Towards a Computational Study of German Book Reviews
A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis

Simone Rebora (Univ. of Verona / Univ. of Bielefeld) simone.rebora@univr.it

Thomas C. Messerli (Univ. of Basel) thomas.messerli@unibas.ch

J. Berenike Herrmann (Univ. of Bielefeld) berenike.herrmann@uni-bielefeld.de

The Goal

Computational study of evaluative practices in German language book reviews.

Main focus:
LovelyBooks
Social Reading Platform

www.lovelybooks.de
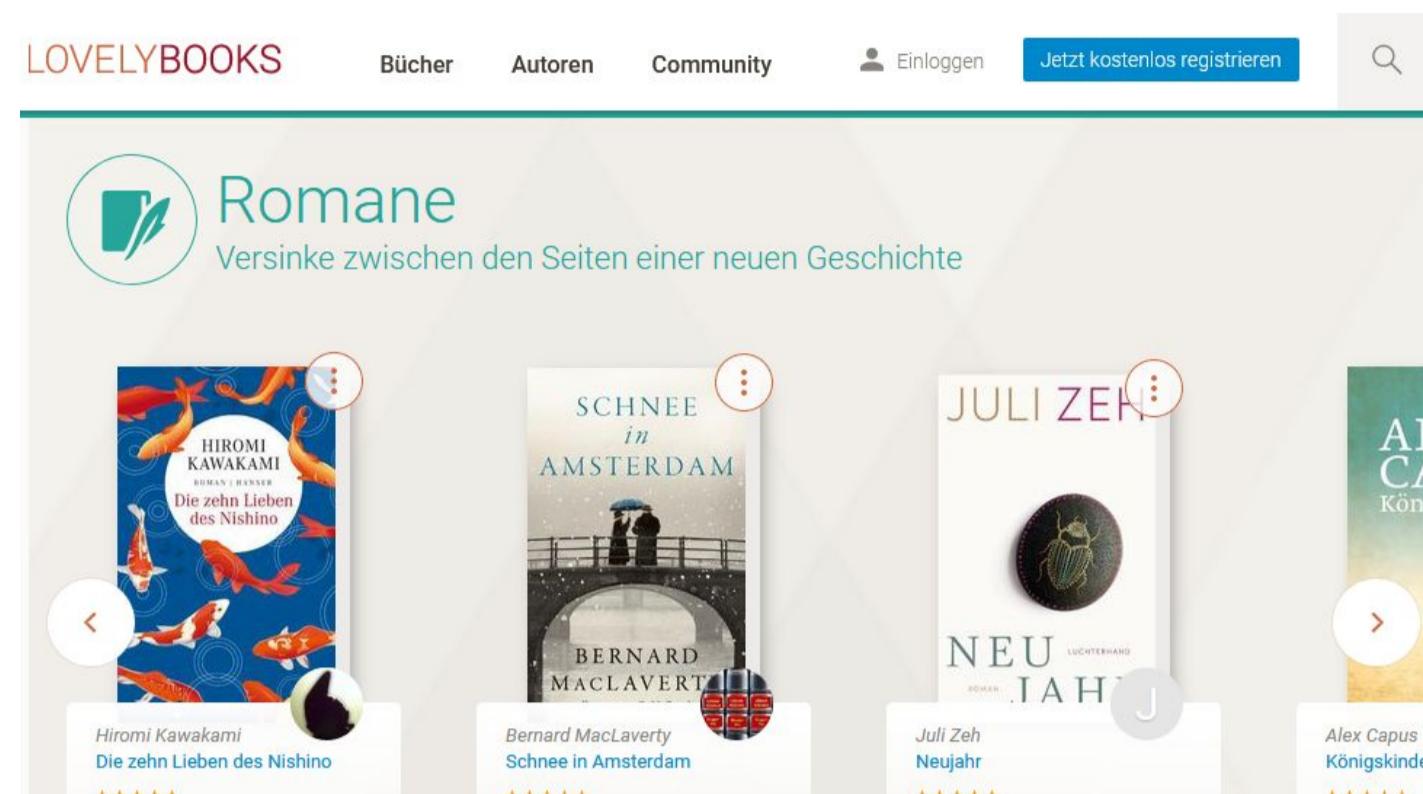
The LOBO Corpus

Genre	No of Reviews
young adult	243,731
crime_thriller	240,132
novel	213,112
fantasy	131,102
romance novel	130,257
children's book	64,875
non-fiction	62,237
erotic_literature	46,368
historical_novel	45,363
biography	19,475
humour	18,463
science_fiction	16,649
classic	13,561
comic_book	12,629
poetry_drama	3,199
N/A	66,302
Total Count	**1,327,455**

