and we manually annotated 441 label descriptions (Figure 2).

To evaluate ZSC, we need to define what counts as a task that the model has not seen during training time. While prior work considers different notions of "unseen" by disallowing the same label or the same dataset to appear during training, our work defines "unseen" more harshly by disallowing the training and the evaluation datasets that are too similar. For example, we consider AG News topic classification dataset (Zhang et al., 2015) and the topic classification dataset from Yin et al. (2019) to be similar, even though their sources and label spaces are different.

We compare the ZSC performance of our meta-tuned model against the UnifiedQA model by calculating the AUC-ROC score for each label description. Figure 1 (c) shows the results: for most label descriptions, our meta-tuned model is better and beats the 0.5 random baseline by a robust margin. Additionally, larger models perform better, and hence we forecast that meta-tuning would also work for even larger models.

We list several implications of our work in Section 5 and 6. First, since GPT-3 is not specialized in answering prompts, current measurements on its intelligence (Hendrycks et al., 2021a,b) might significantly underestimate its true capability. Second, our work is still far from exhausting all NLP classification tasks, and the community might benefit from a shared effort of aggregating datasets and unifying their formats. At last, the meta-tuning approach might incentivize providers of language model inference APIs to collect prompts from users, potentially leading to larger security, privacy and fairness concerns.

2 Data

We gather a wide range of classification datasets and unify them into a "Yes"/"No" question answering format for binary classification. Then we group similar datasets together to decide what counts as unseen tasks during evaluation.

Gathering Classification Datasets We collect classification datasets from Kaggle[1], Huggingface (Wolf et al., 2020), SemEval[2], and other papers. We looked through these sources and only considered English classification datasets. We also skipped the tasks that we felt were already better represented by other data sets in our collection. Then we manually examined a few examples in each remaining dataset to make sure the dataset seemed plausibly clean.

The goals of these classification dataset include, but are not limited to sentiment classification (IMDB Reviews, Maas et al. (2011a)), topic classification (AG News, Zhang et al. (2015)),

[1] https://www.kaggle.com
[2] https://semeval.github.io

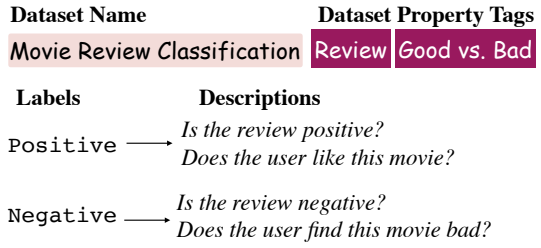


Figure 2: For each dataset, we annotate 2-3 descriptions for each label in the form of a question, and associate it with a set of property tags. The question answering format can be seen in Figure 1 (a).

grammaticality judgement (CoLA, Warstadt et al. (2018)), paraphrase detection[3], detecting definitions (SemEval 2020 Task 6, Spala et al. (2019)), etc. The genre includes academic papers, reviews, tweets, posts, messages, articles, and text books. See the comprehensive list in the Appendix A.

However, some of them are noisy and not peer reviewed, or contain tasks that are too complicated (e.g. Multi-NLI (Williams et al., 2018)) for ZSC. To make our evaluation more informative, we only include them for training but not testing. We make these decisions before running our experiments in Section 4 to prevent selection bias.

Unifying the Dataset Format We convert each classification dataset into a "Yes"/"No" question answering format and provide label information in the question. For each label, we annotate 1-3 questions. If the label is null (for example, a text that does not express a particular emotion in an emotion classification dataset), we skip this label. Three of the authors manually annotated 441 questions for 204 unique labels, and each question is proofread by another author. See Figure 2 for a concrete example, and Figure 3 for a representative subset of label descriptions across different datasets.

Additionally, some datasets contain thousands of labels (Chalkidis et al., 2019; Allaway and McKeown, 2020). In this case we use templates to automatically synthesize label descriptions and exclude them for evaluation.

Grouping Similar Tasks Our goal is to test our models' ability to generalize to tasks that are different enough from the training tasks. Therefore, at test time, we need to exclude not only the same dataset that appeared in the meta-tuning phase, but also ones that are too similar.

[3] https://www.kaggle.com/c/quora-question-pairs

Are these two questions asking for the same thing?
Does the tweet contain irony?
Is this news about world events?
Does the text contain a definition?
Is the tweet an offensive tweet?
Is the text objective?
Does the question ask for a numerical answer?
Is the tweet against environmentalist initiatives?
Is this abstract about Physics?
Does the tweet express anger?
Does the user dislike this movie?
Is the sentence ungrammatical?
Is this text expressing a need for evacuation?
Is this text about Society and Culture?
Is this a spam?
Does this text express no emotion?

Figure 3: Some example manually annotated label descriptions (questions). Three of the authors manually wrote 441 questions in total, and each of them is proofread by at least another author.

This poses a challenge: whether two datasets perform the same task involves subjective opinion, and there is no universally agreed definition. On one extreme, most datasets can be considered dissimilar tasks, since they have different label spaces and input distributions; on the other extreme, all datasets can be considered the same task, since they can all be unified into the question answering task.

To tackle this challenge, we create a set of tags, each describing a dataset property. The set of tags includes *domain classification*, *article*, *emotion*, *social-media*, etc, and the full set of them can be seen in Appendix B. Then we define the task of two datasets to be the same if they are associated with the same set of tags, and prohibit the model to learn from one and test on the other. For example, our work considers AG News topic classification (Zhang et al., 2015) and the topic classification dataset from Yin et al. (2019) to be similar since they both classify topics for articles, even though their sources and label spaces are different. Some example dataset groups can be seen in Figure 4.

