Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media

Xiang Dai[1,2] **Sarvnaz Karimi**[1] **Ben Hachey**[3] **Cecile Paris**[1]
[1]CSIRO Data61, Sydney, Australia
[2]University of Sydney, Sydney, Australia
[3]Harrison ai, Sydney, Australia
{dai.dai,sarvnaz.karimi,cecile.paris}@csiro.au
ben.hachey@gmail.com

Abstract

Recent studies on domain-specific BERT models show that effectiveness on downstream tasks can be improved when models are pretrained on in-domain data. Often, the pretraining data used in these models are selected based on their subject matter, e.g., biology or computer science. Given the range of applications using social media text, and its unique language variety, we pretrain two models on tweets and forum text respectively, and empirically demonstrate the effectiveness of these two resources. In addition, we investigate how similarity measures can be used to nominate in-domain pretraining data. We publicly release our pretrained models at https://bit.ly/35RpTf0.

1 Introduction

Sequence transfer learning (Ruder, 2019), that pretrains language representations on unlabeled text (*source*) and then adapts these representations to a supervised task (*target*), has demonstrated its effectiveness on a range of NLP tasks (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019). Approaches vary in model, pretraining objective, pretraining data and adaptation strategy. We consider a widely used method, BERT (Devlin et al., 2019). It pretrains a transformer-based model using a masked language model objective and then fine-tunes the model on the target task. We investigate the impact of the domain (i.e., the similarity between the underlying distribution of source and target data) of pretraining data on the effectiveness of pretrained models. We also propose a cost-effective way to select pretraining data.

Recent studies on domain-specific BERT models, which are pretrained on specialty source data, empirically show that, when in-domain data is used for pretraining, target task performance can be improved (Lee et al., 2019; Alsentzer et al., 2019; Huang et al., 2019; Beltagy et al., 2019). These publicly available domain-specific BERT models are valuable to the NLP community. However, the selection of in-domain data usually resorts to intuition, which varies across NLP practitioners (Dai et al., 2019). According to Halliday and Hasan (1989), the context specific usage of language is affected by three factors: *field* (the subject matter being discussed), *tenor* (the relationship between the participants in the discourse and their purpose) and *mode* (communication medium, e.g., 'spoken' or 'written').[1] Generally, the selection of pretraining data in existing domain-specific BERT models is based on the field rather than the tenor. For example, BioBERT (Lee et al., 2019) and SciBERT (Beltagy et al., 2019) are both pretrained on scholar articles, but on different fields (biology and computer science).

We conduct a case study of pretraining BERT on social media text which has very different tenor from existing domain-specific BERT models. Our contributions are two-fold: (1) We release two pretrained BERT models trained on tweets and forum text, and we demonstrate the effectiveness of these two resources on a range of NLP data sets using social media text; and, (2) we investigate the correlation of source-target similarity and task accuracy using different domain-specific BERT models. We find that simple similarity measures can be used to nominate in-domain pretraining data (Figure 1).

2 Related Work

Selecting data to pretrain BERT There are two known strategies: (1) collecting very large generic data, such as web crawl and news (Radford et al., 2019; Liu et al., 2019; Baevski et al., 2019); and, (2) selecting in-domain data, which we refer to as

[1] We do not explicitly consider mode in this study, because all data used are written text.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1675–1681
November 16 - 20, 2020.

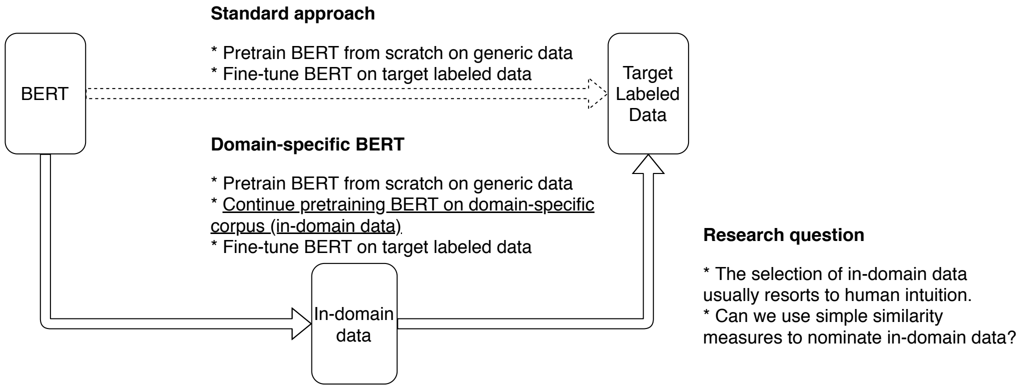
Standard approach
* Pretrain BERT from scratch on generic data
* Fine-tune BERT on target labeled data
BERT
Target Labeled Data
Domain-specific BERT
* Pretrain BERT from scratch on generic data
* Continue pretraining BERT on domain-specific corpus (in-domain data)
* Fine-tune BERT on target labeled data
In-domain data
Research question
* The selection of in-domain data usually resorts to human intuition.
* Can we use simple similarity measures to nominate in-domain data?

