

You Shall Know a Tool by the Traces it Leaves: The Predictability of Sentiment Analysis Tools

Daniel Baumartz and **Mevlüt Bagci** and **Alexander Henlein** and **Maxim Konca**
and **Andy Lücking** and **Alexander Mehler**

Text Technology Lab

Goethe University Frankfurt, Germany

{baumartz,bagci,henlein,konca,luecking,mehler}@em.uni-frankfurt.de

Abstract

If sentiment analysis tools were valid classifiers, one would expect them to provide comparable results for sentiment classification on different kinds of corpora and for different languages. In line with results of previous studies we show that sentiment analysis tools disagree on the same dataset. Going beyond previous studies we show that the sentiment tool used for sentiment annotation can even be predicted from its outcome, revealing an *algorithmic bias* of sentiment analysis. Based on Twitter, Wikipedia and different news corpora from the English, German and French languages, our classifiers separate sentiment tools with an averaged F_1 -score of 0.89 (for the English corpora). We therefore warn against taking sentiment annotations as face value and argue for the need of more and systematic NLP evaluation studies.

1 Introduction

Mining opinions and valuations – sentiment – is an important tool in the repertoire of commercial and non-commercial social media analysts (Taboada, 2016). Educational data mining, for instance, is employed for evaluating microblogging of educational institutions (Kimmons et al., 2017) and the ranking of universities (Abdelrazeq et al., 2016). Sentiment classification on tweets is part of predicting stock market events (Bollen et al., 2011). Understanding elections refers, among others, to mood as communicated on social platforms (Mohammad et al., 2015). All of these approaches have in common that they use single or multiple NLP tools (e.g., from the field of sentiment analysis) to conduct a study in a research area (e.g., computational sociology, psychology, education) outside NLP. For this purpose, they use tools whose authors usually report F-scores above the corresponding SOTA at the time of publication. It is a truism that different tools can perform differently for the same task on the same text corpus (due to different training

datasets, training data domains, etc.) (King, 1996; Ortmann et al., 2019; Wiegreffe and Marasovic, 2021). However, this becomes problematic if the tools are not systematically compared to examine their impact on reported outcomes. In this case, there is a risk that, for example, a sociological, psychological or educational statement is based on the results of one or a few tools without knowing whether these are valid and whether they are not rather biased. This risk becomes apparent when different tools provide very different sentiment results for the same data. As a result of this diversity, the tools can ultimately become recognizable and thus distinguishable from one another through the sentiment analyses they output. The sentiment analyses are then *algorithmically biased*, so to speak: one cannot claim to have determined valid sentiment values with a particular tool, but only those for which the biasing algorithm being applied was decisive. In short, it is not only since the success of neural networks that many scientific disciplines use NLP tools to conduct text-based studies with the aim of substantiating their research (e.g. Thessen et al., 2012). Our hypothesis is that these tools, although they may produce comparable F-scores in evaluations, tend to produce different distributions of output values that may ultimately cover the entire range of (e.g., of sentiment) values. This results in the predictability of the tools based on the output patterns they produce. And the better this predictability, the higher the algorithmic bias caused by the use of these tools.

In this work, we investigate tool predictability in the area of sentiment analysis using corpora from three languages: English, French and German. We test 9 tools, which, depending on methodology and language results in 9 instances for English and 4 for German and French each. We consider four areas of test data – Twitter, Wikipedia, newspaper and Europarl corpus data – to make our analysis less genre-dependent. Our basic finding is that

tool predictability does indeed exist to a relevant extent: In most cases, it is sufficient to examine some statistical moments (e.g., mean and standard deviation) of the sentiment value distribution generated by a tool for the analyzed texts to know which tool it is. To show this, the paper is organized as follows: In section 2 we present related work on sentiment analysis. In section 3 we introduce the data and tools we cover in our experiments. Section 4 presents our experiments and results, which we discuss in section 5. We conclude in section 6.

2 Background and related work

Twitter is a particularly useful corpus for sentiment studies (Feng et al., 2013). However, there is a lot of evidence that sentiment classification is influenced by linguistic and extra-linguistic factors such as the information within an explicit sentiment lexicon (Agarwal et al., 2011), (non-)recognition of implicit sentiment (Van Hee et al., 2021), and specific syntactic constructions (Verma et al., 2018), but not by translated tweets (Salameh et al., 2015; Lu et al., 2011). The choice of the tool used for assessing sentiment has a major effect on the outcome of any classification, too. A systematic comparison of 20 sentiment recognition tools (15 stand-alone, 5 workbenches) on texts from five tweet genres – Pharma, Retail, Security, Tech, Telco – has been carried out by Abbasi et al. (2014). Crucially, the authors found out that the tools not only performed differently on average, but also varied for each text genre. The main error the tools are prone to is ascribed to a lack of recognizing user intentions. Little agreement of tools to manually labeled datasets, and even less agreement among the tools themselves, has also been observed with software engineering data (Jongeling et al., 2015); the best (that is, the least worst) performing tools (NLTK and SentiStrength) even come to disagreeing classifications for different kinds of datasets. Although combining various sentiment classifiers leads to a wider coverage, it does not necessarily lead to a better F-score (Gonçalves et al., 2013). It has to be kept in mind, however, that any comparisons can be influenced by differences in annotation guidelines and annotators (Mozetič et al., 2016) or varying decisions on stop word lists (Maynard and Bontcheva, 2016). The biases bound up with tool choice are amplified by biases of the language models used by the tools (e.g. Fokkens et al., 2013) and their size (Bender et al., 2021). Experimenting

with random seeds and stochastic weight averaging (SWA) from the BERT-based ALBERTA model on sentiment data (compared with a check list evaluation), Khurana et al. (2021), e.g., found that the error rate of random seeds is reduced by SWA but the still leaves a margin of instability. Thus, though important for commercial and non-commercial applications, from recent related work it is known that sentiment classification of tweets and on other kinds of corpora appears to be a fragile affair. In what follows, we show that we are able to identify sentiment analysis tools based solely on their produced output values – shedding light on another aspect of the susceptibility of sentiment analysis.

3 Data and tools

The following corpora and sentiment tools have been selected for our experiments.

3.1 Data

We consider three languages: English (EN), French (FR) and German (DE). For each, we built three corpora based on tweets taken from Twitter and a selection of Wikipedia and newspaper articles. In addition, we used the parallel Europarl corpus to compare sentences in these three languages with the same underlying meaning. Table 1 in the appendix provides an overview of the corpora.

Twitter Using the Full-Archive-Search of their API, we downloaded tweets from Twitter querying for the following hashtags related to the COVID-19 pandemic, that is discussions in the German speaking community: #Aufschrei, #allemalneschichtmachen, #allenichtganzdicht, #allesdichtmachen, #lockdownfuerimmer, #niewiederaufmachen and #TatortBojkott, as well as the language independent topic about the Sputnik V vaccine: #SputnikV (Abrami and Mehler, 2021). The tweets were created between 2013-06-03 and 2021-04-24 (see Table 1 for more details).

Wikipedia We randomly selected a sample of about 20 000 Wikipedia articles independently from the English, French and German Wikipedia, based on the 2021-06-20 dump.

News corpora We analyzed around 20 000 newspaper articles from the New York Times (NYT) (New York Times, 2019), the Süddeutsche Zeitung (SDZ) (Süddeutscher Verlag, 2014) and 10 000 articles from the French 2020

10k Newscrawl by the Leipzig Corpora Collection (Goldhahn et al., 2012). The articles from NYT and SDZ were randomly selected to be evenly distributed by percentage over the years of publication, ranging from 1987 to 2019 in case of the NYT and 1992 to 2014 in case of SDZ.

Europarl Using the Europarl (Koehn, 2005) corpus (released 2012 in version 7), we created a parallel mapping of sentences of all three languages.

The corpora have all been preprocessed by spaCy¹ (using the small/efficient models) running inside TextImager² (Hemati et al., 2016), an NLP platform based on Apache UIMA. We integrated the sentiment tools described in the next section into our platform by mapping their inputs and sentiment output to UIMA.

3.2 Sentiment Analysis Tools

We computed coarse-grained polarity sentiments for all corpora of the three languages using tools in two ways: Firstly, the entire text (i.e., the full tweet or Wikipedia article) is used as input, and secondly, each of its sentences is processed individually. To always generate a single sentiment value per text, we followed the recommendation of the tools and average over the sentiments of all sentences. In this way, we generate two sentiment scores for each language per corpus, per tool and text.

We chose tools that operate differently to capture and compare a broader range. This includes tools that utilize heuristics build upon lexica as well as model-based algorithms. To unify the usage of different tools, we map their respective outputs to a sentiment range of -1 (negative) to $+1$ (positive). The following tools are used (for an overview of them, the data on which their sentiment detection is based, and links to the implementations we used in our experiments, see Table 2 in the appendix):

TextBlob offers two methods for sentiment detection that are lexicon- and model-based. The lexicon is being used by a pattern analyzer based on the pattern library (Smedt and Daelemans, 2012). The analyzer produces sentiment polarity values in the range of -1 (negative) to $+1$ (positive). The model-based method uses a naive Bayes classifier. Sentiments are output by classification (*pos* or *neg*)

¹<https://spacy.io>

²The TextImager platform has since been further developed into the *Docker Unified UIMA Interface* (DUUI) (Leonhardt et al., 2023).

together with two values denoting positive and negative scores in the interval $[0, 1]$. We use the positive score if the classification yields *pos*, otherwise the negative score multiplied by -1 to comply with our sentiment range. We use *textblob-de*, which provides a German sentiment lexicon mainly based on the German Polarity Lexicon (Clematide and Klenner, 2010), and *textblob-fr* for French data.

Stanza Qi et al. (2020) provide sentiment detection for Chinese, English and German by means of a CNN-based classifier based on (Kim, 2014). It produces a discrete value of 0 , 1 or 2 denoting negative, neutral or positive sentiment, allowing for a direct mapping to our schema. Multiple datasets are used by the English model, that is the Stanford Sentiment Treebank³, MELD (Poria et al., 2019; Zahiri and Choi, 2018), Sentiment Labelled Sentences Data Set (Kotzias et al., 2015), ArguAna TripAdvisor Corpus (Wachsmuth et al., 2014) and Twitter US Airline Sentiment⁴. The German model is trained on the German Tweet corpus SB-10k (Cieliebak et al., 2017).

