Language Resources and Evaluation Conference 2022 | 20-25 June 2022, Marseille, France

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]

What Is This All About?

This publication is part of the **DYSEN Project** which stands for:
„Dynamic Sentiment Analysis as Emotional Compass for the Digital Media Landscape"

Research Question: How do print media report about Viennese Policitians?
Aim of this project: Develop a tool that can detect change of emotional polarization of politicians in Austrian Newspapers

Problem

✖ Currently there is no dictionary based on Austrian-German in the domain of news media and politics

To resolve that research gap the Austrian Language Polarity in Newspapers (ALPIN) sentiment dictionary is introduced

Data Sources

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/kultur/archiv/politik/

AMC

https://amc.acdh.oeaw.ac.at/
- 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

Standard Posts (STP)

https://www.derstandard.at/
1 Million Posts Corpus (Schabus et al., 2017)
- Forum posts from 2015 to 2016
- 3599 posts labelled for sentiment by professional forum moderators

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
= 1600 words

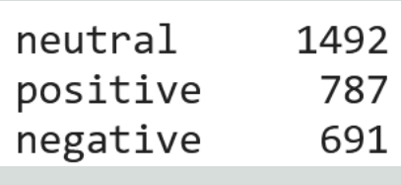
[1] https://de.wikipedia.org/wiki/Liste_von_austriacismen

Crowd Sourcing: AMC

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

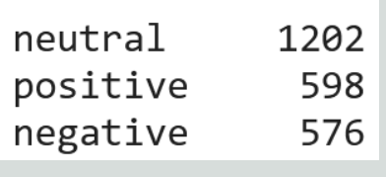
Restricted annotators by:
- Current Country of Residence (Germany, Austria, Switzerland)
- Nationality (Germany, Austria, Switzerland)
- First Language (German)

1st annotation run
(70 annotators after excluding the 14 bad ones)
2376 items
Fleiss-Kappa: 0.295 (fair inter-annotator agreement)



```
neutral     1492
positive     787
negative     691
```

2nd annotation run
(88 annotators after excluding the 15 bad ones)
2970 items
Fleiss-Kappa: 0.283 (fair inter-annotator agreement)



```
neutral     1202
positive     598
negative     576
```

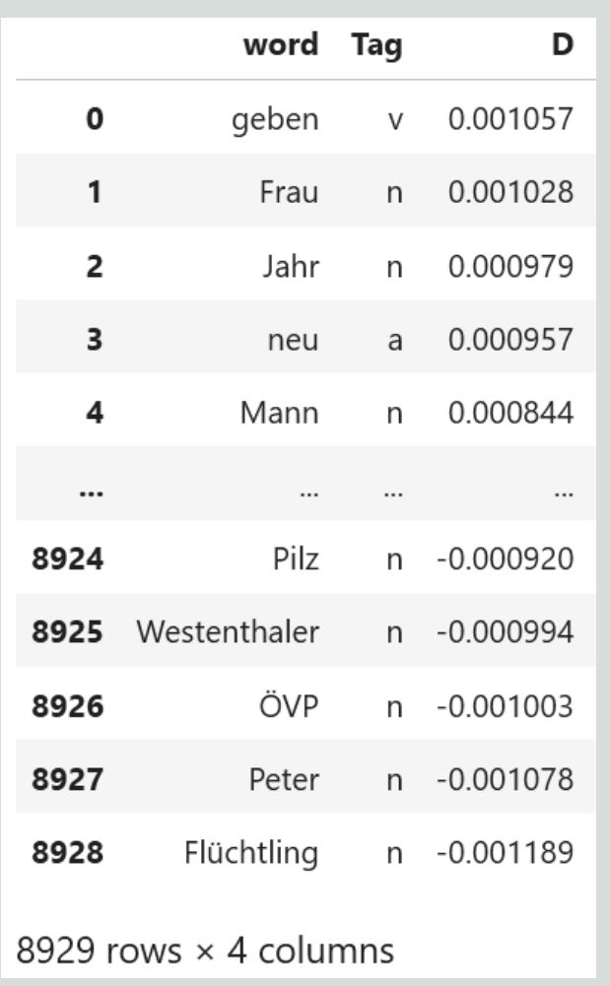
Output: 5346 labelled text snippets including Viennese politicians

Methods: AMC & STP

SPLM method
(Almatarneh & Gamallo, 2018)

