INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization

Fajri Koto   **Jey Han Lau**   **Timothy Baldwin**
School of Computing and Information Systems
The University of Melbourne
ffajri@student.unimelb.edu.au, jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

We present INDOBERTWEET, the first large-scale pretrained model for Indonesian Twitter that is trained by extending a monolingually-trained Indonesian BERT model with additive domain-specific vocabulary. We focus in particular on efficient model adaptation under vocabulary mismatch, and benchmark different ways of initializing the BERT embedding layer for new word types. We find that initializing with the average BERT subword embedding makes pretraining five times faster, and is more effective than proposed methods for vocabulary adaptation in terms of extrinsic evaluation over seven Twitter-based datasets.[1]

1 Introduction

Transformer-based pretrained language models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019) have become the backbone of modern NLP systems, due to their success across various languages and tasks. However, obtaining high-quality contextualized representations for specific domains/data sources such as biomedical, social media, and legal, remains a challenge.

Previous studies (Alsentzer et al., 2019; Chalkidis et al., 2020; Nguyen et al., 2020) have shown that for domain-specific text, pretraining from scratch outperforms off-the-shelf BERT. As an alternative approach with lower cost, Gururangan et al. (2020) demonstrated that domain adaptive pretraining (i.e. pretraining the model on target domain text before task fine-tuning) is effective, although still not as good as training from scratch.

The main drawback of domain-adaptive pretraining is that domain-specific words that are not in the pretrained vocabulary are often tokenized poorly. For instance, in BIOBERT (Lee et al., 2019), *Immunoglobulin* is tokenized into {*I*, *##mm*, *##uno*, *##g*, *##lo*, *##bul*, *##in*}, despite being a common

term in biology. To tackle this problem, Poerner et al. (2020); Tai et al. (2020) proposed simple methods to domain-extend the BERT vocabulary: Poerner et al. (2020) initialize new vocabulary using a learned projection from word2vec (Mikolov et al., 2013), while Tai et al. (2020) use random initialization with weight augmentation, substantially increasing the number of model parameters.

New vocabulary augmentation has been also conducted for language-adaptive pretraining, mainly based on multilingual BERT (MBERT). For instance, Chau et al. (2020) replace 99 "unused" WordPiece tokens of MBERT with new common tokens in the target language, while Wang et al. (2020) extend MBERT vocabulary with non-overlapping tokens ($|\mathbb{V}_{\text{MBERT}} - \mathbb{V}_{new}|$). These two approaches use random initialization for new WordPiece token embeddings.

In this paper, we focus on the task of learning an Indonesian BERT model for Twitter, and show that initializing domain-specific vocabulary with average-pooling of BERT subword embeddings is more efficient than pretraining from scratch, and more effective than initializing based on word2vec projections (Poerner et al., 2020). We use IN-DOBERT (Koto et al., 2020b), a monolingual BERT for Indonesian as the domain-general model to develop a pretrained domain-specific model IN-DOBERTWEET for Indonesian Twitter.

There are two primary reasons to experiment with Indonesian Twitter. First, despite being the official language of the 5th most populous nation, Indonesian is underrepresented in NLP (notwithstanding recent Indonesian benchmarks and datasets (Wilie et al., 2020; Koto et al., 2020a,b)). Second, with a large user base, Twitter is often utilized to support policymakers, business (Fiarni et al., 2016), or to monitor elections (Suciati et al., 2019) or health issues (Prastyo et al., 2020). Note that most previous studies that target Indonesian Twitter tend to use traditional machine learning models

---

[1]Code and models can be accessed at https://github.com/indolem/IndoBERTweet

(e.g. n-gram and recurrent models (Fiarni et al., 2016; Koto and Rahmaningtyas, 2017)).

To summarize our contributions: (1) we release INDOBERTWEET, the first large-scale pretrained Indonesian language model for social media data; and (2) through extensive experimentation, we compare a range of approaches to domain-specific vocabulary initialization over a domain-general BERT model, and find that a simple average of sub-word embeddings is more effective than previously-proposed methods and reduces the overhead for domain-adaptive pretraining by 80%.