VADER (Hutto and Gilbert, 2014) uses a manually validated lexicon based on the sentiment word-banks Linguistic Inquiry Word Count (Andrei et al., 2014), General Inquirer (Stone and Hunt, 1963) and Affective Norms for English Words (Bradley and Lang, 1999), enhanced by expressions commonly used in social media posts. Its rules are based on five heuristics that incorporate word-order sensitive relationships. We use the continuous compound score of VADER as sentiment value, which lies in the interval of -1 (negative) to $+1$ (positive). To process German text we utilize GerVADER (Tyman et al., 2019) which provides a new lexicon, mostly based on SentiWS (Remus et al., 2010) with additions by Langenscheidt⁵ and CoolSlang⁶, and updated heuristics. VADER-FR is used for French text; it contains a manually translated lexicon.

We use six different single- and multi-language models based on language models like BERT (Devlin et al., 2019) and derivates. All these models produce discrete sentiment values that we mapped to -1 , 0 , $+1$ and interim values, if needed.

³<https://github.com/stanfordnlp/sentiment-treebank>

⁴<https://www.kaggle.com/crowdflower/twitter-airline-sentiment/data>

⁵<https://www.langenscheidt.com/jugendwort-des-jahres>

⁶<https://www.coolslang.com>

cardiffnlp/twitter-roberta-base-sentiment (Barbieri et al., 2020) is trained on English tweets and finetuned on the evaluation framework TweetEval, with sentiment data from SemEval 2017 Task 4 Subtask A (Rosenthal et al., 2019).

cardiffnlp/twitter-xlm-roberta-base-sentiment (Barbieri et al., 2021) supports multiple languages. We used it to process English and French texts. It is trained on Twitter and finetuned on the UMSAB dataset, which is based on SemEval 2017 Task 4 Subtask A (Rosenthal et al., 2019) (EN), SB-10k (Cieliebak et al., 2017) (DE) and Deft 2017 (Benamara et al., 2017) (FR).

finiteautomata/bertweet-base-sentiment-analysis (Pérez et al., 2021) is based on BERTweet (Nguyen et al., 2020), a language model of English tweets, and trained with TASS 2020 Task 1 (Vega et al., 2020) and SemEval 2017 Task 4 Subtask A (Rosenthal et al., 2019).

nlptown/bert-base-multilingual-uncased-sentiment was used to detect sentiments in our English and French corpora. It is trained on product reviews.

oliverguhr/german-sentiment-bert Guhr et al. (2020) trained a model with German texts from various domains including Twitter: PotTS (Sidarenka, 2016), SB-10k (Cieliebak et al., 2017), GermEval-2017 (Wojatzki et al., 2017), Scare (Sänger et al., 2016), leipzig-wikipedia (Goldhahn et al., 2012), crawled data from Filmstarts⁷ and HolidayCheck⁸ and Emotions based on experiments with service robots (Guhr et al., 2020).

siebert/sentiment-roberta-large-english Heitmann et al. (2020) provide a binary sentiment detection model for English trained on a total of 15 datasets by Blitzer et al. (2007); Pang and Lee (2005); McAuley and Leskovec (2013); Speriosu et al. (2011); Hartmann et al. (2019); Maas et al. (2011); Nakov et al. (2013); Shamma et al. (2009); Pang et al. (2002), as well as the Yelp Academic⁹ and Kaggle sentiment140¹⁰ datasets. Note that while this model cannot produce neutral sentiments, such values can still appear in case of the variant where we consider sentences as inputs and average over them.

4 Experiments

We evaluate the sentiment values computed for the corpora of subsection 3.1 and find a large variance between the different tools. This becomes apparent when the mean and standard deviation are calculated across all sentiments – see Figure 1 for an overview of the values on the EN Twitter #SputnikV corpus and the EN Wikipedia. Figure 6 in the appendix exemplifies these analyses for DE and FR. Figure 2 shows the distance correlation of the sentiment predictions for all EN corpora except Europarl (see Figure 7 in the appendix for DE and FR). The correlations are low (indicated by light green and white, respectively), suggesting that the tools make rather heterogeneous predictions. *But does that finding already make them predictable?*

Figure 3 shows how often the tools for EN returned a result equal to a majority vote. For each document, this vote is calculated using the original sentiment tool results, normalized in four different ways (as described in subsection 4.1). As expected, the agreement decreases with more possible output margins due to normalization. We observe the same output for DE and FR (s. Figure 8 in the appendix). 3 of the 5 transformer models have almost the same agreement rate. This could indicate that they produce similar analyses. So are they indistinguishable? We will now show that they are.

Our main hypothesis about tool predictability is that we are able to identify the tool used to perform sentiment analysis by the statistical signature of its output. To test this hypothesis, we train several classifiers that recognize tools based on this signature (in terms of *mean*, *std* etc.). We use the tools as target labels and perform a binary or multi-class classification, depending on the underlying tool combination. In addition to experiments that consider all data and tools, we explore several combinations based on the properties of the tools and the domains of our corpora.

4.1 Neural network classifier

Using PyTorch (Paszke et al., 2019) and scikit-learn (Pedregosa et al., 2011) we train a shallow neural network (NN) with one hidden linear layer, a ReLU activation function and softmax output. We generate training samples by first randomly collecting subsets of documents. Then we build for each of these subsets and each tool a so called chunk as the multiset of all sentiment values that this tool produces for the documents included in

⁷<https://www.filmstarts.de>

⁸<https://www.holidaycheck.de>

⁹<https://www.yelp.com/dataset>

¹⁰<https://www.kaggle.com/kazanova/sentiment140>

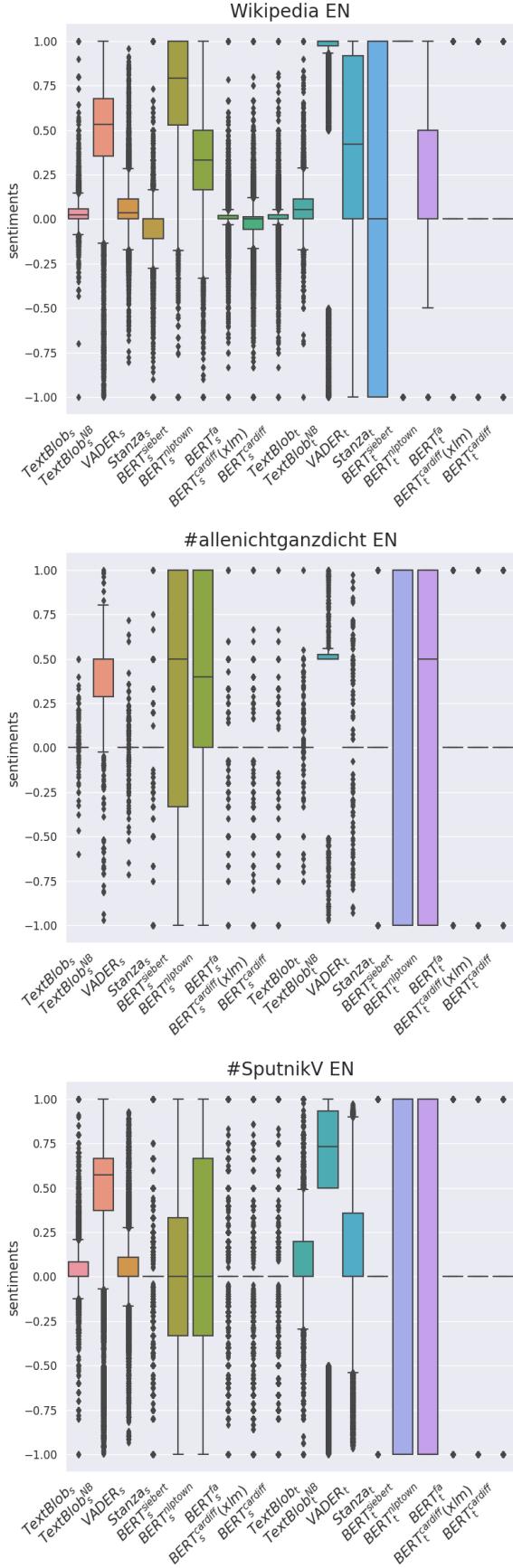


Figure 1: Sentiments of English Wikipedia and Twitter hashtags #allenichtganzdicht and #SputnikV.

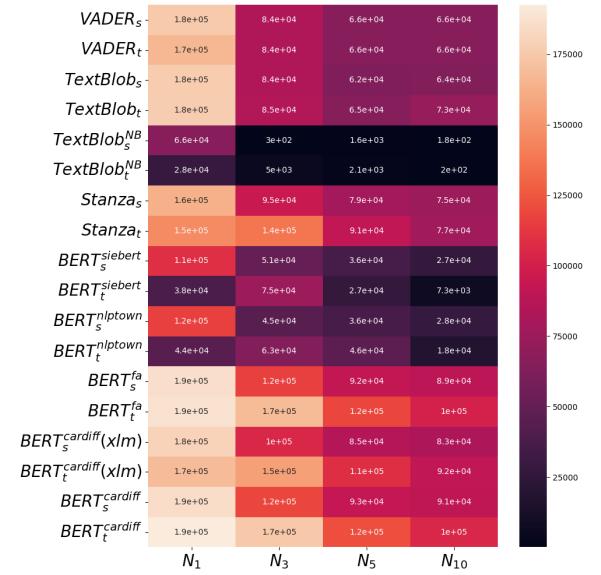
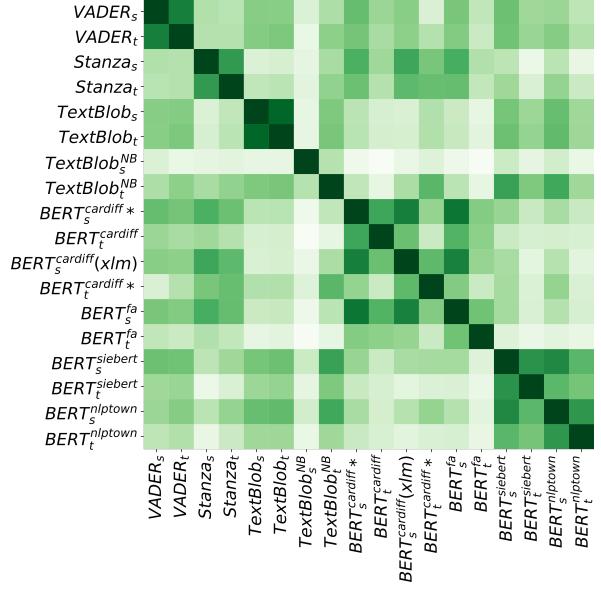


Figure 3: Per-tool rate of agreement with majority vote for the EN C3 corpus using 4 normalization methods.

the subset. Given a tool, we then compute for each chunk 13 different statistical moments using numpy (Harris et al., 2020) (i.e., mean, std, var, median, min, max and the 5-, 10-, 25-, 50-, 75-, 90- and 95-percentiles). In this way, for each of the n tools, we obtain a feature vector consisting of 13 times the number of chunks many elements as input to classification. We vary chunk size in the range of 50–1000 documents; in subsection 4.3 we additionally show effects of varying the sampling method. The training dataset is randomly split in a train set (70 %), a development set and a test set (15 % each).