CWB The IMS Open Corpus Workbench

-> 1,327,455 Texts
-> 54,061 Users
-> 438,227,032 Tokens
-> 187,275 Lemmas
-> 169,759 Books
-> 63,058 Authors
-> 15 Genres

Manual Annotation

**Evaluative language**

tag
- evaluation
- other
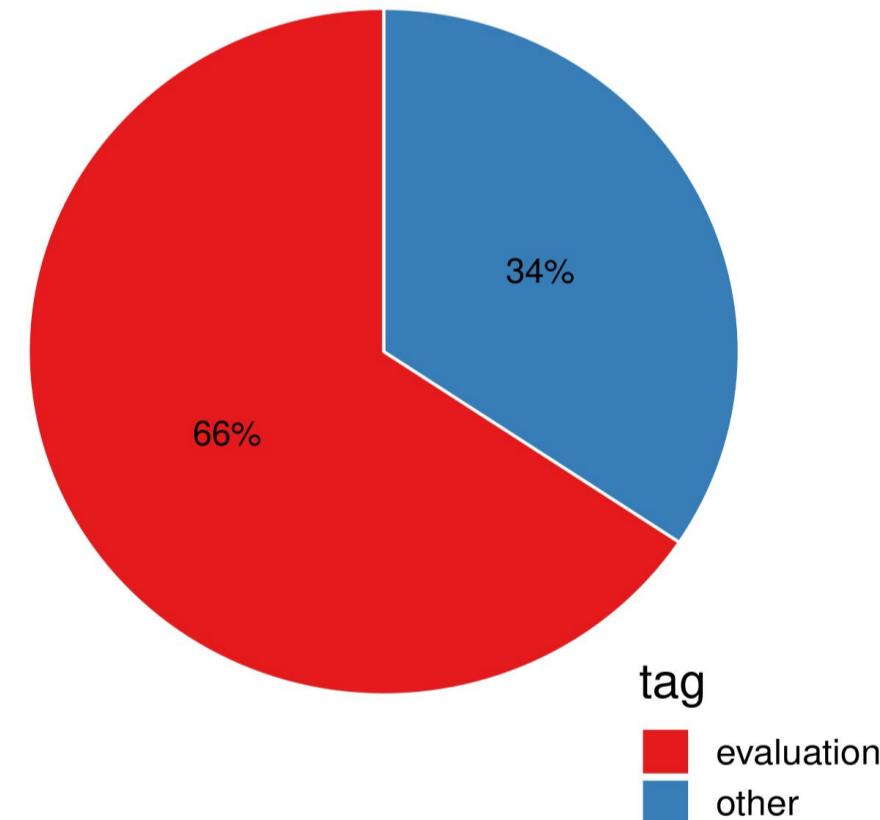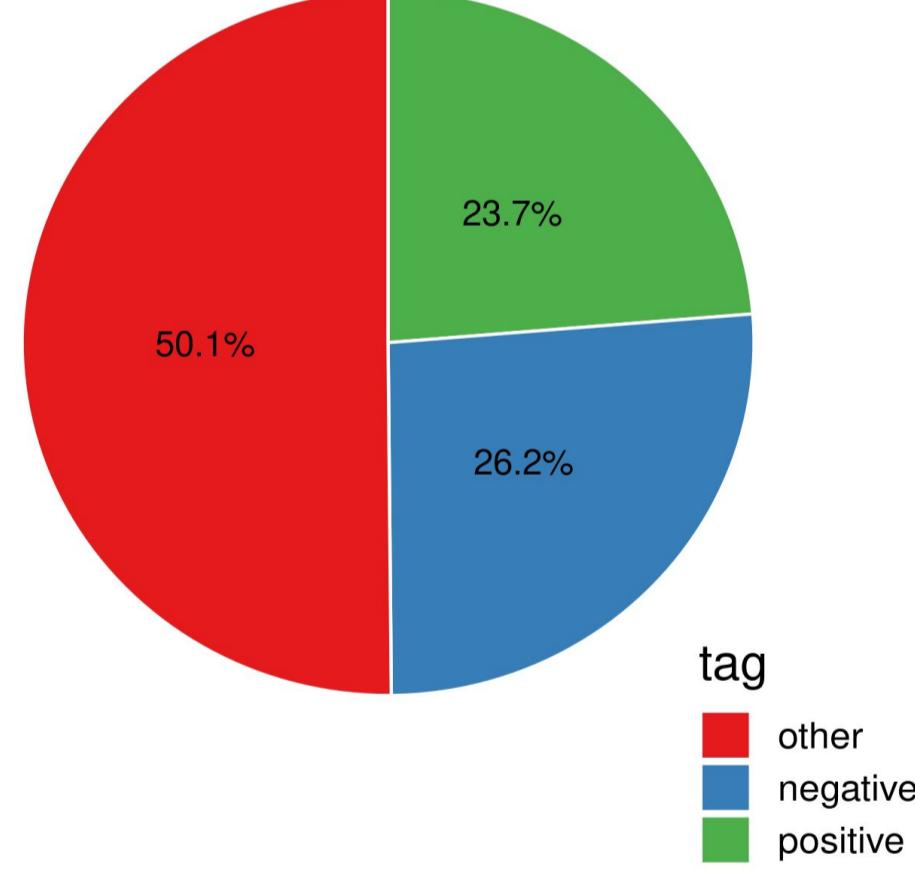
-> 2 annotators
-> 2,000 reviews
-> 21,735 sentences
-> 0.76 Cohen's Kappa

