*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1376–1380
Marseille, 20-25 June 2022


# Polar Quantification of Actor Noun Phrases for German

**Anne Göhring, Manfred Klenner**
Department of Computational Linguistics
University of Zurich
{goehring,klenner}@cl.uzh.ch

**Abstract**

In this paper, we discuss work that strives to measure the degree of negativity - the negative polar load - of noun phrases, especially those denoting actors. Since no gold standard data is available for German for this quantification task, we generated a silver standard and used it to fine-tune a BERT-based intensity regressor. We evaluated the quality of the silver standard empirically and found that our lexicon-based quantification metric showed a strong correlation with human annotators.

**Keywords:** polarity intensity, quantification metric, BERT-based regressor

## 1. Introduction

The polar intensity of words, noun phrases, sentences and texts has been focused on for quite some time, e.g. Polanyi and Zaenen (2006),Taboada et al. (2011) and more recently Huang et al. (2020). In contrast to previous approaches, that are just interested in a proper prediction of the intensity, we also are interested in the discourse function of a particular type of noun phrases, namely noun phrases that could be used to refer to an agent of some action, an actor. Besides our theoretical interest in such an exploration, there is also an application-oriented aspect. One of the strategies of hate speech is defamation and vilification. Verbally, we could use various means to conceptualize somebody as a negative person, a negative actor. Among them are definite descriptions ('the foolish Merkel'), predicative statements ('Merkel is a fool'), or role fillers of particular verbs ('Merkel cheats us'). If a text casts an actor as highly negative, then it might be an instance of hate speech. Our quantification approach, thus, could be used to identify candidates of hate speech.

Since no gold standard data are available for German, we generated a lexicon-based silver standard (Manfred Klenner, Anne Göhring, 2022) and carefully evaluated the resulting data set. Although the lexicon-based metric turned out to strongly correlate with human annotations, the intention was to get rid of the need for lexicon-based modelling, since this is limited due to lexicon gaps. The research question then was: how well does the learned model generalize.

The main contribution of this work is: we introduce the first approach to quantification (possibly not only for German) where a general model is derived in a bootstrapping manner from existing lexical resources.

## 2. Actor Noun Phrase Quantification

For the detection of hate speech the identification of highly negatively conceptualized actors could be useful. Thus, a regression model for the quantification of negativity (of actor noun phrases) is needed. Currently, there is no gold standard with quantified German noun phrases available. The annotation of negative noun phrases with concrete strength values might turn out to be challenging.

| | actor noun phrase |
|---|---|
| 1 | die lügnerische Merkel (*mendacious*) |
| 2 | die sturre Merkel (*stubborn*) |
| 3 | die ungerechte Merkel (*unjust*) |
| 4 | die emotionslose Merkel (*unemotional*) |
| 5 | die übereifrige Merkel (*overzealous*) |
| 6 | die befangene Merkel (*timid*) |
| 7 | die untüchtige Merkel (*inefficient*) |

Table 1: Ranked Negativity

All phrases in Table 1 appeared in Facebook posts of a German right-wing party (called AfD) and conceptualize the former German chancellor Angela Merkel as a negative actor. There are clear differences: number 1 is stronger than number 5 and 6, but comparable to number 2. Maybe number 1 is slightly stronger than 2, but we cannot hope to consistently annotate such subtle differences. Thus the concrete magnitude (-1 or -2 or . . . ) won't be helpful and would be hard to justify. What a resource should accomplish is the ordering from highly negative to highly positive. Every scaling that produces such an ordering would do.

To learn a reliable model, many examples are needed. In order to avoid the time-consuming (and unreliable) manual annotation of (meaningless) numerical strength values we suggest a lexicon-based method, a heuristic strength metric that might be regarded as an approximate model for how humans do rate the negativity strength of words and phrases. We cannot carry out the necessary psychological experiments in order to fully support this claim, but we believe that the strong results of our correlation (Spearman) study with 9 human annotators could be regarded as providing some evidence for this statement (see end of section 2.2 and Table 4 in section 3).