2 INDOBERTWEET

2.1 Twitter Dataset

We crawl Indonesian tweets over a 1-year period using the official Twitter API,[2] from December 2019 to December 2020, with 60 keywords covering 4 main topics: economy, health, education, and government. We found that the Twitter language identifier is reasonably accurate for Indonesian, and so use it to filter out non-Indonesian tweets. From 100 randomly-sampled tweets, we found a majority of them (87) to be Indonesian, with a small number being Malay (12) and Swahili (1).[3]

After removing redundant tweets (with the same ID), we obtain 26M tweets with 409M word tokens, two times larger than the training data used to pre-train INDOBERT (Koto et al., 2020b). We set aside 230K tweets for development, and extract a vocabulary of 31,984 types based on WordPiece (Wu et al., 2016). We lower-case all words and follow the same preprocessing steps as English BERTWEET (Nguyen et al., 2020): (1) converting user mentions and URLs into `@USER` and `HTTPURL`, respectively; and (2) translating emoticons into text using the `emoji` package.[4]

2.2 INDOBERTWEET Model

INDOBERTWEET is trained based on a masked language model objective (Devlin et al., 2019) following the same procedure as the `indobert-base-uncased` (INDOBERT) model.[5] It is a transformer encoder with 12 hidden layers (dimension=768), 12 attention heads, and 3

feed-forward hidden layers (dimension=3,072). The only difference is the maximum sequence length, which we set to 128 tokens based on the average number of words per document in our Twitter corpus.

In this work, we train 5 INDOBERTWEET models. The first model is pretrained from scratch based on the aforementioned configuration. The remaining four models are based on domain-adaptive pretraining with different vocabulary adaptation strategies, as discussed in Section 2.3.

2.3 Domain-Adaptive Pretraining with Domain-Specific Vocabulary Initialization

We apply domain-adaptive pretraining on the domain-general INDOBERT (Koto et al., 2020b), which is trained over Indonesian Wikipedia, news articles, and an Indonesian web corpus (Medved and Suchomel, 2017). Our goal is to fully replace INDOBERT's vocabulary ($\mathbb{V}_{IB}$) of 31,923 types with INDOBERTWEET's vocabulary ($\mathbb{V}_{IBT}$) (31,984 types). In INDOBERTWEET, there are 14,584 (46%) new types, and 17,400 (54%) Word-Piece types which are shared with INDOBERT.[6]

To initialize the domain-specific vocabulary, we use INDOBERT embeddings for the 17,400 shared types, and explore four initialization strategies for new word types: (1) random initialization from $U(-1, 1)$; (2) random initialization from $\mathcal{N}(\mu, \sigma)$, where μ and σ are learned from INDOBERT embeddings; (3) linear projection via `fastText` embeddings (Poerner et al., 2020); and (4) averaging INDOBERT subword embeddings.

For the linear projection strategy (Method 3), we train 300d `fastText` embeddings (Bojanowski et al., 2017) over the tokenized Indonesian Twitter corpus. Following Poerner et al. (2020), we use the shared types ($\mathbb{V}_{IB} \cap \mathbb{V}_{IBT}$) to train a linear transformation from `fastText` embeddings E_{FT} to INDOBERT embeddings E_{IB} as follows:

$$\underset{\mathbf{W}}{\arg\min} \sum_{x \in \mathbb{V}_{IB} \cap \mathbb{V}_{IBT}} \|E_{FT}(x)\,\mathbf{W} - E_{IB}(x)\|_2^2$$

where $\mathbf{W}$ is a $\dim(E_{FT}) \times \dim(E_{IB})$ matrix.

To average subword embeddings of $x \in \mathbb{V}_{IBT}$

Task	Data	#labels	#train	#dev	#test	5-Fold	Evaluation
Sentiment Analysis	IndoLEM (Koto et al., 2020b)	2	3,638	399	1,011	Yes	$F1_{pos}$
	SmSA (Wilie et al., 2020)	3	11,000	1,260	500	No	$F1_{macro}$
Emotion Classification	EmoT (Wilie et al., 2020)	5	3,521	440	442	No	$F1_{macro}$
Hate Speech Detection	HS1 (Alfina et al., 2017)	2	499	72	142	Yes	$F1_{pos}$
	HS2 (Ibrohim and Budi, 2019)	2	9,219	2,633	1,317	Yes	$F1_{pos}$
Named Entity Recognition	Formal (Munarko et al., 2018)	3	6,500	657	1,122	No	$F1_{entity}$
	Informal (Munarko et al., 2018)	3	6,500	657	1,227	No	$F1_{entity}$

Table 1: Summary of Indonesian Twitter datasets used in our experiments.

(Method 4), we compute:

$$E_{IBT}(x) = \frac{1}{|T_{IB}(x)|} \sum_{y \in T_{IB}(x)} E_{IB}(y)$$

where $T_{IB}(x)$ is the set of WordPiece tokens for word x produced by INDOBERT's tokenizer.

