BERTweet: A pre-trained language model for English Tweets

Dat Quoc Nguyen[1], **Thanh Vu**[2,*] and **Anh Tuan Nguyen**[3,†]
[1]VinAI Research, Vietnam; [2]Oracle Digital Assistant, Oracle, Australia; [3]NVIDIA, USA
v.datnq9@vinai.io; thanh.v.vu@oracle.com; tuananhn@nvidia.com

Abstract

We present **BERTweet**, the *first* public large-scale pre-trained language model for English Tweets. Our BERTweet, having the same architecture as BERT$_{\text{base}}$ (Devlin et al., 2019), is trained using the RoBERTa pre-training procedure (Liu et al., 2019). Experiments show that BERTweet outperforms strong baselines RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$ (Conneau et al., 2020), producing better performance results than the previous state-of-the-art models on three Tweet NLP tasks: Part-of-speech tagging, Named-entity recognition and text classification. We release BERTweet under the MIT License to facilitate future research and applications on Tweet data. Our BERTweet is available at: https://github.com/VinAIResearch/BERTweet.

1 Introduction

The language model BERT (Devlin et al., 2019)—the Bidirectional Encoder Representations from Transformers (Vaswani et al., 2017)—and its variants have successfully helped produce new state-of-the-art performance results for various NLP tasks. Their success has largely covered the common English domains such as Wikipedia, news and books. For specific domains such as biomedical or scientific, we could retrain a domain-specific model using the BERTology architecture (Beltagy et al., 2019; Lee et al., 2019; Gururangan et al., 2020).

Twitter has been one of the most popular micro-blogging platforms where users can share real-time information related to all kinds of topics and events. The enormous and plentiful Tweet data has been proven to be a widely-used and real-time source of information in various important analytic tasks (Ghani et al., 2019). Note that the characteristics of Tweets are generally different from those of traditional written text such as Wikipedia and news articles, due to the typical short length of Tweets and frequent use of informal grammar as well as irregular vocabulary e.g. abbreviations, typographical errors and hashtags (Eisenstein, 2013; Han et al., 2013). Thus this might lead to a challenge in applying existing language models pre-trained on large-scale conventional text corpora with formal grammar and regular vocabulary to handle text analytic tasks on Tweet data. To the best of our knowledge, there is not an existing language model pre-trained on a large-scale corpus of English Tweets.

To fill the gap, we train the *first* large-scale language model for English Tweets using a 80GB corpus of 850M English Tweets. Our model uses the BERT$_{\text{base}}$ model configuration, trained based on the RoBERTa pre-training procedure (Liu et al., 2019). We evaluate our model and compare it with strong competitors, i.e. RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$ (Conneau et al., 2020), on three downstream Tweet NLP tasks: Part-of-speech (POS) tagging, Named-entity recognition (NER) and text classification. Experiments show that our model outperforms RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$ as well as the previous state-of-the-art (SOTA) models on all these tasks. Our contributions are as follows:

- We present the first large-scale pre-trained language model for English Tweets.
- Our model does better than its competitors RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$ and outperforms previous SOTA models on three downstream Tweet NLP tasks of POS tagging, NER and text classification, thus confirming the effectiveness of the large-scale and domain-specific language model pre-trained for English Tweets.
- We also provide the first set of experiments investigating whether a commonly used approach of applying lexical normalization dictionaries on Tweets (Han et al., 2012) would help im-

[*] Most of the work done when Thanh Vu was at the Australian e-Health Research Centre, CSIRO, Australia.

[†] Work done during internship at VinAI Research.

Proceedings of the 2020 EMNLP (Systems Demonstrations), pages 9–14
November 16-20, 2020.

prove the performance of the pre-trained language models on the downstream tasks.

- We publicly release our model under the name BERTweet which can be used with `fairseq` (Ott et al., 2019) and `transformers` (Wolf et al., 2019). We hope that BERTweet can serve as a strong baseline for future research and applications of Tweet analytic tasks.

2 BERTweet

In this section, we outline the architecture, and describe the pre-training data and optimization setup that we use for BERTweet.

Architecture

Our BERTweet uses the same architecture as BERT$_{\text{base}}$, which is trained with a masked language modeling objective (Devlin et al., 2019). BERTweet pre-training procedure is based on RoBERTa (Liu et al., 2019) which optimizes the BERT pre-training approach for more robust performance. Given the widespread usage of BERT and RoBERTa, we do not detail the architecture here. See Devlin et al. (2019) and Liu et al. (2019) for more details.