Because the tools output different ranges of sentiment, we conducted additional experiments in which we normalized all values and classified the tools into different groups based on their methodology. We normalize sentiment values in four ways: firstly, we discretize them by returning *positive* for values ≥ 0.5 , *negative* for ≤ -0.5 , and *neutral* otherwise; this results in N_1 . Secondly, we generate normalization N_3 according to the membership of sentiment values in one of the three intervals $(-\infty, -0.333]$, $(-0.333, 0.333]$, and $[0.333, \infty)$. Finally, we divide the ranges of sentiment scores into five (N_5) or ten (N_{10}) equally sized intervals to account for more sentiment gradations. The tools are independently divided into the following groups: a group containing all tools that provide discrete sentiment outputs (i.e., the transformer-based models and Stanza), and the group of all remaining tools (i.e., TextBlob and VADER) that output numerical values. This is to investigate the difficulty of tool prediction when only a few different sentiment values are produced, as is the case with discrete outputs. To prevent that differences in processing the inputs at sentence or text level make the classification too easy, we form two additional groups along this criterion, which are studied in a separate classification. This also aligns with the fact that tools actually operate sentence-wise. Including the group of all tools, this results in five different groups in our study.

We train the classifiers for the latter groups using our datasets: the first training scenario uses our Twitter data, called corpus C1, by considering the tweets of all hashtags mentioned in subsection 3.1. The Wikipedia and the newspaper corpora of the languages are used to generate a second dataset (called C2). Thirdly, we combine both datasets to form a third dataset (C3) for training. We consider

C1	0.83	0.95	0.60
C2	0.91	0.80	0.64
C3	0.89	0.95	0.65
C4	0.38	0.88	0.64
EN			
DE			
FR			

Figure 4: Results of the NN-based classifier: mean F_1 -scores over all languages and corpora, averaged over all chunk sizes and then over all normalization variants.

the Europarl corpus (C4) separately.

Using the development sets, we perform a parameter study on the size of the hidden layer (5–300, the optimizer (SGD and Adam), the learning rate (0.001–0.1) and the number of training epochs (5–100). This brings the total number of experiments to 432 000, that is 180 per model, for each of the 800 combinations described above for each of the three languages. We double the number of experiments with the German corpus to compare the differences when using a scaler on the feature data; however, we find no significant differences when using the NN. Table 3 in the appendix shows a subset of the F_1 -scores of the development and test sets for English (based on scikit-learn); Table 4 shows statistics of the distributions of all scores.

If all tools predicted the same sentiment distributions for the same input texts, the trained classifier would not be able to distinguish these tools based on the statistical moments of these distributions. However, we detect high F_1 -scores across almost all experiments, combinations, and languages, as shown by the mean F_1 -scores in Figure 4.

English Considering corpus C3 without normalization of the sentiment values, we reach an F_1 -score of 0.927 averaging over the results using different chunk sizes, with the smallest F_1 -score of 0.867 at a chunk size of 50 and the largest of 0.988 at a chunk size of 1 000. Normalizing the sentiment values according to N_3 lowers performance to a still high F_1 -score of 0.797, mainly due to effects of small chunk size (i.e., 0.684 at size 50; at size 1 000 the F_1 -score reaches 0.927). The other normalization methods, which leave more room for different sentiment values, result in higher average scores of 0.824, 0.903, and 0.918.

Looking closer at different corpora, we discover an average F_1 -score of 0.86 on Twitter (max 0.959)

and 0.959 on Wikipedia and the NYT (here even reaching 1.0). This suggests that the tweets are more difficult to process. In the case of grouping tools according to whether they produce discrete or continuous sentiment values, we obtain mostly similar results, with higher scores for the latter group. We attribute this to the fact that they leave more room for different sentiments. If we consider the group that emerges when sentiment values are generated from averaging over sentences, we see a decrease from 0.911 to 0.883 for tools with discrete sentiment values and an improvement from 0.978 to 0.992 for those with continuous values. This is expectable, since in the case of discrete sentiments we only perform a binary classification of more difficult-to-separate data.

These results support our claim that one can predict sentiment analysis tools based on simple statistics of the value distributions they produce, regardless of the text genre and the underlying method. Our results on German and French show that this is predominantly true for these languages as well:

German As with the English corpora, we see a quite high average F_1 -score of 0.982 on C3, score of 0.985 on Twitter and of 0.933 on C2. The classifier works perfectly with discrete sentiments on C3 and on Twitter data (F_1 -score: 1.0) and a score close to 0.997 in the case of C2. Results on the sentence only data is equally high. Scores on continuous data is a bit lower, with the sentence-based experiments performing on the same level again.

French The values for the French corpora are generally lower, with an average of 0.694 for C3, 0.648 for Twitter, and 0.683 for Wikipedia and News crawl. We find that the classifier has difficulties separating the tools that provide discrete sentiment scores (F_1 score: 0.415). This can be seen in the boxplots in appendix Figure 6, where the BERT-based models produce very similar statistical distributions. However, the tools producing continuous sentiment values are well separated (average F_1 -score: 0.996).

Europarl While we obtain very good results on the Twitter, Wikipedia, and news corpora, our classifier is not able to distinguish the tools based on the sentiments they generate for the English Europarl corpus. The F_1 value drops to 0.169 for most configurations and groups, except for the continuous data based on sentences, where we reach a value

of 0.967. However, the values for the German and French corpora are on par with those for the other corpora (0.954 and 0.696, respectively). This implies indistinguishable distributions of sentiment values in the Europarl corpus, at least when considering English sentiment tools, which we have previously shown to be distinguishable in the other corpora. Given the high scores for the German Europarl corpus, we can probably rule out the possibility that this result is due to the simplicity and brevity of the Europarl texts; rather, it is likely a result of too little training data.

4.2 Other classifiers

We alternatively experimented with *Support Vector Machines* (SVM) and *Decision Trees* (DT) to see if they lead to similar results. More specifically, we trained k -nearest neighbors (KNN, $k=5$), SVM with linear and rbf kernels, and a DT with a maximum depth of 5. The best results were obtained with KNN with an average F_1 -score of 0.903 on C3 (with 0.981 max). Tools operating on Twitter are also well separated with an average score of 0.845; the C2 corpus yields an F_1 -score of 0.94. Consistent with the results based on the NN classifier, Europarl is problematic with a low average score of 0.213. The other tools perform on average on C3 as follows: DT: 0.551 (0.594 max), SVM-linear: 0.831 (0.881 max), SVM-rbf: 0.811 (0.843 max). On the German corpora, the KNN and SVM classifier perform extremely well, often reaching F_1 -scores of or close to 1.0, over all configurations and groups analyzed. In line with our experiments using a NN, the scores on the French corpora are lower overall, due to the results on the discrete sentiment values. On the C3 corpus, however, the SVM-linear still achieves an average F_1 -score of 0.717 (0.831 max). The appendix shows visualizations of the DTs trained on C3 (Figure 10) and the Europarl corpus (Figure 11), each for EN, DE and FR.

4.3 Monte Carlo sampling

To investigate, whether the results obtained above hold under different sampling regimes, we perform a Monte Carlo simulation, where we draw with replacement a 1%-sized subset of the total sentiment predictions. For each tool we generate m such samples and calculate sample statistics, that is, two variants of entropy and the other moments (see subsection 4.1), for each. That way we produce $n \times m$ feature vectors of length l , where n is

the number of tools analyzed in the experiment, $m \in \{10, 50, 100, 500, 1000\}$, and $l \in [1, 15]$ is the number of features. Subsequently, we train an SVM by means of scikit-learn, that predicts a tool based on the generated feature vectors. Ideally, if the tools produce similar predictions, the SVM would be unable to find a separating hyperplane, resulting in low F_1 -scores. However, the F_1 -scores are significantly higher than expected, see Figure 5 (and Figure 9 in the appendix for results on DE and FR) for the C3 dataset, reaching 1.0 for German, ≈ 1.0 for English and ≈ 0.86 for French. These results agree with low distance correlation scores of Figure 2 and the obtained results so far. As in the previous sections, the only experiment, where the SVM struggles, is the English Europarl corpus.

5 Discussion

Using microblogging data and data from three other genres, our results show that we are able to distinguish sentiment analysis tools based on simple statistical features of their sentiment value distributions. Of course, this result depends on the tools we choose and the corpora we use for evaluation. In particular, the size of the corpora and the subsequently generated training and test samples have influence on our findings. However, looking at the mean F -scores averaged over the different normalization methods and sample sizes, we see that our hypothesis is not falsified. That is, different sentiment analysis tools produce different sentiment distributions that can be distinguished *and* identified after the fact. In other words, *you shall know a sentiment tool by the traces it leaves*. In light of our introduction, this means that the increasing use of NLP tools to support empirical claims in, e.g., the social sciences, education, or the humanities, faces the difficulty of substantiating the validity of its findings. At the very least, these approaches should clarify the extent to which their results are algorithmically biased by the chosen tool and how those results change by using other tools.

