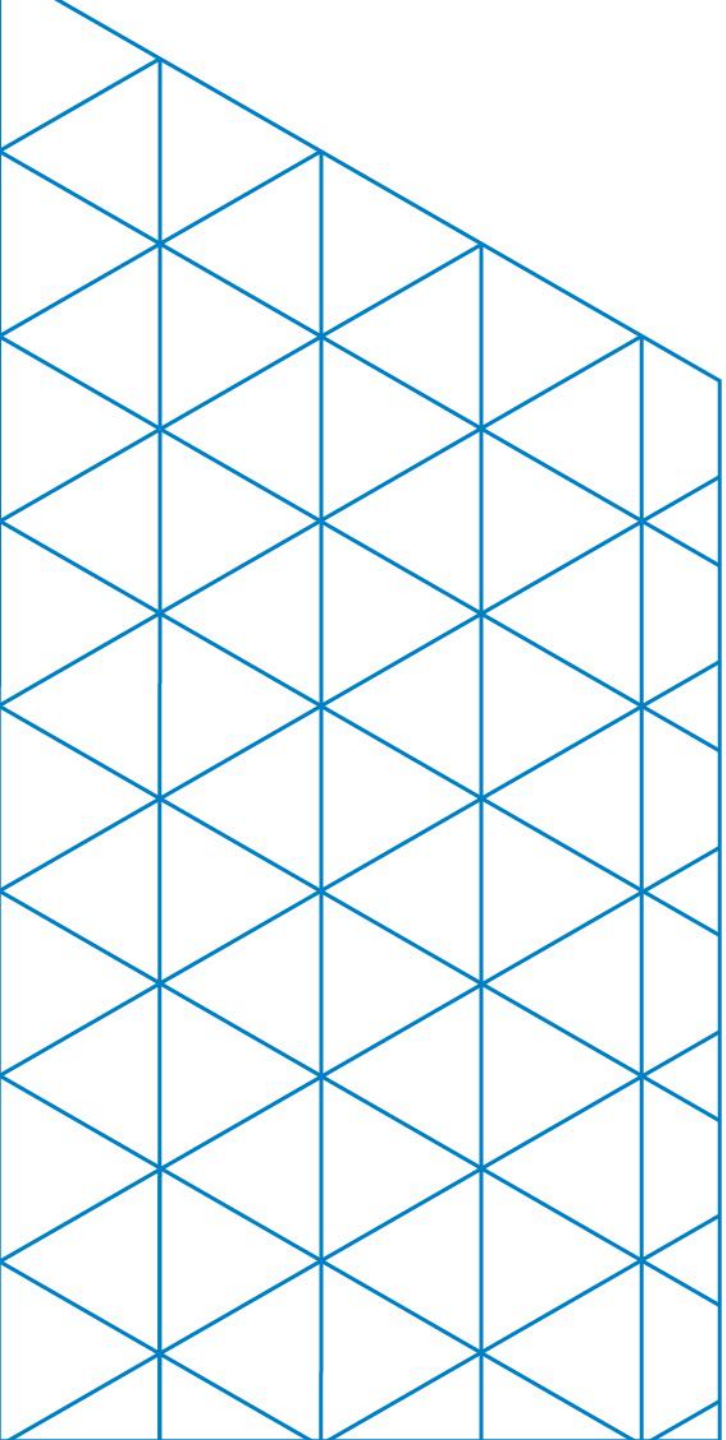


UFG
UNIVERSIDADE
FEDERAL DE GOIÁS

DEEP LEARNING
BRASIL NLP

MultiFiT:Multilingual Fine-Tuning

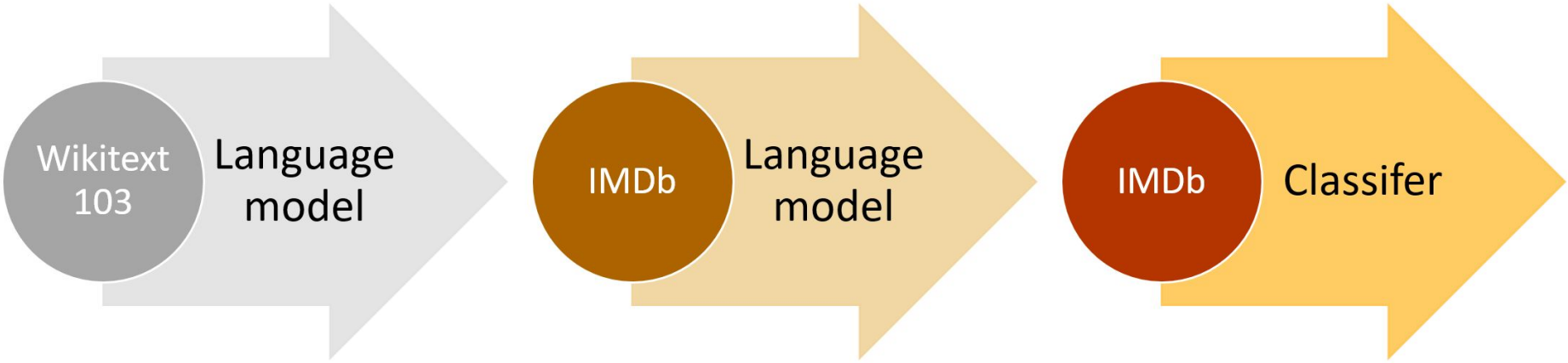
*Do **Paper** até a **Produção***

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Introdução

O MultiFiT é uma evolução do ULMFiT (Universal Language Model Fine-tuning for Text Classification).

ULMFiT



1. Pré-treino em um domínio genérico
2. Fine-tuning no domínio da tarefa
3. Fine-tuning na tarefa (classificação)

O Problema

Modelos de linguagens pré-treinados são promissores principalmente para línguas de baixo recurso, contudo treinar esse modelos requer um alto poder computacional.

As opções de modelos multilíngües pré-treinados geralmente apresentam baixo desempenho em linguagens de baixo recurso.

Abordagens cross-lingual

- Dados paralelos entre idiomas - ou seja, um corpus de documentos com exatamente o mesmo conteúdo, mas escritos em idiomas diferentes.
- Um vocabulário compartilhado - isto é, um vocabulário comum em vários idiomas. Essa abordagem substitui linguagens com muitos dados.

A proposta

- A ideia do MultFiT é uma democratização de modelos de linguagem, principalmente para línguas de baixo recurso.
- Isso é possível pois é facilmente treinável em qualquer língua além de ser possível um zero shot em córpus multilíngua.

