Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4708–4716
Marseille, 20-25 June 2022

The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers

Thomas E. Kolb[1], Katharina Sekanina[2], Bettina M. J. Kern[3], Julia Neidhardt[1], Tanja Wissik[2], Andreas Baumann[3]
[1]TU Wien,[2]Austrian Academy of Sciences,[3]University of Vienna
Vienna, Austria
[1]{thomas.kolb, julia.neidhardt}@tuwien.ac.at
[2]tanja.wissik@oeaw.ac.at
[3]{katharina.sekanina,bettina2.kern, andreas.baumann}@univie.ac.at

Abstract

This paper introduces the Austrian German sentiment dictionary ALPIN to account for the lack of resources for dictionary-based sentiment analysis in this specific variety of German, which is characterized by lexical idiosyncrasies that also affect word sentiment. The proposed language resource is based on Austrian news media in the field of politics, an austriacism list based on different resources and a posting data set based on a popular Austrian news media. Different resources are used to increase the diversity of the resulting language resource. Extensive crowd-sourcing is performed followed by evaluation and automatic conversion into sentiment scores. We show that crowd-sourcing enables the creation of a sentiment dictionary for the Austrian German domain. Additionally, the different parts of the sentiment dictionary are evaluated to show their impact on the resulting resource. Furthermore, the proposed dictionary is utilized in a web application and available for future research and free to use for anyone.

Keywords: Collaborative Resource Construction & Crowdsourcing, Digital Humanities, Document Classification, Text categorisation, Information Extraction, Information Retrieval, Statistical and Machine Learning Methods, Tools, Systems, Applications

1. Introduction

Computational sentiment analysis is a popular method to harness large amounts of textual data and extract information from them. Two main approaches exist for this. The dictionary approach relies on word lists containing sentiment annotations which can be a categorical label (e.g., negative vs. positive) or a numerical value on a continuous scale (from most negative to most positive). The words in a text are then checked against the dictionary to deduce the sentiment of the text. The machine learning approach utilises manually annotated text snippets as training data to extrapolate the sentiment of new, unseen texts (van Atteveldt et al., 2021). A variety of different algorithms can be employed, from relatively simple support vector machines to neural networks (van Atteveldt et al., 2021; Taboada et al., 2011; Liu, 2020; Siegel and Alexa, 2020). Their performance relies on the availability and quality of annotated training data. Attaining high-quality sentiment annotations on large amounts of texts is resource-intensive: Having the texts annotated by a few individuals is time-consuming; relying on crowd-sourcing is expensive. Therefore, dictionaries are widely used, although machine learning approaches have been shown to out-perform dictionary approaches (van Atteveldt et al., 2021). Resorting to dictionary approaches is often less costly in terms of time and money. While dictionary-based sentiment analysis is comparatively easy to implement and more transparent than, e.g., neural networks, it poses the challenge of finding or creating an appropriate dictionary. Furthermore, it is possible to use sentiment dictionaries to create weak labels. This allows the application of supervised methods based on these labels, which is called "weak supervised machine learning" (Zhou, 2017).

This paper introduces the "Austrian Language Polarity in Newspapers" (ALPIN) Sentiment Dictionary for analysing the polarity of newspapers and political contents in Austrian German. The dictionary was developed and implemented in the framework of the DYSEN project[1] which aims to model the polarity in language used in newspaper articles mentioning Viennese politicians over the time span from 1996 to 2017.
In German, there are several smaller sentiment dictionaries that were created using a plethora of different methods and for different purposes (Kern et al., 2021). None of them, though, was specifically created for the domain of political topics and newspaper articles. More strikingly, there is currently no sentiment dictionary for Austrian German.

Why is this problematic? German is spoken in different countries and regions whereas the status of German differs from country to country. In all these countries and regions, the German language is subject to variation on different levels and has formed different varieties. The characteristics of Austrian German can be seen at different levels, e.g., the morphologic, morphosyntactic, pragmatic or lexical level. For this study especially the

[1] Dynamic Sentiment Analysis as Emotional Compass for the Digital Media Landscape: `https://dylen.acdh.oeaw.ac.at/dysen/`

characteristics at the lexical level are of interest. An overview of lexical characteristics can be found e.g. in Ammon (1995). According to Wiesinger (2008) most variants are found in the category of food and dishes. For example, in Austria an apricot is called *Marille* and in Germany it is called *Aprikose*. But variants are not only found in general language but also in specialized language, especially in institutional languages such as administrative language (Markhardt, 2006) or legal language (Wissik, 2014) because institutions play an important role in forming lexical characteristics (Kloss, 1978) e.g. an execution is called in Austria *Exekution* and in Germany *Zwangsvollstreckung* or a Ph.D candidate is called *Dissertant* in Austria and *Promovend* in Germany. Words like this are typically not covered by standard sentiment dictionaries of German.