The heuristic is based on an available polarity lexicon[1], the PolArt lexicon (Clematide and Klenner, 2010), that comes in its base version with three strength values at the word level (e.g. 'tired' has 0.5, 'incautious' has value 0.7 and 'hate' has 1 as a strength value). In the augmented version of the lexicon, the words are also classified according to the appraisal theory (Martin and White, 2005) and the words of type emotion additionally are specified with respect to the base emotion (they took Plutchik's distinctions (Plutchik, 1980)). Finally, the words carry a label related to the behavioral status they express: active, passive, none (for unclear cases).

## 2.1. Lexicon-based Metric

The first step was to produce a more fine-grained strength value at the word level using the various dimensions of PolArt. The polarity strength of phrases can be modelled then as a function of the word level strength of its words, following the principle of compositionality. Given this, we generated a silver standard of quantified noun phrases and used it for training and evaluation. In order to assure the quality of the data, we let 9 raters classify 200 adjectives and compared the ratings to the strength values assigned by our metric.
Each word of PolArt comes with a basic strength value (0.5, 0.7, 1), the appraisal category (appreciation, emotion, moral judgement), for emotion words the base emotion, and the behavioral label (active, passive, none). Some examples of adjectives are shown in Table 2.

| word | str | appr | emo | beh |
|---|---|---|---|---|
| zornig (*angry*) | 1 | E | anger | a |
| betrügerisch (*cheating*) | 1 | M | - | a |
| begriffsstutzig (*obtuse*) | 1 | A | - | p |
| zerstritten (*quarreling*) | 1 | A | - | a |
| müde (*tired*) | 0.7 | A | - | p |
| provinziell (*provincial*) | 0.5 | A | - | none |

Table 2: Strength (str), appraisal (appr) category (A: appreciation; M: moral judgement; E: emotion), base emotion (emo) , and behavioral (beh) tag (a: active; p:passive) of PolArt entries

If we compare e.g. 'cheating' (moral) with 'quarreling' (appreciation), both having 1 as a basic strength value, it becomes clear that the appraisal dimensions should be taken into account in order to scale the strength values. An unethical attribute of a person is more negative than just some factual negative one (a cheating versus quarreling person). The same is true with respect to the behavioral labels (active/passive/none). A negative attribution that indicates an active part is more negative than just a passive one (an aggressive versus a reserved person).
Within the emotional dimension, the various base emotions also partially induce a ranking: disgust is stronger than anger. Please keep in mind that we are interested in the extrinsic negativity of actors ('a cheating person') as compared to an intrinsic negatively affected actor ('a frightened person'). Thus, we exclude words classified as fear or sadness: words from these emotions (most of the time) describe intrinsic attributes.
We defined a partial ordering of negativity strength[2] and assigned factors to the labels: Table 3 shows some examples of the determination of the strength values. The straightforward formula for quantification is:

(1) $pol\_strength(w) = str(w) * ap(w) * mode(w)$

where $str(w)$ denotes the word (w) basic strength as provided by the polarity lexicon; $ap(w)$ is the value associated with the appraisal type of the word or, if the type is emotion, the scaling of the base emotion; $mode(w)$ refers to active - passive - none (which scales by 1). The values range lies between 0.5 and 3.

| word | (s,a,b) | calc | val |
|---|---|---|---|
| ungebildet (*uneducated*) | .5,A,p | .5*1*1 | 0.5 |
| Verweigerung (*refusing*) | .7,A,a | .7*1*1.5 | 1.05 |
| böse (*evil*) | 1,M,a | 1*1.5*1.5 | 2.25 |
| eklig (*disgusting*) | 1,E,p | 1*2*1 | 2 |
| Hass (*hate*) | 1,E,a | 1*2*1.5 | 3 |

Table 3: Calculation of new strength values (val) based on (s,a,b): PolArt strength, appraisal category (A: appreciation; M: moral judgement; E: emotion), and behavioral label (p: passive; a: active)

We quantified all words from the PolArt lexicon that way. The next step was an evaluation of the appropriateness of this metric.