Nevertheless, our procedure is not bullet-proof and one can argue that our notion of unseen tasks, though harsher than prior works (Yin et al.,

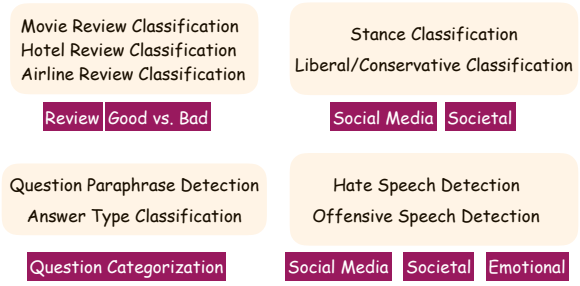


Figure 4: Example dataset groups based on tags. We never train and test on datasets from the same group, e.g. train on hotel review but test on movie review.

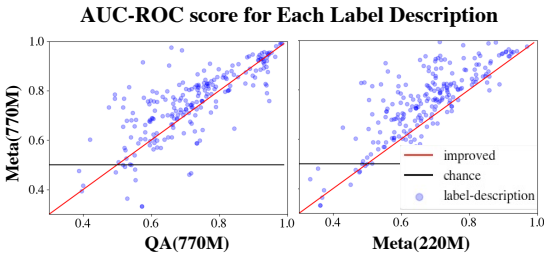


Figure 5: Each dot represents a label description, its x-value represents the AUC-ROC score of one model, and its y-value represents that of the other model. The left figure compares QA (x) vs. Meta-tuning (y) with the same model size (770M), and the right figure compares Meta-tuning on different model sizes: 220M (x) vs. 770M (y). Since most dot is above the red line $y = x$ for both the left and the right figure, larger meta-tuned models are better.

2019; Pushp and Srivastava, 2017), is still lenient. Therefore, we additionally provide concrete examples of how we define unseen tasks, and give the readers the freedom to interpret the results. For each dataset we are evaluating, in Appendix D we manually identify and list the most relevant dataset that is allowed during training. For example, the most relevant task to the IMDB review sentiment classification dataset is the emotion classification dataset from Yin et al. (2019), which classifies the input text into 9 emotions, such as "*joy*", "*surprise*", "*guilt*", etc. We consider the emotion classification dataset to be relevant, since sentiment classification often involves identifying emotions, but one can also reasonably argue that they are different tasks, since their input and label spaces are different, and sad emotion can be caused by a great tragedy, or a trash movie that wastes the users' time.

Our dataset collection, unseen splits, and annotated label descriptions will be released if the paper is accepted.

3 Model

Notation We denote a model with its "{fine-tuning task}({model size})". For example, a UnifiedQA model initialized with the 770 Million parameter T5 model will be denoted as "QA(770M)". If it is meta-tuned, we denote it as "Meta(770M)".

Architecture We format the context and the question in the same way as UnifiedQA (Khashabi et al., 2020), which concatenates the context with the question and a separation token in between. Then we encode the concatenated input with the T5 architecture and produce the answer score by normalizing the "*Yes*"/"*No*" probability of the first decoded token.

Meta-tuning We create a training distribution that balances between datasets, label descriptions, and "*Yes*"/"*No*" answers. To create the next training datapoint for meta-tuning, we select a dataset from the training split uniformly at random (u.a.r.); then we select a label description (question) u.a.r. and with 50% probability select a textual input with the answer "*Yes*"/"*No*". To prevent over-fitting, we do not train on any combination of label description and textual input twice. We use batch size 32 and meta-tune the model for 5000 steps. To evaluate ZSC performance for each dataset, we meta-tune the model on all the other datasets that perform a different task.

4 Results

We evaluate a model by computing the AUC-ROC score for each label descriptions (Figure 2), or calculating its performance on the three benchmark datasets provided by Yin et al. (2019). Both evaluation strategies lead to the same two conclusions:

- Meta(770M) outperforms QA(770M). Further specialization in ZSC is effective.
- Meta(770M) outperforms Meta(220M). Bigger models are better.

4.1 Description-wise AUC-ROC

For each label description (question), we calculate the AUC-ROC score by treating the "Yes" answer as the positive class. We do not evaluate F-score or accuracy, since they are very sensitive to the decision cutoff threshold, and usually additional calibration is needed Zhao et al. (2021). We use a

Model	emotion	situation	topic
Prior(335M)	25.2	38.0	52.1
QA(770M)	21.7	21.5	46.3
Meta(220M)	26.2	33.8	51.4
Meta(770M)	**27.0**	**43.2**	**58.9**

Table 1: Evaluating on the 3 datasets from Yin et al. (2019): emotion, situation, and topic. The first two uses weighted F-score and the last uses accuracy as the evaluation metrics. “Prior" means the best performing system from Yin et al. (2019) for each dataset. We observe that meta-tuned larger models are better.

scatter plot to compare two models, where each dot represents a label description, its x-value represents the AUC-ROC score of one model, and its y-value represents that of the other model.

We generally do not try to aggregate a model's performances on different datasets into a single number and compare this number only. The reasons are as follows: 1) it is unclear a priori how to aggregate AUC-ROC scores; for example, even if dataset A and dataset B are equally important, it is not clear whether scores (0.6, 0.9) or (0.75, 0.8) is better. 2) aggregate statistics remove lower-level information on the description-wise difference between AUC-ROC scores of the two models. We only resort to aggregate statistics when the conclusion is unclear based on visualizations.

The results are in Figure 5. Since most dots are above the red line, we conclude that, for most labels, larger/meta-tuned models are better.

4.2 Benchmarking with Yin et al. (2019)

We compare our models on the zero shot classification datasets proposed by Yin et al. (2019), which consist of 3 multi-label classification datasets: emotion, situation, and topic classification. We use the exact same evaluation metrics and label resolution strategy when the model answers “Yes" for multiple labels.

The results can be seen in Table 1. For reference, we also list the best performing system for each dataset from Yin et al. (2019) (Prior), which uses a natural language inference system based on RoBERTA-large (355M parameters, Liu et al. (2020)), ensembles several different models, and uses other auxillary training objectives such as categorizing Wikipedia articles. Similar to Section 4.1, we find that larger/meta-tuned model are better for all 3 datasets. Additionally, after controlling for the parameter count, our meta-tuned model (220 M) is comparable to the prior best system (335M).

4.3 Robustness Checks

We examine a series of additional results to ensure that our conclusions are robust.