Clearly, lexical differences also affect the semantic level in general and emotional connotations in particular. For instance, the word *ausrasten* only has the single meaning 'to rage' in most varieties of German with a clearly negative sentiment. In Austrian German and other Southern German varieties, however, *ausrasten* can also mean 'to take a rest', which is obviously more positive. Likewise, *gespritzt* has the relatively neutral meaning 'sprayed' in most German varieties but has the additional negative meaning 'snobbish' in Austrian German. Another example is *aufrecht*, which in most German varieties has the relatively neutral meaning of 'upright'. In Austrian German it also means 'legally valid', which is more positive. It is obvious that phenomena like this affect the quality of dictionary-based methods for sentiment analysis in Austrian German texts and illustrate the need for sentiment resources dedicated to this variety of German.

In addition to the problem of linguistic varieties, it is well known that sentiment analysis can be challenging if the text genre and domain are very specific (Han et al., 2018; Taboada et al., 2009). There is also a wide range of different methods for conducting sentiment analysis, but these methods vary in how well they work depending on the domain (van Atteveldt et al., 2021; Hussein, 2018). In the present study, we focus on political discourse represented in Austrian media. While there is a lot of research is done in English regarding social media and news media, there is less research done in German e.g. Siegel and Alexa (2020) and even lesser for the Austria variety of German (Sidarenka, 2019). Crowd-sourcing can help to overcome this research gap by providing a tool for the efficient creation of text annotation which can be used to create a new dictionary for a specific domain. This is shown in a paper series where political communication in election campaigns was analysed (Haselmayer and Jenny, 2017). The sentiment dictionary created by this research team is based on a corpus which contains party press releases, minutes of parliamentary debates and media reports on election campaigns and only contains negative words, while in our research the focus lies on news media texts about Viennese politicians in general which were published by news media. Although the output of their research is highly valuable, it is restricted to the domain of discussions in election campaigns and focuses on negativity only, whereby our work, in contrast, also includes positivity.

Our contribution is thus to generate a new sentiment dictionary, ALPIN, that (a) represents a specific variety of German, namely Austrian German, and that is (b) specific to a particular genre and domain, namely political discourse in news and online media. To do so, we base our dictionary on diverse data: news media in the field of politics, an austriacism list based on different resources and a posting data set based on a popular Austrian news media. We perform extensive crowd-sourcing followed by evaluation and automatic conversion into sentiment scores.

The paper is organized as follows: Section 2.1 first introduces the different data sources which were used to create the sentiment dictionary in question and then explains the steps taken to preprocess 2.2, annotate 2.3, modelling 2.4 and subsequently postprocess 2.5 the data. Section 3 presents the resulting resource which is then evaluated in Section 4. Finally, Section 5 outlines the conclusion, the limitations faced during the process as well as potential further steps and improvements.

2. Sentiment dictionary ALPIN

ALPIN stands for Austrian Language Polarity in Newspapers. This section outlines how the sentiment dictionary is created and which methods were used to calculate the sentiment scores. Figure 1 illustrates the workflow. For the generation of the ALPIN sentiment dictionary the in section 2.1 mentioned data sources were used. For each of the data sources different methods were chosen to calculate sentiment scores. The Austrian Media Corpus (AMC) corpus data extraction based on Viennese politicians was performed by a crowd-sourcing step followed with applying the SPLM algorithm introduced by Almatarneh and Gamallo (2018). The STANDARD posts (STP) corpus is already labeled so that SPLM was applied directly to generate sentiment scores. The austriacism list Austriacisms (AUT) was labeled by performing crowd-sourcing in combination with best-worst-scaling (BWS) to improve inter-annotator agreement (Kiritchenko and Mohammad, 2017b). In a nutshell, the idea behind BWS is to let annotators rank the best and worst word for each tuple and to subsequently compute sentiment scores indirectly rather than collecting direct sentiment ratings. Finally the dictionaries were merged together by using the min-max-abs scaling based on the sklearn framework.