3 Experimental Setup

We accumulate gradients over 4 steps to simulate a batch size of 2048. When pretraining from scratch, we train the model for 1M steps, and use a learning rate of $1e-4$ and the Adam optimizer with a linear scheduler. All pretraining experiments are done using $4 \times$ V100 GPUs (32GB).

For domain-adaptive pretraining (using IN-DOBERT model), we consider three benchmarks: (1) domain-adaptive pretraining without domain-specific vocabulary adaptation ($\mathbb{V}_{IBT} = \mathbb{V}_{IB}$) for 200K steps; (2) applying the new vocabulary adaptation approaches from Section 2.3 without additional domain-adaptive pretraining; and (3) applying the new vocabulary adaptation approaches from Section 2.3 with 200K domain-adaptive pretraining steps.

Downstream tasks. To evaluate the pretrained models, we use 7 Indonesian Twitter datasets, as summarized in Table 1. This includes sentiment analysis (Koto and Rahmaningtyas, 2017; Purwarianti and Crisdayanti, 2019), emotion classification (Saputri et al., 2018), hate speech detection (Alfina et al., 2017; Ibrohim and Budi, 2019), and named entity recognition (Munarko et al., 2018). For emotion classification, the classes are `fear`, `angry`, `sad`, `happy`, and `love`. Named entity recognition (NER) is based on the `PERSON`, `ORGANIZATION`, and `LOCATION` tags. NER has two test set partitions, where the first is formal texts (e.g. news snippets on Twitter) and the second is informal texts. The train and dev partitions are a

mixture of formal and informal tweets, and shared across the two test sets.

Fine-tuning. For sentiment, emotion, and hate speech classification, we add an MLP layer that takes the average pooled output of INDOBER-TWEET as input, while for NER we use the first subword of each word token for tag prediction. We pre-process the tweets as described in Section 2.1, and use a batch size of 30, maximum token length of 128, learning rate of $5e-5$, Adam optimizer with epsilon of $1e-8$, and early stopping with patience of 5. We additionally introduce a canonical split for both hate speech detection tasks with 5-fold cross validation, following Koto et al. (2020b). In Table 1, SmSA, EmoT, and NER use the original held-out evaluation splits.

Baselines. We use the two INDOBERT models from Koto et al. (2020b) and Wilie et al. (2020) as baselines, in addition to multilingual BERT (MBERT, which includes Indonesian) and a monolingual BERT for Malay (MALAYBERT).[7] Our rationale for including MALAYBERT is that we are interested in testing its performance on Indonesian, given that the two languages are closely related and we know that the Twitter training data includes some amount of Malay text.

4 Experimental Results

Table 2 shows the full results across the different pretrained models for the 7 Indonesian Twitter datasets. Note that the first four models are pretrained models without domain-adaptive pre-training (i.e. they are used as purely off-the-shelf models). In terms of baselines, MALAYBERT is a better model for Indonesian than MBERT, consistent with Koto et al. (2020b), and better again are the two different INDOBERT models at al-

---

[7] https://huggingface.co/huseinzol05/bert-base-bahasa-cased

10662

Model	Sentiment		Emotion	Hate Speech		NER		Average
	IndoLEM	SmSA	EmoT	HS1	HS2	Formal	Informal	
MBERT	76.6	84.7	67.5	85.1	75.1	85.2	83.2	79.6
MALAYBERT	82.0	84.1	74.2	85.0	81.9	81.9	81.3	81.5
INDOBERT (Wilie et al., 2020)	84.1	88.7	73.3	86.8	80.4	86.3	84.3	83.4
INDOBERT (Koto et al., 2020b)	84.1	87.9	71.0	86.4	79.3	88.0	86.9	83.4
INDOBERTWEET (1M steps)	86.2	90.4	76.0	**88.8**	**87.5**	88.1	85.4	86.1
INDOBERT (Koto et al., 2020b) + *200K steps of domain-adaptive pretraining*								
Same vocabulary ($\mathbb{V}_{IBT} = \mathbb{V}_{IB}$)	86.4	**92.7**	76.8	88.7	82.2	87.9	86.9	85.9
INDOBERT (Koto et al., 2020b) + *vocabulary adaptation + 0 steps of domain-adaptive pretraining*								
Uniform distribution	82.9	84.6	73.2	84.9	78.2	84.3	84.4	81.8
Normal distribution	83.5	86.7	71.1	85.2	77.4	85.0	86.3	82.2
fastText projection	84.4	83.6	72.2	85.5	80.9	85.4	85.6	82.5
Average of subwords	84.2	88.1	71.6	86.2	78.3	86.4	**87.4**	83.2
INDOBERT (Koto et al., 2020b) + *vocabulary adaptation + 200K steps of domain-adaptive pretraining*								
Uniform distribution	85.6	90.9	75.7	88.4	83.0	87.7	85.9	85.3
Normal distribution	87.1	92.5	75.4	**88.8**	82.5	**88.7**	86.6	85.9
fastText projection	86.4	89.7	78.5	88.7	84.4	88.0	86.6	86.0
Average of subwords	**86.6**	**92.7**	**79.0**	88.4	84.0	87.7	86.9	**86.5**