Sentiment

tag
- other
- negative
- positive

-> 2 annotators
-> 1,100 reviews
-> 13,552 sentences
-> 0.77 Cohen's Kappa

Automated Classification

Emotion dictionaries

Lexicon	Length (words)	Dimensions (categories)	Reference
ADU	26,832	12	(Hölzer et al., 1992)
AffectiveNorms	351,617	4	(Köper and Schulte im Walde, 2016)
BAWLr	2,902	3	(Võ et al., 2009)
Ekman	4,293	7	(Klinger et al., 2016)
Germanlex	8,693	1	(Clematide and Klenner, 2010)
LANG	1,000	3	(Kanske and Kotz, 2010)
MorphComp	9,256	3	(Ruppenhofer et al., 2017)
NRC	4,622	10	(Mohammad and Turney, 2013)
Plutchik	951	8	(Stamm, 2014)
PolarityCues	10,790	3	(Waltinger, 2010)
SentiArt	116,313	7	(Jacobs, 2019)
sentiWS	3,471	1	(Remus et al., 2010)

Transfer learning

fastText
(Mikolov et al., 2017)

Google BERT
(Devlin et al., 2019)

Efficiency (Dictionary-based)

Evaluative language

	Precision	Recall	F1	Support
Evaluation	0.854	0.777	0.813	2852.6
Other	0.636	0.746	0.687	1494.4
Accuracy			0.766	4347
Macro	0.745	0.761	0.750	4347
Weighted	0.779	0.766	0.770	4347

SVM trained on all features (with oversampling)

Sentiment

	Precision	Recall	F1	Support
Positive	0.606	0.630	0.618	642.36
Negative	0.559	0.633	0.593	708.88
Other	0.752	0.686	0.718	1358.76
Accuracy			0.659	2710
Macro	0.639	0.650	0.643	2710
Weighted	0.667	0.659	0.662	2710