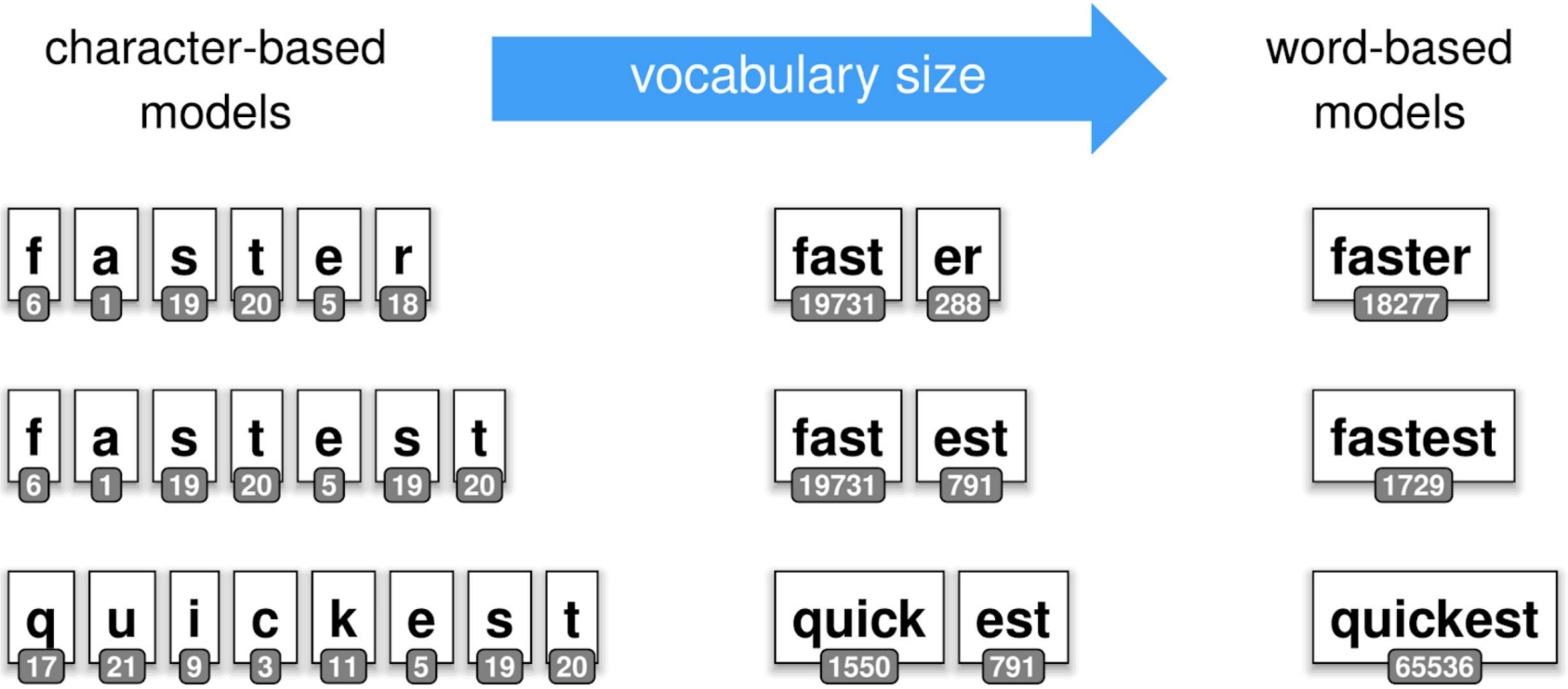
A proposta

- Ser possível ajustar modelos de idiomas monolíngues eficientes e competitivos com o BERT multilíngue, em muitos idiomas, em algumas centenas de exemplos.
- Ser muito mais barato de pré-treinar e mais eficiente em termos de complexidade de espaço e tempo.

Diferenças ULMFiT

1. Tokenização de sub-palavras.
2. Camadas QRNN ao invés das AWD-LSTMs originais.
3. Ciclo único.
4. Suavização de labels.

