Creating an Austrian Language Polarity Dictionary with the Crowd

Thomas E. Kolb, Katharina Sekanina, Bettina M. J. Kern,

Julia Neidhardt, Andreas Baumann and Tanja Wissik

ÖTSI Workshop @ 46. Österreichische Linguistik-Tagung

Funded by:

Stadt Wien *Grant number: MA7-737909/19*

GO!DIGITAL NEXT GENERATION

ÖAW Österreichische Akademie der Wissenschaften

universität wien

TU WIEN Technische Universität Wien

Project Team

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

Katharina Sekanina

Bettina M. J. Kern

Tanja Wissik

DYSEN

universität wien

Andreas Baumann

Julia Neidhardt

Thomas E. Kolb

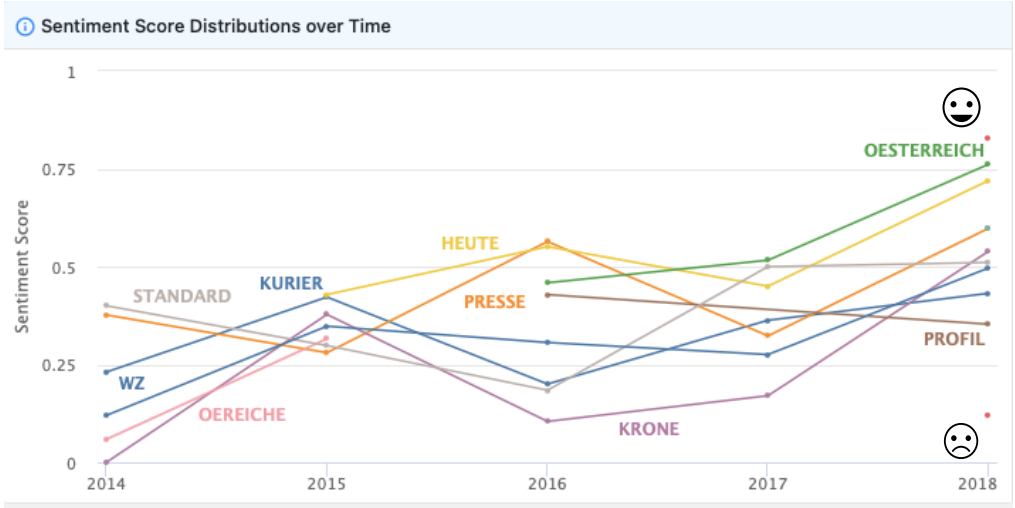
TU WIEN TECHNISCHE UNIVERSITÄT WIEN

DYSEN Project

Dynamic **Sen**timent Analysis as Emotional Compass for the Digital Media Landscape

Research question: How do print media report about the Viennese politicians?

Aim of the project: Develop a tool that can detect change of emotional polarization of politicians in Austrian Newspapers



Funded by:

Stadt Wien *Grant number: MA7-737909/19*

GO!DIGITAL NEXT GENERATION

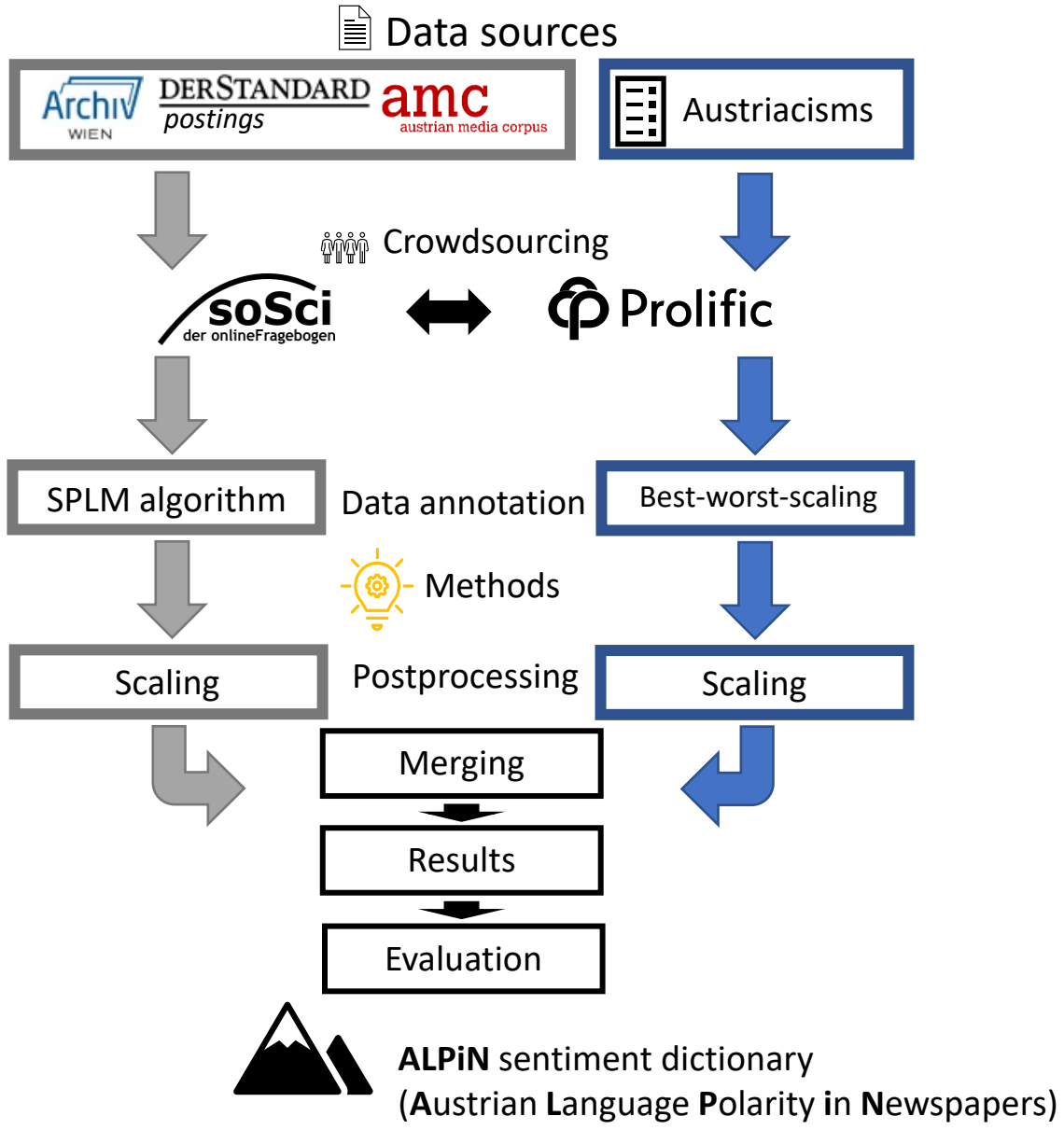
https://dylen.acdh.oeaw.ac.at/dysen/

Problem Statement

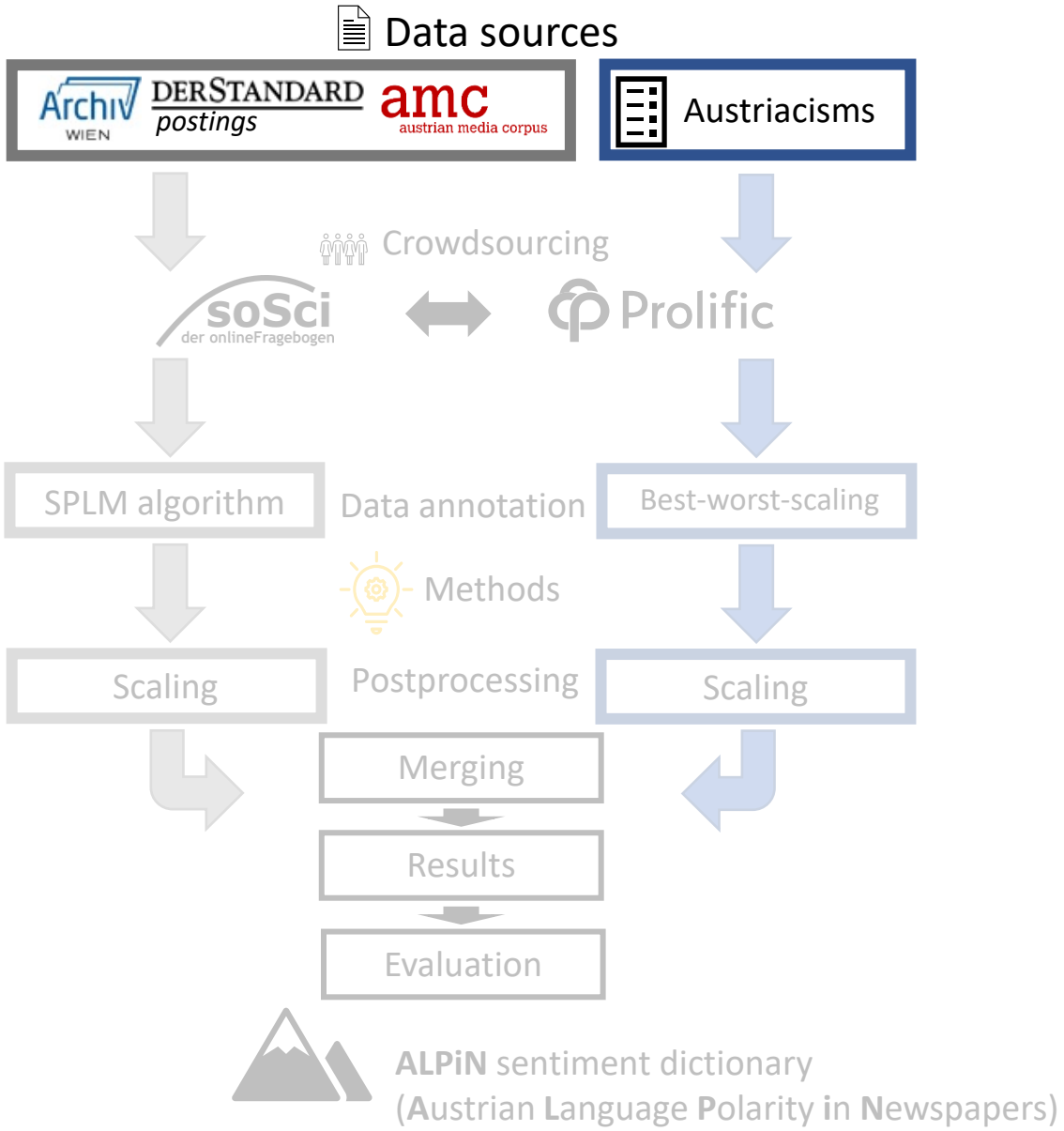
❌ There is no sentiment dictionary for Austrian German in this domain

🎯 **Goal**: Create Austrian German language resource in the domain of news media and politics

Content


Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary
(Austrian Language Polarity in Newspapers)

Content


Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary
(Austrian Language Polarity in Newspapers)

Data sources: Archiv WIEN Viennese politicians

Politician archive of Vienna (*POLAR*[1]) of the Vienna City and State Archives

Members of the

- Vienna City Council
- Vienna City Senate
- Vienna State Parliament
- Vienna State Government

active between the 13th and 20th parliamentary term (1983 to 2020)

= **487 politicians**

[1] https://www.wien.gv.at/advuew/internet/AdvPrSrv.asp?Layout=histpolsuche&Type=S&Hlayout=histpolsuche&HP=Y&RF=02&ICD=2011021810192827

Data sources: DERSTANDARD *postings* (1 Million Posts Corpus)

(Schabus et al., 2017)

- Forum posts from 2015 to 2016
- 3599 posts labelled for sentiment by professional forum moderators

	ID_Post	Body	Category
0	3326	Top qualifizierte Leute verdienen auch viel.	SentimentNeutral
1	5321	Gott sei dank ist für sie eine Umfrage alles, ...	SentimentNegative
2	5590	Sorry, aber die FPÖ tut eigentlich gar nichts ...	SentimentNeutral
3	6015	Weil es dein meisten Leuten verständlicherweis...	SentimentNegative
4	8213	Na wer weis was da vorgefallen ist...	SentimentNeutral

```
SentimentNeutral     1865
SentimentNegative    1691
SentimentPositive      43
```

Data sources: amc austrian media corpus Austrian Media Corpus[1]

- Contains Austrian print media
- Preprocessed and linguistically annotated (Ransmayr et al.,2017)
- Yearly updates

Our data:

- Print media related to Vienna between 1996 and 2017
- No APA and OTS articles *("Presseaussendungen")*

➡ Text snippets of around 60 tokens around the politicians' name were extracted

[1] https://amc.acdh.oeaw.ac.at/

Data sources: Austriacisms

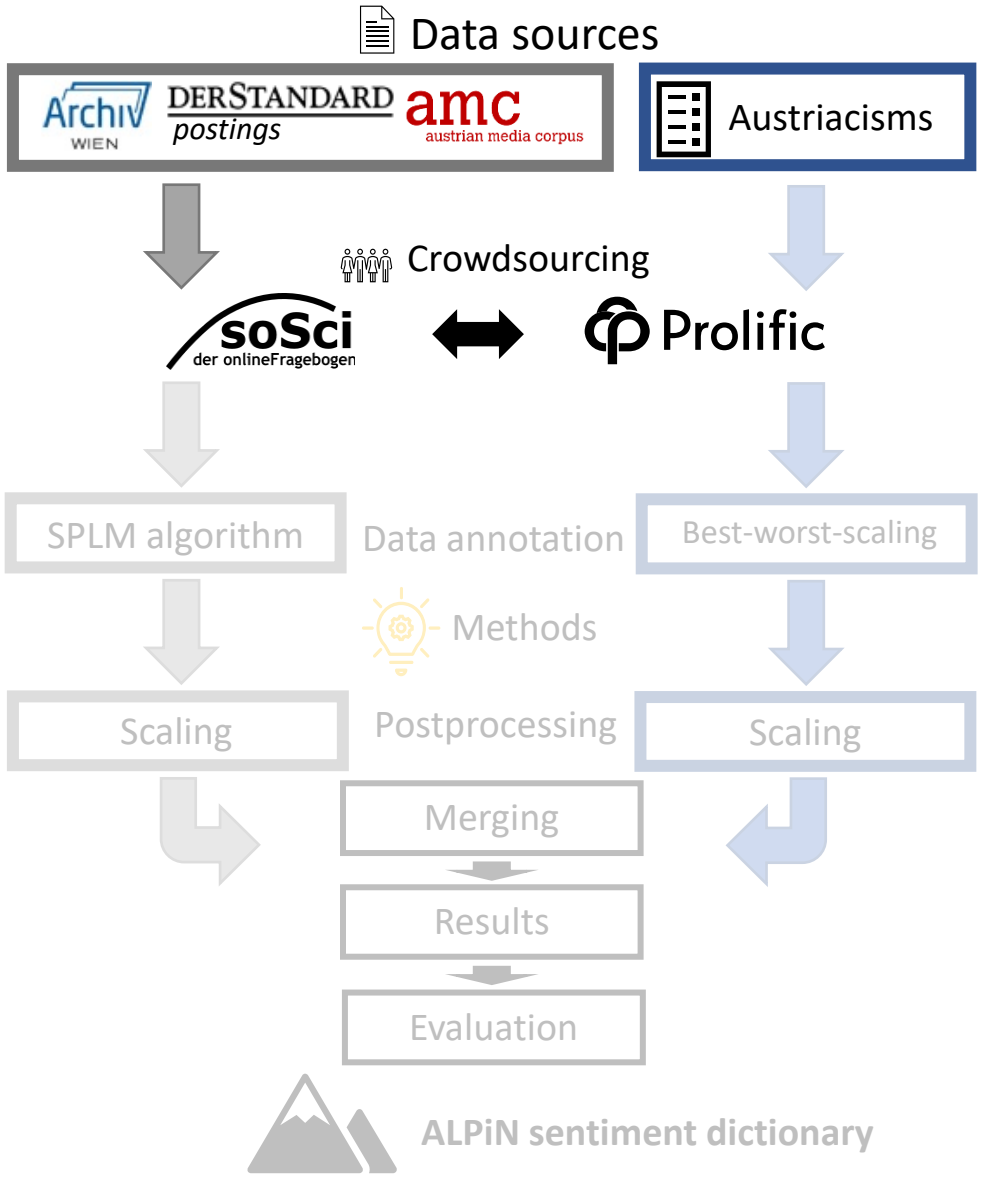
Based on:

- „Variantenwörterbuch des Deutschen“ (VWB; words specific to Austria) (Ammon et al.,2016)
- Austriacism list of Wikipedia[1]

Combined list manually checked by linguist experts of our project team

= <u>1600 words</u>

1. https://de.wikipedia.org/wiki/Liste_von_austriacismen


Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary

Crowd sourcing: amc austrian media corpus Austrian Media Corpus

- Each item labelled ≥ 3 times
- Majority vote (equal number per class = rated as neutral)
- Three classes: positive, neutral, negative
- quality control (≥75% correct)

Restricted annotators by:

- Current Country of Residence (Germany, Austria, Switzerland)
- Nationality (Germany, Austria, Switzerland)
- First Language (German)

Crowd sourcing: amc austrian media corpus

1st annotation run (70 annotators after excluding the 14 bad ones)

- 2376 items
- Fleiss-Kappa: 0.295 (fair inter-annotator agreement)

neutral	1202
positive	598
negative	576

2nd annotation run (88 annotators after excluding the 15 bad ones)

- 2970 items
- Fleiss-Kappa: 0.283 (fair inter-annotator agreement)

neutral	1492
positive	787
negative	691

Output: 5346 labelled text snippets including Viennese politicians

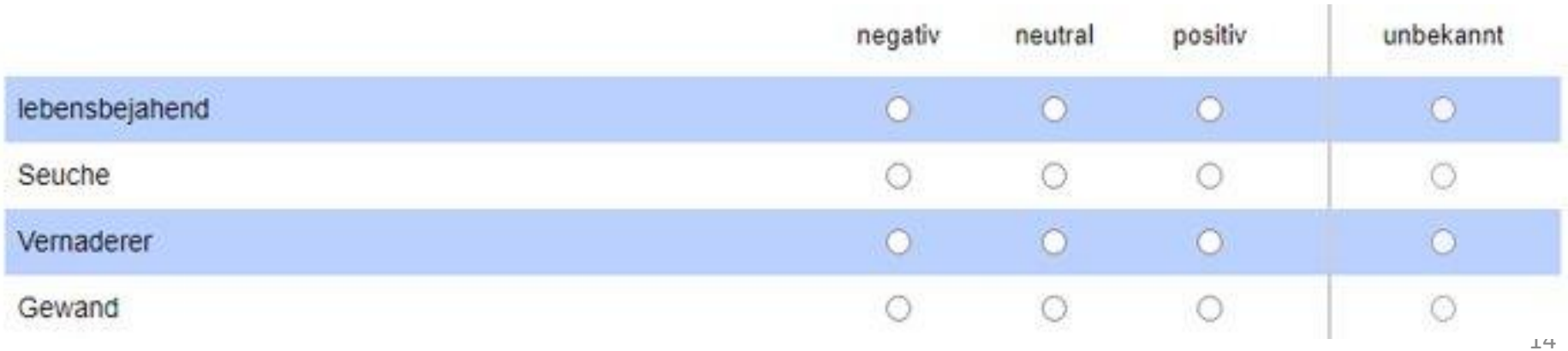
Crowd sourcing: Austriacisms (Survey 1)

Survey 1 (Preselection):

- Over 1 600 words in total
- quality control (≥75% correct)
- Four options (positive, neutral, negative, unknown)

Restricted annotators by:

- Current Country of Residence (Austria)
- Nationality (Austria)
- First Language (German)



	negativ	neutral	positiv	unbekannt
lebensbejahend	○	○	○	○
Seuche	○	○	○	○
Vernaderer	○	○	○	○
Gewand	○	○	○	○

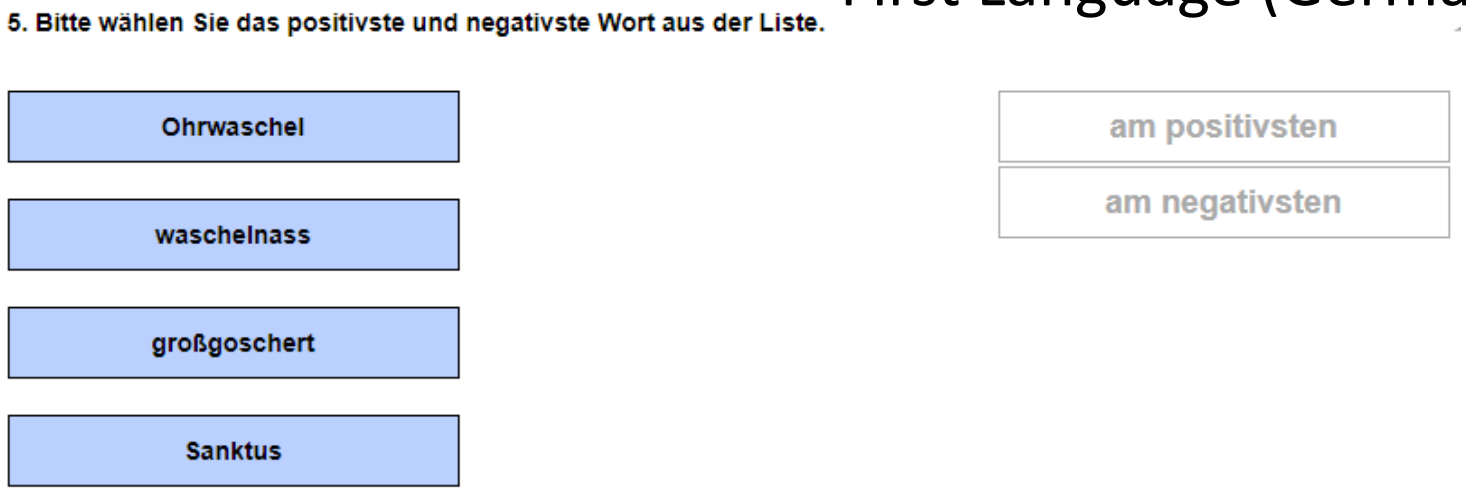
Crowd sourcing: Austriacisms (Survey 2)

