Towards a Computational Study of German Book Reviews

A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis

Rebora, Simone
simone.rebora@uni-bielefeld.de
University of Bielefeld, Germany

Messerli, Thomas
thomas.messerli@unibas.ch
University of Basel, Switzerland

Herrmann, J. Berenike
berenike.herrmann@uni-bielefeld.de
University of Bielefeld, Germany

This poster reports on the groundwork for the computational study of evaluative practices in German language book reviews. We trained classifiers for evaluation and sentiment at sentence level on the LOBO corpus, comprising ~1.3 million book reviews downloaded from the social reading platform LovelyBooks.

For the two classification tasks, we compared performance of dictionary-based and transfer-learning (TL) based sentiment analysis (SA). To use dictionary-based SA systematically, a repository of twelve open-source German-language SA lexicons was created (see Table 1). Lexicon formats were uniformed to automatically annotate reviews for sentiment in a processing pipeline. For the TL approaches, we chose BERT and FastText, both of which based on distributional representations of natural language (see Devlin et al., 2019; Mikolov et al., 2017) .

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)

Tab. 1: Overview of German-language sentiment dictionaries

The dictionary-based and TL approaches were evaluated on two manually annotated datasets, working with two annotators: in the first dataset (~21,000 sentences), the annotation task was that of identifying evaluative language (vs. descriptive language); in the second dataset (~13,500 sentences), the task focused on the distinction between positive and negative sentiment. These two classification tasks form the basis for a large-scale analysis of the LOBO corpus, which segments reviews into evaluative and descriptive passages, to describe differences in evaluation practices across genres (e.g., romance, science fiction) and ratings (1-5 stars).

For the creation of the Gold Standard of Task 1 (evaluation classification), manual annotation reliability was evaluated on a subset of 250 reviews (~4,000 sentences). Cohen's *Kappa* (0.76) indicated a strong agreement between annotators. Overall, 66% of the total sentences were annotated as "evaluation". Training an SVM classifier on the features generated by the 12 sentiment dictionaries rendered a macro *F1* score of 0.75 (see Table 2 for details).

	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

Tab. 2: Efficiency of dictionary-based SVM on Task 1

To compare the efficiency of the dictionaries, the same classifier was trained separately with the single dictionaries. Fig. 1 shows the results, with AffectiveNorms as the best-performing dictionary (macro *F1* score of 0.67).

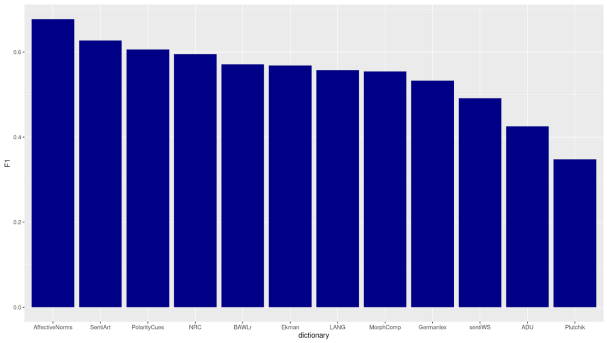

Fig. 1: Efficiency of single German-language SA dictionaries (Task 1)

Yet, by contrast, TL-methods proved substantially more efficient, with macro *F1* scores of 0.83 for FastText and of 0.89 for BERT (results obtained via a 5-fold cross validation, repeated five times to average variance, see Table 3 for details on BERT).

	Precision	Recall	*F1*	Support
Evaluation	0.9206	0.9273	0.9239	2828.52
Other	0.8595	0.8473	0.8533	1482.6
Accuracy			0.8998	4311.12
Macro	0.89	0.8873	0.8886	4311.12
Weighted	0.8996	0.8998	0.8996	4311.12

Tab. 3: Efficiency of BERT on Task 1

The evaluation procedure was repeated on Task 2 (positive vs. negative sentiment). Again, inter-annotator agreement was strong for manual annotation of the Gold Standard (Cohen's Kappa = 0.79). Annotation percentages are shown by Fig. 2 (where the

"other" category indicates both mixed feelings and the absence of evaluation).

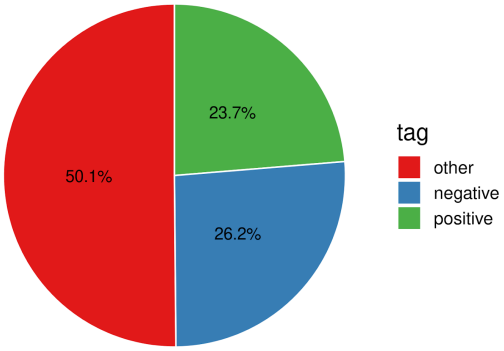

Fig. 2: Percentages of annotations for task 2

The dictionary-based SVM classifier reached a macro *F1* score of 0.64, while the best performance was obtained by SentiArt (see Fig. 3). Efficiency was again higher for FastText (macro *F1* score = 0.72) and best for BERT (macro *F1* score = 0.83). However, the learning curve for BERT shows how there is still room for improvement, with efficiency not fully reaching a plateau (see Fig. 4).