Pre-training data

We use an 80GB pre-training dataset of uncompressed texts, containing 850M Tweets (16B word tokens). Here, each Tweet consists of at least 10 and at most 64 word tokens. In particular, this dataset is a concatenation of two corpora:

- We first download the general Twitter Stream grabbed by the Archive Team,[1] containing 4TB of Tweet data streamed from 01/2012 to 08/2019 on Twitter. To identify English Tweets, we employ the language identification component of fastText (Joulin et al., 2017). We tokenize those English Tweets using "TweetTokenizer" from the NLTK toolkit (Bird et al., 2009) and use the `emoji` package to translate emotion icons into text strings (here, each icon is referred to as a word token).[2] We also normalize the Tweets by converting user mentions and web/url links into special tokens `@USER` and `HTTPURL`, respectively. We filter out retweeted Tweets and the ones shorter than 10 or longer than 64 word tokens. This pre-process results in the first corpus of 845M English Tweets.

- We also stream Tweets related to the COVID-19 pandemic, available from 01/2020 to 03/2020.[3] We apply the same data pre-process step as described above, thus resulting in the second corpus of 5M English Tweets.

We then apply `fastBPE` (Sennrich et al., 2016) to segment all 850M Tweets with subword units, using a vocabulary of 64K subword types. On average there are 25 subword tokens per Tweet.

Optimization

We utilize the RoBERTa implementation in the `fairseq` library (Ott et al., 2019). We set a maximum sequence length at 128, thus generating 850M × 25 / 128 ≈ 166M sequence blocks. Following Liu et al. (2019), we optimize the model using Adam (Kingma and Ba, 2014), and use a batch size of 7K across 8 V100 GPUs (32GB each) and a peak learning rate of 0.0004. We pre-train BERTweet for 40 epochs in about 4 weeks (here, we use the first 2 epochs for warming up the learning rate), equivalent to 166M × 40 / 7K ≈ 950K training steps.

3 Experimental setup

We evaluate and compare the performance of BERTweet with strong baselines on three downstream NLP tasks of POS tagging, NER and text classification, using benchmark Tweet datasets.

Downstream task datasets

For POS tagging, we use three datasets Ritter11-T-POS (Ritter et al., 2011), ARK-Twitter[4] (Gimpel et al., 2011; Owoputi et al., 2013) and TWEEBANK-V2[5] (Liu et al., 2018). For NER, we employ datasets from the WNUT16 NER shared task (Strauss et al., 2016) and the WNUT17 shared task on novel and emerging entity recognition (Derczynski et al., 2017). For text classification, we employ the 3-class sentiment analysis dataset from the SemEval2017 Task 4A (Rosenthal et al., 2017) and the 2-class irony detection dataset from the SemEval2018 Task 3A (Van Hee et al., 2018).

For Ritter11-T-POS, we employ a 70/15/15 training/validation/test pre-split available from Gui et al. (2017).[6] ARK-Twitter contains two

[1] https://archive.org/details/twitterstream

[2] https://pypi.org/project/emoji

[3] We collect Tweets containing at least one of 11 COVID-19 related keywords, e.g. covid19, coronavirus, sars-cov-2.

[4] https://code.google.com/archive/p/ark-tweet-nlp/downloads (twpos-data-v0.3.tgz)

[5] https://github.com/Oneplus/Tweebank

[6] https://github.com/guitaowufeng/TPANN

files daily547.conll and oct27.conll in which oct27.conll is further split into files oct27.traindev and oct27.test. Following Owoputi et al. (2013) and Gui et al. (2017), we employ daily547.conll as a test set. In addition, we use oct27.traindev and oct27.test as training and validation sets, respectively. For the TWEEBANK-V2, WNUT16 and WNUT17 datasets, we use their existing training/validation/test split. The SemEval2017-Task4A and SemEval2018-Task3A datasets are provided with training and test sets only (i.e. there is not a standard split for validation), thus we sample 10% of the training set for validation and use the remaining 90% for training.