Survey 2:

- Best-worst-scaling (BWS) method[1] (Kiritchenko & Mohammad, 2017)
- 1074 tuples
- quality control (≥75% correct)

Restricted annotators by:

- Current Country of Residence (Austria)
- Nationality (Austria)
- First Language (German)



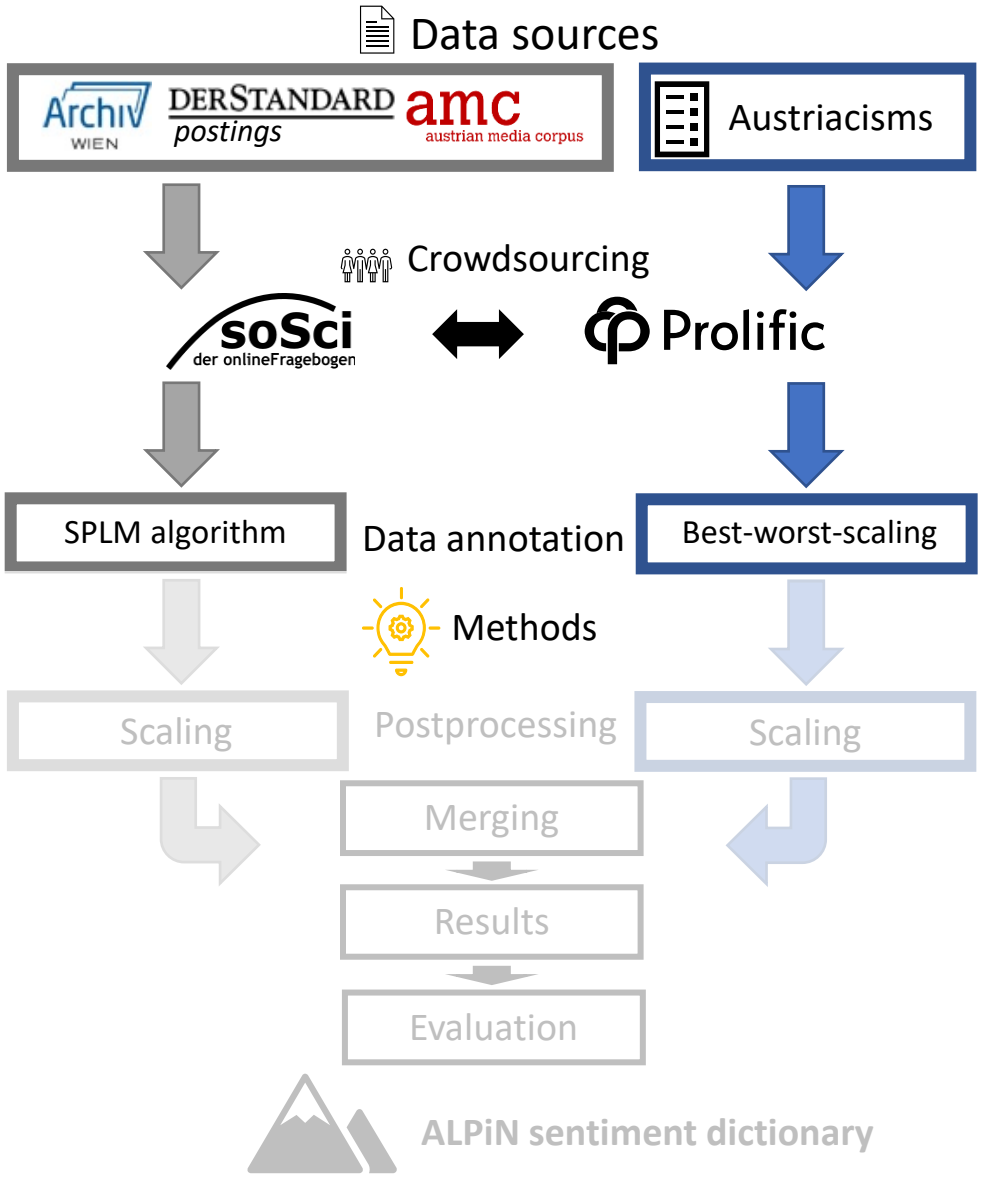
1. Calculation script provided by Mohammad: http://saifmohammad.com/WebPages/BestWorst.html

Crowd sourcing: Austriacisms

- 34 annotators after excluding the 6 bad ones
- **Output:** 4417 tuples (BestItem, WorstItem)

	Item1	Item2	Item3	Item4	BestItem	WorstItem
0	Rodel	Knödelakademie	Keiler	Gelenksbeschwerden	Rodel	Gelenksbeschwerden
1	brennheiß	Stornoversicherung	Scherz(e)l	sich ausgehen	sich ausgehen	brennheiß
2	Steireranzug	Causa	Pönale	Lokalaugenschein	Lokalaugenschein	Steireranzug
3	Alumnat	Beiwagerl	Servus	kiefeln	Servus	kiefeln
4	Patschenkino	Aufnahmestopp	Straßenerhalter	Marmeladinger	Straßenerhalter	Aufnahmestopp
...	...	...	...	...	...	...
4412	ferten	Ermäßigungsausweis	Halbpreispass	versumpern	Ermäßigungsausweis	versumpern
4413	Zuhaus	Bramburi	Mistbauer	Beiwagerl	Zuhaus	Mistbauer
4414	Oja!	ludeln	Rettung	gar	Oja!	ludeln
4415	Stützlehrer	Mascherl	Einspänner	grauslich	Mascherl	grauslich
4416	Jausenbrot	enthaften	versperren	Schubhaft	Jausenbrot	Schubhaft

4417 rows × 6 columns


Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary

Methods: Data Annotation (Autriacisms)

Best-worst-scaling (BWS) method (Kiritchenko & Mohammad, 2017)

split-half reliability:

- Spearman correlation: 0.9159 +/- 0.0051

Output: 538 words

	word	tag	short-tag	score	scaled
0	fesch	ADJ	a	0.882	0.910217
1	Zuckerl	NOUN	n	0.879	0.907121
2	Topfenpalatschinke	NOUN	n	0.857	0.884417
3	leiwand	ADJ	a	0.853	0.880289
4	Ersparnis	NOUN	n	0.844	0.871001
...	...	...	...	...	...
533	Schussattentat	NOUN	n	-0.844	-0.871001
534	Exekution	NOUN	n	-0.848	-0.875129
535	speiben	VERB	v	-0.875	-0.902993
536	Brandleger	NOUN	n	-0.879	-0.907121
537	Fotze	NOUN	n	-0.969	-1.000000

538 rows × 5 columns

Methods: Data Annotation (amc, derStandard postings)

SPLM method (Almatarneh & Gamallo, 2018)

Algorithm to generate a sentiment score based on labelled text items.

Remark: “neutral” sentiment labels of the derStandard dataset were converted to “positive”. This was required to the high imbalance in the dataset.