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

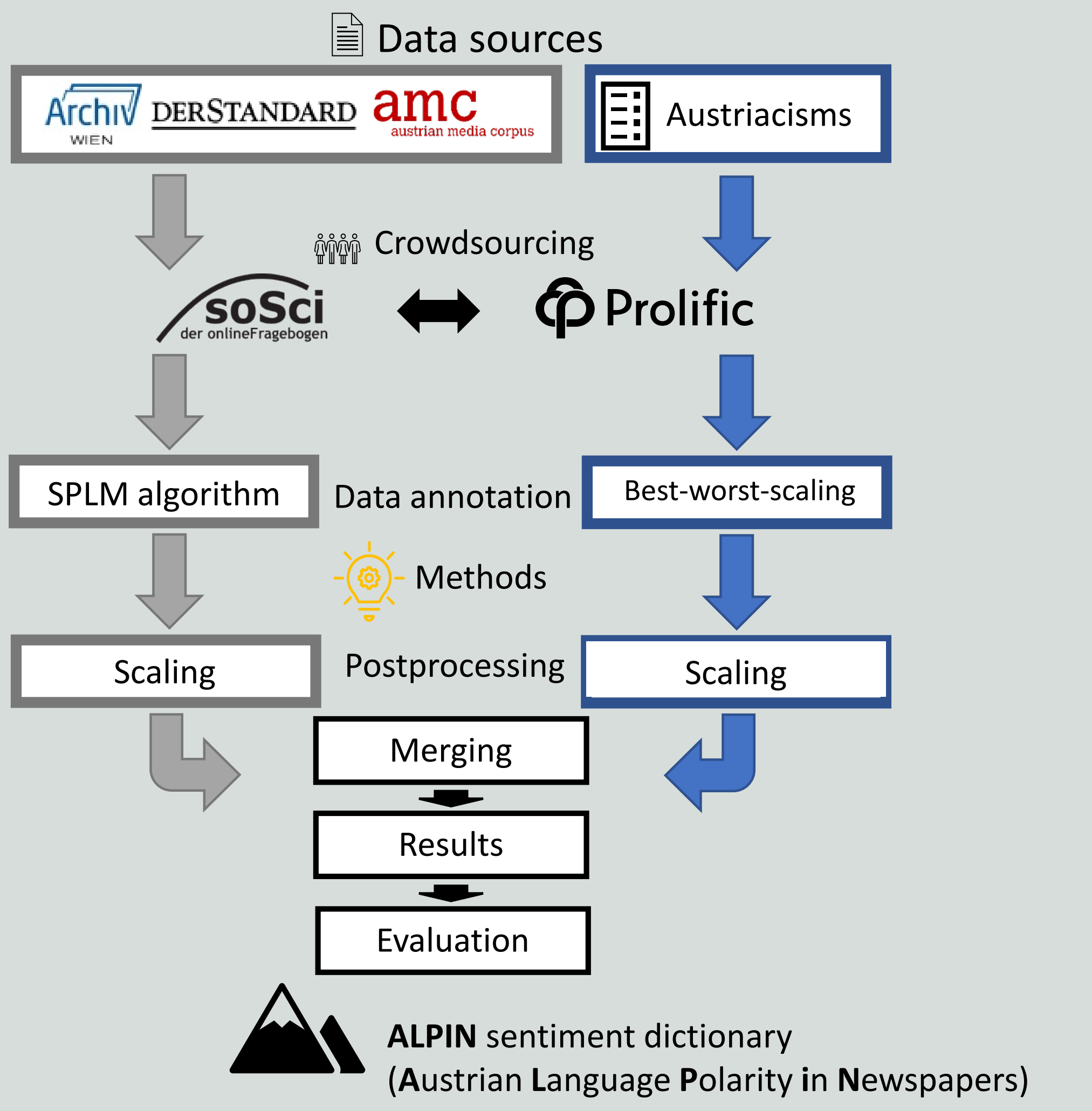
D(w): sentiment score
D(w) [-1;+1]



