

Indonesian Language Model

Using

Universal Language Model Fine-tuning for Text Classification
(ULMFiT)

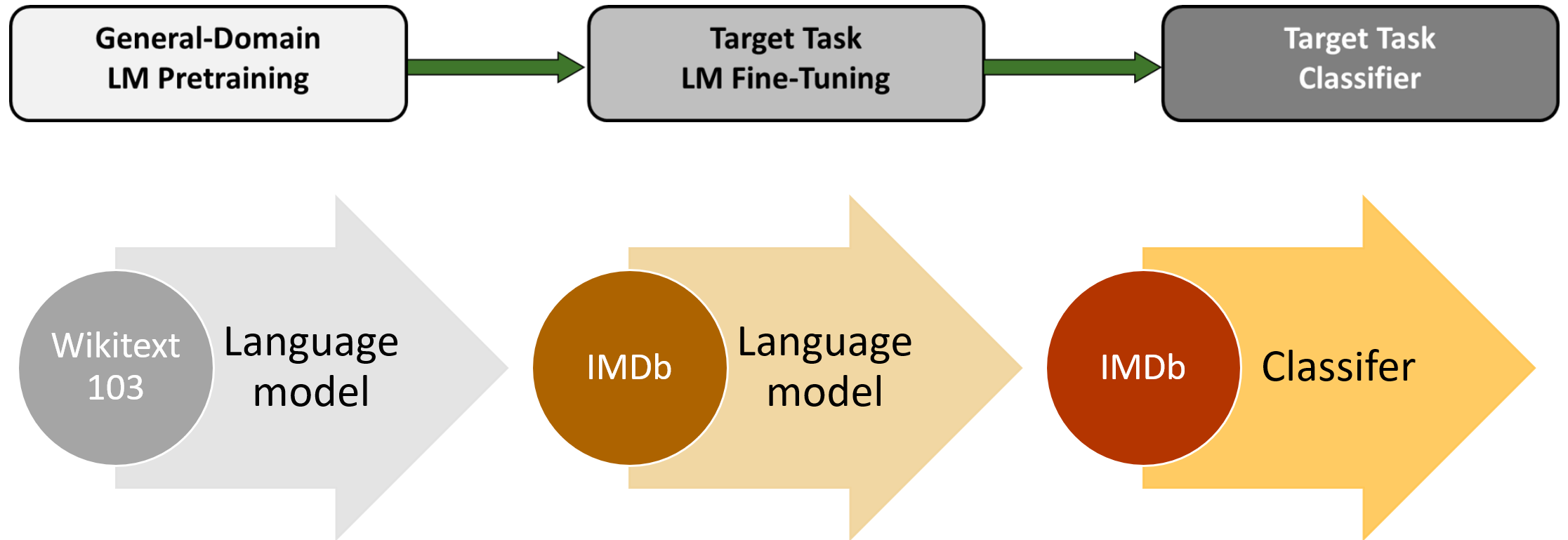
Language Model

- Predicting the next word based on previous words
 - Joko Widodo adalah ...
- One of main researches on NLP
- Applications: speech recognition, text classification, text recognition, machine translation, semantic extraction, etc...