Figure 1: Recent studies have demonstrated the effectiveness of domain-specific BERT models. However, the selection of in-domain data usually resorts to intuition, which varies across NLP practitioners, especially regarding intersecting domains. We investigate the correlation of source-target similarity and the effectiveness of pretrained models. In other words, we aim to use simple similarity measures to nominate in-domain pretraining data.

domain-specific BERT models.

Those following the first strategy intend to build universal language representations that are useful across multiple domains. They also believe that pretraining on larger data leads to better pretrained models. For example, Baevski et al. (2019) empirically show that the average GLUE score (Wang et al., 2019) can increase from lower than 80 to higher than 81 when the size of pretraining data increases from 562 million to 18 billion tokens.

Our study uses the second strategy. However, we select our pretraining data from the tenor perspective rather than the field. A summary of the source data used in these domain-specific BERT models can be found in Table 1.

Model	Source data
Original BERT	Books and encyclopedia articles, various fields
BioBERT (Lee et al., 2019)	Scholar articles on biology
ClinicalBERT (Alsentzer et al., 2019)	Nursing and physician notes on hospital admission
SciBERT (Beltagy et al., 2019)	Scholar articles on biology and computer science
TwitterBERT (this work)	Tweets, various fields
ForumBERT (this work)	Forum text on business review

Table 1: A summary of source data used in the original BERT and several domain-specific BERT models.

Finding in-domain data Our study relates to the literature on investigating domain similarity (Blitzer et al., 2006; Ben-David et al., 2007; Ruder and Plank, 2017) and text similarity (Mihalcea et al., 2006; Pavlick et al., 2015; Kusner et al., 2015). Our work is also inspired by the study by Dai et al. (2019) on the impact of source data on pretrained LSTM-based models (i.e., ELMo) and by Van Asch and Daelemans (2010) on the correlation between similarity and accuracy loss of POS taggers.

3 Pretraining BERT Models

We follow the practices used in other domain-specific BERT models (Lee et al., 2019; Beltagy et al., 2019) to pretrain our BERT models. We use the original vocabulary of BERT-Base as our underlying word piece vocabulary[2] and use the pretrained weights from the original BERT-Base as the initialization weights. Note that all domain-specific models we consider in this study are based on this paradigm,[3] which means these models are supposed to capture both generic (inheriting from original BERT) and domain-specific knowledge.

For pretraining objective, we remove the Next Sentence Prediction (NSP) objective. Social media text, especially tweets, are often too short to sample consecutive sentences. In addition, recent studies observe benefits in removing the NSP objective with sequence-pair training (Liu et al., 2019).

[2] Beltagy et al. (2019) investigated the effect of having an in-domain vocabulary. Their results show that, although an in-domain vocabulary is helpful, the magnitude of improvement is relatively small.

[3] We notice a very recent resource by Nguyen et al. (2020) who pretrain RoBERTa on general English tweets, as well as tweets related to the COVID-19 pandemic. We did not consider this model as it involves more variants: byte pair encoding and initialization weights.

Twitter We use English tweets ranging from Sep 1 to Oct 30, 2018[4] to pretrain our Twitter BERT. There are in total 60 million English tweets, consisting of 0.9B tokens. Although we aim to avoid tailored pre-processing strategies to make a fair comparison with other domain-specific BERT models, we find 44% of these tweets contain url and 78% contain other user names (@, if a tweet replies another tweet, @ is added automatically). We thus employ minimal processing by: (1) replacing tokens starting with '@', referring to a Twitter user's account name, with a special token [TwitterUser]; and, (2) replacing urls as a special token [URL]. We hypothesize that the surface form of these tokens do not contain useful information.

Forum We use local businesses reviews released by Yelp[5] to pretrain our Forum BERT. There are in total five million reviews, consisting of 0.6B tokens. No preprocessing is conducted on the text.