2.1. Data Sources

Three different text sources were used to create the ALPIN dictionary: (i) texts taken from a corpus of Aus-

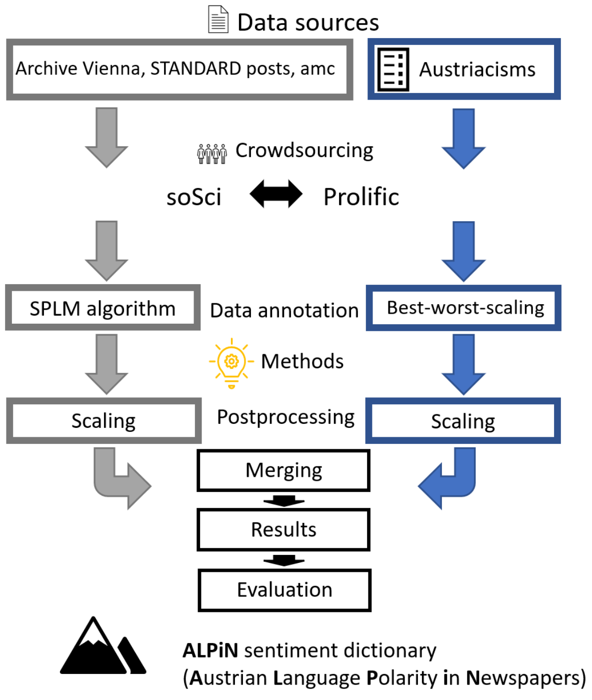


Figure 1: ALPIN sentiment dictionary - workflow

trian print and online media, (ii) postings from a user-forum of an Austrian online newspaper, and (iii) a list of austriacisms. They will be briefly introduced in the upcoming paragraphs.

2.1.1. List of Viennese politicians

Viennese politicians were selected since the focus of our research is on analyzing political discourse in the political landscape of Vienna as defined in the DYSEN project[2]. To retrieve a list of Viennese politicians, the politician archive of Vienna (POLAR)[3] from the Vienna City and State Archives was used. Since the AMC corpus Version 3.1 (Ransmayr et al., 2017) includes media from 1986 to 2018 the result list was limited to politicians which were active in that time frame. Therefore, only politicians that were active between the 13th and the 20th parliamentary term were selected. This list does not function as a text source as such but rather helped us to identify relevant contexts in our media data (see section 2.2.1).

2.1.2. Austrian Media Corpus (AMC)

The AMC[4] is an Austrian text corpus which covers nearly the entire Austrian media landscape of the last decades and it is constantly growing. With a size of around 45 million articles it is one of the larger corpora in the German language. For the creation of the ALPIN dictionary the version 3.1 of the AMC corpus tagged with linguistic information, as described in (Ransmayr et al., 2017) was used.

2.1.3. STANDARD posts (STP)

The data set provided by (Schabus et al., 2017) contains user comments posted to the online portal of the Austrian daily newspaper STANDARD. For our study, we selected the 3599 posts that were labelled for sentiment by professional forum moderators. The distribution is as follows: 43 positive, 1691 negative and 1865 neutral. The extreme imbalance results from the sampling strategy by the original authors focused on posts with a negative sentiment. We will use the shorthand STP to refer to this data set.

2.1.4. Austriacisms (AUT)

In the pluricentric approach Austrian variants of a word related to the state Austria are called austriacisms (Ammon, 1995). The list was collected from the “Variantenwörterbuch des Deutschen” (Ammon et al., 2016) (thereby only selecting those words that only surface in Austrian German and in no other variety of German) and an austriacism list of Wikipedia[5]. The Wikipedia resource was used to enhance the wordlist based on (Ammon et al., 2016) by providing recent changes. Both lists were evaluated and cleaned by native speakers of Austrian German. In total, the list shows 1648 words (shorthand: AUT).

2.2. Preprocessing

Different preprocessing steps were necessary to account for the different structure and information provided by the each of the three data sources;

2.2.1. Preprocessing: AMC

The goal of the preprocessing procedure was to extract text parts consisting of complete sentences featuring politician names that could be labeled for sentiment afterwards (see section 2.3.1). We only selected print media related to the Viennese area. To avoid duplicates in the extracted data set, standardized press releases were excluded from the data extraction as they are usually identical across newspapers. Only paragraphs featuring at least one name of the politician list were selected. The length of the extracted text pieces was limited to 60 tokens as the annotated and processed AMC (Ransmayr et al., 2017) is copy-right protected and those text passages are shown to survey participants. If paragraphs contained more than 60 tokens, they were shortened sentence-wise always removing the last sentence. If the sentence contains the politician’s name, it is not removed. In that case the same procedure is applied at the beginning of the paragraph. If the sentence showing the politician name was longer than 60 tokens it was discarded. A total of 494111 text areas were extracted from 22 different news media, with 5346 texts (each with no

[2] https://dylen.acdh.oeaw.ac.at/dysen/
[3] https://polar.wien.at
[4] https://amc.acdh.oeaw.ac.at/
[5] https://de.wikipedia.org/wiki/Liste_von_Austriazismen

more than 60 tokens) randomly selected to perform the crowd-sourcing step.

