Re-Evaluating GermEval17 Using German Pre-Trained Language Models

SwissText 2021

M. Aßenmacher, A. Corvonato, C. Heumann (LMU)

June 15, 2021

Problem setting

GermEval 2017 Shared Task ▸ Wojatzki et al., 2017

- Social media customer feedback about "Deutsche Bahn" (DB)
- Four different ABSA related subtasks
  - *A:* Relevance classification (binary: `true/false`)
  - *B:* Document-level sentiment classification (multi-class: `pos/neg/neutral`)
  - *C:* Aspect-based Sentiment Analysis (multi-label: `3 sentiments + 20 aspects`)
  - *D:* Opinion Target Extraction (sequence labeling)
- Synchronic (same time period) & diachronic (half a year later) test sets
- Back then, mainly "traditional" ML/DL classifiers were used

Problem setting

English-centric NLP benchmarking:

- Vast amount of benchmark data sets for the English language
  $\rightarrow$ Used for developing/evaluating SOTA pre-trained LMs
- Conclusions are transferred to other languages
- Amount of available non-English pre-trained models grows rapidly
  $\rightarrow$ Lack of standardized resources for benchmarking/evaluation

Research Goals

We set ourselves to ..

- ☒ .. evaluating German pre-trained models on a challenging task (cased vs. uncased, German vs. multilingual, BERT vs. DistilBERT),
- ☒ .. drawing parallels to the development of SOTA performance in English ABSA (at the example of the popular SemEval-2014 data sets),
- ☒ .. comparing pre-BERT to BERT-based approaches.

Model overview

Model variant	Pre-training corpus	Properties
`bert-base-german-cased`	12GB of German text (deepset.ai)	L=12, H=768, A=12, 110M parameters
`bert-base-german-dbmdz-cased`	16GB of German text (dbmdz)	L=12, H=768, A=12, 110M parameters
`bert-base-german-dbmdz-uncased`	16GB of German text (dbmdz)	L=12, H=768, A=12, 110M parameters
`bert-base-multilingual-cased`	Largest Wikipedias (top 104 languages)	L=12, H=768, A=12, 179M parameters
`bert-base-multilingual-uncased`	Largest Wikipedias (top 102 languages)	L=12, H=768, A=12, 168M parameters
`distilbert-base-german-cased`	16GB of German text (dbmdz)	L=6, H=768, A=12, 66M parameters
`distilbert-base-multilingual-cased`	Largest Wikipedias (top 104 languages)	L=6, H=768, A=12, 134M parameters

Overview of the evaluated pre-trained model architectures
(which were available via the huggingface transformers library by the end of 2020)

Model overview II

Model	Authors	A	B	C1	C2	D1	D2
		Subtask					
Models from 2017	▸ Wojatzki et al., 2017 ▸ Ruppert et al., 2017	X	X	X	X	X	X
Our BERT models		X	X	X	X	X	X
CNN	▸ Attia et al., 2018	–	X	–	–	–	–
CNN+FastText	▸ Schmitt et al., 2018	–	–	X	X	–	–
ELMo+GloVe+BCN	▸ Biesialska et al., 2020	–	X	–	–	–	–
ELMo+TSA	▸ Biesialska et al., 2020	–	X	–	–	–	–
FastText	▸ Guhr et al., 2020	–	X	–	–	–	–
`bert-base-german-cased`	▸ Guhr et al., 2020	–	X	–	–	–	–

Overview of the model architectures used for comparison

Results – Subtask A

Language model	$test_{syn}$	$test_{dia}$
XGboost (Best 2017) ▸ Sayyed et al., 2017	0.903	0.906
`bert-base-german-cased`	0.950	0.939
`bert-base-german-dbmdz-cased`	0.951	0.946
`bert-base-german-dbmdz-uncased`	**0.957**	**0.948**
`bert-base-multilingual-cased`	0.942	0.933
`bert-base-multilingual-uncased`	0.944	0.939
`distilbert-base-german-cased`	0.944	0.939
`distilbert-base-multilingual-cased`	0.941	0.932

F1 scores for Subtask A on synchronic and diachronic test sets

Results – Subtask B

Language model	test$_{syn}$	test$_{dia}$
SVM (Best 2017 on **test**$_{syn}$) ▸ Ruppert et al., 2017 XGboost (Best 2017 on **test**$_{dia}$) ▸ Sayyed et al., 2017	0.767	0.750
`bert-base-german-cased`	0.798	0.793
`bert-base-german-dbmdz-cased`	0.799	0.785
`bert-base-german-dbmdz-uncased`	**0.807**	**0.800**
`bert-base-multilingual-cased`	0.790	0.780
`bert-base-multilingual-uncased`	0.784	0.766
`distilbert-base-german-cased`	0.798	0.776
`distilbert-base-multilingual-cased`	0.777	0.770
CNN ▸ Attia et al., 2018	0.755	–
ELMo+GloVe+BCN ▸ Biesialska et al., 2020	0.782	–
ELMo+TSA ▸ Biesialska et al., 2020	0.789	–
FastText ▸ Guhr et al., 2020	0.698†	–
`bert-base-german-cased` ▸ Guhr et al., 2020	0.789†	–