Table 2: A comparison of pretrained models with different adaptive pretraining strategies for Indonesian tweets (%).

most identical performance.[8] INDOBERTWEET — trained from scratch for 1M steps — results in a substantial improvement in terms of average performance (almost +3% absolute), consistent with previous findings that off-the-shelf domain-general pretrained models are sub-optimal for domain-specific tasks (Alsentzer et al., 2019; Chalkidis et al., 2020; Nguyen et al., 2020).

First, we pretrain INDOBERT (Koto et al., 2020b) *without* vocabulary adaptation for 200K steps, and find that the results are slightly lower than INDOBERTWEET. In the next set of experiments, we take INDOBERT (Koto et al., 2020b) and replace the domain-general vocabulary with the domain-specific vocabulary of INDOBERTWEET, without any pretraining ("0 steps"). Results drop overall relative to the original model, with the embedding averaging method ("Average of Subwords") yielding the smallest overall gap of −0.2% absolute.

Finally, we pretrain INDOBERT (Koto et al., 2020b) for 200K steps in the target domain, after performing vocabulary adaptation. We see a strong improvement for all initialization methods, with the embedding averaging method once again per-

forming the best, in fact outperforming the domain-specific INDOBERTWEET when trained for 1M steps from scratch. These findings reveal that we can adapt an off-the-shelf pretrained model very efficiently (5 times faster than training from scratch) *with better average performance*.

5 Discussion

Given these positive results on Indonesian, we conducted a similar experiment in a second language, English: we follow Nguyen et al. (2020) in adapting ROBERTA[9] for Twitter using the embedding averaging method to initialize new vocabulary, and compare ourselves against BERTWEET (trained from scratch on 845M English tweets).

A caveat here is that BERTWEET (Nguyen et al., 2020) and ROBERTA (Liu et al., 2019) use different tokenization methods: byte-level BPE vs. fastBPE (Sennrich et al., 2016). Because of this, rather than replacing ROBERTA's vocabulary with BERTWEET's (like our Indonesian experiments), we train ROBERTA's BPE tokenizer on English Twitter data (described below) to create a domain-specific vocabulary. This means that the two models (BERTWEET and domain-adapted ROBERTA

---

[8]Noting that Wilie et al. (2020)'s version includes 100M words of tweets for pretraining, but Koto et al. (2020b)'s version does not.

[9]The base version.

Model	Average
ROBERTA	72.9
BERTWEET (Nguyen et al., 2020)	**76.3**
ROBERTA + *vocabulary adaptation* + *200K steps of domain-adaptive pretraining*	74.1

Table 3: English results (%). The presented performance is averaged over 7 downstream tasks (Nguyen et al., 2020). Refer to the Appendix for details.

with modified vocabulary) will not be directly comparable.

Following Nguyen et al. (2020), we download 42M tweets from the Internet Archive[10] over the period July 2017 to October 2019 (the first two days of each month), which we use for domain-adaptive pretraining. Note that this pretraining data is an order of magnitude smaller than that of BERTWEET (42M vs. 845M). We use SpaCy[11] to filter English tweets, and follow the same preprocessing steps and downstream tasks as Nguyen et al. (2020) (7 tasks in total; see the Appendix for details). We pretrain ROBERTA for 200K steps using the embedding averaging method.

In Table 3, we see that BERTWEET outperforms ROBERTA (+3.4% absolute). With domain-adaptive pretraining using domain-specific vocabulary, the performance gap narrows to +2.2%, but are not as impressive as our Indonesian experiments. There are two reasons for this: (1) our domain-adaptive pretraining data is an order of magnitude smaller than for BERTWEET; and (2) the difference in tokenization methods between BERTWEET and ROBERTA results in a very different vocabulary.