2.2.2. Preprocessing: STP

The SPLM algorithm requires a wordlist with words and their wordform as input (see section 2.4.1). Therefore, the STP data set was tokenized, POS tagged and lemmatized by using the NLTK framework for tagging and the spaCy "de_core_news_sm" pipeline for POS tagging and lemmatizing. The lemma is used to merge similar words together and reducing the noise through different word variations of the same lemma. The resulting tagged and lemmatized data set is limited to nouns, adjectives, adverbs and verbs. In STP, only 43 texts are labeled as positive. To overcome this high imbalance in the data set, texts labeled as neutral and texts labeled as positive were merged into a single 'non-negative' class. Thus, the final data set consists of 1691 negative and 1908 non-negative texts.

2.2.3. Preprocessing: Austriacisms

To perform the data annotation with the help of crowdsourcing, AUT was evaluated by linguistic experts and native speakers of Austrian German in the project team. Duplicates were removed and wordnet POS tags were assigned. Wordnet POS tags are required to map the austriacism list in the last step with the dictionary based on AMC and STP. For the best-worst-scaling the creation of tuples (of four words each) was required. Kiritchenko and Mohammad (2017a) provided a script[6] to create such a tuple list.

2.3. Data Annotation

All surveys were created using the Germany-based platform SoSciSurvey[7] which is freely available for academic purposes. Once created, the surveys were published on Prolific[8], a UK-based website which aims to connect researchers with potential participants all over the world. Each survey took an estimated 30 minutes and each participant (after passing a fair and previously announced quality check) received a monetary compensation of £3.75. The text passages extracted from the AMC and AUT needed to be labelled; The STP data set already contains sentiment values.

2.3.1. Annotation: AMC

A total of 5346 extracted texts of the AMC were divided into two surveys (2376 phrases in the first one and 2970 in the second one). This division enabled an effective randomization of the items in SoSciSurvey while also reliably ensuring that enough data points are collected for each text. Structurally, both surveys were identical. Participants were asked to rate 126 randomly selected phrases. Three options were given: positive, neutral and negative. An opt-out option was not provided, thus, each participant had to label each and every phrase they were shown. In addition, a "golden sample" of 24 manipulated texts with a clear positive/negative connotation were included in each survey. Positive texts were created by replacing single nouns, verbs or adjectives in existing texts with extremely positive words. To check if the golden sample texts are perceived as clearly positive, all of them were rated by five native speakers of German. Only those texts for which all five ratings were positive were used as golden samples. Negative texts were created, *mutatis mutandis*, analogously. Only if more than 75% of the golden sample were rated correctly by a participant, their ratings were used for the sentiment dictionary, in order to ensure high-quality data.

The prescreening parameters in Prolific were selected as following: current country of residence (Germany, Austria, Switzerland), Nationality (Germany, Austria, Switzerland) and First Language (German).

About 3000 active users on Prolific matched these criteria. In total, 182 people responded to the survey. 24 participants which who did not meet the 75 percent correctness threshold for the golden samples were excluded from the results. The remaining 158 participants, resulting in 5346 labelled data points that received at least three ratings, were used to create the labelled data set.

Both surveys were evaluated and a Fleiss-Kappa of 0.295 and 0.283 was reached for the first and second annotation run, respectively. This can be considered a "fair" inter-annotator agreement in the sense of Landis and Koch (1977). This shows that despite filtering out bad commentators and conducting internal pre-survey tests, the data is quite challenging to label.

2.3.2. Annotation: Austriacisms

To increase the inter-annotator agreement during crowd-sourcing, the AUT data was labeled in two steps as suggested by Rouces et al. (2018). The first step, the direct annotation step, was similar to the previously conducted survey (with the only difference that here, words rather than texts were annotated; see section 2.3.1). Each survey included 500 words and 25 golden samples. In total, 1600 words were labeled at least three times. Since some of these words are only part of certain regional dialects, participants were offered a fourth option, unknown, as well, meaning they could now choose between positive, neutral, negative and unknown.

In the second step best-worst-scaling was applied which is further explained in section 2.4.2.