We visualized with the same opacity for each label description in Figure 5, which implicitly assumes that each description is equally important for evaluation. To ensure that the improvement is not caused by the improvement of a small number of labels that have more annotated descriptions, or by the improvement on a dataset that has more distinct labels, Appendix E performs the same visualization with opacities inverse to the number of descriptions for each label/dataset.

To provide additional supporting evidence for our forecast that larger models are better, Appendix C.1 compares Meta(60M) with Meta(220M) and finds that the later is much better.

Another concern is that our models are initialized with T5 (Raffel et al., 2019), which is trained on the open web and might have seen the datasets we gathered. We address this in Appendix C.2 by showing that larger models are also better with BERT Devlin et al. (2019), which is trained on Wikipedia and Book Corpus (Zhu et al., 2015).

5 Discussion

Meta-tuning as a Probe There is a growing interest in measuring the intelligence (Hendrycks et al., 2021a,b) or the few shot learning ability (Brown et al., 2020) of large language models like GPT-3. However, since these models are not specialized to answer those prompts, we suspect that its true capability is much higher than reported. We can potentially address this issue by first using meta-tuning as a probe to make them specialized in answering prompts before measuring their performance.

Nevertheless, to make this methodology rigorous, interpreting and controlling the strength of the probes will be an important future direction(Hewitt and Liang, 2019). For example, if the meta-tuning procedure is so strong that it enables a non-pretrained model to answer prompts correctly, or the training set contains a prompt that is too similar to the prompt to be tested, our measurement will be meaningless.

Aggregating and Unifying Datasets The main bottle-neck of our research is to manually gather a

wide range of datasets and unify their format. The difficulties are: 1) we need to brainstorm and review the NLP literature extensively to decide what new tasks to look for; 2) different datasets encode their data in different formats, and we need to write programs manually for each of them to convert to the desired format; 3) it is hard to tell the quality of a dataset purely by its provenance, and sometimes we need to examine the dataset manually. Our research would have been much easier if there is a platform that gathers all the NLP classification datasets, unifies them into a single format, and provides information about their individual quality.

Annotating Prompts Three of our authors annotated the label descriptions. Since they are all Computer Science major students who understand machine learning and natural language processing, they might not be representative of the final user population of this zero shot classification application. Annotating prompts that match the target user distribution will be an important research direction.

Optimizing Prompts Our work is complementary to recent works that optimize the prompts to achieve better accuracy. Even if our meta-tuned model is specialized in answering prompts, it might still react very differently towards different prompts. For example, in the stance classification dataset (Barbieri et al., 2020), we annotated two label descriptions (prompts) for the same label: “Does this post support atheism?" and “Is the post against having religious beliefs?". They have similar meanings, but the former has much lower accuracy than the later. We conjecture that this is because the model cannot ground abstract concepts like “atheism".

Other Extensions We can use prompting to perform a more diverse set of tasks beyond 0-shot binary classification (Brown et al., 2020), and the idea of meta-tuning is similarly applicable. To extend to multi-label classification, we need to develop a procedure to resolve the labels when the model predicts positive for more than 1 label. To extend to few-shot learning, we need to increase the context length to fit several training examples into the input, which requires a larger context window and hence more computational resources. To extend to other sequence generation tasks, we need to collect a wide range of sequence generation tasks to meta-tune the model (e.g. machine translation, summarization, free-form question answering, etc).

6 Ethics

Data and Incentives In the existing prompting framework, end users send the natural language descriptions and a few training examples to the large language model inference API to perform few-shot learning (Brown et al., 2020). This becomes a natural source of training data for meta-tuning. Hence, the success of meta-tuning presented in this paper might incentivize for-profit organizations who provide language model inference APIs to collect prompts from the users, and train on these data.

Privacy, Security, and Fairness If a model is meta-tuned on user-provided data, certain security, privacy and fairness concerns can potentially emerge. For example, Carlini et al. (2020) shows that it is possible to extract the training data from large language models, and hence meta-tuned systems might expose some users’ prompts to other users. Wallace et al. (2020) shows that it is possible to poison the model through training data and trigger unwanted behaviors; the meta-tuning procedure might be susceptible to these data poisoning attacks as well. Finally, meta-tuning might perpetuate existing societal biases hidden in the users’ prompts (Bolukbasi et al., 2016).

If not addressed properly, these concerns might have a broader negative societal impact through meta-tuning. Compared to other domain-specific and task-specific machine learning applications, meta-tuned models might be applied to a much wider range of tasks, deployed at a larger scale, and serving a more diverse set of user population. Therefore, biased or poisoned training data for one task from one user population might compromise fairness and performance of another task and harm another user population.

Potential Abuse As shown in Figure 5, the AUC-ROC score for a lot of tasks are still well below 0.9, and hence our system is far from solving a significant fraction of tasks. Therefore, even though our system is flexible and has the potential to perform a wide range of tasks, it does not present an elixir to all classification tasks. Particularly, it should not be applied to higher stake scenarios (e.g. hate speech detection, fake news

detection, etc), since its efficacy, robustness, and fairness properties remain unknown.

References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Tiago Almeida, José María Gómez Hidalgo, and Tiago Pasqualini Silva. 2013. Towards sms spam filtering: Results under a new dataset. *International Journal of Information Security Science*, 2(1):1–18.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI With Shared Human Values. *arXiv e-prints*, page arXiv:2008.02275.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011a. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011b. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Amita Misra, Mansurul Bhuiyan, Jalal Mahmud, and Saurabh Tripathy. 2019. Using structured representation and data: A hybrid model for negation and sentiment in customer service conversations. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 46–56, Minneapolis, USA. Association for Computational Linguistics.

Rishabh Misra. 2018. News category dataset.

Rishabh Misra. 2019. Imdb spoiler dataset.

Rishabh Misra and Prahal Arora. 2019. Sarcasm detection using hybrid neural network. *arXiv preprint arXiv:1908.07414*.

Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 422–426. ACM.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Sasha Spala, Nicholas Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. SemEval-2020 task 6: Definition extraction from free text with the DEFT corpus.