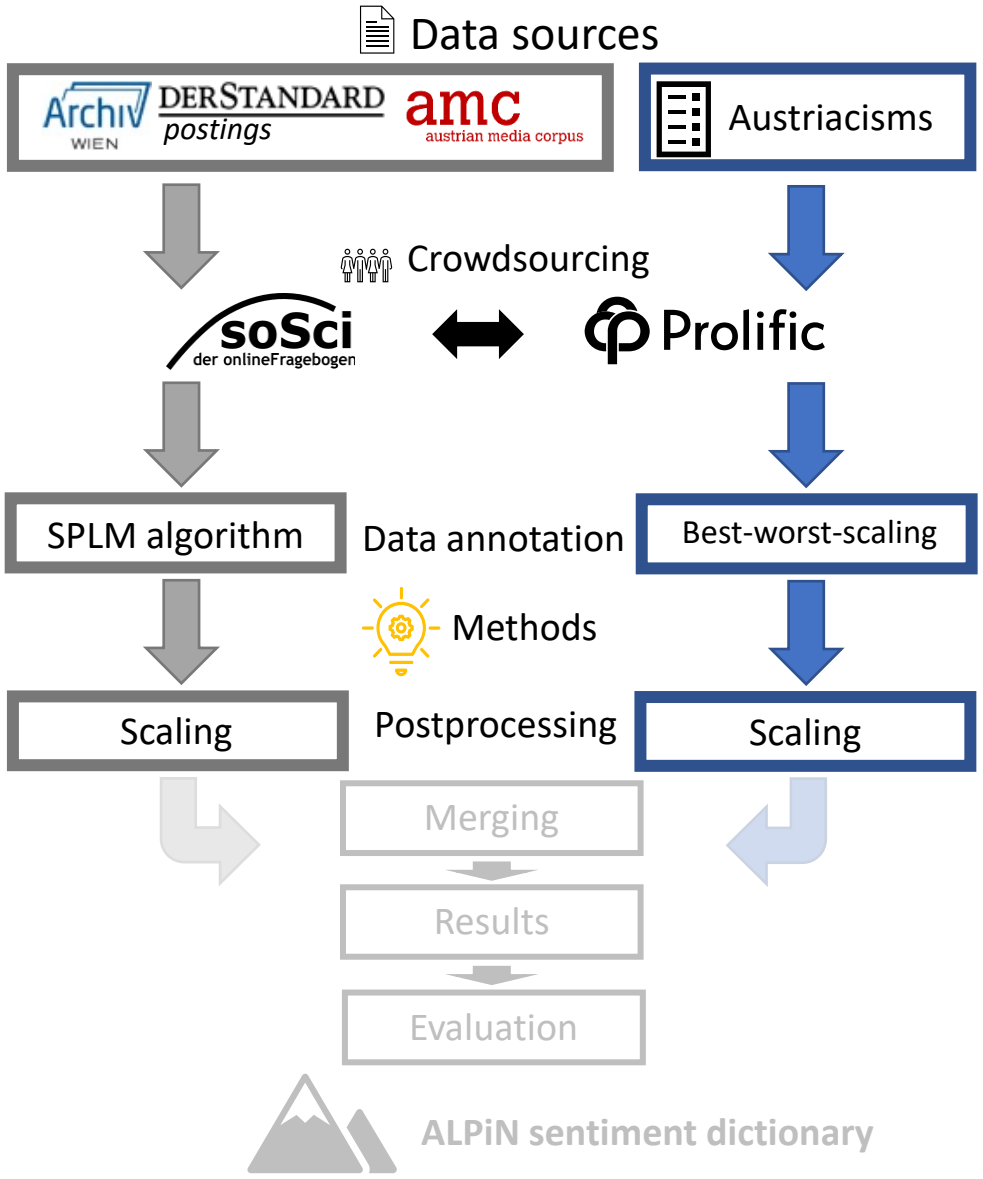
SentimentNeutral	1865
SentimentNegative	1691
SentimentPositive	43

Methods: Data Annotation (amc, derStandard postings) result

	word	Tag	D
0	geben	v	0.001057
1	Frau	n	0.001028
2	Jahr	n	0.000979
3	neu	a	0.000957
4	Mann	n	0.000844
...	...	...	...
8924	Pilz	n	-0.000920
8925	Westenthaler	n	-0.000994
8926	ÖVP	n	-0.001003
8927	Peter	n	-0.001078
8928	Flüchtling	n	-0.001189

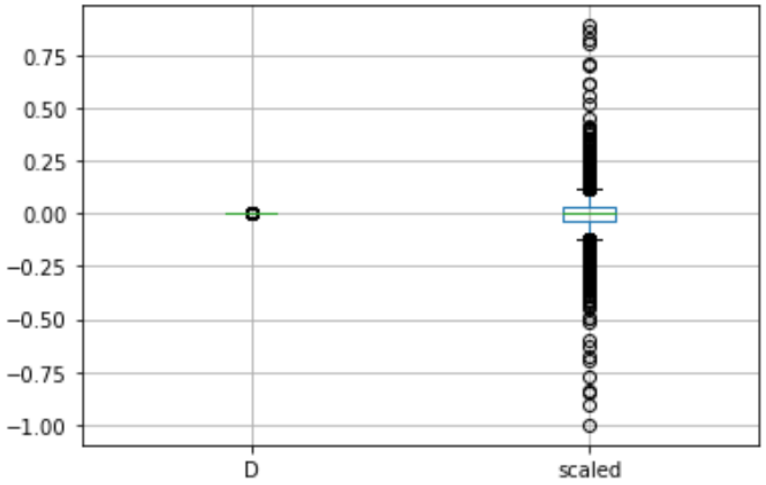
8929 rows × 4 columns

D(w): sentiment score
D(w) [-1;+1]

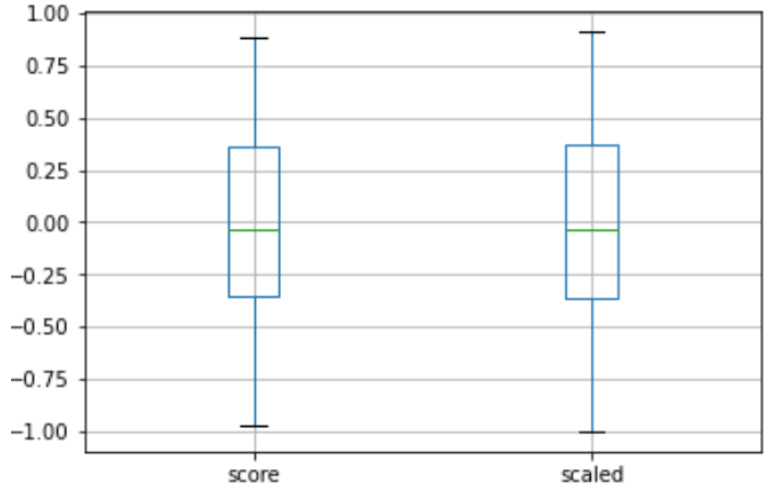

Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary

Methods: Postprocessing (1)

Scaling to [-1,+1] with „max_abs_scaler of sklearn“[1] before merging the dictionaries