Type of Language Model

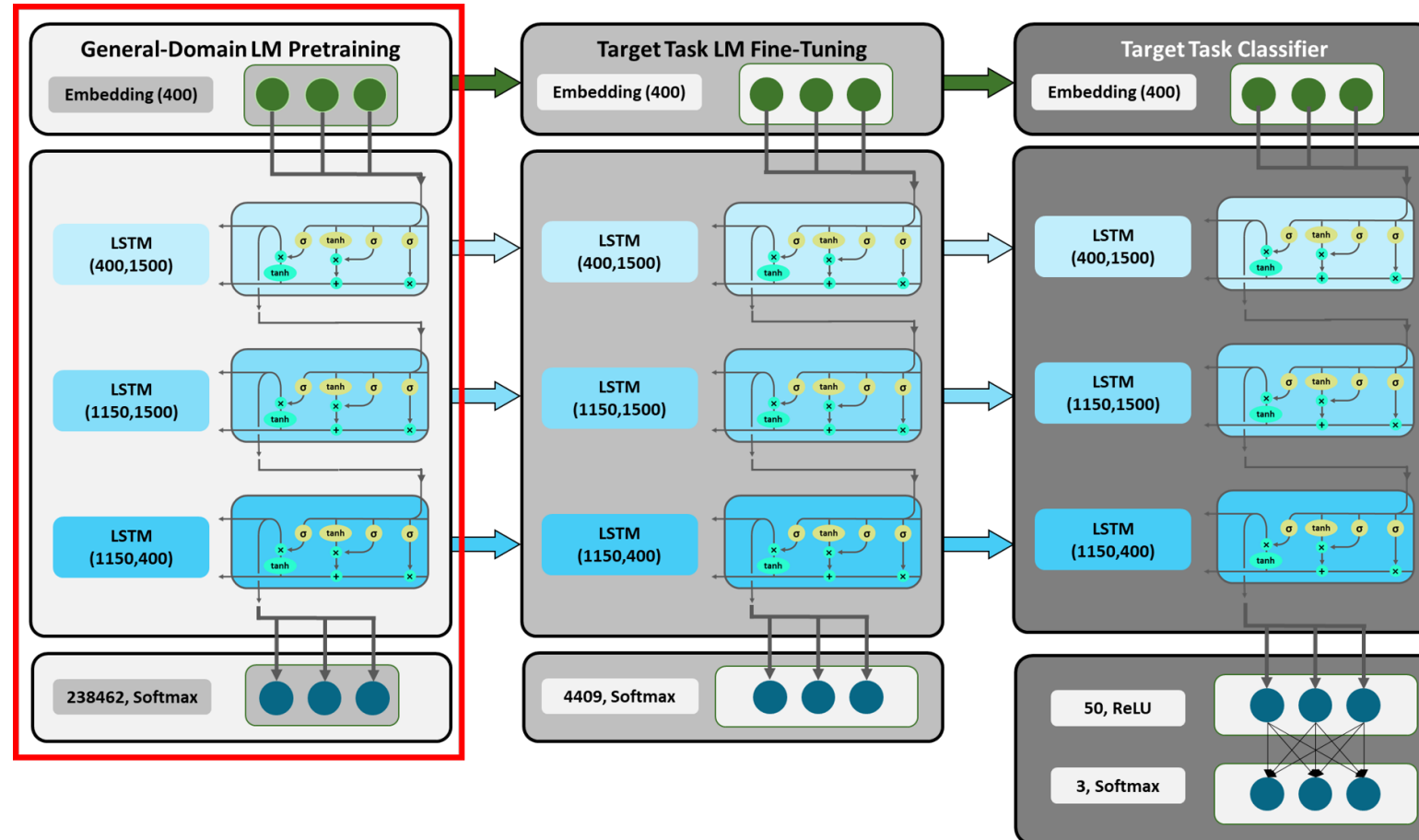
- N-Gram based LM
- Neural network based LM

Universal Language Model Fine-tuning for Text Classification (ULMFiT)