We used four Nvidia P100 GPUs for the pretraining. Training of each model took seven days.

4 Effectiveness of Pretrained BERT Models

To evaluate the effectiveness of our pretrained BERT models, we experiment on a range of classification and Named Entity Recognition (NER) data sets. Both text classification and NER are fundamental NLP tasks that can employ generic architectures on top of BERT. For the classification task, the representation of the first token (i.e., [CLS]) is fed into the output layer for the final prediction. For the NER task, the representations of the first sub-token within each token are taken as input to a token-level classifier to predict the token's tag. We did not explore more complex architectures, such as adding LSTM or CRF on top of BERT (Beltagy et al., 2019; Baevski et al., 2019), because our aim is to demonstrate the efficacy of domain-specific BERT models and to observe the impact of pretraining data, rather than to achieve state-of-the-art performance on these data sets.

Our BERT results follow the standard two-stage approach of finetuning the pretrained model. Domain-specific BERTs add a stage in the middle: finetuning BERT on domain-specific unlabeled data (cf. Figure 1).

4.1 Target Tasks

We use eight target tasks with their text sampled from Twitter and forums, to examine whether our BERT models can lead to improvements, compared to the original BERT. These tasks are **Airline**[6]: classifying sentiment on tweets about major U.S. airlines; **BTC**: identifying location, person, and organization on tweets (Derczynski et al., 2016); **SMM4H-18**: classifying whether the user reports an adverse drug events (task3) (Weissenbacher et al., 2018), or intends to receive a seasonal influenza vaccine (task4) on tweets about health (Joshi et al., 2018); **CADEC**: identifying adverse drug events etc. on reviews about medications (Karimi et al., 2015); **SemEval-14**: identifying product or service attributes on reviews about laptops and restaurants (Pontiki et al., 2014); **SST**: classifying sentiment on movie reviews (Socher et al., 2013).

In addition, we use four tasks that do not use social media text to investigate how our BERT models perform on out-of-domain target tasks: **Paper Field**: classifying the research topic based on the title of scholar articles about various fields (Beltagy et al., 2019); **EBM**: identifying intervention, outcome etc. on scholar articles about clinical trials (Nye et al., 2018); **i2b2-10**: identifying treatment, test and problem on clinical notes about health (Uzuner et al., 2011); **JNLPBA**: identifying RNA, DNA etc. on scholar articles about biology (Kim et al., 2004).

4.2 Results

We observe that our BERT models achieve the highest F1 score on 6 out of 8 target tasks that use social media text (Table 2). On CADEC (medications) and SemEval-14 laptop, SciBERT achieves the highest score due to the overlapping fields (i.e., medication and computer hardware, respectively). We note, however, that our Forum BERT achieves very close results. This demonstrates the effectiveness of our pretrained models on target tasks using social media text. To our surprise, on target tasks using tweets, forum BERT achieves better results than Twitter BERT on 3 classification tasks. On one hand, this may be explained by Baldwin et al. (2013)'s observation that forum text is the 'median' data, which is similar to all other types of

[4] Internet archive, Accessed 1 June 2020.

[5] Yelp Challenge, Accessed 1 June 2020.

[6] Kaggle Twitter US Airline Sentiment Challenge

Target Text type	Corpus	**BERT** (3.3B)	**Bio** (18B)	**Clinical** (0.5B)	**Sci** (3.1B)	**Twitter** (0.9B)	**Forum** (0.6B)
Tweets	Airline (C)	80.5± 0.3	79.0± 0.5	78.8± 0.8	78.8± 0.9	80.8± 0.6	**81.6± 0.5**
	BTC (N)	78.0± 0.5	75.2± 0.3	76.9± 0.5	77.4± 0.4	**79.0± 0.5**	77.0± 0.4
	SMM4H-18 task3 (C)	76.5± 0.9	75.4± 1.1	75.6± 0.7	75.4± 1.0	77.0± 1.0	**77.2± 1.3**
	SMM4H-18 task4 (C)	89.4± 0.5	87.7± 0.4	88.1± 0.8	88.7± 0.8	90.3± 0.3	**91.1± 0.6**
Forum	CADEC (N)	71.9± 0.6	72.1± 0.6	72.1± 0.8	**73.2± 0.4**	72.1± 1.0	72.9± 0.6
	SemEval-14 laptop (N)	81.1± 0.8	79.3± 0.3	78.5± 0.4	**81.6± 1.1**	81.3± 0.6	81.4± 1.1
	SemEval-14 restaurant (N)	87.5± 0.6	84.9± 0.3	85.5± 0.7	86.7± 0.5	87.4± 0.7	**89.3± 0.5**
	SST-2 (C)	92.4± 0.2	91.1± 0.5	90.4± 0.3	91.4± 0.4	92.3± 0.4	**93.4± 0.4**
Non-social media	EBM (N)	41.5± 0.5	42.1± 0.2	41.1± 0.5	**42.4± 0.7**	40.5± 0.5	41.5± 0.5
	i2b2-10 (N)	85.8± 0.1	**87.4± 0.2**	**87.4± 0.1**	87.3± 0.2	84.8± 0.2	85.2± 0.1
	JNLPBA (N)	72.5± 0.3	**74.2± 0.2**	71.9± 0.1	73.6± 0.3	72.2± 0.2	72.5± 0.2
	Paper Field (C)	74.5± 0.1	74.3± 0.1	73.3± 0.1	**75.1± 0.1**	74.1± 0.1	73.3± 0.2