## 2.2. Evaluation of the Lexicon-based Metric

In order to evaluate the resulting polarity strength lexicon and thus our metric, we followed the idea of the Likert scale (Likert, 1932), namely we generated a data set of 200 adjectives with statements about the negativity of these words. We then asked 9 volunteers to tell us their opinion. The 200 adjectives were chosen to represent an evenly distributed sample of (new) strength values: from very negative to only slightly negative. The raters should classify each adjective on the basis of 5 classes: neutral, just a bit negative, negative, strongly negative, extremely negative. Internally, we coded these classes as 0,1,2,3,4 respectively. This allowed us to determine an average strength value per word.
We measured Spearman's correlation for each annotator pair and between each annotator and the strength values of the metric (thus, we have n=10 and there are n(n-1)/2, namely 45 pairs). The correlation between

[1] https://sites.google.com/site/iggsahome/downloads

[2] We are aware of the fact that words might be vague (e.g. 'cowardly') and their strength value depends on the context of their usage. The specifications for the lexical entries might be regarded as worst usage values.

humans ranges from 0.44 up to 0.71, the mean is 0.57 which is a moderate correlation. Between humans and the metric, Spearman's $\rho$ lies between 0.38 and 0.59, the mean is 0.48 which also is moderate, but lower. We noticed that the correlation between humans fluctuates considerably and that the task at hand is not trivial. The reason is that we use a fine-grained distinction with 5 classes. Since the metric produces 13 different strength values (4 are produced more than once) starting with 0.5 to 3, such a fine-grained class inventory is needed in order to have as many ordinal anchor points as possible on both sides of the spectrum: annotator and metric. There are a number of cases where the correlation between the predictions of the metric and a human is higher than those between two humans. But a clear statistical claim, namely that the metric is on par with the human raters cannot be given easily on that basis.
The preferred way to generate a lexicon with strength values of words from crowd-sourced data would be to take the average assignments for each word. A given adjective then gets as a strength value the average of the assignments of the 9 raters:

$$(2) \quad strength(w) = \frac{1}{9} * \sum_{rater} score(rater, w)$$

If we now measure the correlation of the metric with this lexicon, the result might be indicative of the usefulness of the metric.
The value for Spearman's $\rho$ for this was 0.645 which counts as a strong correlation[3]. The p-value (null hypothesis is that the two data sets are uncorrelated) is extremely low. We clearly withdraw $H_0$. We took this result as an indicator of goodness.

## 3. Quantification Experiments

Given a binary classifier for actor classification and a polarity lexicon with sophisticated strength values, we could envisage joint experiments for actor detection and quantification. However, in a first round we created a silver standard of noun phrases on the basis of an actor list[4], a large newspaper corpus and our metric. At that stage we were interested in the performance of a learned regressor, independent from an actor classifier. In a second round, we evaluated a joint model where the noun phrases had to be unknown (with respect to the actor list) actors, reflecting the generalisation capacity of a joint system (section 3.2).

### 3.1. Regression on Silver Standard

In order to quantify the polar strength of a noun phrase (NP), a simple additive projection was used: The polar strength of an NP is the sum of the polar values of its words. For instance, 'a horrible liar' gets 4 because liar has 3 and horrible has 1 as a strength value[5]. We extracted from a large newspaper corpus[6] about 40,000 genuine unique NPs that had an adjective from the polarity quantified lexicon and an actor from the actor list. We used an MLP regressor with FastText embeddings and compared it to a BERT-based[7] regression. In a first experiment, a 10-fold cross-validation setting, MLP had a mean coefficient of determination $R^2$ of 92.3% compared to 95.5% of BERT, we, thus, continued with BERT. High $R^2$ values are given if a model explains most of the variance in the data. Although this cannot be interpreted as a perfect fit at every data point, it indicates that in general the regression comes close to existing data points.

| | explvar | mse | $R^2$ | $\rho$ |
|---|---|---|---|---|
| overlapping | 0.955 | 0.045 | 0.955 | 0.820 |
| exclusive | 0.387 | 0.591 | 0.385 | 0.564 |

Table 4: Correlation metrics applied in two settings: overlapping, where NPs of train and test set might have overlapping words, and the exclusive data set where this in not the case.