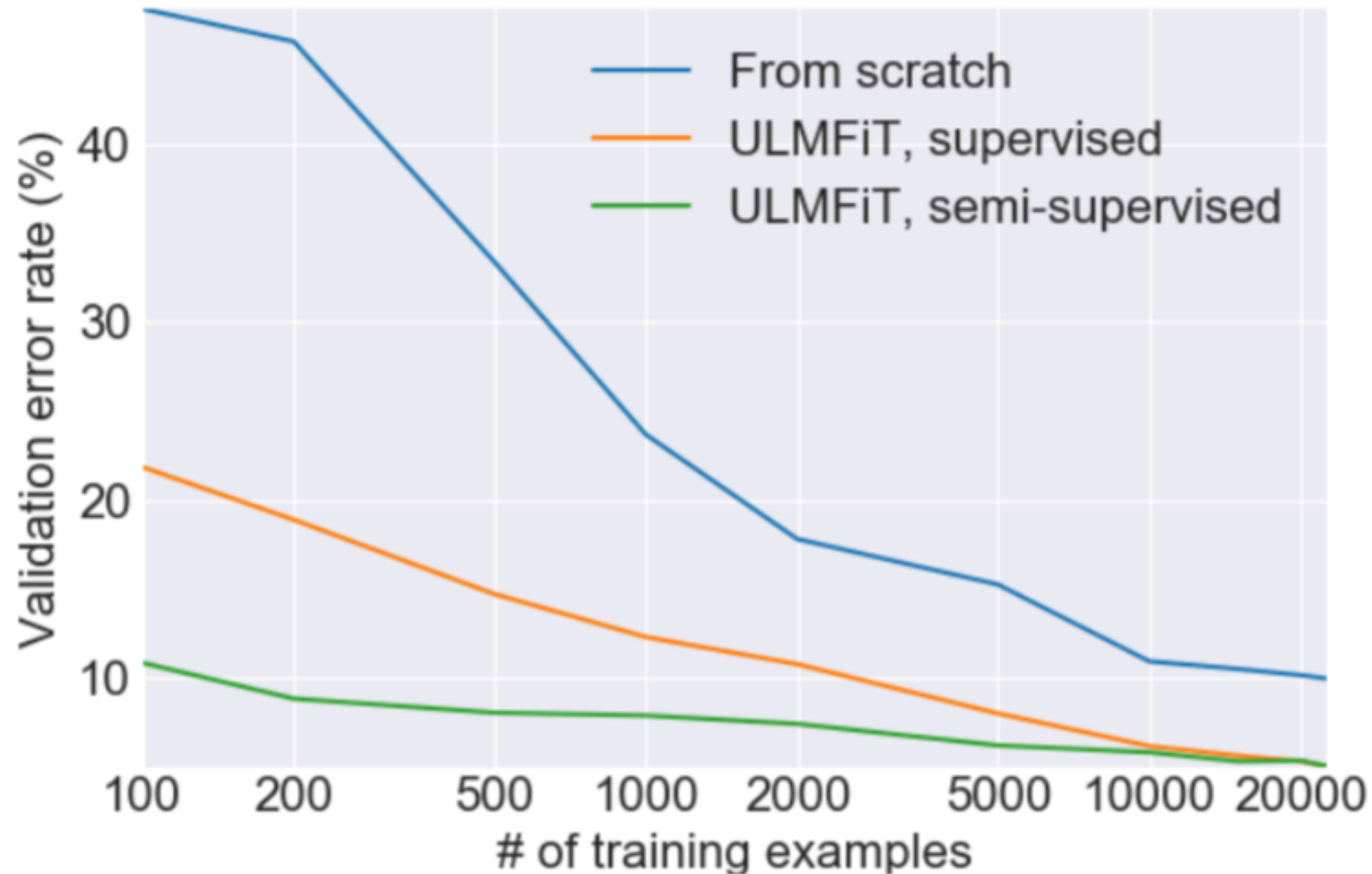


- Humboldt Universität zu Berlin: https://humboldt-wi.github.io/blog/research/information_systems_1819/group4_ulmfit/
- Fast AI: <https://fast.ai>

ULMFiT Architecture



ULMFiT on Text Classification



ULMFiT for Bahasa Indonesia

- Dataset: Indonesian Wikipedia
 - > 400K articles
 - Vocabulary size: 60K
- Perplexity: 27.67
 - Comparison:
 - 400 (?) on N-Gram Indonesian LM
 - 40.68 on English wikitext2 (Yang et al, 2017)
 - 29.2 on English wikitext103 (Rae et al, 2018)

ULMFiT for Bahasa Indonesia

Text Classification

- Dataset: Word Bahasa Indonesia Corpus and Parallel English Translation (BPPT) for PAN Localization
 - indonesian-english translation
 - 500K words
 - Categories:
 - Economy
 - International
 - Science
 - Sport

ULMFiT for Bahasa Indonesia

Text Classification

- Performance Comparison:

Name	Accuracy
NB, Count Vectors	0.9269
NB, WordLevel TF-IDF	0.9162
NB, N-Gram Vectors	0.7822
NB, CharLevel Vectors	0.8433
LC, Count Vectors	0.9265
LC, WordLevel TF-IDF	0.9179
LC, N-Gram Vectors	0.8085
LC, CharLevel Vectors	0.8888
SVM, N-Gram Vectors	0.7970
RF, Count Vectors	0.8392
RF, WordLevel TF-IDF	0.8338
Xgb, Count Vectors	0.8087
Xgb, WordLevel TF-IDF	0.8070
Xgb, CharLevel Vectors	0.8202
CNN	0.9263
Kim Yoon's CNN	0.9163
RNN-LSTM	0.9305
RNN-GRU	0.9296
Bidirectional RNN	0.9267
RCNN	0.9221
FasText(1)	0.8550
ULMFiT	0.9563

ULMFiT for Bahasa Indonesia

Text Generation

<https://lmzoo.uncool.ai/text-generator/>

Future Work for Indonesian Language Model

- Language Model based on OpenAI GPT2 or better
 - Dataset collections
 - Training on 100+ Google TPUs with the help from Tensorflow Research Cloud
<https://www.tensorflow.org/tfrc>

Thank You