Lastly, we argue that the different tokenization settings between INDOBERTWEET and BERTWEET (ours) may also contribute to the difference in results. The differences include: (1) uncased vs. cased; (2) WordPiece vs. `fastBPE` tokenizer; and (3) vocabulary size (32K vs. 50K) between both models. In Figure 1, we present the frequency distribution of #subword of new types in both models after tokenizing by each general-domain tokenizer. Interestingly, we find that BERTWEET has more new types than INDOBERTWEET, with #subword after tokenization being more varied (average length of #subword of
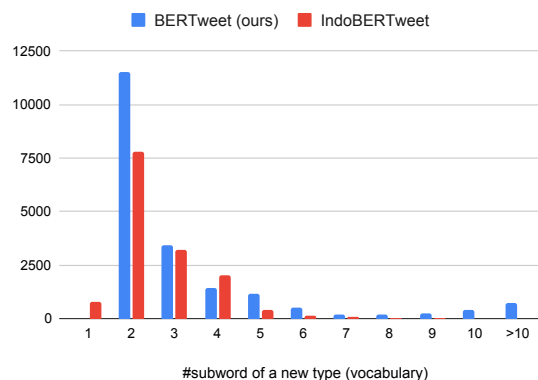
Figure 1: Frequency of #subword of new types in BERTWEET (ours) and INDOBERTWEET, tokenized by ROBERTA and INDOBERT tokenizers, respectively. #subword = 1 means the new type is tokenized as "[UNK]".

new types are 2.6 and 3.4 for INDOBERTWEET and BERTWEET, respectively).

6 Conclusion

We present the first large-scale pretrained model for Indonesian Twitter. We explored domain-adaptive pretraining with domain-specific vocabulary adaptation using several strategies, and found that the best method — averaging of subword embeddings from the original model — achieved the best average performance across 7 tasks, and is five times faster than the dominant paradigm of pretraining from scratch.

Acknowledgements

References

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

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 the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. 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.

Cut Fiarni, Herastia Maharani, and Rino Pratama. 2016. Sentiment analysis system for Indonesia online retail shop review using hierarchy naive Bayes technique. In *2016 4th international conference on information and communication technology (ICoICT)*, pages 1–6. IEEE.

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 the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fajri Koto and Gemala Y Rahmaningtyas. 2017. InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*, pages 391–394. IEEE.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. 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 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

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*.

Marek Medved and Vít Suchomel. 2017. Indonesian web corpus (idWac). In *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Yuda Munarko, MS Sutrisno, WAI Mahardika, Ilyas Nuryasin, and Yufis Azhar. 2018. Named entity recognition model for Indonesian tweet using CRF classifier. In *IOP Conference Series: Materials Science and Engineering*, volume 403. IOP Publishing.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Pulung Hendro Prastyo, Amin Siddiq Sumi, Ade Widyatama Dian, and Adhistya Erna Permanasari. 2020. Tweets responding to the Indonesian government's handling of COVID-19: Sentiment analysis using SVM with normalized poly kernel. *Journal of Information Systems Engineering and Business Intelligence*, 6(2):112–122.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving Bi-LSTM performance for Indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

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.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on Indonesian Twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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 the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Andi Suciati, Ari Wibisono, and Petrus Mursanto. 2019. Twitter buzzer detection for Indonesian presidential election. In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–5. IEEE.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 5998–6008.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey,

Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

A   Results with English BERTWEET

Model	POS Tagging			SemEval2017	SemEval2018	NER		Avg.
	Ritter11	ARK	TB-v2	Sentiment an.	Irony det.	WNUT2016	WNUT2017	
ROBERTA	88.6	91.0	93.4	70.8	71.6	45.5	49.9	72.9
BERTWEET	**90.5**	**93.2**	94.8	**72.6**	**75.5**	**51.9**	**55.9**	**76.3**
steps = 0								
ROBERTA w/ BERTWEET tokenizer	87.7	90.2	92.5	65.7	66.1	39.2	39.2	68.7
ROBERTA w/ new tokenizer	87.4	89.6	92.8	66.5	70.7	42.5	42.4	70.3
steps = 200K								
ROBERTA w/ new tokenizer	90.1	90.9	**94.9**	72.1	73.3	47.2	50.0	74.1

Table 4: English Results (%) over the test sets. All data, metrics, and splits are based off the experiments of Nguyen et al. (2020). We re-ran all experiments and found slightly lower performance for some models as compared to BERTWEET. For evaluation, the POS tagging datasets (Ritter et al., 2011; Gimpel et al., 2011; Liu et al., 2018) use accuracy, SemEval2017 (Rosenthal et al., 2017) uses Avg_{Rec}, SemEval2018 (Van Hee et al., 2018) uses F1_{pos}, and NER (Strauss et al., 2016; Derczynski et al., 2017) uses $\text{F1}_{\text{entity}}$.