†Guhr et al., 2020 created their own (balanced & unbalanced) data splits, which limits comparability.
(We compare to the performance on the unbalanced data since it more likely resembles the original data splits)

Micro-averaged F1 scores for Subtask B on synchronic and diachronic test sets

Results – Subtask C

Language model	*Aspect only* test$_{syn}$	*Aspect only* test$_{dia}$	*Aspect+Sentiment* test$_{syn}$	*Aspect+Sentiment* test$_{dia}$
SVM (Best 2017) ‣ Ruppert et al., 2017	0.537	0.556	0.396	0.424
`bert-base-german-cased`	0.756	0.762	0.634	0.663
`bert-base-german-dbmdz-cased`	0.756	0.781	0.628	0.663
`bert-base-german-dbmdz-uncased`	**0.761**	**0.791**	**0.655**	**0.689**
`bert-base-multilingual-cased`	0.706	0.734	0.571	0.634
`bert-base-multilingual-uncased`	0.723	0.752	0.553	0.631
`distilbert-base-german-cased`	0.738	0.768	0.629	0.663
`distilbert-base-multilingual-cased`	0.716	0.744	0.589	0.642
CNN+FastText ‣ Schmitt et al., 2018	0.523	0.557	0.423	0.465

Micro-averaged F1 scores for Subtask C1 & C2 on synchronic and diachronic test sets

Results – Subtask D (all models *with* CRF layer)

	Exact match		*Overlapping match*	
Language model	**test**$_{syn}$	**test**$_{dia}$	**test**$_{syn}$	**test**$_{dia}$
CRF (Best 2017) ▸ Ruppert et al., 2017	0.229	0.301	0.348	0.365
`bert-base-german-cased`	0.446	0.443	0.455	0.457
`bert-base-german-dbmdz-cased`	0.466	0.444	0.476	0.469
`bert-base-german-dbmdz-uncased`	**0.515**	**0.518**	**0.523**	**0.533**
`bert-base-multilingual-cased`	0.472	0.466	0.476	0.474
`bert-base-multilingual-uncased`	0.477	0.452	0.484	0.464
`distilbert-base-german-cased`	0.424	0.403	0.433	0.423
`distilbert-base-multilingual-cased`	0.436	0.418	0.442	0.427

Entity-level micro-averaged F1 scores for Subtask D1 & D2 on synchronic and diachronic test sets

Main Takeaways

We observed that ..

- ☒ .. uncased models have a tendency of outperforming their cased counterparts for the monolingual models, for multilingual models this cannot be clearly confirmed.
- ☒ .. monolingual models outperform the multilingual ones.
- ☒ .. there are no large performance differences between the two cased BERT models.
  $\rightarrow$ Suggests only a minor influence of the different corpora, which the models were pre-trained on.
- ☒ .. the monolingual DistilBERT model is pretty competitive.
  It consistently outperforms its multilingual counterpart as well as the mBERT models on the subtasks A – C and is at least competitive to the monolingual BERT models.

Discussion – SemEval-2014

SemEval-2014 Shared Task ‣ Pontiki et al., 2014

- English data set on Restaurant & Laptop reviews
- Different ABSA related subtasks
  - SB2: Aspect term polarity (Laptops and Restaurants)
  - SB3: Aspect category extraction (Restaurants only; 5 categories)
  - SB4: Aspect category polarity (Restaurants only; 3 sentiments + 5 categories)
- SB3 & SB4 *similar* to C1 & C2; SB2 only *related*

Discussion – SemEval-2014

	Language model	Restaurants SB3	Restaurants SB4
pre-BERT	Best model SemEval-2014 ▸ Pontiki et al., 2014	0.8857	0.8292
pre-BERT	ATAE-LSTM ▸ Wang et al., 2016	—	0.840
BERT-based	BERT-pair ▸ Sun et al., 2019	0.9218	0.899
BERT-based	CG-BERT ▸ Wu et al., 2020	0.9162†	0.901†
BERT-based	QACG-BERT ▸ Wu et al., 2020	0.9264	0.904†

†Additional auxiliary sentences were used.

SOTA F1 scores for Subtask SB3 & SB4 (SemEval-2014)