[6] http://saifmohammad.com/WebPages/BestWorst.html

[7] https://soscisurvey.de

[8] https://prolific.co

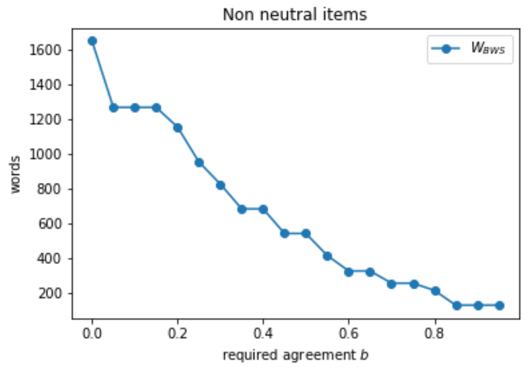


Figure 2: Non neutral items per required agreement

2.4. Models

After finishing the crowd-sourcing the resulting labeled datasets were used to calculate sentiment scores by using the SPLM algorithm for the AMC list combined with the STP list and the best-worst-scaling approach applied on the labeled AUT list. In addition to the methodology specified in Sharma and Dutta (2021) (SPLM) and the Kiritchenko and Mohammad (2017a; Kiritchenko and Mohammad (2017b) (BWS), crowd-sourcing was carried out instead of using a smaller set of fixed annotators. This was done to achieve greater diversity during the annotation step.

2.4.1. SPLM

The SPLM algorithm introduced in Almatarneh and Gamallo (2018) and later shown by Sharma and Dutta (2021) as effective for the automatic construction of sentiment dictionaries is used to generate sentiment scores out of the labeled AMC and STP data. The algorithm is based on four equations whereby a polarity weight is calculated for each of the words (Almatarneh and Gamallo, 2018). The performance of this algorithm is evaluated in section 4.

2.4.2. Best-Worst-scaling (BWS)

Best-Worst-scaling (BWS) as proposed by Kiritchenko and Mohammad (2017a; Kiritchenko and Mohammad (2017b) is a method to improve the quality of labels obtained during crowd-sourcing. This systematic approach was used to obtain the labeled AUT list. The BWS method is performed in two steps first by generating tuples which need to be labeled by a team of annotators in this work done by utilizing crowd-sourcing and in the second step by calculating scores based on the collected annotations. In this variation of the BWS method score calculation is performed using “Counts Analysis” (Orme, 2009). Counts analysis computes a score for each item by subtracting the percentage how often this item was chosen as worst from how often this item was chosen as best. The resulting score lies within [-1,+1].

First, all items labeled as unknown and items without a label were removed. Second, non-neutral items were selected. This was done by first computing a word’s sentiment with the formula

$$sen_{DA}(w) = \frac{\sum_{a \in A_{DA}} l_{DA}(a, w)}{|A_{DA}|} \quad (1)$$

introduced by Rouces et al. (2018), where $l_{DA}(w)$ is 1 if a annotated w as positive, 0 if a annotated w as neutral and -1 if a annotated w as negative. Here, A_{DA} is the set of annotators and W_{DA} is the set of words, so that $|A_{DA}|$ is number of annotators who labeled a specific item w and $sen_{DA}(w)$ is the resulting sentiment of that item. Rouces et al. (2018) suggest the set of non-neutral items to be defined as

$$W_{BWS} = \{w : w \in W_{DA} \wedge |sen_{DA}(w)| \geq b\}. \quad (2)$$

In Rouces et al. (2018), the agreement threshold was set to $b = 2/3$. After examining the size of W_{BWS} depending on b, as shown in Figure 2, the threshold was set to $b = 1/2$ in our study. This lead to a non-neutral count of 538 words out of the pre-survey.

Using the script provided by Kiritchenko and Mohammad (2017a), we generated a set of 4417 tuples of 4 items each (employing a scaling factor of 2, cf. Rouces et al. (2018)) that had to be labeled. For both austriacism surveys, the prescreening parameters were narrowed to exclusively Austrian (Nationality and current country of residence) speakers of German (First Language). Overall we needed 34 annotators after excluding 6 bad ones.

Ratings for AUT were assessed by calculating the split-half reliability as suggested by Kiritchenko and Mohammad (2017a), resulting in a Spearman correlation of: 0.9159 (+/- 0.0051), indicating a very high reliability of our resulting austracism list.

2.5. Postprocessing

To combine the results of the different datasets alignment of the word lists was required. This was necessary since the usage of SPLM for the AMC and STP and BWS for AUT resulted in different scales. The alignment is done by scaling the wordlists to a range from [-1,+1] by using the MaxAbsScaler of the python package sklearn (Pedregosa et al., 2011).