In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 336–345, Barcelona (online). International Committee for Computational Linguistics.

Sasha Spala, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. 2019. DEFT: A corpus for definition extraction in free- and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 124–131, Florence, Italy. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. 2020. Customizing triggers with concealed data poisoning. *arXiv preprint arXiv:2010.12563*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Datasets

IMDB movie review sentiment classification (Maas et al., 2011b). Classifies whether the user likes the movie.

POSITIVE: "'My favourite police series of all time turns to a TV-film. Does it work? Yes. ..."

NEGATIVE: " "Stupid! Stupid! Stupid! I can not stand Ben stiller anymore."

Zero Shot Emotion Classification (Yin et al., 2019). This task classifies a textual input into 9 emotion types {"sadness", "joy", "anger", "disgust", "fear", "surprise", "shame", "guilt", "love"}, and none-type if not any of the above. For example,

JOY: "*Making new friends is always fun , specially when playing dress up*"

ANGER: "*People that smoke cigarettes irritate my soul.*"

Zero Shot topic Classification (Yin et al., 2019). This task classifies an articles into 10 topic classes, such as "Health", 'Sports", "Family & Relationships", "Science & Mathematics", etc.

SCIENCE: "*How the human species evolved ? How the human species evolved ? A tough question as it overlaps science and theology ...*"

SPORTS: "*Number one formula one driver ? Depends on your question.Kimi Raikkonen won more races ...*"

Zero Shot Situation Classification (Yin et al., 2019). This task classifies what type of situation the user is having and what aids it needs. The classes include "food", "terrorism", etc.

FOOD: "livestock , crops , rural and urban housing , roads , river and canal embankments , and other infrastructure suffered extensive damage"

TERRORISM: "Daish attacked Tahurmahu village with chemical gas."

SemEval 2017 Task 4 (Barbieri et al., 2020) (Rosenthal et al., 2017). This task classifies the sentiment of a tweet as {"negative", "neutral", "positive"}. For example,

NEGATIVE: "*That sucks if you have to take the SATs tomorrow*"

NEUTRAL: "*@user Well said on HMW. Can you now address why Texans fans file out of the stadium midway through the 4th qtr of every game?*"

POSITIVE: "*"This is the last Sunday w/o football .....,NFL is back baby"*"

ETHICS Justice (Hendrycks et al., 2020). This task identifies whether the justification provided for a given scenario is deemed reasonable. For example,

REASONABLE: "*I deserve to be paid by my boss because I keep her house clean daily.*"

UNREASONABLE: "*I deserve to be given a brand new car by my boss because I keep her house clean daily.*"

ETHICS Deontology (Hendrycks et al., 2020). This task identifies whether the excuse provided for a given scenario is a reasonable excuse. For example,

SCENARIO: "*Aren't you supposed to be preparing for the meeting?*"

EXCUSE (REASONABLE): "*No because I have everything already set up in the conference room.*"

EXCUSE (UNREASONABLE): "*No because you don't have a meeting*"

ETHICS Common Sense (Hendrycks et al., 2020). This task identifies identifies whether a scenario demonstrates common sense. For example,

COMMON SENSE: "*I went to the principal's office to change my records before going to a different school.*"

NOT COMMON SENSE: "*I secured the loan because I would make the payments.*"

EURLEX57K (Chalkidis et al., 2019). This task classifies European legislation.

NATIONAL CURRENCY: "*Council Regulation (EC) No 2595/2000 of 27 November 2000 amending Regulation (EC) No 1103/97 on certain provisions relating to the introduction of the euro*"

SOUTHERN AFRICA: "*95/458/EC: Commission Regulation (EC) No 302/2006 of 20 February 2006 on import licences in respect of beef and veal products originating in Botswana, Kenya, Madagascar, Swaziland, Zimbabwe and Namibia*"

SemEval 2019 Task 6 (Barbieri et al., 2020) (Zampieri et al., 2019). This task classifies the tweet as either offensive or not offensive. For example,

OFFENSIVE: "*@user She has become a parody unto herself? She has certainly taken some heat for being such an....well idiot. Could be optic too Who know with Liberals They're all optics. No substance*"

NOT OFFENSIVE: "*@user @user She is great. Hi Fiona!*"

Click Bait Detection [4] This task detects whether a news title is a click bait.

CLICK BAIT: "*Can You Pass This Basic Trigonometry Quiz*"

NON CLICK BAIT: "*NASCAR driver Kyle Busch wins 2011 Jeff Byrd 500*".

Abstract Domain Classification [5] This classifies the abstract into 4 domains: "Physcis", "Maths", "Computer Science", "Statistics". For example,

PHYSICS: "*a ever-growing datasets inside observational astronomy have challenged scientists inside many aspects, including an efficient and interactive data exploration and visualization. many tools have been developed to confront this challenge ...*"

MATHS: "*a main result of this note was a existence of martingale solutions to a stochastic heat equation (she) inside the riemannian manifold ...*"

SemEval 2019 Task 5 (Barbieri et al., 2020) (Basile et al., 2019). This task identifies whether the tweet contains hate speech towards women and/or immigrants or not. For example,

HATE SPEECH: "*This account was temporarily inactive due to an irrational woman reporting us to Twitter. What a lack of judgement, shocking. #YesAllMen*"

NO HATE SPEECH: "*@user nice new signage. Are you not concerned by Beatlemania -style hysterical crowds crongregating on you. . .* "

SemEval 2019 Task 8 (Mihaylova et al., 2019). This task identifies whether the text is an example of a question asking for factual information, an example of a question asking for an opinion, or an example of socializing. For example,

FACTUAL: "*is there any place i can find scented massage oils in qatar?*"

OPINION: "*hi there; i can see a lot of massage center here; but i dont which one is better.*

[4] https://www.kaggle.com/c/clickbait-news-detection

[5] https://www.kaggle.com/abisheksudarshan/topic-modeling-for-research-articles?select=Train.csv

can someone help me which massage center is good...and how much will it cost me? thanks"