We selected Wikipedia and news corpora to contrast the short, rather subjective microblogging texts with texts from other genres and to test whether our hypothesis is also confirmed by them. The use of Europarl data served the comparison under the condition of considering texts with the same meaning for the targeted languages. The differences in the results for Europarl suggest that additional languages should be investigated.

We considered a number of tools to cover different aspects of sentiment detection methods, i.e., we evaluated lexicon-based and model-based approaches. Since the use of SOTA transformer-based models has been shown to be very effective in various NLP tasks, we have also included them. We show that these tools produce partially similar results. This could be due to the fact that they generate discrete sentiment values or use the same model architecture. To control for the first of these candidates, we used different normalization variants to discretize the results of all tools. We found that classifiers trained with normalizations that partition their range of values to produce more discrete values have a higher F_1 -score than classifiers based on normalizations that result in less discrete sentiment values. To further substantiate this research, more tools for more genres should be considered.

6 Conclusion

We examined for three languages, with multiple corpora each, whether different sentiment tools are distinguishable with regards to their outputs. The sentiment tools analyzed mostly use lexicons or transformer-based models trained with data from Twitter or reviews. In several experiments, we trained classifiers that could identify a tool based on statistical moments of the sentiment value distribution it produced. This research shows that empirical studies relying on a single sentiment analysis tool may be algorithmically biased by the choice of that tool: other tools are likely to produce different, readily identifiable results for the same data. To support this finding, we argue for an expansion of research on algorithmic biasing. In future work, we plan to extend our experiments to more tools and corpora from other genres and languages. This will include experimenting with pre-processing, sampling, and feature selection methods.

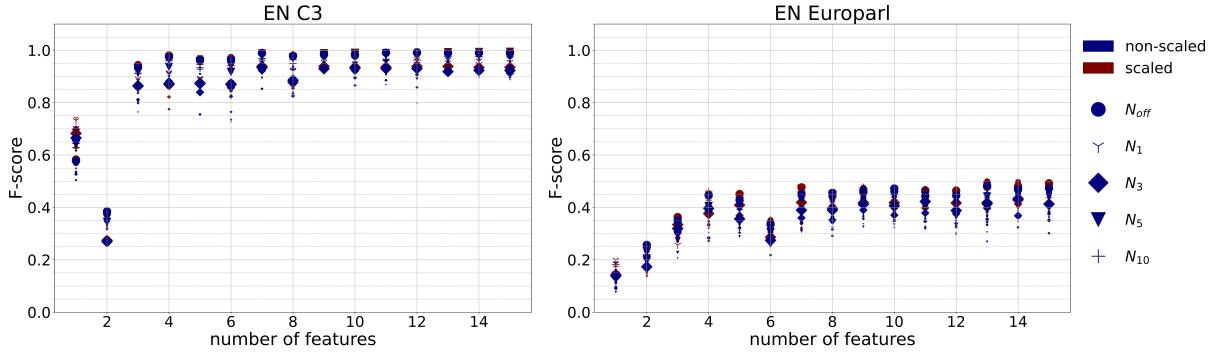


Figure 5: F_1 -scores of the SVM classifier for the C3 and Europarl corpora. The x-axis indicates the number of features, which are selected randomly for each feature vector size independently.

References

- Ahmed Abbasi, Ammar Hassan, and Milan Dhar. 2014. Benchmarking Twitter sentiment analysis tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 823–829, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Anas Abdelrazeq, Daniela Janßen, Christian Tummel, Sabina Jeschke, and Anja Richert. 2016. Sentiment analysis of social media for evaluating universities. In Sabina Jeschke, Ingrid Isenhardt, Frank Hees, and Klaus Henning, editors, *Automation, Communication and Cybernetics in Science and Engineering 2015/2016*, pages 233–251. Springer International Publishing, Cham.
- Giuseppe Abrami and Alexander Mehler. 2021. Collection of Tweets covid-19 pandemic, 2013-06-03 – 2021-04-24. Downloaded with TTLab Twitter API.
- Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon. Association for Computational Linguistics.
- Amanda Andrei, Alison Dingwall, Theresa Dillon, and Jennifer Mathieu. 2014. Developing a tagalog linguistic inquiry and word count (LIWC) ‘disaster’ dictionary for understanding mixed language social media: A work-in-progress paper. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@EACL 2014, April 26, 2014, Gothenburg, Sweden*, pages 91–94. The Association for Computer Linguistics.
- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for Twitter. <https://arxiv.org/abs/2104.12250>.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.
- Farah Benamara, Cyril Grouin, Jihen Karoui, Véronique Moriceau, and Isabelle Robba. 2017. Analyse d’opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017. In *Atelier TALN 2017 : Défi Fouille de Textes (DEFT 2017)*, pages pp. 1–12, Orléans, FR. Association pour le Traitement Automatique des Langues (ATALA).
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.
- Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.
- Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
- Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.
- S Clematide and M Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In

- Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 7–13, Lisbon, Portugal.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Shi Feng, Le Zhang, Binyang Li, Daling Wang, Ge Yu, and Kam-Fai Wong. 2013. Is Twitter a better corpus for measuring sentiment similarity? In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 897–902, Seattle, Washington, USA. Association for Computational Linguistics.
- Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701. Association for Computational Linguistics.
- Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23–25, 2012*, pages 759–765. European Language Resources Association (ELRA).
- Pollyanna Gonçalves, Matheus Araújo, Fabrício Benvenuto, and Meeyoung Cha. 2013. Comparing and combining sentiment analysis methods. In *Proceedings of the First ACM Conference on Online Social Networks, COSN ’13*, page 27–38.
- Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a broad-coverage German sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France. European Language Resources Association.
- Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585:357–362.
- Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38.
- Mark Heitmann, Christian Siebert, Jochen Hartmann, and Christina Schamp. 2020. More than a feeling: Benchmarks for sentiment analysis accuracy. Available at SSRN 3489963.
- Wahed Hemati, Tolga Uslu, and Alexander Mehler. 2016. Textimager: a distributed uima-based system for nlp. In *Proceedings of the COLING 2016 System Demonstrations*. Federated Conference on Computer Science and Information Systems.
- Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014*. The AAAI Press.
- Robbert Jongeling, Subhajit Datta, and Alexander Serebrenik. 2015. Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 531–535.
- Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2021. How emotionally stable is ALBERT? Testing robustness with stochastic weight averaging on a sentiment analysis task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
- Royce Kimmons, George Veletsianos, and Scott Woodward. 2017. Institutional uses of Twitter in U.S. higher education. *Innov High Educ*, 42:97–111.
- Margaret King. 1996. Evaluating natural language processing systems. *Commun. ACM*, 39(1):73–79.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13–15, 2005*, pages 79–86.
- Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10–13, 2015*, pages 597–606. ACM.

- Alexander Leonhardt, Giuseppe Abrami, Daniel Baumart, and Alexander Mehler. 2023. Unlocking the heterogeneous landscape of big data NLP with DUUI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 385–399, Singapore. Association for Computational Linguistics.
- Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 320–330, Portland, Oregon, USA. Association for Computational Linguistics.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Diana Maynard and Kalina Bontcheva. 2016. Challenges of evaluating sentiment analysis tools on social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).
- Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, page 165–172, New York, NY, USA. Association for Computing Machinery.
- Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.
- Igor Mozetič, Miha Grčar, and Jasmina Smilović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5):1–26.
- Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.
- New York Times. 2019. New York Times. <https://developer.nytimes.com/apis>. Accessed: 2019; Data provided by The New York Times.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.
- Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating off-the-shelf NLP tools for German. In *KONVENS*, pages 212–222.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.
- Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A Python toolkit for sentiment analysis and SocialNLP tasks. <https://arxiv.org/abs/2106.09462>.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – A publicly available German-language resource for sentiment analysis. In *Proceedings of the International Conference on Language*

- Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.* European Language Resources Association.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in Twitter. *CoRR*, abs/1912.00741.
- Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics.
- Mario Sänger, Ulf Leser, Steffen Kemmerer, Peter Adolphs, and Roman Klinger. 2016. SCARE – the sentiment corpus of app reviews with fine-grained annotations in German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).
- David Shamma, Lyndon Kennedy, and Elizabeth Churchill. 2009. Tweet the debates: Understanding community annotation of uncollected sources. *Tweet the debates: understanding community annotation of uncollected sources*, pages 3–10.
- Uladzimir Sidarenka. 2016. PotTS: The Potsdam Twitter Sentiment Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).
- Tom De Smedt and Walter Daelemans. 2012. Pattern for Python. *J. Mach. Learn. Res.*, 13:2063–2067.
- Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldwin. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63, Edinburgh, Scotland. Association for Computational Linguistics.
- Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the General Inquirer system. In *Proceedings of the 1963 spring joint computer conference, AFIPS 1963 (Spring), Detroit, Michigan, USA, May 21-23, 1963*, pages 241–256. ACM.
- Süddeutscher Verlag. 2014. Süddeutsche Zeitung. Süddeutscher Verlag.
- Maite Taboada. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347.
- Anne E. Thessen, Hong Cui, and Dmitry Mozzherin. 2012. Review article: Applications of natural language processing in biodiversity science. *Advances in Bioinformatics*, 2012:1–17.
- Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. 2019. GerVADER – A German adaptation of the VADER sentiment analysis tool for social media texts. In *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen", Berlin, Germany, September 30 - October 2, 2019*, volume 2454 of *CEUR Workshop Proceedings*, pages 178–189. CEUR-WS.org.
- Cynthia Van Hee, Orphee De Clercq, and Veronique Hoste. 2021. Exploring implicit sentiment evoked by fine-grained news events. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 138–148, Online. Association for Computational Linguistics.
- Manuel García Vega, Manuel Carlos Díaz-Galiano, Miguel Ángel García Cumbreiras, Flor Miriam Plaza del Arco, Arturo Montejo-Ráez, Salud María Jiménez Zafra, Eugenio Martínez Cámará, César Antonio Aguilar, Marco Antonio Sobrevilla Cabezudo, Luis Chiruzzo, and Daniela Moctezuma. 2020. Overview of TASS 2020: Introducing emotion detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*, Málaga, Spain, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, pages 163–170. CEUR-WS.org.
- Rohil Verma, Samuel Kim, and David Walter. 2018. Syntactical analysis of the weaknesses of sentiment analyzers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1122–1127, Brussels, Belgium. Association for Computational Linguistics.
- Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 115–127, Berlin Heidelberg New York. Springer.
- Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, pages 1–12.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, volume WS-18 of *AAAI Workshops*, pages 44–52. AAAI Press.