3. Results

The resulting sentiment dictionary ALPIN is only one part of the results. Additionally the three different parts “STP” (labels based on STANDARD posts only), “AMC” (labels based on AMC texts only), “AUT” (labels based on austriacism list only) are published separately. This allows further research by using the individual dictionary components before they are merged into the ALPIN. A brief summary for each of the three components and the full dictionary is shown in table 1.

dictionary	words	algorithm	data source
AMC	4816	SPLM 2.4.1	section: 2.1.2
STP	5117	SPLM 2.4.1	section: 2.1.3
AUT	538	BWS 2.4.2	section: 2.1.4
ALPIN	9435	SPLM 2.4.1,BWS 2.4.2	sections: 2.1.2, 2.1.3, 2.1.4

Table 1: Sentiment dictionaries

Tables 2 and 3 show the top and bottom 10-word-overlaps of the AMC+STP word list compared to the AUT list before the merging step. Comparing AMC+STP with AUT, there is an intersection of 32 words in total. 16 words from the AMC+STP list out of 21 positive words in AUT also have a positive sentiment score in AUT. Among the words with negative sentiment, only 1 out of 11 words was not negatively annotated in AMC+STP.

word	AUT	AMC+STP
Wiese	0.750	0.027
Karenz	0.742	0.040
Angelobung	0.729	-0.051
Ehrenzeichen	0.710	0.067
Gehalt	0.645	0.211
aufrecht	0.625	-0.031
maturieren	0.625	0.027
ÖAMTC	0.562	-0.031
einbringen	0.548	-0.123
Team	0.516	0.166

Table 2: Top 10 word intersections: AUT score, AMC+STP score

The sentiment dictionaries are available at Zenodo[9].

4. Evaluation

For the evaluation two already existing dedicated sentiment dictionaries were used. “German Polarity Clues” (GPL) is a dictionary created to perform sentiment analysis in the German language which consists of 10141 polarity features (Waltinger, 2010a; Waltinger, 2010b). The second dictionary called “Affective Norms” (AN) consists of 350000 German lemmatized words (Köper and Schulte im Walde, 2016). AN provides sentiment values for four different emotional dimensions; for the purpose of our comparison we used only the sentiment ratings for valence.

word	AUT	AMC+STP
klagen	-0.312	-0.054
angreifen	-0.344	-0.005
Fleck	-0.376	-0.031
Einvernahme	-0.387	-0.031
Freunderlwirtschaft	-0.438	-0.031
versperren	-0.486	-0.031
Mist	-0.594	-0.031
sekkieren	-0.688	-0.031
exekutieren	-0.838	-0.004
Exekution	-0.875	0.027

Table 3: Bottom 10 word intersections: AUT score, AMC+STP score

We perform an evaluation approach similar to the one proposed by (Almatarneh and Gamallo, 2018). Here, the corresponding sentiment dictionary is used to count the positive and negative words, respectively, as well as the proportion of positive against negative words for each labeled text. These three features are used to train a support vector machine (SVM) with a linear kernel (Pedregosa et al., 2011) predicting the overall sentiment for each text. The table 4 and 5 show (averaged) performance metrics computed through five-fold cross-validation.

Table 4 shows how well the sentiment dictionaries work by predicting the sentiment of the labeled STP data set that were introduced in section 2.2.2. Table 5 shows the performance of the different dictionaries against the labeled AMC data introduced in section 2.3.1. In both cases, ALPIN shows the best performance. Note, though, that this is partially due to the fact that AMC and STP were used to derive ALPIN, so that this dictionary and the test data are not fully independent.

dict.	accuracy	precision	recall	f1
GPL	0.526	0.528	0.986	0.688
AN	0.530	0.530	**1.0**	0.693
ALPIN	**0.768**	**0.778**	0.794	**0.783**

Table 4: Sentiment dicts. against STP

To overcome this restriction and to evaluate the impact of each of the text resources on the quality of the resulting dictionary, a second evaluation was performed in which the labeled data set (AMC and STP) was split into a train and test data set. Crucially, this was done before deriving sentiment dictionaries through SPLM. Only the train data set was employed together with SPLM to obtain sentiment estimates for words. Subse-

[9] https://doi.org/10.5281/zenodo.5857151

dict.	accuracy	precision	recall	f1
GPL	0.647	0.641	0.738	0.686
AN	0.657	0.675	0.660	0.670
ALPIN	**0.822**	**0.828**	**0.840**	**0.830**

Table 5: Sentiment dicts. against AMC

quently, the test data set was used for evaluation by applying the same methodology as in the first evaluation round described above (calculating count of positive words, count of negative words, proportion pos/neg, training an SVM that predicts text sentiment, 5-fold cross-validation, calculating metrics). In this way it is ensured that the sentiment dictionary to be evaluated is not based on the labeled text data it is tested against. The procedure was done for different train-test ratios, defining how much of the given labeled data set is used for the creation of the sentiment dictionary. The result of this evaluation is shown in table 6 and table 7. Due to the limited amount of labeled data the accuracy is not as high as it would be by using the whole data set for the sentiment dictionary creation. By comparing the results in Tables 6 and 7 one can see that the AMC part of the ALPIN dictionary has a greater positive impact on the metrics than the STP based data.