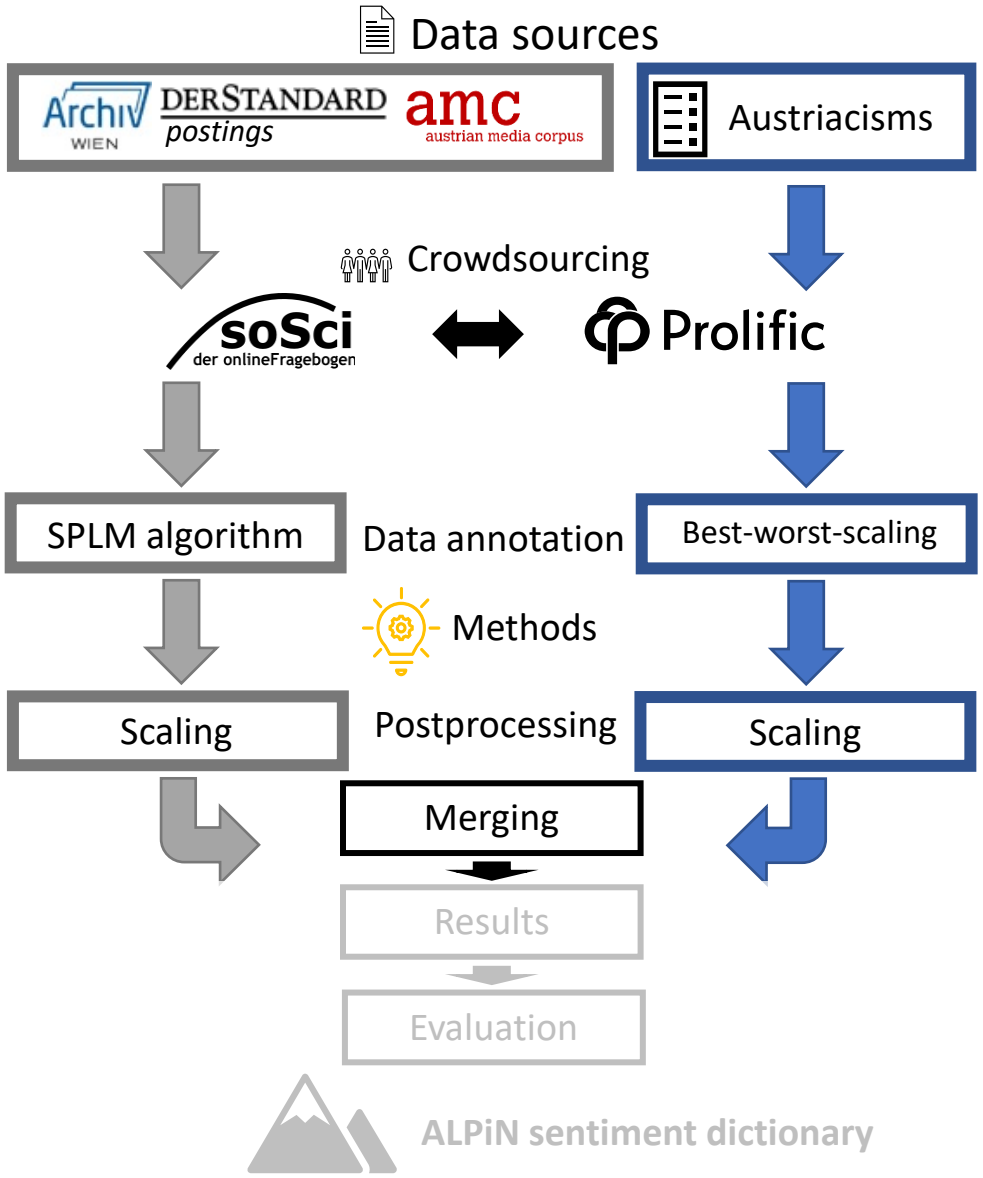


amc with derStandard postings after applying SPLM



Austriacisms after applying BWS

1. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html


Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary

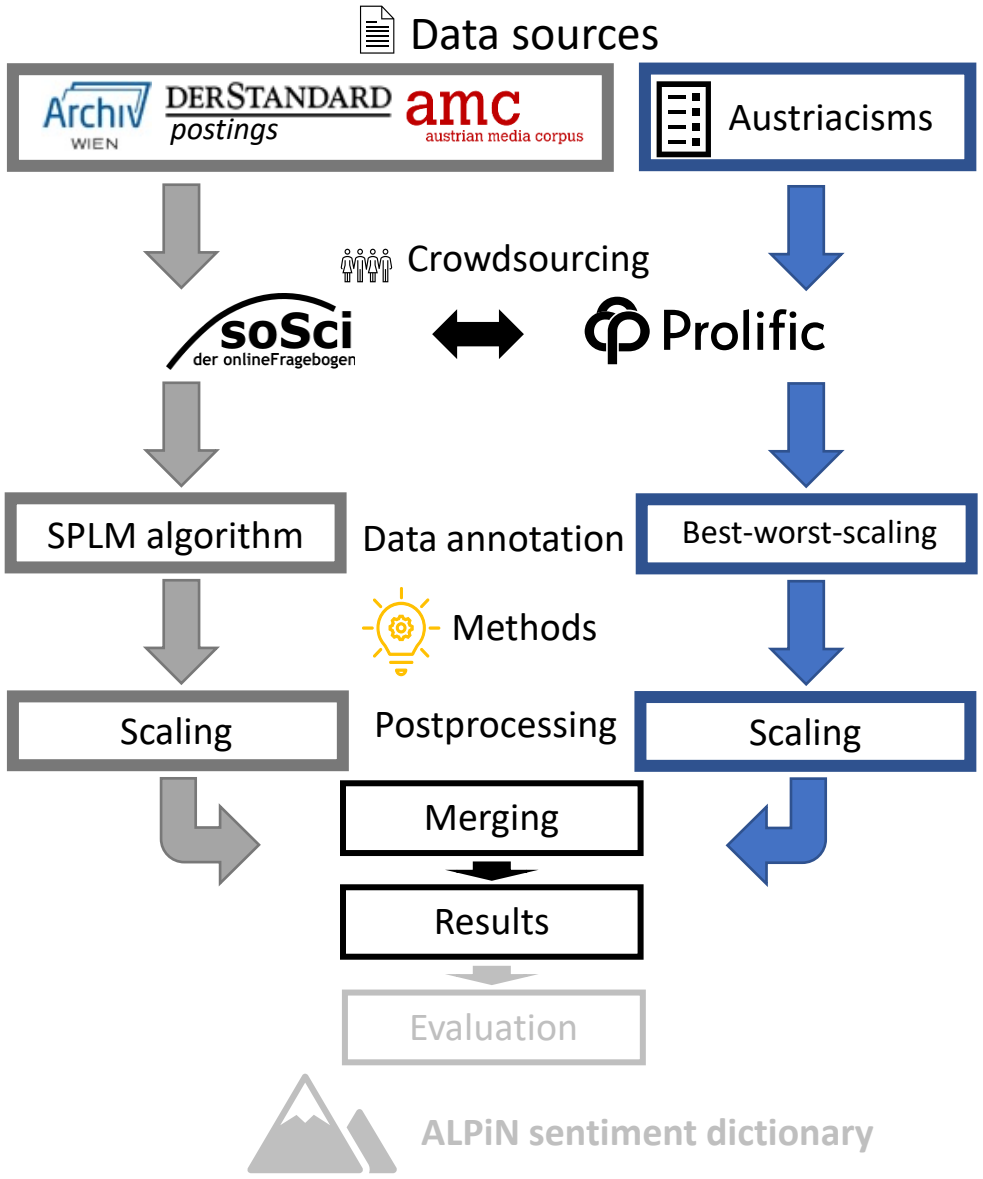
Methods: Postprocessing (2)

Comparison of words which occur in both dictionaries (austriacisms vs amc+derStandard postings):

	word	short-tag	sentiment_austriazism	sentiment_dysen_with_derstandard
0	Wiese	n	0.750258	0.026873
1	Karenz	n	0.742002	0.040310
2	Angelobung	n	0.728586	-0.050719
3	Ehrenzeichen	n	0.710010	0.067183
4	Gehalt	n	0.644995	0.210822
5	aufrecht	a	0.625387	-0.031037
6	maturieren	v	0.625387	0.026873
7	ÖAMTC	n	0.562436	-0.031037
8	einbringen	v	0.547988	-0.123202
9	Team	n	0.515996	0.166348
20	Abgang	n	0.000000	-0.077592
21	Klappe	n	-0.226006	-0.031037
22	klagen	v	-0.312693	-0.054883
23	angreifen	v	-0.343653	-0.005300
24	Fleck	n	-0.375645	-0.031037
25	Einvernahme	n	-0.386997	-0.031037
26	Freunderlwirtschaft	n	-0.437564	-0.031037
27	versperren	v	-0.486068	-0.031037
28	Mist	n	-0.594427	-0.031037
29	sekkieren	v	-0.688338	-0.031037
30	exekutieren	v	-0.837977	-0.004164
31	Exekution	n	-0.875129	0.026873

Restrictions:

During merging duplicates will be removed by using the Austriacism words prioritized.