A Appendix

We provide additional visualizations and data in the appendix on the next pages, mainly showing results on German and French corpora as well as an more detailed overview of the neural network based classifier results on English corpora and the decision trees.

Table 1 gives an overview of the corpora, with indication of its size and covered time spans. We show on what data domain the tools we experimented with are trained in Table 2, which also provides links to the implementation on GitHub and Hugging Face. Additional visualizations of the German and French corpora are given in the figures 6, 7, 8 and 9. Figure 10 and 11 show cutouts of the decision trees for all languages. We add tables showing samples of the experiment results on the English corpora. That is Table 3, which shows macro and weighted F_1 -scores for the development and test set for a selection of groups and corpora, and Table 4 which provides statistical data about the scores collected over all chunk sizes. Lastly we provide a 3D visualization of Table 4 in Figure 12.

Lang	Corpus	Included in corpus	# Docs	# Token	Date
EN	Wikipedia	C2, C3	21 997	10 430 671	Dump 2021-06-20
EN	New York Times	C2, C3	21 960	13 873 994	1987 – 2019
EN	Europarl	C4	11 000	313 112	Release v7 2012
EN	#Aufschrei	C1, C3	777	20 465	2013-06-03 – 2021-05-21
EN	#allemalneschichtmachen	C1, C3	752	7 324	2021-04-24 – 2021-04-26
EN	#allenichtganzdicht	C1, C3	857	11 423	2020-04-14 – 2021-04-23
EN	#allesdichtmachmen	C1, C3	1 135	14 807	2021-03-11 – 2021-04-23
EN	#lockdownfuerimmer	C1, C3	1 498	19 726	2020-06-23 – 2021-04-23
EN	#niewiederaufmachen	C1, C3	1 514	18 702	2021-04-22 – 2021-04-23
EN	#SputnikV	C1, C3	164 504	3 332 801	2020-08-11 – 2021-05-03
EN	#TatortBoykott	C1, C3	164	1 123	2017-12-15 – 2021-05-03
DE	Wikipedia	C2, C3	21 999	10 707 115	Dump 2021-06-20
DE	Süddeutsche Zeitung	C2, C3	22 001	9 365 670	1992 – 2014
DE	Europarl	C4	11 000	350 864	Release v7 2012
DE	#Aufschrei	C1, C3	197 713	5 356 038	2013-06-03 – 2021-05-21
DE	#allemalneschichtmachen	C1, C3	9 602	221 448	2021-04-24 – 2021-04-26
DE	#allenichtganzdicht	C1, C3	36 189	836 361	2020-04-14 – 2021-04-23
DE	#allesdichtmachmen	C1, C3	24 826	565 868	2021-03-11 – 2021-04-23
DE	#lockdownfuerimmer	C1, C3	17 113	336 476	2020-06-23 – 2021-04-23
DE	#niewiederaufmachen	C1, C3	19 280	394 170	2021-04-22 – 2021-04-23
DE	#SputnikV	C1, C3	14 873	374 925	2020-08-11 – 2021-05-03
DE	#TatortBoykott	C1, C3	4 061	93 634	2017-12-15 – 2021-05-03
FR	Wikipedia	C2, C3	21 993	9 342 397	Dump 2021-06-20
FR	Newscrawl 2020	C2, C3	10 000	240 548	2020
FR	Europarl	C4	11 000	354 870	Release v7 2012
FR	#SputnikV	C1, C3	14 566	411 391	2020-08-11 – 2021-05-03

Table 1: Overview of the corpora used in our study, the number of documents (articles or tweets), and the date or version in which these data were obtained.

Tool	Implementation	①	②	③	④	⑤	⑥
TextBlob EN	[1] sloria/textblob						✓
TextBlob NB EN	[1] sloria/textblob						✓
TextBlob DE	[1] markuskiller/textblob-de					✓	
TextBlog FR	[1] sloria/textblob-fr						✓
Stanza EN	[1] stanfordnlp/stanza	✓		✓	✓	✓	
Stanza DE	[1] stanfordnlp/stanza	✓					
VADER EN	[1] cjhutto/vaderSentiment					✓	
GerVADER	[1] KarstenAMF/GerVADER					✓	
VADER FR	[1] thomas7lieues/vader_FR					✓	
twitter-roberta	[2] cardiffnlp/twitter-roberta-base-sentiment	✓					
twitter-xlm-roberta	[2] cardiffnlp/twitter-xlm-roberta-base-sentiment	✓					
finiteautomata	[2] finiteautomata/bertweet-base-sentiment-analysis	✓					
nlptown	[2] nlptown/bert-base-multilingual-uncased-sentiment			✓			
german-sentiment	[2] oliverguhr/german-sentiment-bert	✓	✓	✓	✓		
siebert	[2] siebert/sentiment-roberta-large-english	✓		✓			

Table 2: Overview of the datasets grouped by domain used by the selected sentiment detection tools and the implementation used by our experiments. The data is grouped in the following categories: ① Social media posts like Twitter, ② specialised texts based on books or Wikipedia, ③ text taken from ratings or reviews, ④ emoticons, ⑤ entries from semantic lexicons and ⑥ unknown training data. [1] denotes a link to GitHub and [2] to the Hugging Face repositories.

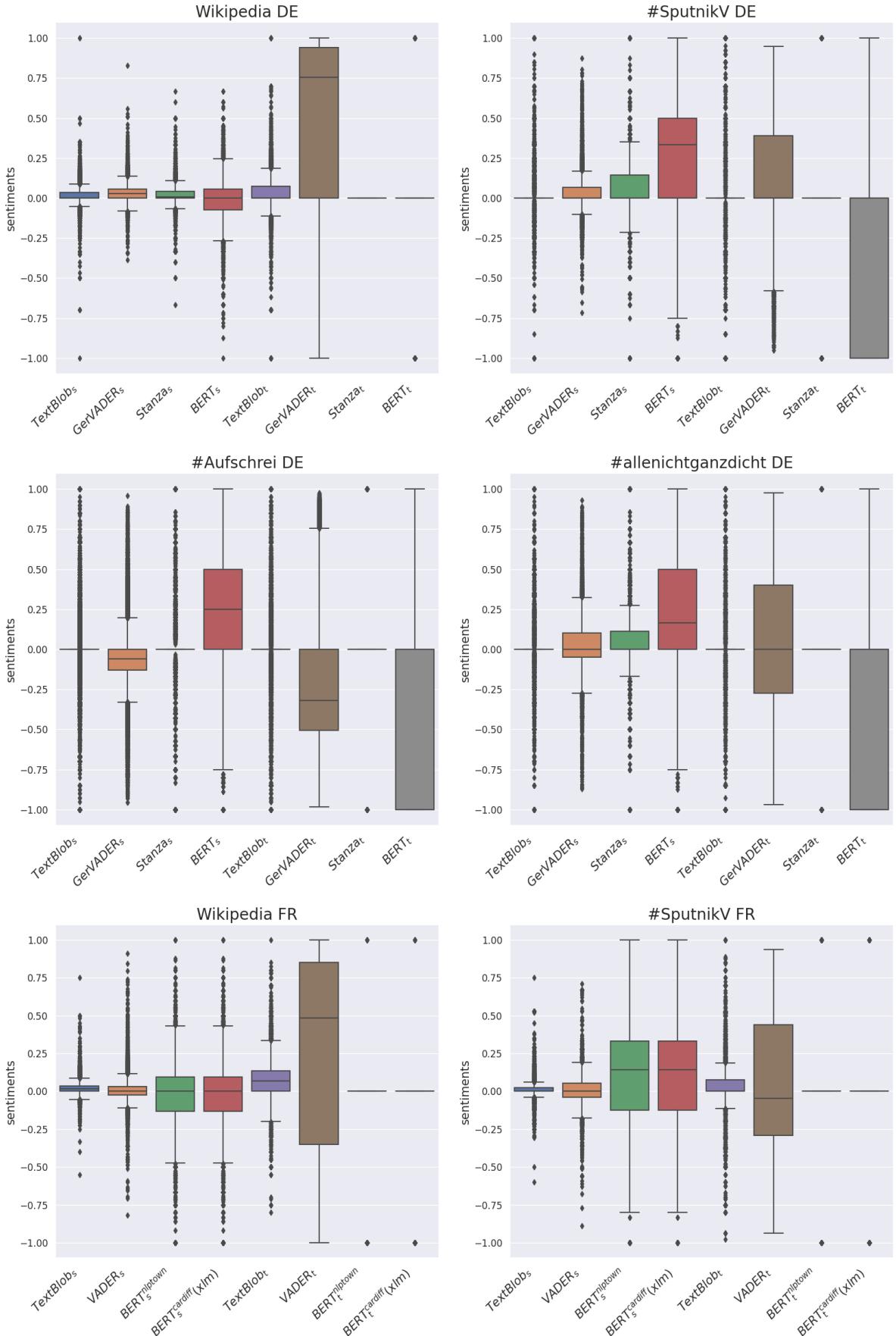


Figure 6: Sentiments of Wikipedia and Twitter hashtags #SputnikV, #Aufschrei and #allenichtganzdicht for German and French language.