SVM trained on single dictionaries
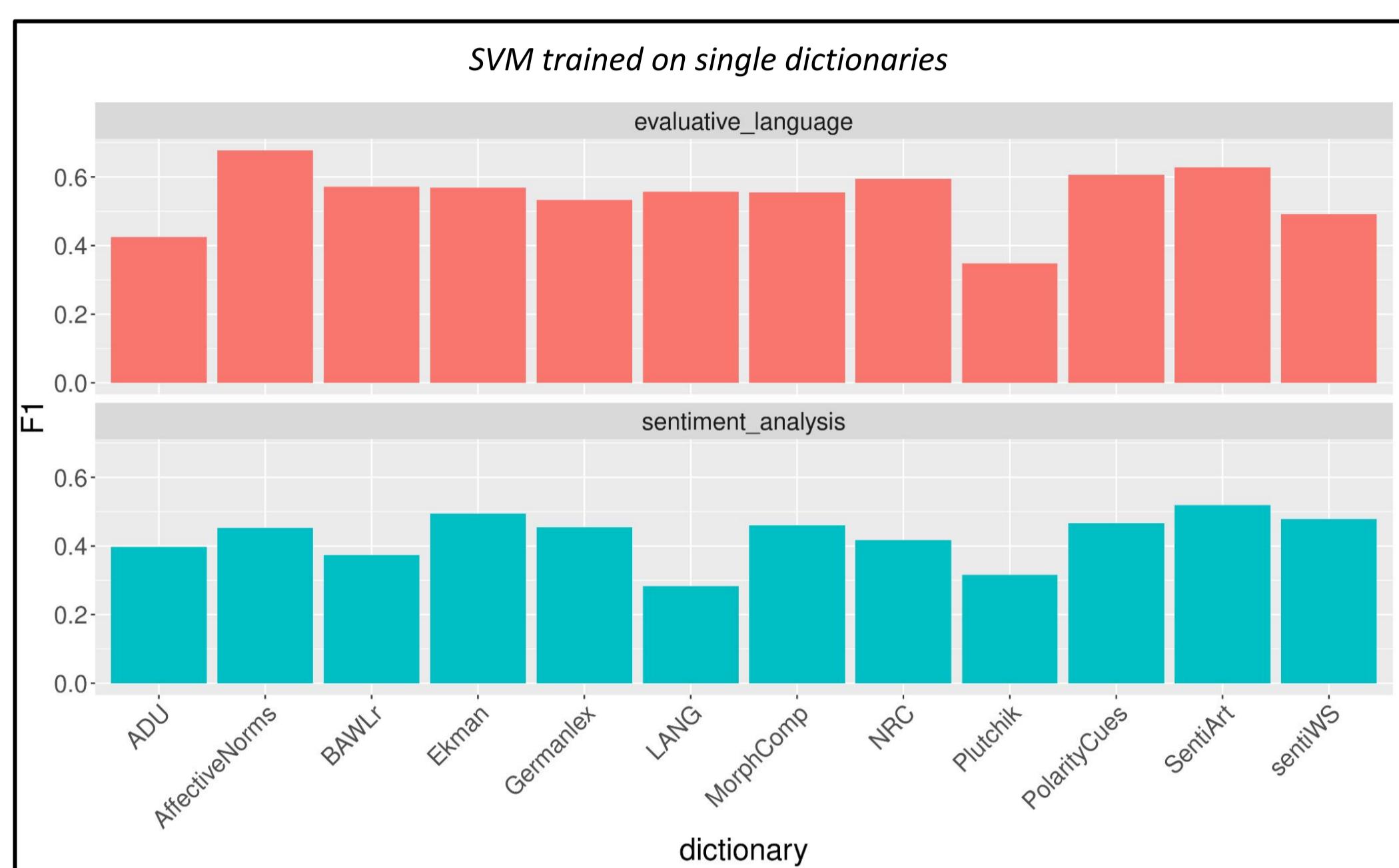
Efficiency (Transfer learning)

Evaluative language

	Precision	Recall	F1	Support
Evaluation	0.921	0.927	0.924	2828.52
Other	0.860	0.847	0.853	1482.6
Accuracy			0.900	4311.12
Macro	0.890	0.887	0.889	4311.12
Weighted	0.900	0.900	0.900	4311.12