Tokenização de sub-palavras



UFG
UNIVERSIDADE FEDERAL DE GOIÁS

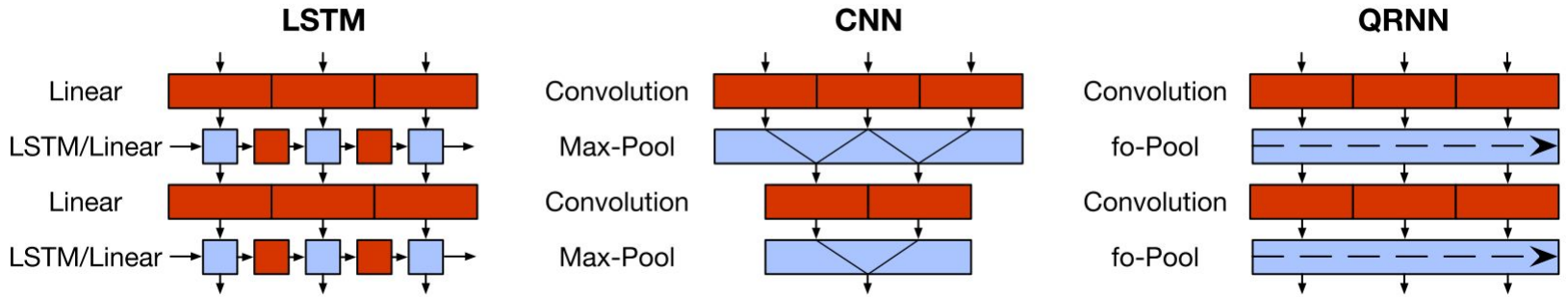
Tokenização de sub-palavras

1. As subpalavras representam mais facilmente inflexões, incluindo prefixos e sufixos comuns e, portanto, são adequadas para linguagens morfologicamente ricas.

2. A tokenização de subpalavras é uma boa opção para problemas de vocabulário aberto e elimina os tokens fora do vocabulário, pois a cobertura é próxima de 100%.

QRNN

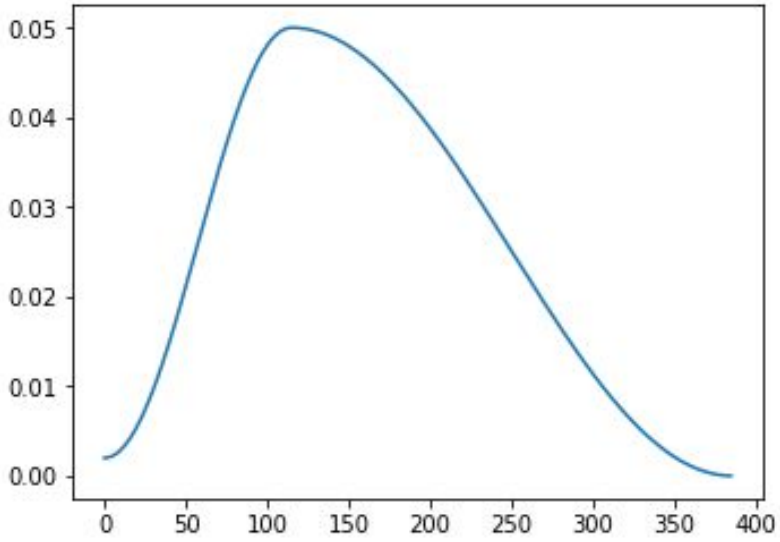
1. QRNN apresenta um balanço entre CNN e LSTM.
2. Apresenta paralelismo e minibatchs da CNN.
3. Também apresenta a estrutura sequencial e dependência de ordenação da LSTM.



UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Ciclo único

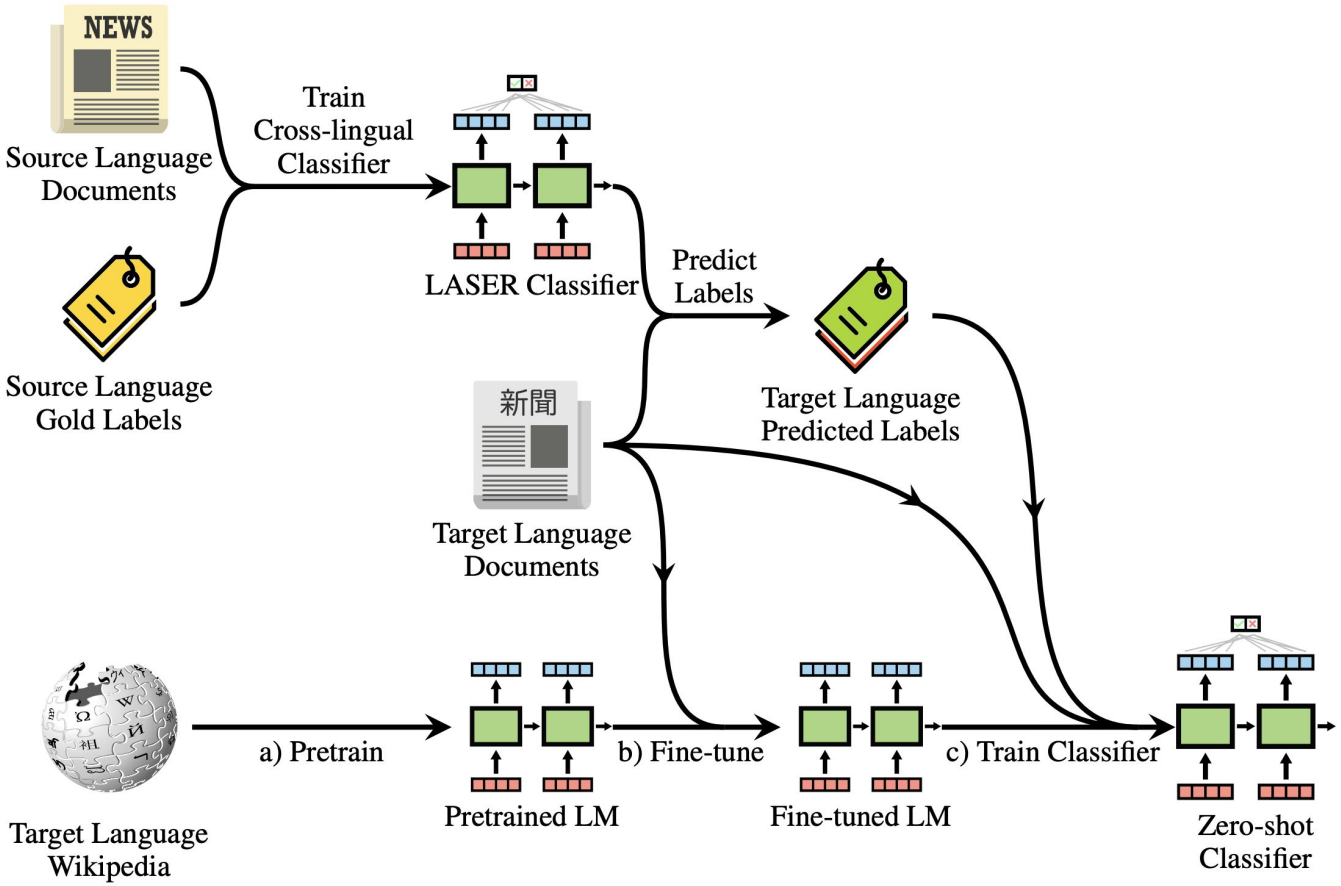
- MultiFit não utiliza a taxa de aprendizado do ULMFiT nem o descongelamento gradual.
- No lugar, o MultiFit usa uma técnica chamada de "ciclo único".
- Técnica consiste em fixar a taxa de aprendizado duas vezes durante o treino, formando um ciclo.