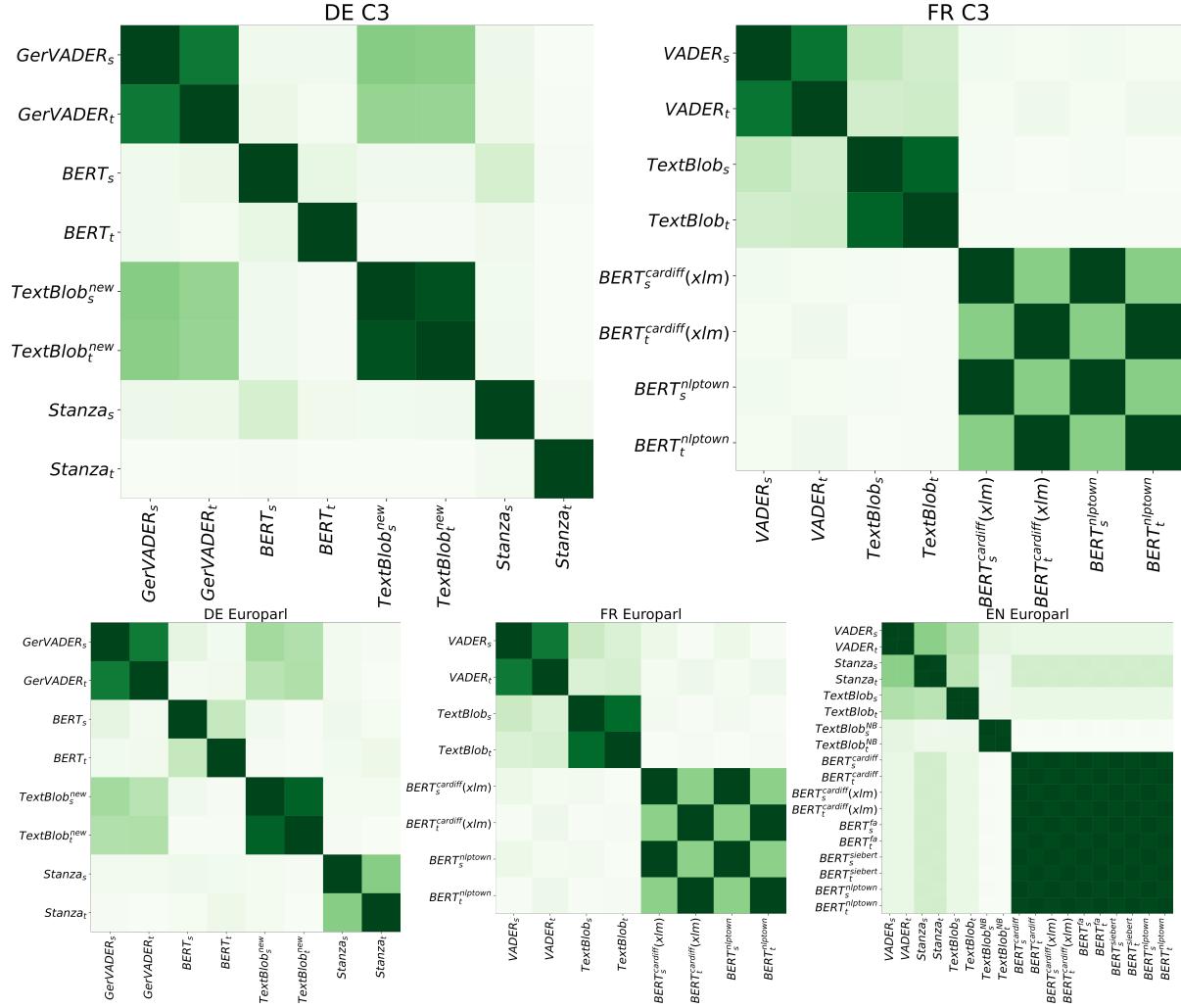


Figure 7: Distance correlation of sentiment scores for different tools on the C3 and Europarl corpora, with darker color indicating a higher correlation.

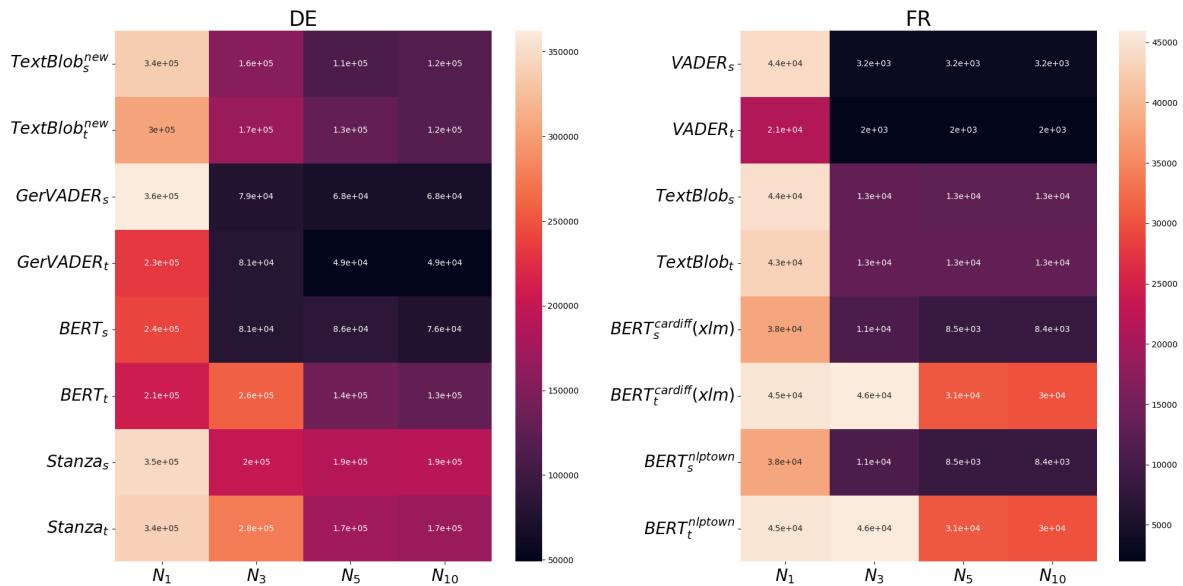


Figure 8: Agreement counts to the majority vote of the sentiment tools for German and French C3 corpus.

Group	Corpus	Chunks	Dev m F_1	Dev w F_1	Test m F_1	Test w F_1	# Train	# Test
all	C3	50	0.869	0.868	0.868	0.867	54 116	11 597
all	C3	70	0.888	0.889	0.891	0.889	38 656	8 284
all	C3	90	0.899	0.898	0.893	0.894	30 076	6 445
all	C3	100	0.909	0.909	0.909	0.908	27 064	5 800
all	C3	200	0.937	0.941	0.94	0.939	13 532	2 900
all	C3	400	0.961	0.961	0.962	0.96	6 766	1 450
all	C3	800	0.991	0.992	0.974	0.972	3 388	727
all	C3	1000	0.996	0.997	0.987	0.988	2 708	581
all	C1	50	0.76	0.759	0.753	0.751	43 054	9 226
all	C1	100	0.814	0.814	0.817	0.825	21 532	4 615
all	C1	1000	0.975	0.974	0.958	0.959	2 154	462
all	C2	50	0.921	0.921	0.909	0.907	11 074	2 374
all	C2	100	0.948	0.947	0.955	0.957	5 544	1 188
all	C2	1000	1.0	1.0	0.978	0.974	554	119
all	C4	50	0.238	0.25	0.195	0.198	2 772	594
all	C4	100	0.223	0.235	0.163	0.174	1 386	297
all	C4	1000	0.243	0.317	0.092	0.074	138	30
discrete	C1	50	0.668	0.667	0.665	0.666	28 702	6 151
discrete	C1	1000	0.968	0.968	0.968	0.968	1 436	308
discrete	C2	50	0.909	0.907	0.906	0.91	7 382	1 583
discrete	C2	1000	1.0	1.0	0.972	0.975	368	80
discrete	C3	50	0.839	0.838	0.841	0.84	36 077	7 731
discrete	C3	1000	0.981	0.979	0.993	0.992	1 806	387
discrete	C4	50	0.127	0.122	0.128	0.126	1 848	396
discrete	C4	1000	0.17	0.136	0.078	0.086	92	20
continuous	C1	50	0.953	0.953	0.952	0.952	14 350	3 076
continuous	C1	1000	1.0	1.0	1.0	1.0	718	154
continuous	C2	50	0.973	0.971	0.952	0.954	3 691	792
continuous	C2	1000	1.0	1.0	0.982	0.975	184	40
continuous	C3	50	0.943	0.942	0.943	0.945	18 038	3 866
continuous	C3	1000	1.0	1.0	0.994	0.995	902	194
continuous	C4	50	0.506	0.508	0.472	0.463	924	198
continuous	C4	1000	0.511	0.627	0.25	0.15	46	10
disc. sentences	C1	50	0.666	0.667	0.677	0.675	14 350	3 076
disc. sentences	C1	1000	0.979	0.981	0.955	0.954	718	154
disc. sentences	C2	50	0.884	0.886	0.867	0.873	3 691	791
disc. sentences	C2	1000	1.0	1.0	0.943	0.923	230	50
disc. sentences	C3	50	0.804	0.806	0.8	0.8	18 038	3 866
disc. sentences	C3	1000	0.974	0.975	0.975	0.974	902	194
disc. sentences	C4	50	0.291	0.277	0.267	0.266	924	198
disc. sentences	C4	1000	0.337	0.337	0.167	0.2	46	10
cont. sentences	C1	50	0.961	0.962	0.968	0.968	7 175	1 538
cont. sentences	C1	1000	1.0	1.0	1.0	1.0	359	77
cont. sentences	C2	50	1.0	1.0	0.995	0.995	1 845	396
cont. sentences	C2	1000	1.0	1.0	0.943	0.949	92	20
cont. sentences	C3	50	0.972	0.972	0.968	0.969	9 019	1 933
cont. sentences	C3	1000	1.0	1.0	1.0	1.0	451	97
cont. sentences	C4	50	1.0	1.0	1.0	1.0	462	99
cont. sentences	C4	1000	1.0	1.0	1.0	1.0	23	5

Table 3: This table shows a selection of the raw results produced by the NN experiment described in subsection 4.1. The selection contains non-normalized data for different corpora, groupings and chunk sizes. We provide macro and weighted F_1 -scores for the test and development set calculated by scikit-learn.