Table 2: Effectiveness of different BERT models, evaluated on downstream tasks. # tokens in each pretraining data are listed in brackets. C: Classification task, for which we report macro-F1; N: NER task, for which we report span-level micro-F1. We repeat all experiments five times with different random seeds. Mean values are reported. underline: the best result is significantly better than the second best result (paired student's t-test, p: 0.05).

		ForumBERT	
		✓	✗
SciBERT	✓	159	36
	✗	43	161

Table 3: False positives by the BERT model on CADEC. ✓ represents the number of errors which are fixed by the domain-specific BERT. ✗ indicates errors are not fixed.

		ForumBERT	
		✓	✗
SciBERT	✓	41	22
	✗	34	258

Table 4: False negatives by the BERT model on CADEC.

social media text. On the other hand, it also reveals the challenge of pretraining contextual language representations on short tweets.

We also observe that, when domain-specific models are applied on a target task with out-of-domain data, they achieve much lower results than the original BERT. For example, BioBERT achieves lower results than the original BERT on 7 out of 8 target social media tasks. It only achieves a better result on CADEC, which is about medications. Recall that all these domain-specific BERT models use the pretrained weights of the original BERT as initialization. On one hand, we argue that this observation may challenge the conventional wisdom that the larger the pretraining data is, the better the pretrained model is. Training on out-of-domain source data may cause negative impact, at least for the two-stage pretraining approach we consider. On the other hand, this observation reinforces recent work showing the importance of task-adaptive pretraining (Gururangan et al., 2020).

Error analysis on CADEC We conduct an error analysis on CADEC, because it is at the intersection between social media tenor (online posts) and medication field (adverse drug events), and thus could be similar to multiple sources. We compare the error predictions by the two best performing BERT models, ForumBERT and SciBERT, as well as the baseline BERT model. In Table 3, we observe that both domain-specific BERT models can reduce greatly the number of false positives made by the baseline BERT. Specifically, 159 false positives made by the baseline BERT are fixed by the domain-specific BERT models. However, domain-specific BERT models do not reduce much of the number of false negatives. There are 258 gold mentions recognized by none of three models, and only 41 false negatives by the baseline BERT are fixed by the domain-specific BERT models (Table 4).

5 Analysis

After we empirically show the importance of selecting in-domain source data, the next question is: can we find a cost-effective way to nominate in-domain source data?

5.1 Measuring Similarity

We use three measures of the similarity between source and target data. We then observe whether

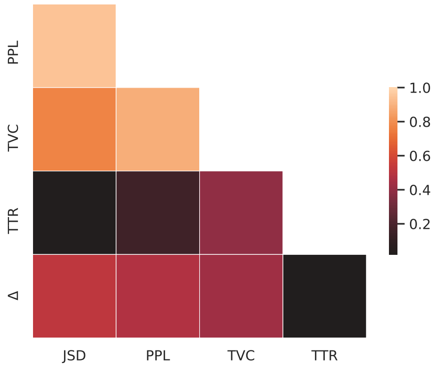

Figure 2: Correlation between different similarity measures and diversity measure and the improvement (Δ) due to domain-specific BERT models.

these similarity values correlate with the usefulness of pretrained models in § 5.2.

Language model perplexity (PPL) has been used to provide a proxy to estimate corpus similarity (Baldwin et al., 2013). We construct Kneser-Ney smoothed 3-gram models (Heafield, 2011) on source data and use the perplexity of target data relative to these language models as the similarity between source and target data.