We use a "soft" normalization strategy to all of the experimental datasets by translating word tokens of user mentions and web/url links into special tokens @USER and HTTPURL, respectively, and converting emotion icon tokens into corresponding strings. We also apply a "hard" strategy by further applying lexical normalization dictionaries (Aramaki, 2010; Liu et al., 2012; Han et al., 2012) to normalize word tokens in Tweets.

Fine-tuning

Following Devlin et al. (2019), for POS tagging and NER, we append a linear prediction layer on top of the last Transformer layer of BERTweet with regards to the first subword of each word token, while for text classification we append a linear prediction layer on top of the pooled output.

We employ the transformers library (Wolf et al., 2019) to independently fine-tune BERTweet for each task and each dataset in 30 training epochs. We use AdamW (Loshchilov and Hutter, 2019) with a fixed learning rate of 1.e-5 and a batch size of 32 (Liu et al., 2019). We compute the task performance after each training epoch on the validation set (here, we apply early stopping when no improvement is observed after 5 continuous epochs), and select the best model checkpoint to compute the performance score on the test set.

We repeat this fine-tuning process 5 times with different random seeds, i.e. 5 runs for each task and each dataset. We report each final test result as an average over the test scores from the 5 runs.

Baselines

Our main competitors are the pre-trained language models RoBERTa$_{\text{base}}$ (Liu et al., 2019) and XLM-R$_{\text{base}}$ (Conneau et al., 2020), which have the same architecture configuration as our BERTweet. In addition, we also evaluate the pre-trained RoBERTa$_{\text{large}}$ and XLM-R$_{\text{large}}$ although it is not a fair comparison due to their significantly larger model configurations.

	Model	Ritter11		ARK		TB-v2	
		soft	hard	soft	hard	soft	hard
Our results	RoBERTa$_{\text{large}}$	91.7	91.5	93.7	93.2	94.9	94.6
	XLM-R$_{\text{large}}$	92.6	92.1	94.2	93.8	95.5	95.1
	RoBERTa$_{\text{base}}$	88.7	88.3	91.8	91.6	93.7	93.5
	XLM-R$_{\text{base}}$	**90.4**	**90.3**	92.8	92.6	94.7	94.3
	BERTweet	90.1	89.5	**94.1**	**93.4**	**95.2**	**94.7**
	DCNN (Gui et al.)	89.9		–		–	
	DCNN (Gui et al.)	91.2 [+a]		92.4 [+a+b]		–	
	TPANN	90.9 [+a]		92.8 [+a+b]		–	
	ARKtagger	90.4		93.2 [+b]		94.6 [+c]	
	BiLSTM-CNN-CRF	–		–		92.5 [+c]	

Table 1: POS tagging accuracy results on the Ritter11-T-POS (Ritter11), ARK-Twitter (ARK) and TWEEBANK-V2 (TB-v2) test sets. Result of ARKtagger (Owoputi et al., 2013) on Ritter11 is reported in the TPANN paper (Gui et al., 2017). Note that Ritter11 uses Twitter-specific POS tags for retweeted (RT), user-account, hashtag and url word tokens which can be tagged perfectly using some simple regular expressions. Therefore, we follow Gui et al. (2017) and Gui et al. (2018) to tag those words appropriately for all models. Results of ARKtagger and BiLSTM-CNN-CRF (Ma and Hovy, 2016) on TB-v2 are reported by Liu et al. (2018). Also note that "+a", "+b" and "+c" denote the additional use of extra training data, i.e. models trained on bigger training data. "+a": additional use of the POS annotated data from the English WSJ Penn treebank sections 00-24 (Marcus et al., 1993). "+b": the use of both training and validation sets for learning models. "+c": additional use of the POS annotated data from the UD_English-EWT training set (Silveira et al., 2014).

The pre-trained RoBERTa is a strong language model for English, learned from 160GB of texts covering books, Wikipedia, CommonCrawl news, CommonCrawl stories, and web text contents. XLM-R is a cross-lingual variant of RoBERTa, trained on a 2.5TB multilingual corpus which contains 301GB of English CommonCrawl texts.

We fine-tune RoBERTa and XLM-R using the same fine-tuning approach we use for BERTweet.