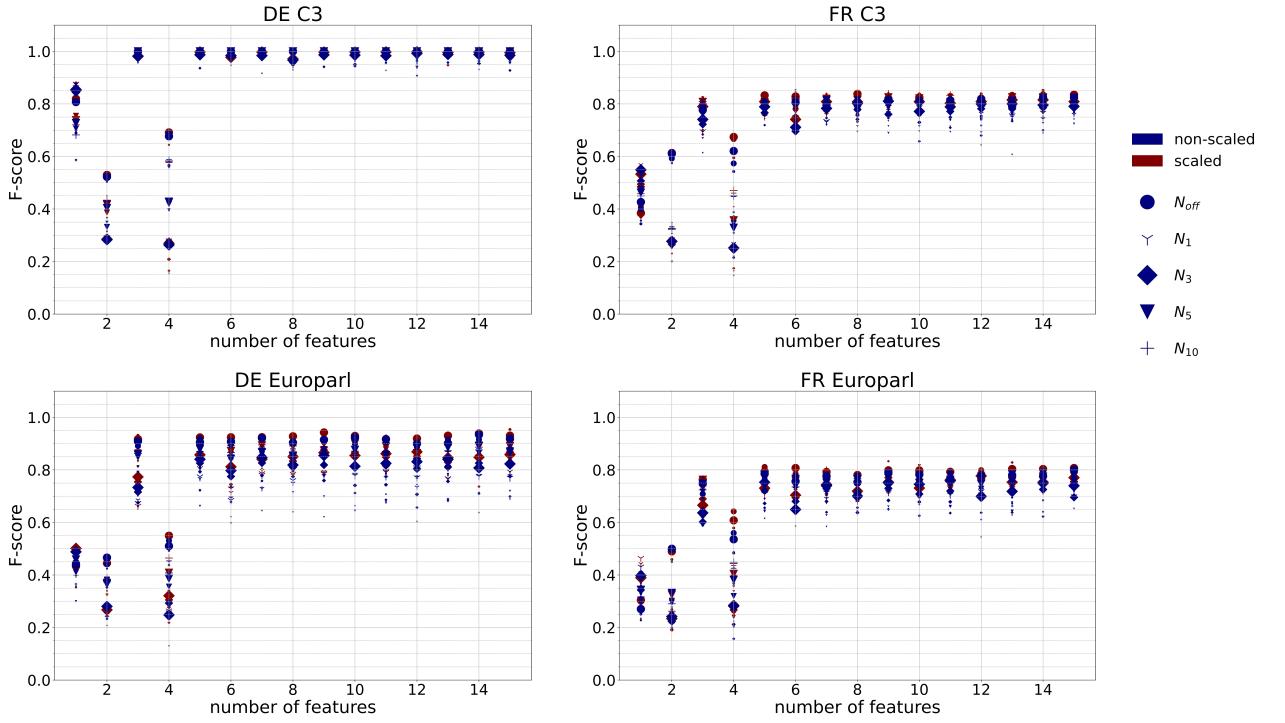
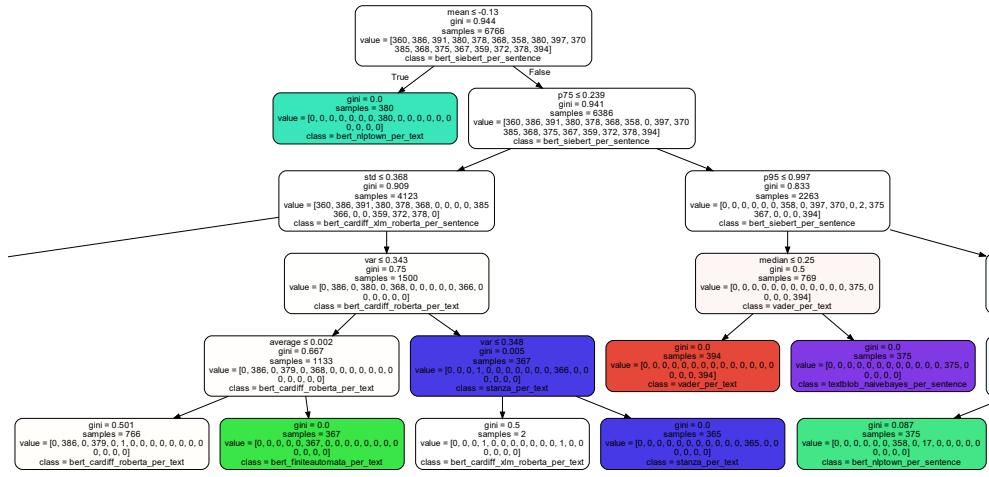


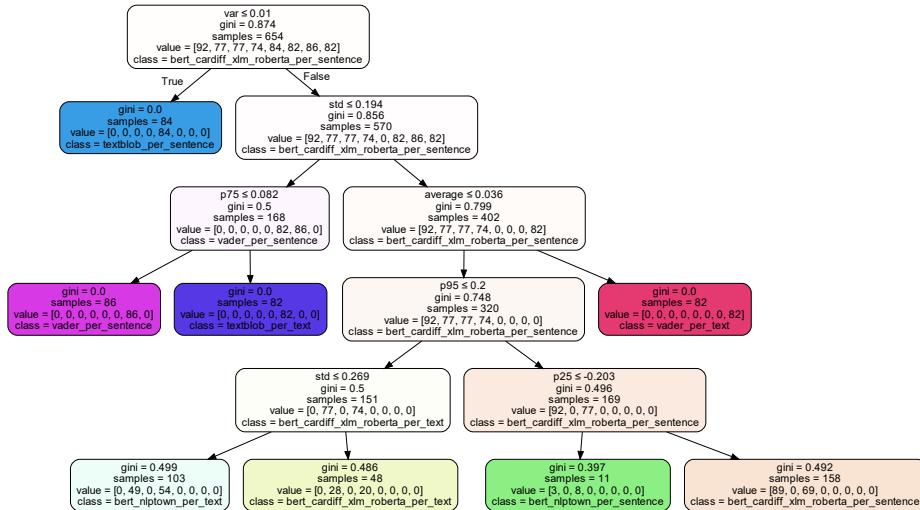
Figure 9: F_1 -scores of the SVM classifier for the C3 and Europarl corpora. The x-axis indicates the number of features, which are selected randomly for each feature vector size independently.

Group	Corpus	Dev mean F_1	Dev std F_1	Test mean F_1	Test std F_1	Test min F_1	Test max F_1
all	C3	0.932	0.048	0.927	0.044	0.867	0.988
all	C1	0.865	0.087	0.86	0.079	0.751	0.959
all	C2	0.964	0.033	0.959	0.033	0.907	1.0
all	C4	0.253	0.031	0.169	0.048	0.074	0.227
discrete	C3	0.912	0.058	0.911	0.055	0.84	0.992
discrete	C1	0.812	0.118	0.8	0.114	0.666	0.968
discrete	C2	0.964	0.035	0.953	0.028	0.91	0.99
discrete	C4	0.136	0.013	0.099	0.032	0.029	0.132
continuous	C3	0.978	0.023	0.978	0.02	0.945	1.0
continuous	C1	0.981	0.018	0.982	0.018	0.952	1.0
continuous	C2	0.992	0.011	0.982	0.017	0.954	1.0
continuous	C4	0.555	0.045	0.37	0.106	0.15	0.478
disc. sentences	C3	0.886	0.063	0.883	0.063	0.8	0.974
disc. sentences	C1	0.807	0.112	0.808	0.118	0.675	0.969
disc. sentences	C2	0.959	0.044	0.932	0.044	0.873	1.0
disc. sentences	C4	0.29	0.036	0.204	0.099	0.038	0.384
cont. sentences	C3	0.992	0.01	0.992	0.011	0.969	1.0
cont. sentences	C1	0.99	0.013	0.988	0.014	0.968	1.0
cont. sentences	C2	1.0	0.0	0.993	0.018	0.949	1.0
cont. sentences	C4	1.0	0.0	0.967	0.053	0.857	1.0

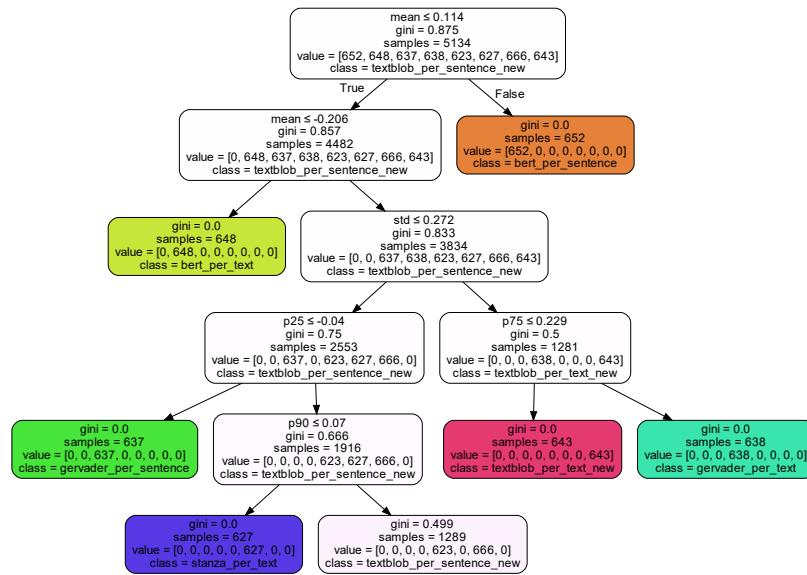
Table 4: Sample of the results of the experiments on the EN corpora from experiment in subsection 4.1. The results show the non-normalized experiments for all corpora and groupings, the weighted F_1 -scores, as reported by scikit-mean, are averaged over the different chunk sizes.