ratio	accuracy	precision	recall	f1
0.6	0.542	0.543	0.783	0.633
0.7	0.549	0.563	0.626	0.5770
0.8	0.567	0.557	**0.912**	0.692
0.9	**0.614**	**0.596**	0.831	**0.693**

Table 6: Train-test by using the STP data set 2.2.2

ratio	accuracy	precision	recall	f1
0.6	0.644	0.655	**0.716**	0.682
0.7	0.647	0.670	0.689	0.671
0.8	0.678	0.726	0.643	0.675
0.9	**0.704**	**0.739**	0.702	**0.719**

Table 7: Train-test by using the AMC data set 2.3.1

train-test ratio	accuracy	precision	recall	f1
0.6	0.588	0.586	**0.777**	0.664
0.7	0.607	0.609	0.727	0.659
0.8	0.577	0.583	0.707	0.634
0.9	**0.615**	**0.618**	0.774	**0.671**

Table 8: Train-test by using the AMC 2.3.1 and STP data set 2.2.2

5. Conclusion

Our paper shows that incorporating crowd-sourcing instead of few, professional annotators and more intricate methods like best-worst-scaling can improve the inter-annotator agreement. In addition a new language resource is created, filling the research gap of Austrian German in the domain of news media and politics. The scope around Viennese politicians captures the unique language used in the second biggest German speaking city. In addition to the full ALPIN dictionary the individual components of the ALPIN dictionary (i.e., sentiment entries from the AMC, STP and AUT) are also provided separately to give interested users maximal flexibility.

The resulting dictionary ALPIN is implemented into the interactive web tool that we developed in the DYSEN project. The tool allows to track the sentiment tendency with which Austrian print media report about Viennese politicians in the time span from 1990 to 2018. The tool is free to use and can be found on the website of the DYSEN project[10]. A screenshot of the application is shown in figure 3.

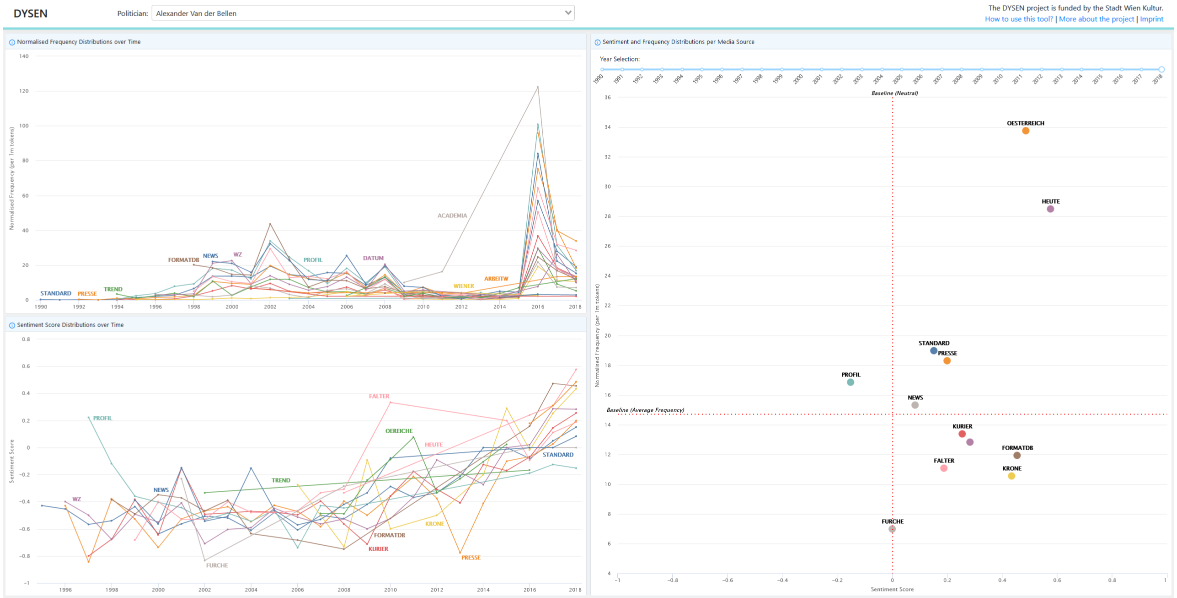
It is important to mention that there were several limitations which needed to be taken into account. There is a legal limitation by a maximum token-size per item due to the copyright of the AMC corpus. News media are in general difficult to label due to their mostly neutral sentiment which resulted in a high amount of neutral labeled items during crowd-sourcing. There is currently no similar sentiment dictionary in this specific domain which makes external evaluation difficult.