BERT (bert-base-german-cased) with Adam optimizer and 5x5-fold cross validation

FastText with German model and Automatic hyperparameter optimization
F1 Macro = 0.83

Sentiment

	Precision	Recall	F1	Support
Positive	0.839	0.809	0.824	630.4
Negative	0.820	0.793	0.806	693.84
Other	0.850	0.879	0.864	1332.76
Accuracy			0.840	2657
Macro	0.836	0.827	0.831	2657
Weighted	0.840	0.840	0.839	2657

F1 Macro = 0.72

Evaluative language
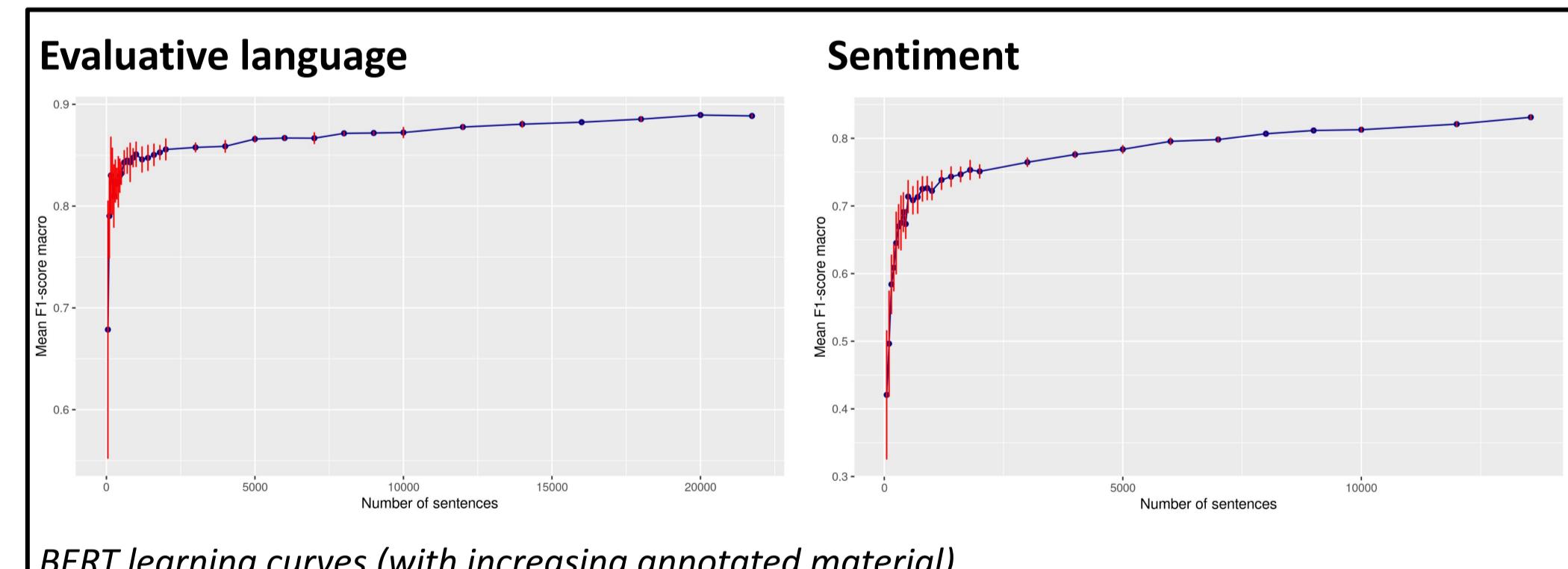
Sentiment

BERT learning curves (with increasing annotated material)

Summary

Task	Lexicon-based sentiment analysis	TL-based sentiment analysis
Evaluative language (21,735 sentences)	SVMs trained on features generated by SA dictionaries: macro *F1*-score .75	BERT: macro *F1*-score .89 FastText: macro *F1*-score .83
Valence (13,552 sentences)	SVMs trained on features generated by SA dictionaries: macro *F1*-score .64	BERT: macro *F1*-score .85 FastText: macro *F1*-score .72