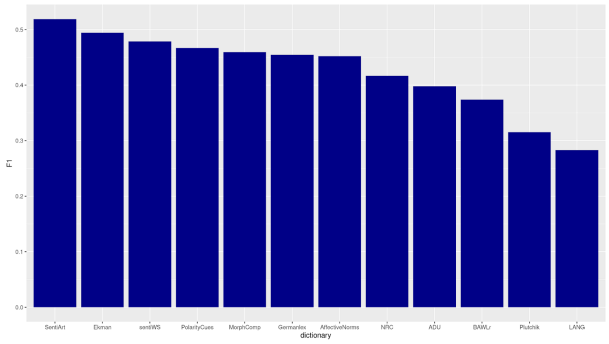

Fig. 3: Efficiency of single German-language SA dictionaries on Task 2

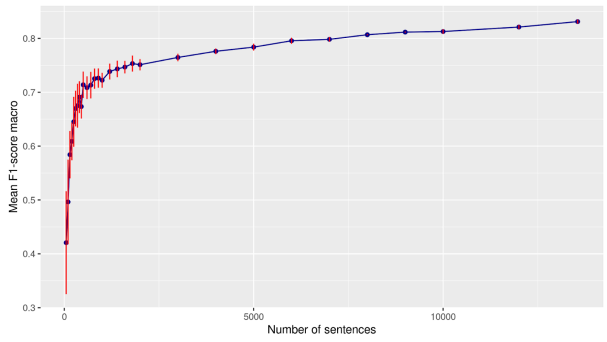

Fig. 4: Learning curve (with increasing amount of training material) of BERT for Task 2

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

Tab. 4: Overview of results for lexicon-based and TL-based approache

Our results highlight the higher efficiency of TL-methods (see Table 4) and of dictionaries based on vector space models (like SentiArt and AffectiveNorms). They show that computational methods can reliably identify sentiment of book reviews in German. In order to fruitfully use similar methodology to identify types of engagement by reviewers with literature beyond the descriptive/evaluative and positive/negative dichotomies, a useful next step will be to attempt the design of TL-tasks for the identification of more fine-grained evaluative practices. These include the construction of and orientation to particular evaluative scales (e.g. reading pleasure, literary quality) and particular subjects of evaluation (e.g. novels, authors, characters).

Bibliography

Clematide, S. and Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. s.n. doi:10.5167/UZH-45506. https://www.zora.uzh.ch/id/eprint/45506 (accessed 12 July 2021).

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]* http://arxiv.org/abs/1810.04805 (accessed 13 July 2021).

Hölzer, M., Scheytt, N. and Kächele, H. (1992). Das „Affektive Diktionär Ulm" als eine Methode der quantitativen Vokabularbestimmung. In Züll, C. and Mohler, P. Ph. (eds), *Textanalyse: Anwendungen der computerunterstützten Inhaltsanalyse. Beiträge zur 1. TEXTPACK-Anwenderkonferenz* . (ZUMA-Publikationen). Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 131–54 doi:10.1007/978-3-322-94229-6_7. https://doi.org/10.1007/978-3-322-94229-6_7 (accessed 12 July 2021).

Jacobs, A. M. (2019). Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI* , **6** doi:10.3389/frobt.2019.00053. https://www.frontiersin.org/article/10.3389/frobt.2019.00053/full (accessed 8 September 2019).

Kanske, P. and Kotz, S. A. (2010). Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods* , **42** (4): 987–91 doi:10.3758/BRM.42.4.987.

Klinger, R., Suliya, S. S. and Reiter, N. (2016). Automatic Emotion Detection for Quantitative Literary Studies: A case study based on Franz Kafka's 'Das Schloss' und 'Amerika'. *DH2016 Book of Abstracts* . Kraków: ADHO https://dh2016.adho.org/abstracts/318.

Köper, M. and Schulte im Walde, S. (2016). Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* . Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2595–98 https://aclanthology.org/L16-1413 (accessed 12 July 2021).

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. and Joulin, A. (2017). Advances in Pre-Training Distributed

Word Representations. *ArXiv:1712.09405 [Cs]* http://arxiv.org/abs/1712.09405 (accessed 13 July 2021).

Mohammad, S. M. and Turney, P. D. (2013). CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON. *Computational Intelligence* , **29** (3): 436–65 doi:10.1111/j.1467-8640.2012.00460.x.

Remus, R., Quasthoff, U. and Heyer, G. (2010). SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* . Valletta, Malta: European Language Resources Association (ELRA) http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf (accessed 12 July 2021).

Ruppenhofer, J., Steiner, P. and Wiegand, M. (2017). Evaluating the Morphological Compositionality of Polarity. *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning* . Incoma Ltd. Shoumen, Bulgaria, pp. 625–33 doi:10.26615/978-954-452-049-6_081. http://www.acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP081.pdf (accessed 12 July 2021).

Stamm, N. (2014). Klassifikation und Analyse von Emotionswörtern in Tweets für die Sentimentanalyse.

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J. and Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods* , **41** (2): 534–38 doi:10.3758/BRM.41.2.534.

Waltinger, U. (2010). GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* . Valletta, Malta: European Language Resources Association (ELRA) http://www.lrec-conf.org/proceedings/lrec2010/pdf/91_Paper.pdf (accessed 12 July 2021).