Jensen-Shannon divergence (JSD), based on term distributions, has been successfully used for domain adaptation (Ruder and Plank, 2017). We first measure the probability of each term (up to 3-gram) in source and target data, separately. Then, we use the Jensen-Shannon divergence between these two probability distributions as the similarity between source and target data.

Target vocabulary covered (TVC) measures the percentage of the target vocabulary present in the source data, where only content words (nouns, verbs, adjectives) are counted. Dai et al. (2019) show that it is very informative in predicting the effectiveness of pretrained word vectors.

In addition, Ruder and Plank (2017) show that the diversity of source data is as important as domain similarity for domain adaptation. Inspired by this, we also explore a very simple diversity measure: type token ratio (**TTR**, $\frac{\#\,\text{unique tokens}}{\#\,\text{tokens}}$), that measures the lexical diversity of the source data.

To mitigate the impact of source data size on these measurements, for each source data, we sample five sub-corpora, each of which contains 10M tokens. Then we measure the similarity of source and target data and the diversity of source data as the average values of these sub-corpora.

5.2 Correlation Analysis

To analyze how the effectiveness of domain-specific BERT models correlate to the similarity between source and target data, we employ the Pearson correlation analysis to find out the relationships between improvements due to domain-specific BERT models and similarity between source and target data. For example, considering the BTC task, we use the performance of the original BERT as baseline, and measure the improvement due to Twitter BERT as 1.0, whereas the corresponding value using BioBERT is -2.9. Note that we repeat all the experiments five times; therefore, we collect 300 source-target data points in total.

The correlation results are visualized in Figure 2. JSD has the strongest correlation (0.519) with the improvement due to domain-specific models, while the other two measures also have modest correlation (0.481 for PPL and 0.436 for TVC). Recall that the calculation of JSD takes uni-grams, bi-grams and tri-grams into consideration, whereas PPL considers tri-grams only and the TVC considers uni-grams only. Correlations between different measures indicate that these measures are able to reach agreement on whether source and target are similar. We find no correlation between the TTR of source data and the improvement.

6 Summary

We conduct a case study of pretraining BERT on social media text. Through extensive experiments, we show the importance of selecting in-domain source data. Based on empirical analysis, we recommend measures to help select pretraining data for best performance on new applications.

Acknowledgments

We would like to thank anonymous reviewers for their helpful comments. We thank Shubin Du and Ying Zhou (University of Sydney) for an early investigation of this work. Xiang Dai is supported by University of Sydney's Engineering and Information Technologies Research Scholarship and a CSIRO Data61 top up scholarship.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *ClinicalNLP@NAACL*, pages 72–78, Minneapolis, Minnesota.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. In *EMNLP-IJCNLP*, pages 5359–5368, Hong Kong, China.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *IJCNLP*, pages 356–364, Nagoya, Japan.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP*, pages 3613–3618, Hong Kong, China.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137–144. Vancouver, Canada.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*, pages 120–128, Sydney, Australia.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. In *NAACL*, pages 1460–1470, Minneapolis, Minnesota.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *COLING*, pages 1169–1179, Osaka, Japan.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, Online.

Michael A.K. Halliday and Ruqaiya Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford University Press.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *WMT*, pages 187–197, Edinburgh, Scotland.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *CoRR abs/1904.05342*.

Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In *SMM4H@EMNLP*, pages 43–47, Brussels, Belgium.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A corpus of adverse drug event annotations. *J Biomed Inform*, 55:73–81.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *BioNLP*, Geneva, Switzerland.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*, pages 957–966, Lille, France.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Levis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692*.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pages 775–780, Boston, Massachusetts.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for English Tweets. *CoRR abs/2005.10200*.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL*, pages 197–207, Melbourne, Australia.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *ACL-IJCNLP*, pages 425–430, Beijing, China.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *SemEval*, pages 27–35, Dublin, Ireland.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Iyya Sutskever. 2018. Improving language understanding with unsupervised learning.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Iyya Sutskever. 2019. Language models are unsupervised multitask learners.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *EMNLP*, pages 372–382, Copenhagen, Denmark.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642, Seattle, Washington.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.*, 18(5):552–556.

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *DANLP@ACL*, pages 31–36, Uppsala, Sweden.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, New Orleans, Louisiana.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *SMM4H@EMNLP*, pages 13–16, Brussels, Belgium.