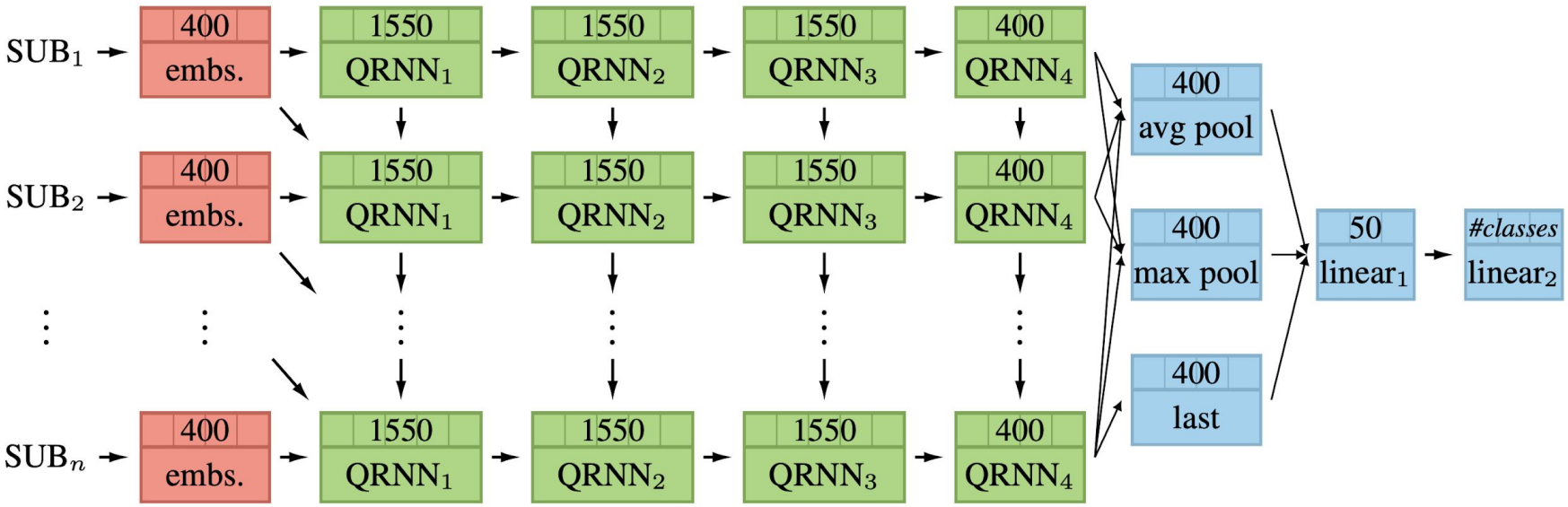
Suavização de labels

- No mundo real, datasets não são anotados 100% corretamente
- Anotações erradas de labels podem afetar resultado final do modelo
- Solução: utilizar suavizadores de label
  - 0 para falso e 1 para verdadeiro → 0.1 para falso e 0.9 para verdadeiro

Zero-shot com um professor multilíngue



Arquitetura



Resultados

- **Supervionado** - Nessa configuração, todos os modelos são ajustados e avaliados em exemplos do idioma de destino.
- **Zero-shot** - Cada modelo é ajustado nos exemplos do idioma de origem e avaliado nos exemplos do idioma de destino. Em todos os casos, foi usado o inglês como idioma de origem.

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Resultados

	LSTM	QRNN
Language model pretraining	143	**71**
Classifier fine-tuning	467	**156**

Table 4: Comparison of LSTM and QRNN per-batch training speed on a Tesla V100 (in ms) in MultiFiT.

Resultados

	DE	ES	ZH
ULMFiT	94.19	95.23	66.82
MultiFiT, no wiki	95.23	95.07	90.03
MultiFiT, small Wiki	95.37	95.30	89.80
MultiFiT	**95.90**	**96.07**	**92.52**

Table 5: Comparison of MultiFiT results with different pretraining corpora and ULMFiT, fine-tuned with 1k labels on MLDoc.

Resultados

IMDb

The IMDb dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet Movie Database (IMDb) labeled as positive or negative. The dataset contains an even number of positive and negative reviews. Only highly polarizing reviews are considered. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. No more than 30 reviews are included per movie. Models are evaluated based on accuracy.

Model	Accuracy	Paper / Source
XLNet (Yang et al., 2019)	96.21	XLNet: Generalized Autoregressive Pretraining for Language Understanding
BERT_large+ITPT (Sun et al., 2019)	95.79	How to Fine-Tune BERT for Text Classification?
BERT_base+ITPT (Sun et al., 2019)	95.63	How to Fine-Tune BERT for Text Classification?
ULMFiT (Howard and Ruder, 2018)	95.4	Universal Language Model Fine-tuning for Text Classification
Block-sparse LSTM (Gray et al., 2017)	94.99	GPU Kernels for Block-Sparse Weights
oh-LSTM (Johnson and Zhang, 2016)	94.1	Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings
Virtual adversarial training (Miyato et al., 2016)	94.1	Adversarial Training Methods for Semi-Supervised Text Classification
BCN+Char+CoVe (McCann et al., 2017)	91.8	Learned in Translation: Contextualized Word Vectors