	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

Methodology

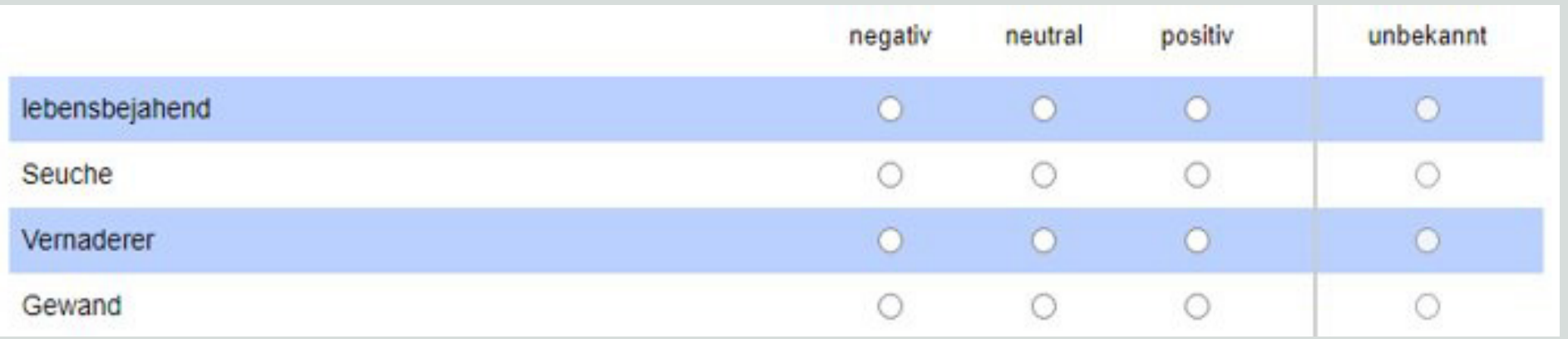


Data sources
Austriacisms
Crowdsourcing
SPLM algorithm
Data annotation
Best-worst-scaling
Methods
Scaling
Postprocessing
Scaling
Merging
Results
Evaluation
ALPIN sentiment dictionary
(**A**ustrian **L**anguage **P**olarity **in** **N**ewspapers)

Crowd Sourcing: Austriacisms

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

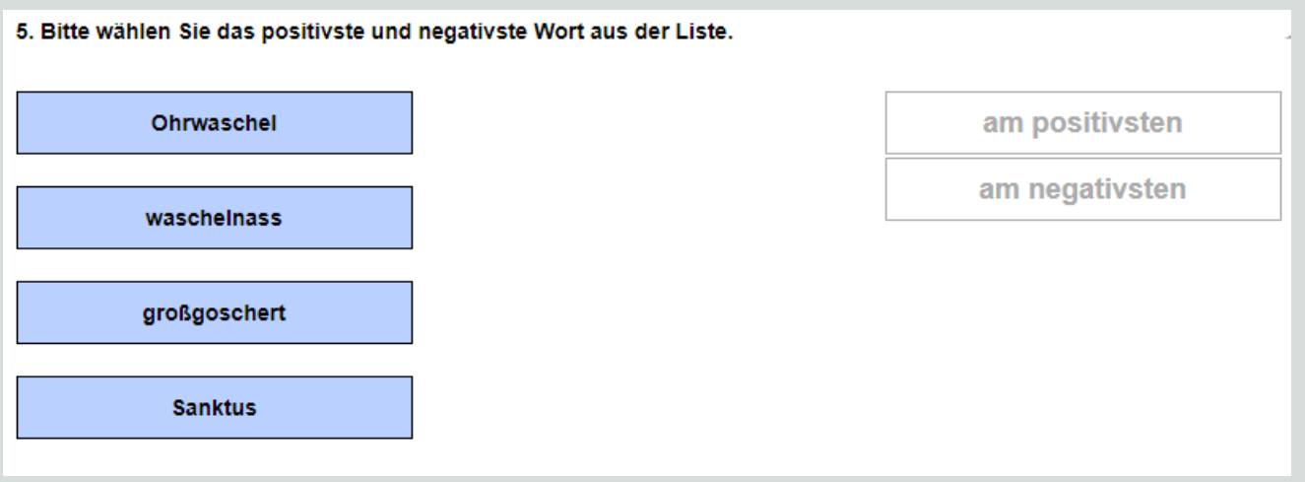
Restricted annotators by:
- Current Country of Residence (Austria)
- Nationality (Austria)
- First Language (German)



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

Survey 2:
- Best-worst-scaling (BWS) method[1] (Kiritchenko & Mohammad, 2017)
- 1074 tuples
- quality control (≥75% correct test items)