4 Experimental results

Main results

Tables 1, 2, 3 and 4 present our obtained scores for BERTweet and baselines regarding both "soft" and "hard" normalization strategies. We find that for

	Model	WNUT16		WNUT17			
		soft	hard	entity		surface	
				soft	hard	soft	hard
Our results	RoBERTa$_{\text{large}}$	55.4	54.8	56.9	57.0	55.6	55.6
	XLM-R$_{\text{large}}$	55.8	55.3	57.1	57.5	55.9	56.4
	RoBERTa$_{\text{base}}$	49.7	49.2	52.2	52.0	51.2	51.0
	XLM-R$_{\text{base}}$	49.9	49.4	53.5	53.0	51.9	51.6
	BERTweet	**52.1**	**51.3**	**56.5**	**55.6**	**55.1**	**54.1**
	CambridgeLTL	52.4 [+b]		–		–	
	DATNet (Zhou et al.)	53.0 [+b]		42.3		–	
	Aguilar et al. (2017)	–		41.9		40.2	

Table 2: F1 scores on the WNUT16 and WNUT17 test sets. CambridgeLTL result is reported by Limsopatham and Collier (2016). "entity" and "surface" denote the scores computed for the standard entity level and the surface level (Derczynski et al., 2017), respectively.

	Model	AvgRec		F_1^{NP}		Accuracy	
		soft	hard	soft	hard	soft	hard
Our results	RoBERTa$_{\text{large}}$	72.5	72.2	72.0	71.8	70.7	71.3
	XLM-R$_{\text{large}}$	71.7	71.7	71.1	70.9	70.7	70.6
	RoBERTa$_{\text{base}}$	71.6	71.8	71.2	71.2	71.6	70.9
	XLM-R$_{\text{base}}$	70.3	70.3	69.4	69.6	69.3	69.7
	BERTweet	**73.2**	**72.8**	**72.8**	**72.5**	**71.7**	**72.0**
	Cliche (2017)	68.1		68.5		65.8	
	Baziotis et al. (2017)	68.1		67.7		65.1	

Table 3: Performance scores on the SemEval2017-Task4A test set. See Rosenthal et al. (2017) for the definitions of the AvgRec and F_1^{NP} metrics, in which AvgRec is the main ranking metric.

	Model	F_1^{pos}		Accuracy	
		soft	hard	soft	hard
Our results	RoBERTa$_{\text{large}}$	73.2	71.9	76.5	75.1
	XLM-R$_{\text{large}}$	70.8	69.7	74.2	73.2
	RoBERTa$_{\text{base}}$	71.0	71.2	74.0	74.0
	XLM-R$_{\text{base}}$	66.6	66.2	70.8	70.8
	BERTweet	**74.6**	**74.3**	**78.2**	**78.2**
	Wu et al. (2018)	70.5		73.5	
	Baziotis et al. (2018)	67.2		73.2	

Table 4: Performance scores on the SemEval2018-Task3A test set. F_1^{pos}—the main ranking metric—denotes the F_1 score computed for the positive label.

each pre-trained language model the "soft" scores are generally higher than the corresponding "hard" scores, i.e. applying lexical normalization dictionaries to normalize word tokens in Tweets generally does not help improve the performance of the pre-trained language models on downstream tasks.

Our BERTweet outperforms its main competitors RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$ on all experimental datasets (with only one exception that XLM-R$_{\text{base}}$ does slightly better than BERTweet on Ritter11-T-POS). Compared to RoBERTa$_{\text{large}}$ and XLM-R$_{\text{large}}$ which use significantly larger model configurations, we find that they obtain better POS tagging and NER scores than BERTweet. However, BERTweet performs better than those large models on the two text classification datasets.

Tables 1, 2, 3 and 4 also compare our obtained scores with the previous highest reported results on the same test sets. Clearly, the pre-trained language models help achieve new SOTA results on all experimental datasets. Specifically, BERTweet improves the previous SOTA in the novel and emerging entity recognition by absolute 14+% on the WNUT17 dataset, and in text classification by 5% and 4% on the SemEval2017-Task4A and SemEval2018-Task3A test sets, respectively. Our results confirm the effectiveness of the large-scale BERTweet for Tweet NLP.

Discussion

Our results comparing the "soft" and "hard" normalization strategies with regards to the pre-trained language models confirm the previous view that lexical normalization on Tweets is a lossy translation task (Owoputi et al., 2013). We find that RoBERTa outperforms XLM-R on the text classification datasets. This finding is similar to what is found in the XLM-R paper (Conneau et al., 2020) where XLM-R obtains lower performance scores than RoBERTa for sequence classification tasks on traditional written English corpora.