SOCIALIZING: "*Hello people...let's play this game...you have to write something good about the person whose 'post' is above you on QL.You can write anything and you can write multiple times.*"

SemEval 2018 Task 3 (Barbieri et al., 2020) (Van Hee et al., 2018). This task identifies whether the tweet contains irony or not. For example,

IRONY: "*seeing ppl walking w/ crutches makes me really excited for the next 3 weeks of my life*"

NO IRONY: "*@user on stage at #flzjingleball at the @user in #Tampa #iheartradio*"

SemEval 2018 Task 1 (Barbieri et al., 2020; Mohammad et al., 2018) This task classifies a tweet as one of 4 emotion types {"sadness", "joy", "anger", "optimism"}. For example,

SADNESS: "*@user I so wish you could someday come to Spain with the play, I can't believe I'm not going to see it #sad*"

JOY: "*#ThisIsUs has messed with my mind & now I'm anticipating the next episode with #apprehension & #delight! #istherea-helplineforthis*"

ANGER: "*@user Haters!!! You are low in self worth. Self righteous in your delusions. You cower at the thought of change. Change is inevitable.*"

OPTIMISM: "*Don't be #afraid of the space between your #dreams and #reality. If you can #dream it, you can #make it so*"

SemEval 2016 Task 6 (Mohammad et al., 2016; Barbieri et al., 2020) This task classifies a tweet's stance as {"neutral", "against", "favor"}. Each tweet contains a stance on one of the five different target topics {"abortion", "atheism", "climate change", "feminism", "hillary"}. For example,

NEUTRAL: "*@user maybe that's what he wants #SemST*"

AGAINST: "*Life is #precious & so are babies, mothers, & fathers. Please support the sanctity of Human Life. Think #SemST*"

FAVOUR: "*@user @user Nothing to do with me. It's not my choice, nor is it yours, to dictate what another woman chooses. #feminism #SemST*"

SemEval 2020 Task 6 (Spala et al., 2020). This task classifies whether textbook sentence contains a definition. For example,

CONTAINS DEFINITION: "*Since 2005, automated sequencing techniques used by laboratories are under the umbrella of next-generation sequencing, which is a group of automated techniques used for rapid DNA sequencing*"

DOESN'T CONTAIN DEFINITION: "*These automated low-cost sequencers can generate sequences of hundreds of thousands or millions of short fragments (25 to 500 base pairs) in the span of one day.*"

TREC (Li and Roth, 2002). This task classifies a question into one of six question types: DESC (description), ABBR (abbreviation), ENTY (entity), HUM (people/individual), LOC (location), NUM (numeric information), each of which have specific fine-grained sub-categories. For example,

DESC: "*How did serfdom develop in and then leave Russia?*"

ABBR: "*What is the full form of .com?*"

ENTY: "*What films featured the character Popeye Doyle?*"

HUM: "*What contemptible scoundrel stole the cork from my lunch?*"

LOC: "*What sprawling U.S. state boasts the most airports?*"

NUM: "*How many Jews were executed in concentration camps during WWII?*"

SUBJ (Pang and Lee, 2004). This task classifies a sentence as being subjective or objective. For example,

SUBJECTIVE: "*smart and alert, thirteen conversations about one thing is a small gem.*"

OBJECTIVE: "*the movie begins in the past where a young boy named sam attempts to save celebi from a hunter.*"

The Corpus of Linguistic Acceptability (Warstadt et al., 2018).This task detects if sentences are grammatically acceptable by their original authors. For example,

GRAMMATICALLY ACCEPTABLE: "*Her little sister will disagree with her.*"

GRAMMATICALLY NOT ACCEPTABLE: "*Has not Henri studied for his exam?*"

The Multi-Genre NLI Corpus (Williams et al., 2018). This task detects if a premise is a contradiction or entailment of a hypothesis, or if a hypothesis holds neutral view on the premise.. For example,

NEUTRAL: "*Premise: Exoatmospheric Kill Vehicles orbiting Earth would be programmed to collide with warheads. Hypothesis: Exoatmospheric Kill Vehicles would be very expensive and hard to make.*"

ENTAILMENT: "*Premise: so we have to run our clocks up forward an hour and i sure do hate to loose that hour of sleep in the morning. Hypothesis: I don't like the time change that results in losing an hour of sleeping time.*"

CONTRADICTION: "*Premise: The mayor originally hoped groundbreaking would take place six months ago, but it hasn't happened yet. Hypothesis: The mayor doesn't want groundbreaking to happen at all.*"

Metaphor as a Medium for Emotion: An Empirical Study (**?**). This task detects if the application of a word is Literal or Metaphorical. For example,

WORD: ABUSE

LITERAL: "*This boss abuses his workers.*"

METAPHORICAL: "*Her husband often abuses alcohol.*"

Political Preference Classification (Allaway and McKeown, 2020). This task predicts a comment's stand point on a political topic. For example,

TOPIC: COMPANIES REGULATION

CON: "*Regulation of corporations has been subverted by corporations. States that incorporate corporations are not equipped to regulate corporations that are rich enough to influence elections, are rich enough to muster a legal team that can bankrupt the state. Money from corporations and their principals cannot be permitted in the political process if democracy is to survive.*"

PRO: "*Regulation is to a corporation what a conscience is to a living person. Without a conscience, we would all be sociopaths. Corporations do not have a conscience, thus they need regulation to make sure they are focused on benefiting society instead on merely benefiting themselves.*"

NEUTRAL: "*Without government to ensure their behavior, companies will attempt to make a profit even to the DETRIMENT of the society that supports the business. We have seen this in the environment, in finances, in their treatment of workers and customers. Enough.*"

Airline Service Review [6] This task classifies if an airline review has a positive or negative sentiment. For example,

POSITIVE: "*This is such a great deal! Already thinking about my 2nd trip to Australia; I haven't even gone on my 1st trip yet!*"

NEGATIVE: "*amazing to me that we can't get any cold air from the vents.*"

Covid-19 Tweets Sentiment Analysis [7] This task classifies if a tweet has a positive or negative sentiment. For example,