Restricted annotators by:
- Current Country of Residence (Austria)
- Nationality (Austria)
- First Language (German)

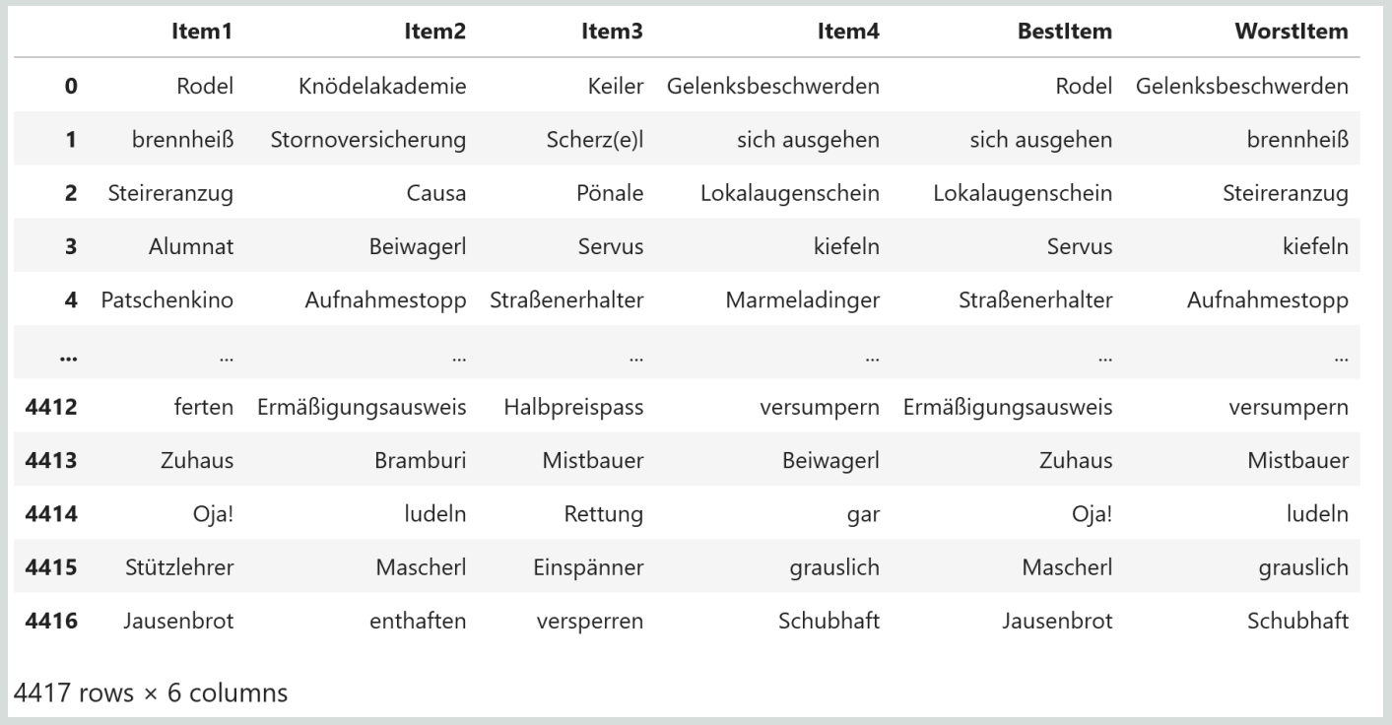


5. Bitte wählen Sie das positivste und negativste Wort aus der Liste.

Ohrwaschel
waschelnass
großgoschert
Sanktus

am positivsten
am negativsten

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



	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

34 annotators after excluding the 6 bad ones