Further work is possible in different areas. The scope could be expanded to more politicians. Different methods e.g. aspect-based sentiment analysis could be used to further increase the accuracy of the dictionary. As shown in section 2.4.2 BWS scaling is very effective and could also be used for labeling text items instead of single words. Investing more money to label a bigger data set is another possibility to improve the performance. In future work bias during data collection and its impact on the model will be addressed in more depth. Some of these improvements come with higher cost which is not always feasible.

6. Acknowledgements

This research was supported by the DYSEN project, funded by the City of Vienna (MA7-737909/19) and by the DYLEN project, funded by the ÖAW go!digital Next Generation grant (GDNG 2018-02).

[10] `https://dysen-tool.acdh.oeaw.ac.at/`



Politician: Alexander Van der Bellen

The DYSEN project is funded by the Stadt Wien Kultur.
How to use this tool? | More about the project | Imprint

Figure 3: DYSEN web application

7. Bibliographical References

Almatarneh, S. and Gamallo, P. (2018). Automatic construction of domain-specific sentiment lexicons for polarity classification. In Fernando De la Prieta, et al., editors, *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017*, pages 175–182, Cham. Springer International Publishing.

Ammon, U. (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. de Gruyter.

Han, H., Zhang, J., Yang, J., Shen, Y., and Zhang, Y. (2018). Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools and Applications*, 77(16):21265–21280.

Haselmayer, M. and Jenny, M. (2017). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity*, 51(6):2623–2646, Nov.

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338.

Kern, B. M. J., Baumann, A., Kolb, T. E., Sekanina, K., Hofmann, K., Wissik, T., and Neidhardt, J. (2021). A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis. In Dagmar Gromann, et al., editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 37:1–37:17, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Kiritchenko, S. and Mohammad, S. (2017a). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada, July. Association for Computational Linguistics.

Kiritchenko, S. and Mohammad, S. M. (2017b). Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling.

Kloss, H. (1978). *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. Pädagogischer Verlag Schwann, Düsseldorf.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Markhardt, H. (2006). *Wörterbuch der österreichischen Rechts-, Wirtschafts-, und Verwaltungsterminologie*. Peter Lang, Frankfurt am Main.

Orme, B. K. (2009). Maxdiff analysis : Simple counting , individual-level logit , and hb.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rouces, J., Tahmasebi, N., Borin, L., and Eide, S. R. (2018). Generating a gold standard for a swedish sentiment lexicon. In *LREC*.

Sharma, S. S. and Dutta, G. (2021). Sentidraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. *Inf. Process. Manag.*, 58:102412.

Sidarenka, U. (2019). *Sentiment analysis of German Twitter*. doctoralthesis, Universität Potsdam.

Siegel, M. and Alexa, M. (2020). *Sentiment-Analyse deutschsprachiger Meinungsäußerungen*. Springer.

Taboada, M., Brooke, J., and Stede, M. (2009). Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, page 62–70, USA. Association for Computational Linguistics.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307, 06.

van Atteveldt, W., van der Velden, M. A. C. G., and Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140.

Wiesinger, P. (2008). *Das Österreichische Deutsch in Gegenwart und Geschichte*. LIT, Wien, 2 edition.

Wissik, T. (2014). *Terminologische Variation in der Rechts- und Verwaltungssprache Deutschland – Österreich – Schweiz*. Frank & Timme, Berlin.

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 08.

8. Language Resource References

Ammon, U., Bickel, H., and Ebner, J. (2016). *Variantenwörterbuch des Deutschen : die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. Walter de Gruyter, Berlin.

Köper, M. and Schulte im Walde, S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Ransmayr, J., Mörth, K., and Ďurčo, M. (2017). AMC (Austrian Media Corpus). In *Korpusbasierte Forschungen zum österreichischen Deutsch. In Digitale Methoden der Korpusforschung in Österreich (= Veröffentlichungen zur Linguistik und Kommunikationsforschung Nr. 30)*, pages 27–38. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Schabus, D., Skowron, M., and Trapp, M. (2017). One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1241–1244, New York, NY, USA. Association for Computing Machinery.

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

Waltinger, U. (2010b). Sentiment analysis reloaded: A comparative study on sentiment polarity identification combining machine learning and subjectivity features. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10)*, Valencia, Spain, April.