(a) EN Decision Tree

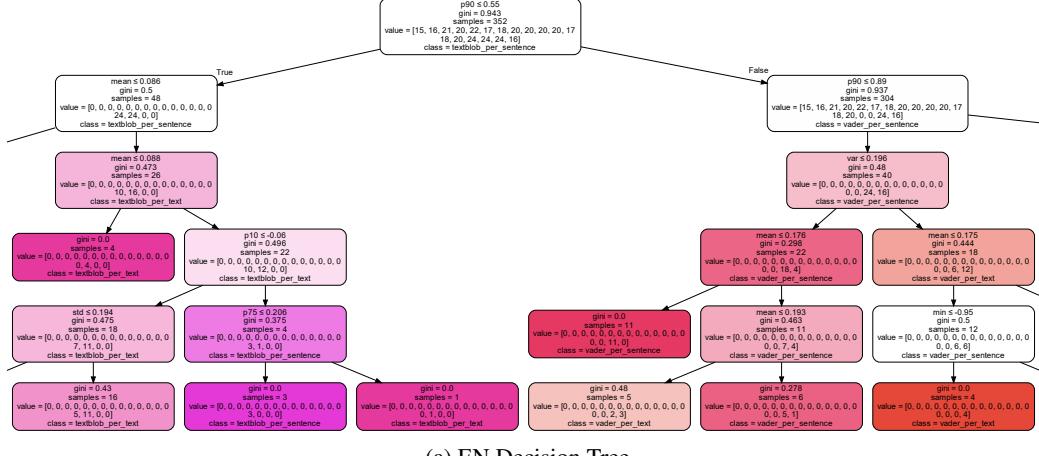


(b) FR Decision Tree

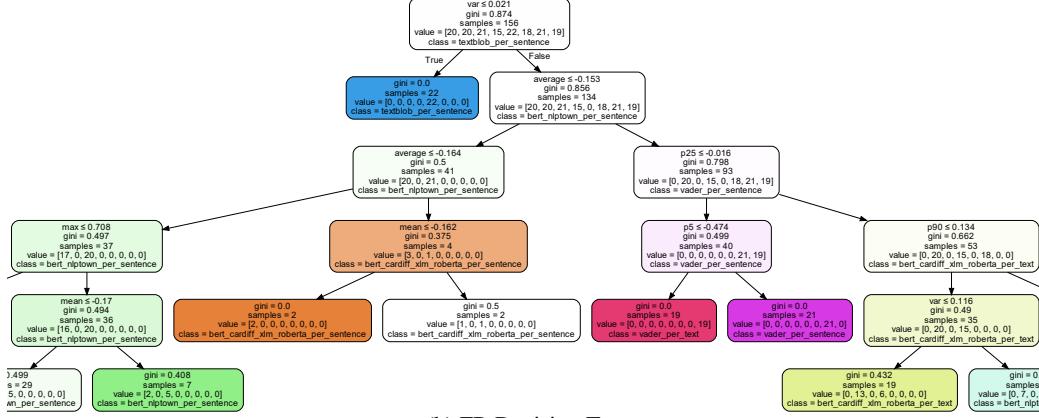


(c) DE Decision Tree

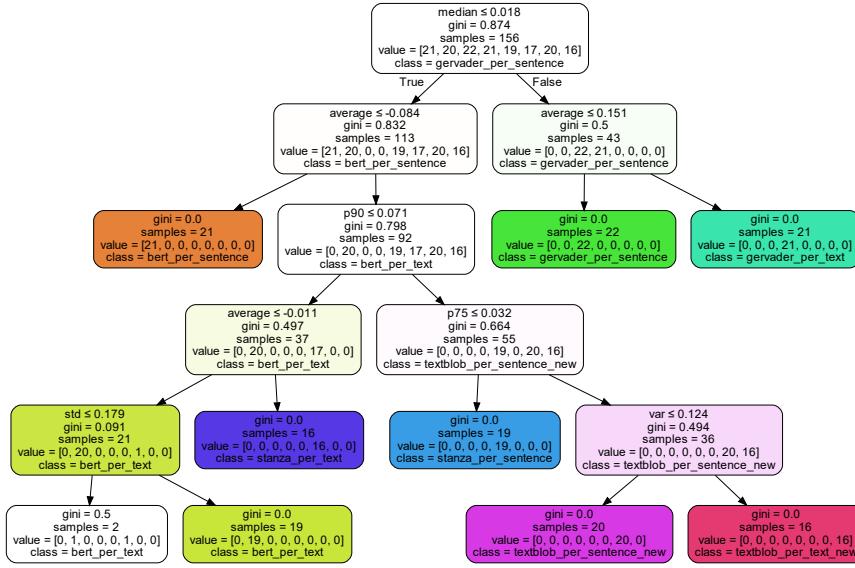
Figure 10: Cutouts of the Decision Trees on the C3 corpus with chunk size 400. Especially for French and German, very clear rules can be identified. Structure of the node from top to bottom: (1) feature by which the data is divided (2) gini: strength of the feature (3) number of data in the node (4) division into classes (5) class: strongest class in the node (class decision that would be made). More information see subsection 4.2.



(a) EN Decision Tree



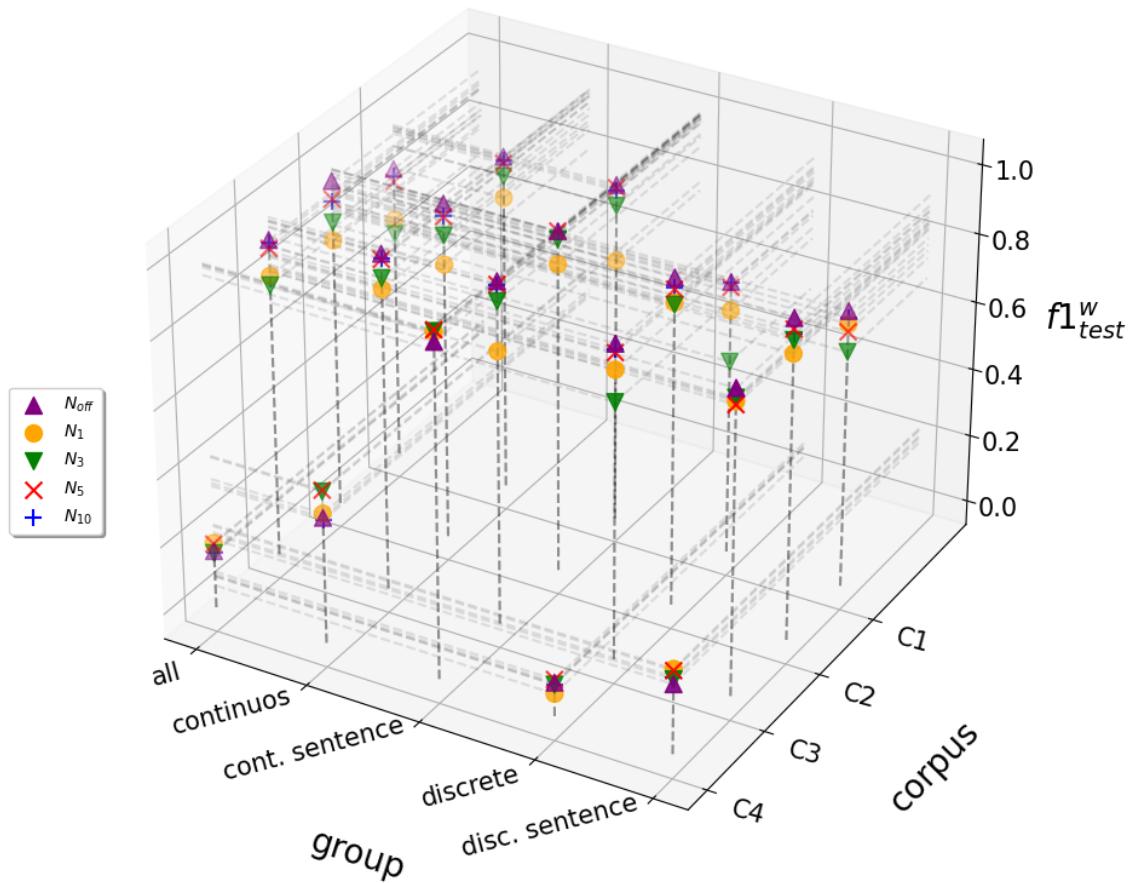
(b) FR Decision Tree



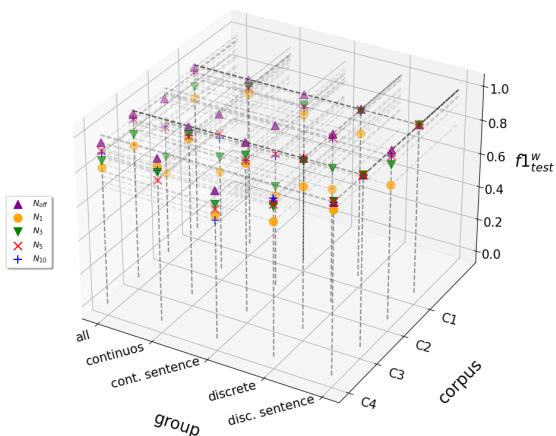
(c) DE Decision Tree

Figure 11: Cutouts of the Decision Trees on the Europarl corpus with chunk size 400. Especially for German, very clear rules can be identified. Structure of the node from top to bottom: (1) feature by which the data is divided (2) gini: strength of the feature (3) number of data in the node (4) division into classes (5) class: strongest class in the node (class decision that would be made). More information see subsection 4.2.

EN



DE



FR

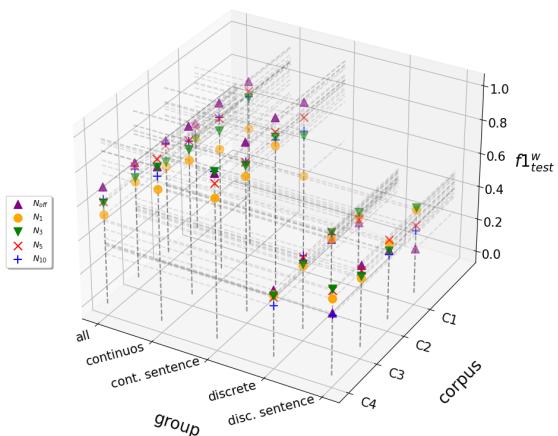


Figure 12: 3D visualization of Table 4, including different normalization results.