Data sources
Archiv WIEN
DERSTANDARD postings
amc austrian media corpus
Austriacisms
Crowdsourcing
soSci der onlineFragebogen
Prolific
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPiN sentiment dictionary

Results

amc + derStandard postings

+ austriacisms

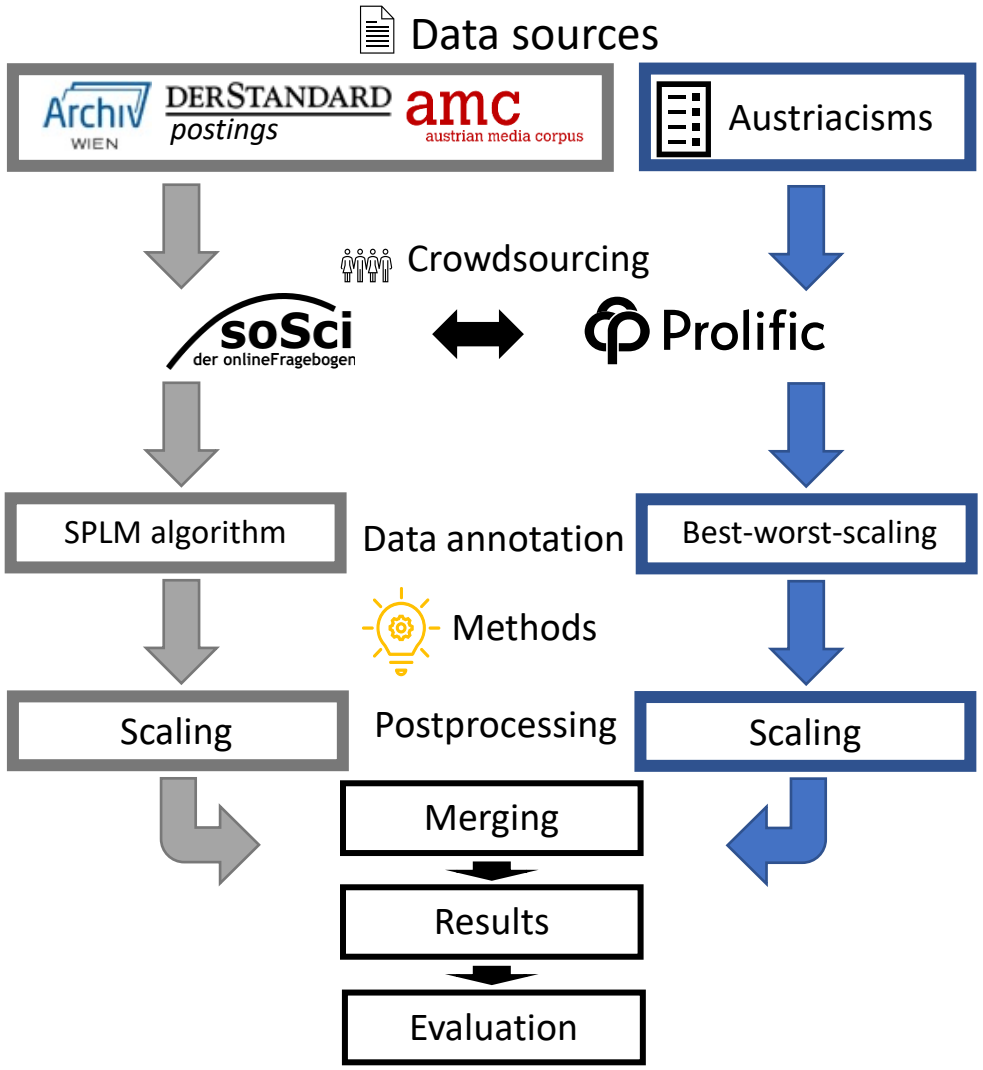
Scaled to [-1,+1] with

„max_abs_scaler of sklearn“[1]

	word	short-tag	scaled
0	fesch	a	0.910217
1	Zuckerl	n	0.907121
2	geben	v	0.888855
3	Topfenpalatschinke	n	0.884417
4	leiwand	a	0.880289
...	...	...	...
9430	speiben	v	-0.902993
9431	Peter	n	-0.906709
9432	Brandleger	n	-0.907121
9433	Fotze	n	-1.000000
9434	Flüchtling	n	-1.000000

9435 rows × 3 columns

1. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html



ALPiN sentiment dictionary
(**A**ustrian **L**anguage **P**olarity **i**n **N**ewspapers)

Evaluation (2)

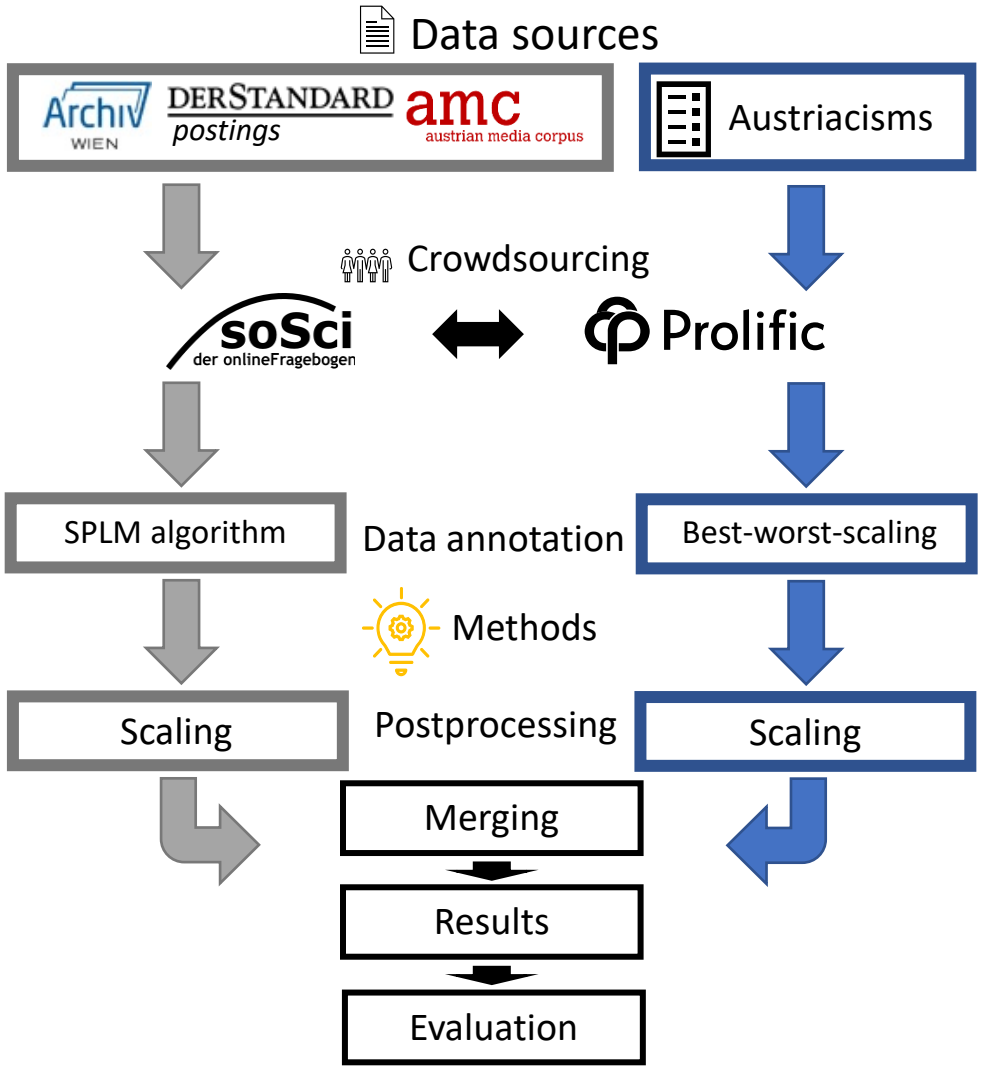
Evaluate the dictionary which is based on amc, derStandard postings and the austriacism list against "derStandard postings" and "DYSEN":