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Resultados

	DE	ES	FR	IT	JA	RU	ZH
Zero-shot (1,000 source language examples)							
MultiCCA	81.20	72.50	72.38	69.38	67.63	60.80	74.73
LASER, paper	86.25	**79.30**	78.30	70.20	60.95	67.25	70.98
LASER, code	87.65	75.48	84.00	71.18	64.58	66.58	76.65
MultiBERT	82.35	74.98	83.03	68.27	64.58	**71.58**	66.17
MultiFiT, pseudo	**91.62**	79.10	**89.42**	**76.02**	**69.57**	67.83	**82.48**
Supervised (100 target language examples)							
MultiFit	90.90	89.00	85.03	80.12	80.55	73.55	88.02
Supervised (1,000 target language examples)							
MultiCCA	93.70	94.45	92.05	85.55	85.35	85.65	87.30
LASER, paper	92.70	88.75	90.80	85.93	85.15	84.65	88.98
MultiBERT	94.00	95.15	93.20	85.82	87.48	86.85	90.72
Monolingual BERT	94.93	-	-	-	-	-	92.17
MultiFiT, no wiki	95.23	95.07	94.65	89.30	88.63	87.52	90.03
MultiFiT	**95.90**	**96.07**	**94.75**	**90.25**	**90.03**	**87.65**	**92.52**

Table 2: Comparison of zero-shot and supervised methods on MLDoc.

UFG UNIVERSIDADE FEDERAL DE GOIÁS

Resultados

Comparação no dataset Cross-Lingual Sentiment dataset **CLS** constituído por reviews da Amazon em quarto línguas.

		DE			FR			JA		
		Books	DVD	Music	Books	DVD	Music	Books	DVD	Music
Zero-shot	LASER, code	84.15	78.00	79.15	83.90	83.40	80.75	74.99	74.55	76.30
	MultiBERT	72.15	70.05	73.80	75.50	74.70	76.05	65.41	64.90	70.33
	MultiFiT, pseudo	**89.60**	**81.80**	**84.40**	**87.84**	**83.50**	**85.60**	**80.45**	**77.65**	**81.50**
Translat.	MT-BOW	79.68	77.92	77.22	80.76	78.83	75.78	70.22	71.30	72.02
	CL-SCL	79.50	76.92	77.79	78.49	78.80	77.92	73.09	71.07	75.11
	BiDRL	84.14	84.05	84.67	84.39	83.60	82.52	73.15	76.78	78.77
Super.	MultiBERT	86.05	84.90	82.00	86.15	86.90	86.65	80.87	82.83	79.95
	MultiFiT	**93.19**	**90.54**	**93.00**	**91.25**	**89.55**	**93.40**	**86.29**	**85.75**	**86.59**

Table 3: Comparison of zero-shot, translation-based and supervised methods (with 2k training examples) on all domains of CLS. MT-BOW and CL-SCL results are from (Zhou et al., 2016).

UFG UNIVERSIDADE FEDERAL DE GOIÁS

Resultados

Model	F1	P	R	Acc
Ensemble	**0.727**	**0.729**	**0.726**	**0.723**
XLNet	0.679	0.696	0.692	0.690
ALBERT	0.679	0.684	0.676	0.675
BERT	0.675	0.680	0.672	0.670
MultiFiT	0.665	0.665	0.669	0.662

Table 2: Result Semeval-2020.

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Conclusões

- De fato é um modelo simples de treinar do zero.
- É um modelo "leve" com um bom tempo de inferência, mesmo rodando em CPU.
- Só funciona para tarefa de classificação de texto.

Obrigado(a)!

UFG
UNIVERSIDADE
FEDERAL DE GOIÁS

Manoel Veríssimo
verissimo.manoel@gmail.com