POSITIVE: "*Taken by Henk Zwoferink on Saturday in Wargl, our black beauty hauled a train bringing the last tourists home. Our colleagues are #workinghard to keep supply chains running while respecting the measures to ensure everyone's #safety. A pleasure to work with such #DedicatedPeople!*"

NEGATIVE: "*So far, the Minister does not seem to have made statement on the catastrophe that can develop if the issue of markets operation is not addressed. Food insecurity has potential to make current Covid-19 panic look like a kindergarten and could lead to riots. I submit.*"

Hotel Review [8] This task predicts if a hotel review is a positive or negative review. For example,

NEGATIVE: "*The single rooms like hospital rooms single rooms hotel sparse intentional know ugly like trapped hospital white walls sink basin room small rectangle shape.the beds hard rocks blankets rough really noisy.this overrated hotel stayed fans type hotels*"

POSITIVE: "*loved stay, stayed univ, inn 10 days april 2005 thoroughly enjoyed, free parking clean spacious room friendly staff great breakfast snack, loved location, definitely stay,* "

Stock Market Sentiment [9] This task predicts if a comment holds a positive or negative view on the performance of the stock market. For example,

NEGATIVE: "*GPS wow that wa s a fast fast fade...*"

POSITIVE: "*user Maykiljil posted that: I agree that MSFT is going higher & possibly north of 30*"

[6] https://www.kaggle.com/welkin10/airline-sentiment

[7] https://www.kaggle.com/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_test.csv

[8] https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews

[9] https://www.kaggle.com/yash612/stockmarket-sentiment-dataset

AG-News (Zhang et al., 2015). This task classifies the topic of news based on their contents. For example,

WORLD NEWS: "*Greek duo could miss drugs hearing*"

SPORTS NEWS: "*AL Wrap: Olerud Cheers Yankees by Sinking Ex-Team*"

BUSINESS NEWS: "*Lowe's Second-Quarter Profit Rises*"

TECH NEWS: "*Satellite boosts Olympic security*"

Real and Fake News [10] This task classifies if a news is fake or real. For example,

REAL: "*WASHINGTON (Reuters) - Alabama Secretary of State John Merrill said he will certify Democratic Senator-elect Doug Jones as winner on Thursday despite opponent Roy Mooreâ x80 x99s challenge, in a phone call on CNN. Moore, a conservative who had faced allegations of groping teenage girls when he was in his 30s, filed a court challenge late on Wednesday to the outcome of a U.S. Senate election he unexpectedly lost.*"

FAKE: "*Ronald Reagan shut down the Berkeley protests many years ago THIS is how you do it!*"

Disaster Tweets [11] This task detects if a tweet announces an emergency or a disaster. For example,

CONTAINS DISASTER: "*Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all.*"

DOES NOT CONTAIN DISASTER: "*My dog attacked me for my food #pugprobs.*"

Obama vs Trump Tweets [12] This task detects if a tweet was send by Obama or Trump. For example,

OBAMA: "*Michelle and I are delighted to congratulate Prince Harry and Meghan Markle on their engagement. We wish you a lifetime of joy and happiness together.*"

TRUMP: "*Together, we dream of a Korea that is free, a peninsula that is safe, and families that are reunited once again!*"

Kaggle Sexually Explicit Tweets [13] This dataset provides positive examples of profane comments. For example,

EXPLICIT"*What do guys say when you get naked in front of them for the first time?*"

Democratic vs Republican Tweets [14] This task detects if a tweet was send by the Democratic or Republican Party. For example,

DEMOCRATIC: "*#YuccaMountain would require moving tens of thousands of metric tons of radioactive waste across the country and through Southern Nevada.*"

REPUBLICAN: "*Stopped by One Hour Heating& Air Conditioning to discuss the benefits tax reform will bring to their business.*"

Women E-commerce Clothing Reviews (Misra et al., 2019)[15] This task predicts if the buyer likes or recommends a product base on its review. For example,

LIKE: "*After reading the previous reviews, i ordered a size larger. i am so glad i did it! it fits perfectly! i am 5'4"/115/32dd and went with the s regular. so beautiful! i can't wait to wear it!*"

DISLIKE: "*The zipper broke on this piece the first time i wore it. very disappointing since i love the design. I'm actually going to try to replace the zipper myself with something stronger, but annoying that it's come to that.*"

Quora Question Pairs [16] This task predicts if a pair of Quora question is asking for the same thing. For example,

SAME: "*Question 1: How many months does it take to gain knowledge in developing Android apps from scratch?; Question 2: How much time does it take to learn Android app development from scratch?*"

DIFFERENT: "*Question 1: How would you review the site Waveclues? ; Question 2: Is there a good pay for reviews site out there?*"

Headline Sarcasm Detection (Misra and Arora,

[10] https://www.kaggle.com/amananandrai/ag-news-classification-dataset?select=train.csv

[11] https://www.kaggle.com/c/nlp-getting-started/data?select=train.csv

[12] https://www.kaggle.com/shaharz/classifying-tweets-of-trump-and-obama

[13] https://www.kaggle.com/harsh03/sexually-explicit-comments

[14] https://www.kaggle.com/kapastor/democratvsrepublicantweets?select=ExtractedTweets.csv

[15] https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

[16] https://www.kaggle.com/c/quora-question-pairs/data

2019)[17] This task detects if is a news headline contains scarcasm. For example,

SARCASM: "*guy who just wiped out immediately claims he's fine*"

NO SARCASM: "*Donald trump effigies burn across Mexico in Easter ritual*"

Company Account Tweets [18] This task detects whether the tweet is targeted towards a company account. For example,

YES: "*@VirginTrains Oh, that's nice. What are you doing about it? What are you targets next year?*"

NO: "*@115738 That's the best kind of trick-or-treating. All treats, my friend. -Becky*"

SMS Spam Detection (Almeida et al., 2013) This task detects whether the SMS is a spam message. For example,

SPAM: "*Thank you, winner notified by sms. Good Luck! No future marketing reply STOP to 84122 customer services 08450542832*"

HAM: "*Lol great now I am getting hungry.*"

Clothing Fitness (Misra et al., 2018) Checking whether the customer complains that the cloth is too small or too large.