Table 4 shows various metrics: explained variance (explvar), mean squared error (mse), $R^2$ and Spearman's $\rho$. Most important is $\rho$, least important mse, since $\rho$ takes the ordinal nature of the data into account, while mse measures the error in terms of the squared strength value difference. As already discussed, the relative order is important, not the predicted magnitude.
The data split in this cross-validation setting produces unique NPs, but does not prevent overlapping NPs (first line in Table 4): an NP from the test set might share an adjective or the noun with an NP from the training set. Thus, our first experiment - though it represents the expected application conditions - could not reveal much about the real generalization impact of the model. Instead of a 10-fold cross-validation, we, thus, created a single, though random, train/test split, where the vocabulary of the test set is exclusive with those of the training set. That is, the model has never seen a word embedding as part of an NP sequence before when applied to the test set. This clearly is a worst case scenario, but it reveals us, whether we have been successful. The performance dropped, as expected, but the model still performed reasonably: MLP regression under this condition (second line of Table 4: exclusive) resulted in a $R^2$ of 0.302, BERT regression delivered

[3]However, it depends on the scientific field, see e.g. (Akoglu, 2018)

[4]List of 5,600 common nouns compiled from different sources as potential actors, e.g. professions, groups of persons.

[5]Currently, we haven't considered NPs with opposing polarities (Kiritchenko and Mohammad, 2016) or negation (e.g. 'kein Lügner', Engl. 'not a liar' (Socher et al., 2013)), and thus we don't need to compare to different composition baselines.

[6]The corpus comprises about 1,000,000 articles, we used the dependency parsed version.

[7]We used the pretrained BERT model `huggingface.co/dbmdz/bert-base-german-cased`

0.385. But the more important correlation coefficient $\rho$ can be characterized as moderate, namely 0.564. We can, thus, replace the lexicon-based metric prone to lexicon gaps by a regressor that performs very well on cases our metric could handle as well, but also is useful for cases the metric could not deal with at all.

### 3.2. Manual Evaluation of a Joint Model

In order to evaluate whether the model generalizes also on new potential actors, we generated 670 noun phrases where the head noun was not on the actor list, but where a binary classifier labeled it as 'actor'. We then applied our learned BERT regressor to these phrases in order to quantify their polarity strength to be able to rank them. One evaluation scenario would be to use our metric in order to quantify the NPs and then determine the performance of the BERT regressor with respect to this silver standard. But we wanted to have a gold standard evaluation. It is however not so clear how to accomplish this. Should we manually bring the generated noun phrases in a negativity-based ordering and compare this ranked list to the regressor output? As discussed previously, this set up is bound to fail. We do not have a clear intuition or otherwise sharp decision criterion for evaluating slightly different strength values of two NPs.
Take for example the three NPs:

1. eine grausame Lügnerin (*a horrible liar*)
2. ein wütender Schwätzer (*an angry chatterbox*)
3. ein unglaublicher Lügner (*an incredible liar*)

The strength value for 1) is 4 and that of 3) is 4.25. How could we argue in favour of some ordering, especially out of its sentence context. The annotation task, thus, cannot be to order successive pairs. But what we might be able to detect are NPs that do not fit in well in a window of $N$ predecessors and successors. For instance, 2), 'an angry chatterbox', ranked in between 1) and 3). It is misplaced, because it is obvious that it is a magnitude less negative than the NPs before and after.
The task then was to identify in the ranked list those NPs that are either less or more negative than their neighbours within a window of 5 positions. Two annotators inspected altogether the 670 NPs. The agreement in terms of Cohen's Kappa is 0.75 with an observed agreement of 93.2%. We conclude that the task we set in order to evaluate the results can be carried out reliably. The noise rate measured as the relative frequency of misplacement is 18.1% (annotator A) and 15.7% (annotator B). On that basis, we might conclude that the quality of the ranking is good (81.9% and 84.4%, respectively).
We also carried out a small qualitative evaluation in order to find the kind of errors we have to expect (only the false positives of annotator A). 35% of the false positives are phrases describing victims ('die misshandelte Bevölkerung': 'the maltreated population') rather than negative actors, 10% (presumably) are cases of irony ('der schlechte Retter': 'the bad saviour'). The remaining 55% are errors of the actor classification ('die gute Botschaft': 'the good message').