We also recall that although RoBERTa and XLM-R use 160 / 80 = 2 times and 301 / 80 ≈ 3.75 times bigger English data than our BERTweet, respectively, BERTweet does better than its competitors RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$. Thus this confirms the effectiveness of a large-scale and domain-specific pre-trained language model for English Tweets. In future work, we will release a "large" version of BERTweet, which possibly performs better than RoBERTa$_{\text{large}}$ and XLM-R$_{\text{large}}$ on all three evaluation tasks.

5 Conclusion

We have presented the first large-scale language model BERTweet pre-trained for English Tweets. We demonstrate the usefulness of BERTweet by showing that BERTweet outperforms its baselines RoBERTa$_{\text{base}}$ and XLM-R$_{\text{base}}$ and helps produce better performances than the previous SOTA models for three downstream Tweet NLP tasks of POS tagging, NER, and text classification (i.e. sentiment analysis & irony detection).

As of September 2020, we have collected a corpus of about 23M "cased" COVID-19 English Tweets consisting of at least 10 and at most 64 word tokens. In addition, we also create an "uncased" version of this corpus. Then we continue pre-training from our pre-trained BERTweet on each of the "cased" and "uncased" corpora of 23M Tweets for 40 additional epochs, resulting in two BERTweet variants of pre-trained "cased" and "uncased" *BERTweet-COVID19* models, respectively. By publicly releasing BERTweet and its two variants, we hope that they can foster future research and applications of Tweet analytic tasks, such as identifying informative COVID-19 Tweets (Nguyen et al., 2020) or extracting COVID-19 events from Tweets (Zong et al., 2020).

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A Multi-task Approach for Named Entity Recognition in Social Media Data. In *Proceedings of WNUT*, pages 148–153.

Eiji Aramaki. 2010. TYPO CORPUS. http://luululu.com/tweet/.

Christos Baziotis, Athanasiou Nikolaos, Pinelopi Papalampidi, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive RNNs. In *Proceedings of SemEval*, pages 613–621.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of SemEval*, pages 747–754.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*, pages 3615–3620.

Steven Bird, Ewan Klein, and Edward Loper, editors. 2009. *Natural language processing with Python*. O'Reilly.

Mathieu Cliche. 2017. BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. In *Proceedings of SemEval*, pages 573–580.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*, page to appear.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings of WNUT*, pages 140–147.

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

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. 2019. Social media big data analytics: A survey. *Comput. Hum. Behav.*, 101:417–428.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of ACL-HLT*, pages 42–47.

Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, Di Liang, Keyu Ding, and Xuanjing Huang. 2018. Transferring from Formal Newswire Domain with Hypernet for Twitter POS Tagging. In *Proceedings of EMNLP*, pages 2540–2549.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. 2017. Part-of-Speech Tagging for Twitter with Adversarial Neural Networks. In *Proceedings of EMNLP*, pages 2411–2420.

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 *Proceedings of ACL*, pages 8342–8360.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically Constructing a Normalisation Dictionary for Microblogs. In *Proceedings of EMNLP-CoNLL*, pages 421–432.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical Normalization for Social Media Text. *ACM Transactions on Intelligent Systems and Technology*, 4(1).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of EACL*, pages 427–431.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint*, arXiv:1412.6980.

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*, page btz682.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for Named Entity Recognition in Twitter Messages. In *Proceedings of WNUT*, pages 145–152.

Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A Broad-Coverage Normalization System for Social Media Language. In *Proceedings of ACL*, pages 1035–1044.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing Tweets into Universal Dependencies. In *Proceedings of NAACL-HLT*, pages 965–975.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*, pages 1064–1074.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of WNUT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, pages 48–53.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of EMNLP*, pages 1524–1534.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of SemEval*, pages 502–518.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, pages 1715–1725.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of LREC*.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 Named Entity Recognition Shared Task. In *Proceedings of WNUT*, pages 138–144.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of SemEval*, pages 39–50.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*, arXiv:1910.03771.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of SemEval*, pages 51–56.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition. In *Proceedings of ACL*, pages 3461–3471.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. Extracting COVID-19 Events from Twitter. *arXiv preprint*, arXiv:2006.02567.