SMALL: "runs a bit small. wish it fit".

LARGE: "too big".

Water Problem Topic Classification [19] Classifying the topic of a report on water problems. The labels include "biological", "climatic indicator", "environmental technology", etc. For example,

BIOLOGICAL: "*Mineralization of organic phosphorus in bottom sediments reaches 40–80% and as we found out during the project implementation it intensified in autumn-winter period.*"

CLIMATIC INDICATOR: "*The average amount of precipitation in the lower part of the basin makes 470 mm to 540 mm. The relative average annual air humidity makes 60-65%*".

ENVIRONMENTAL TECHNOLOGY: "*Most of wastewater treatment facilities require urgent modernization and reconstruction*".

Sexist Statement Detection [20] This task classifies whether the statement is sexist. For example,

SEXIST: "*It's impossible for a girl to be faithful.*"

NON SEXIST: "*Without strength, can we work to create wealth?*"

Movie Spoiler Detection (Misra, 2019) [21] This task classifies whether the movie review is a spoiler. For example,

SPOILER: "*I must say that this movie was good but several things were left unsaid. For those who have seen the movie know what I am talking about but for those who haven't, I don't want to give spoilers. I was also impressed by Vin Diesel's acting skills. Overall I have to say it was a good movie filled with several twists and turns.*"

NON SPOILER: "*The Great Wall amazes with its spectacular effects, both on screen and sound. Usually I do not appreciate 3D movies, but in this case I felt like it worth it.However, being honest, the storytelling and the story itself had its weaknesses. There were many logical lapses, and for me, many details are still waiting to be answered.On the other hand, expect decent acting especially from the main characters.All in all, The Great Wall is a solid popcorn-movie, but I expected a more elaborated unfolding of the legend it tells about.*"

News Summary/headline Topic Classification (Misra, 2018) [22] This task classifies the topic of the summary of a news. For example,

POLITICS: "*City and state officials said they received little advance warning of the decision.*"

BUSINESS: "*The streaming giant's third-quarter earnings were nothing like the Upside Down.*"

B Dataset Property Tags

Here we list all the dataset property tags (Section 2). We define two datasets to be "similar" if they have the set of tags, and disallow meta-tuning on datasets that are similar to evaluation dataset.

social media: whether the source is from social media (e.g. tweets).

[17] https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection?select=Sarcasm_Headlines_Dataset_v2.json

[18] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

[19] https://www.kaggle.com/vbmokin/nlp-reports-news-classification?select=water_problem_nlp_en_for_Kaggle_100.csv

[20] https://www.kaggle.com/dgrosz/sexist-workplace-statements

[21] https://www.kaggle.com/rmisra/imdb-spoiler-dataset?select=IMDB_reviews.json

[22] https://www.kaggle.com/rmisra/news-category-dataset

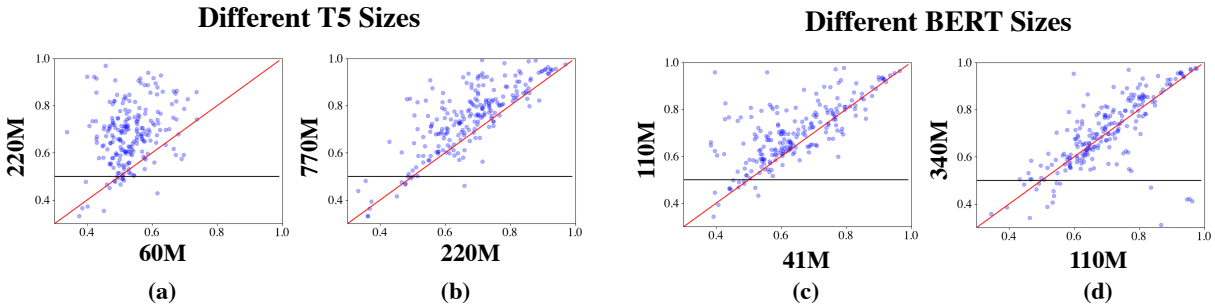


Figure 6: The detailed explanation of the figures is in the caption of Figure 5. We compare different model sizes of T5 (a, b) and BERT after meta-tuning (c, d). For most label descriptions, larger models are better (above the red line); additionally, the ROC-AUC score is above the random guess baseline (above the black line) by a robust margin.

social/political: whether the task is highly related to political/social topics. Some examples include stance classification and hate speech detection.

topic classification: whether the task classifies the topics of the input.

good vs. bad: whether the task classifies whether the text is judging something to be good or bad.

paper: whether input text comes from a paper.

review: whether the input text is a review of a product (e.g. movie, hotel).

questions: whether the input texts are questions. Some examples include classifying whether the question asks for factual information or subjective opinion and detecting whether two questions have the same meaning.

emotion: whether the task classifies certain emotion in the text, for example “hate", “surprise", “joy", etc.

Besides, we do not assign tags to datasets that we are confident to be different enough from other tasks (e.g. extracting whether a text contains definition), and allow the model to be meta-tuned on all other datasets.

C Comparing Different Sizes with BERT and T5

C.1 Larger T5 Models are Better

Figure 6 (a) compares a meta-tuned model with 60 M parameters against one with 220M parameters, and (b) compares 220M parameters against 770M parameters. For most label descriptions, larger models are better, and the ROC-AUC score is above the random guess baseline (above the black line) by a robust margin.

C.2 Larger BERT Models are Better

We concatenate the textual input to be classified and the label description (question) and feed them into the BERT model; to classify between “Yes" and “No", we add a linear layer on top of the pooled output (BERT-For-Sequence-Classification, Wolf et al. (2020)).

Figure 6 (c) compares a meta-tuned model with 41 M parameters against one with 110M parameters, and (d) compares 110 parameters against 340M parameters. For most label descriptions, larger models are better, and the ROC-AUC score is above the random guess baseline (above the black line) by a robust margin.

