



POLISH NLP MEETUP #1

14.05.2020

To spotkanie jest nagrywane | This meeting is being recorded

Livestream: <https://youtu.be/PyLRJwJFIBQ>

AGENDA

- KLEJ Benchmark (Piotr Rybak, Allegro)
- Introduction to BERT and learnings from training Polish BERT (Darek Kłeczek)
- Training a Polish language model and using it in a classifier: Polberta (Marcin Zabłocki)
- Question answering in Polish (Henryk Borzymowski)



INTRO TO BERT AND POLBERT

Darek Kłeczek

ABOUT ME

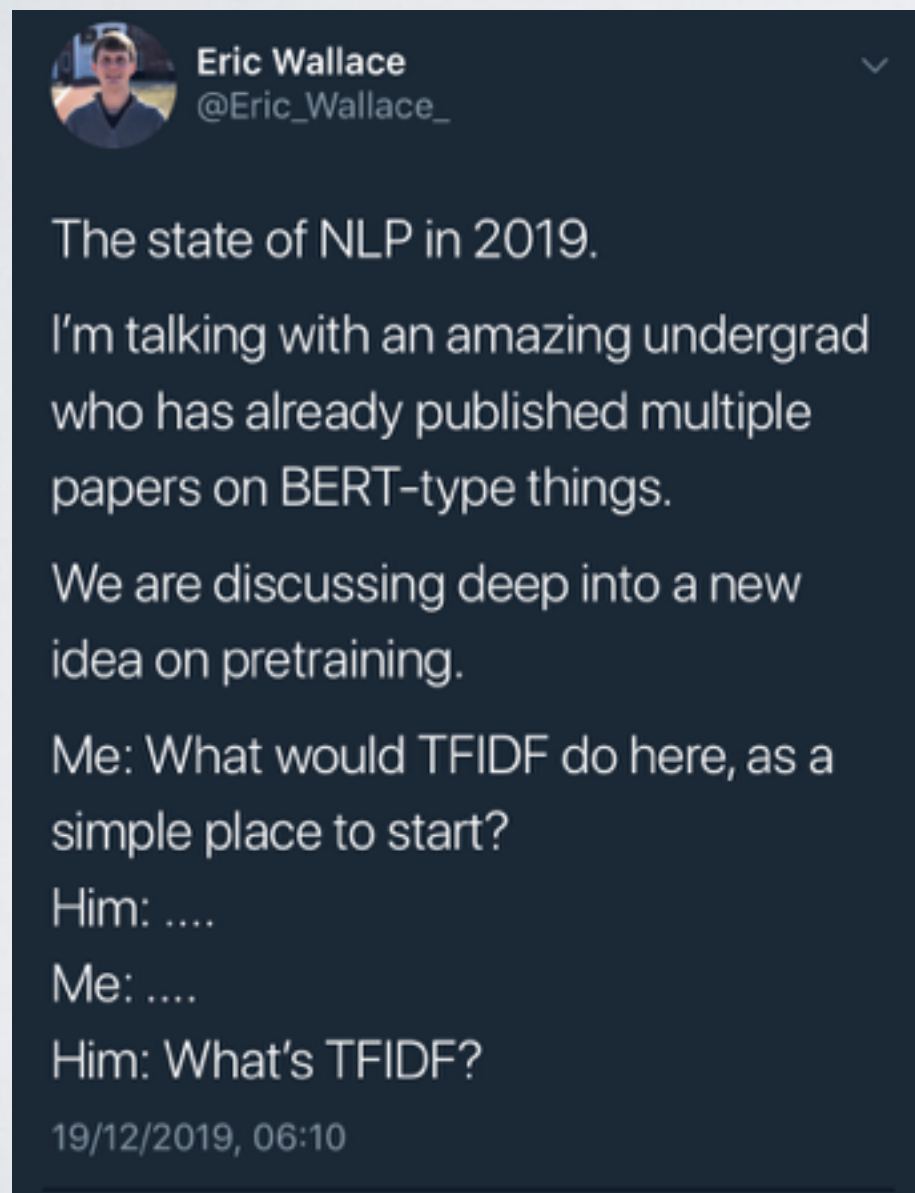
- Intelligent Automation at P&G
- Married, with a 3-year old daughter
- ML | Kaggle | Kung Fu in the remaining time
- Online educated in ML:
 - Andrew Ng / Coursera
 - Jeremy Howard / Fast.ai
- Projects: Polish NLP
 - Generate nursery rhymes
 - Compare press coverage from different sources
 - Polish BERT (first to make publicly available?)



Twitter: @dk21

Blog: skok.ai

SEISMIC SHIFT IN NLP



- Transformers winning every recent Kaggle NLP competition
- Transformers lead every (?) NLP benchmark
- Models getting bigger and bigger
- Some LSTM resistance (ULMFiT, SHA-RNN)

+ Code + Text

[] !pip install transformers -q

645kB 2.7MB/s

890kB 9.2MB/s

3.8MB 19.1MB/s

1.0MB 46.5MB/s

Building wheel for sacremoses (setup.py) ... done

[] from transformers import pipeline, BertTokenizer, BertModel, BertForNextSentencePrediction, BertConfig
import torch

▼ What can I use BERT for?

▼ Text classification

Probably the most popular use case for BERT is text classification. This means that we are dealing with sequences of text and want to classify them into discrete categories.

Here are some examples of text sequences and categories:

- Movie Review - Sentiment: positive, negative
- Product Review - Rating: one to five stars
- Email - Intent: product question, pricing question, complaint, other

Below is a code example of sentiment classification use case.

INTRO TO BERT

<https://skok.ai/2020/05/11/Top-Down-Introduction-to-BERT.html>

TRAINING POLBERT

1. Code: Google, HuggingFace
2. Corpus!!!!
3. Tokenization and accents
4. Hardware / TPU
5. Evaluation (thank you for KLEJ!)

<https://github.com/kldarek/polbert>