## 4. Related Work

There are quite some papers on the quantification of polarity strength or intensity for English, both on the word level and on the phrase level as in (Socher et al., 2013). The situation is different for German, where only few approaches exist. One of the first papers dealing with this topic is (Taboada et al., 2011). The authors define a lexicon-based strength metric for the determination of phrase, sentence and text level polarity. In contrast to our metric, the lexicon entries are directly quantified on a scale from -5 to 5. Our intention is to learn a model that replaces the need for a limited (since incomplete) lexicon-based metric. Another resource is SentiWordnet (Esuli and Sebastiani, 2006), where word senses do have positive, negative and neutral strength values. This resource can be used for English if either a word sense disambiguation method was used, or - like Gatti and Guerini (2012) suggest - some prior word level strength values were determined from the senses values. Another approach (Rill et al., 2012) uses star ratings from reviews in order to determine the polarity strength. The assumption there is that the number of stars correlates with the strength of the words being used. Although this might be given partially, it certainly introduces noise as well. In 2016, a SemEval shared task was suggested (Kiritchenko et al., 2016). The gold standard word and phrase-level intensity data was crowd-sourced and calculated on the basis of a metric called best-worst scaling (BWS). This is a very interesting method for the creation of a gold standard and clearly is an option for our future work. Finally, in (Nielsen, 2011) a new version of the ANEW resources was created suited for microblogging data. No such resource for German is available. (Waltinger, 2010) base their strength value determination for German on star ratings, which is only in part reliable.
In contrast to previous work our goal was not to create a lexicon but to learn a model that generalizes in order to capture unseen words and noun phrases. Currently, intensifiers/diminishers as well as negation are not considered. This is future work.

## 5. Conclusions

In this paper, a so-far uncovered NLP task for German has received a first consideration and empirical evaluation: polarity strength quantification of actor denoting noun phrases. Since no annotated corpus is available for this task in German, we applied a bootstrapping approach. Based on lexical resources, we generated a silver standard, evaluated it and learned a regression model that shows enough generalization capacity to yield good results on unseen and non-overlapping data. The ranking of noun phrases that denote actors conceptualized as highly negative by the writer of a text is the main contribution of this paper.

## 6. Acknowledgements

Our work is supported by the Swiss National Foundation (SNF) under the project number 105215_179302.

## 7. References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18, 08.

Clematide, S. and Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 7–13.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Gatti, L. and Guerini, M. (2012). Assessing sentiment strength in words prior polarities. In *Proceedings of COLING 2012: Posters*, pages 361–370, Mumbai, India, December. The COLING 2012 Organizing Committee.

Huang, M., Xie, H., Rao, Y., Feng, J., and Wang, F. L. (2020). Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Inf. Sci.*, 520:389–399.

Kiritchenko, S. and Mohammad, S. M. (2016). Sentiment composition of words with opposing polarities. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1108, San Diego, California, June. Association for Computational Linguistics.

Kiritchenko, S., Mohammad, S., and Salameh, M. (2016). SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 42–51, San Diego, California, June. Association for Computational Linguistics.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 (140).

Martin, J. R. and White, P. R. R. (2005). *Appraisal in English*. Palgrave Macmillan, London, England.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

Plutchik, R. (1980). *A general psychoevolutionary theory of emotion*. Academic press, NewYork.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*, 20:1–10, 01.

Rill, S., Adolph, S., Drescher, J., Reinel, D., Scheidt, J., Schütz, O., Wogenstein, F., Zicari, R. V., and Korfiatis, N. (2012). A phrase-based opinion list for the german language. In Jeremy Jancsary, editor, *11th Conference on Natural Language Processing, KONVENS*, pages 305–313. ÖGAI, Vienna, Austria.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Waltinger, U. (2010). Germanpolarityclues: A lexical resource for german sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May.

## 8. Language Resource References

Manfred Klenner, Anne Göhring. (2022). *Silver Polar Intensity of German NP*. ISLRN n/a.