1st against derStandard only

Accuracy: 0,77
Precision: 0,78
Recall: 0,79
F1: 0,78

2nd against amc only

Accuracy: 0,82
Precision: 0,83
Recall: 0,84
F1: 0,83



ALPiN sentiment dictionary
(**A**ustrian **L**anguage **P**olarity **i**n **N**ewspapers)

https://dylen.acdh.oeaw.ac.at/dysen/

Discussion (1)

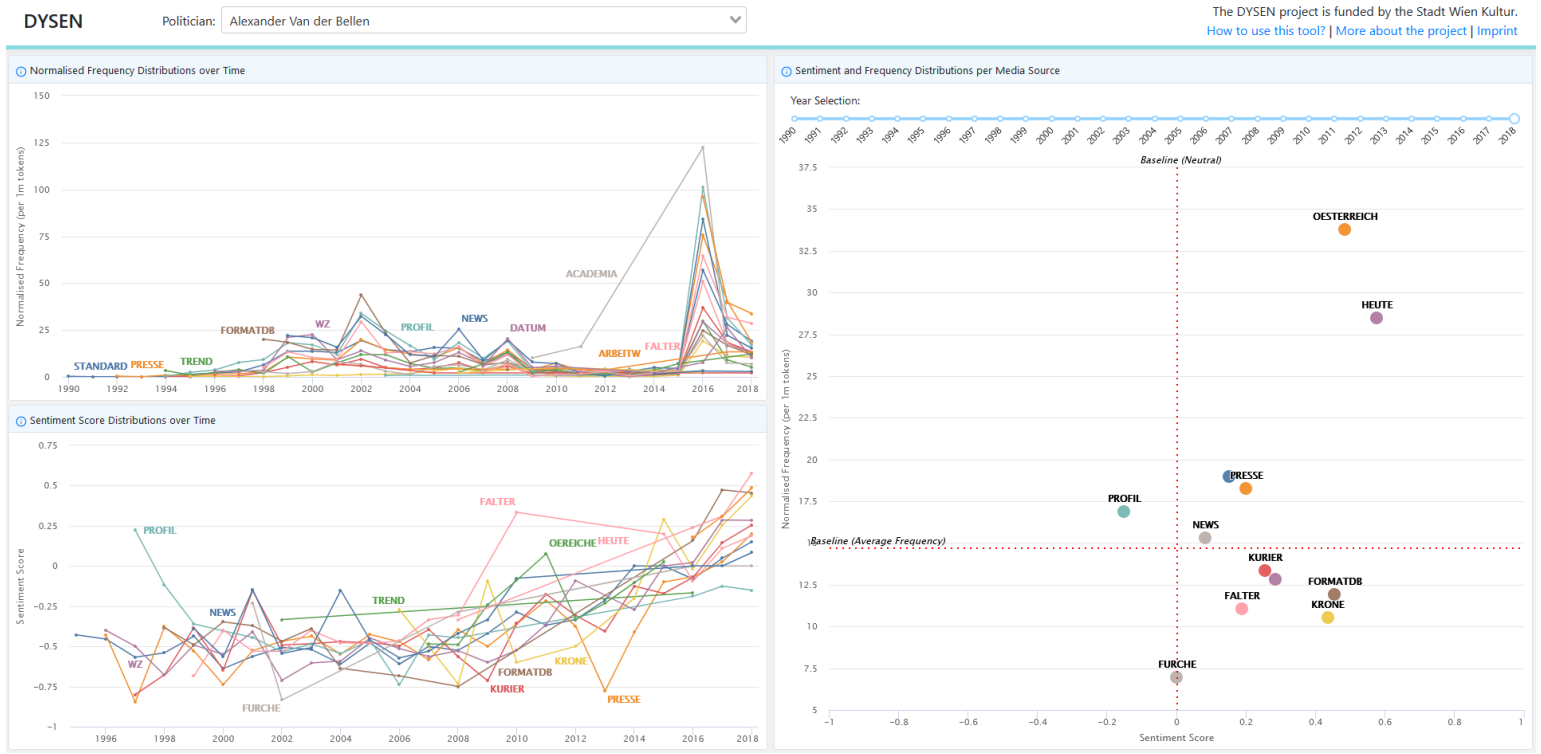
- Difficult to label news media (mainly “neutral” texts)
- Limited text length
- No external dataset for evaluation
- Potential bias during labelling e.g. words like “Flüchtling” negatively annotated

Future work:

- Mitigate the potential bias due to labelling
- Improvement of the text extraction by using Aspect-based sentiment analysis
- Investing more money to label a bigger dataset
- Expanding the scope of the project to all politicians and media in Austria

Discussion (2)

Tool created as part of the DYSEN project which uses the ALPiN dict.:



https://dylen.acdh.oeaw.ac.at/dysen/

Thank you very much!

thomas.kolb@tuwien.ac.at

Funded by:

Stadt Wien

Grant number:
MA7-737909/19

GO!DIGITAL NEXT GENERATION

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

universität wien

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

References

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 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). Association for Computing Machinery, New York, NY, USA, 1241–1244. DOI:https://doi.org/10.1145/3077136.3080711

Ransmayr, Jutta, Karlheinz Mörth, und Matej Ďurčo (2017): AMC (Austrian Media Corpus) – Korpusbasierte Forschungen zum österreichischen Deutsch. In Digitale Methoden der Korpusforschung in Österreich (= Veröffentlichungen zur Linguistik und Kommunikationsforschung Nr. 30), Hrsg. C. Resch und W. U. Dressler, 27-38. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Bickel, H., Hofer, L., & Suter, S. (2015). 22. Variantenwörterbuch des Deutschen (VWB)–NEU. In Regionale Variation des Deutschen (pp. 541-562). De Gruyter.

Kiritchenko, S., & Mohammad, S. M. (2017). Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling.

Kiritchenko, S., & Mohammad, S. (2017). Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 465–470. https://doi.org/10.18653/v1/P17-2074

Almatarneh, S., & Gamallo, P. (2018). Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification. 175–182. https://doi.org/10.1007/978-3-319-61578-3_17

Rouces, J., Tahmasebi, N., Borin, L., & Eide, S. R. (2018). Generating a Gold Standard for a Swedish Sentiment Lexicon. LREC.