D Most Relevant Datasets

To ensure that we are testing the models’ ability to generalize to an unseen tasks, we disallow both training and testing on datasets that are too similar, which is defined as “having the same set of dataset property tags" (Section 2). To help interpret how we define unseen tasks, for each dataset that we evaluate on, we try to find the “most relevant" dataset that the model has seen during the meta-tuning phase, and list it in Table 2.

E Visualization with Different Opacity

We plot the same figure as in Section 4.1, except that now the opacity is inversely proportional to the number of annotations for each label (i.e. each label has the same weight). We still find that larger meta-tuned models are better.

We may also plot the same figure with opacity inversely proportional to the number of label annotations for each dataset (i.e. each dataset has the same total weight). We can clearly see that larger models are better from the visualization (right),

Evaluation Dataset	Most Relevant Training Dataset
SemEval 2016 Task 6, stance classifications on issues like feminism, atheism, etc	SemEval 2019 Task 5, detecting hate speech against women and immigrants
SemEval 2019 Task 6, classifying whether the text is offensive	A dataset from Kaggle that classifies sexually explicit comments
SemEval 2019 Task 5, detecting hate speech against women and immigrants	SemEval 2016 Task 6, stance classifications on issues like feminism, atheism, etc
TREC, classifying the type the question is asking about (e.g. numbers, acronyms, human/occupations, etc)	AG News, which classifies news into different categories (e.g. sports, world events).
SemEval 2019 Task 8, classifying whether the question is asking for subjective opinion, factual information, or simply having a conversation	N/A
SUBJ, classifying whether the text contains subjective or objective information	N/A
QQP, classifying whether two questions have the same meaning	N/A
Yin et al. (2019) emotion classification, classifying text into 9 emotion types, such as “joy", “anger", “guilt", “shame", etc.	Classifying whether an IMDB movie review is positive.
Yin et al. (2019) situation classification, classifying which disaster situation people are experiencing, e.g. “regime change", “crime and violence", and what resource they need, e.g. “food and water", “search and rescue".	Classifying (binary) whether a tweet is related to a natural disaster.
Yin et al. (2019) topic classification, classifying the domain of an article into domains such as “family and relationship", “education", “business", “sports"	classifying the domain of a paper abstract into physics, maths, computer sciences, and statistics.
AG News, which classifies news into different categories (e.g. sports, world events).	Abstract Domain classification, classifying the domain of a paper abstract into physics, maths, computer sciences, and statistics.
Abstract Domain classification, classifying the domain of a paper abstract into physics, maths, computer sciences, and statistics.	AG News, which classifies news into different categories (e.g. sports, world events).
IMDB movie reviews, classifying whether the user feels positive about the movie	Stock market sentiment, classifying whether a comment is optimistic about the market.
Women Clothing Recommendation, classifying whether the user feels positive about the women clothing	Stock market sentiment, classifying whether a comment is optimistic about the market.
Hotel Reviews, classifying whether the user feels positive about the hotel	Stock market sentiment, classifying whether a comment is optimistic about the market.
Stock market sentiment, classifying whether a comment is optimistic about the market.	IMDB movie reviews, classifying whether the user feels positive about the movie.

Table 2: For each dataset that we evaluate on, we list the task in the training split that we consider to be the most relevant. We list “N/A" if we think that none of the training dataset is particularly relevant. Due to space constraint of a page, this table is continued by 3.

Evaluation Dataset	Most Relevant Training Dataset
CoLA, classifying whether a sentence is grammatical	N/A
SemEval 2020 Task 6, classifying whether a sentence contains a definition	N/A
Clothing Fitness, classifying whether the review thinks the cloth is too small or large for him/her	Women Clothing Recommendation, classifying whether the user feels positive about the women clothing
Spam classification, classifying whether a text message is a spam	click-bait classification, classifying whether the title of an article is a clickbait.

Table 3: Continuation of Table 2

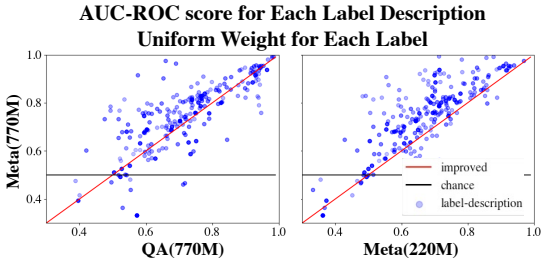


Figure 7: Same figure as Figure 5, except that now the opacity is inversely proportional to the number of annotated descriptions for each label (i.e. each label has the same weight). We still find that larger meta-tuned models are better.

but it is not clear whether the meta-tuned model is better than the QA model (left), and hence we have to rely on additional aggregate statistics.

The weighted average of the AUC-ROC scores of the meta-tuned model is 71.5, compared to 70.4 for the QA model. 60% of the total weight is above the red line, while 40% of them is below. 30.7% of the total weight is at least 0.05 above the red line (i.e. the meta-tuned model improves the AUC-ROC score over the QA model by at least 0.05), while 18% is at least 0.05 below the red line. 13% of the total weight is at least 0.1 above the red line, while 12% of the total weight is at least 0.1 below the red line. Since we only have 25 distinct datasets in total, we cannot claim any statistical significance for all these results; however, overall we see positive evidence that meta-tuned model is better at 0SC than the UnifiedQA model.

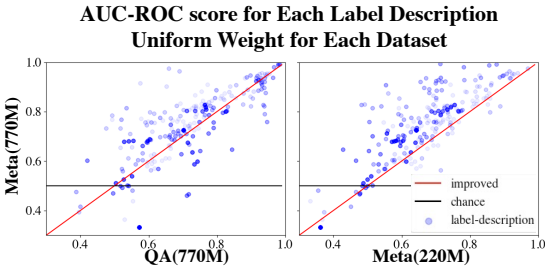


Figure 8: Same figure as Figure 5, except that now the opacity is inversely proportional to the number of annotated label descriptions for each dataset (i.e. each dataset has the same weight).