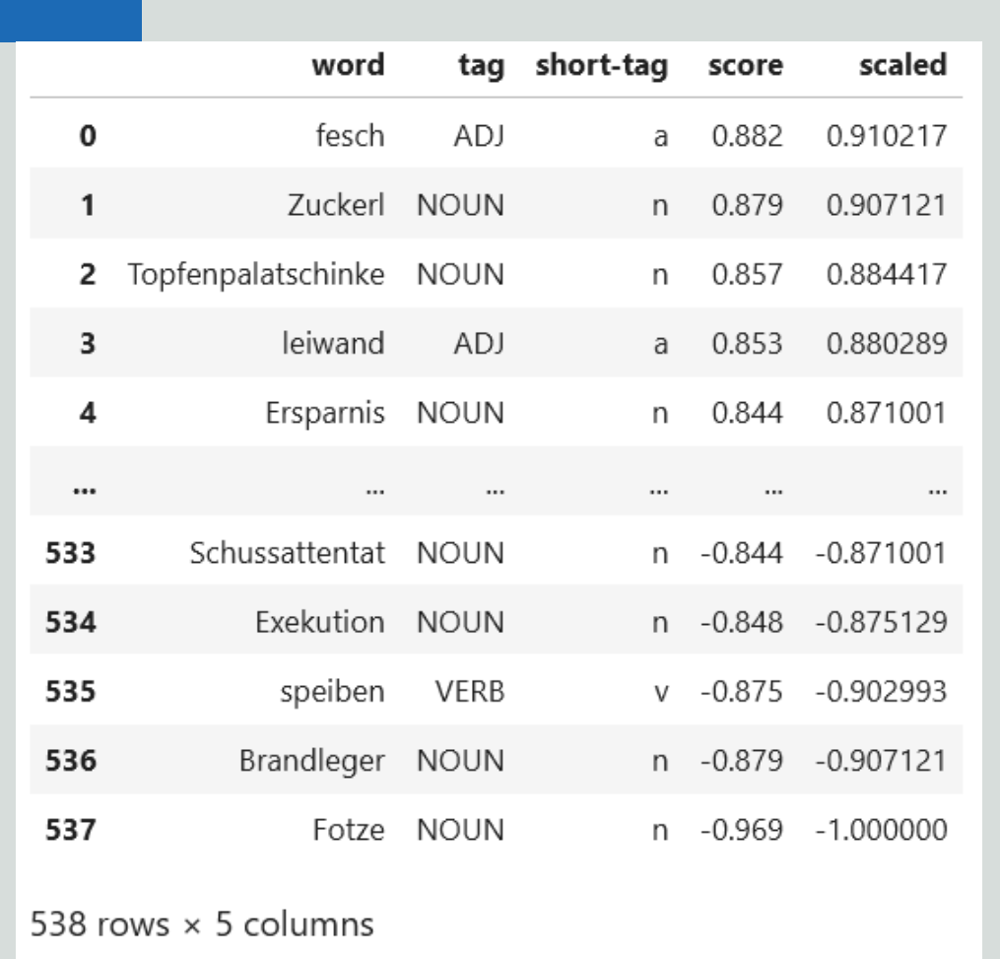
Output: 4417 tuples (BestItem, WorstItem)

Methods: Austriacisms

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

Split-half reliability:
Spearman correlation: 0.9159 +/- 0.0051

Output: 538 words

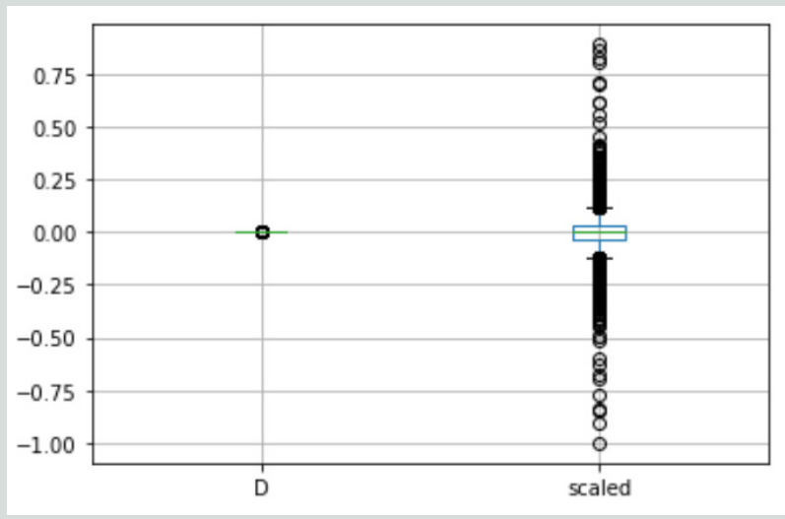


	word	tag	short-tag	score	scaled
0	fesch	ADJ	a	0.882	0.910217
1	Zuckerl	NOUN	n	0.879	0.907121
2	Topfenpalatschinke	NOUN	n	0.857	0.884417
3	leiwand	ADJ	a	0.853	0.880289
4	Ersparnis	NOUN	n	0.844	0.871001
...	...	...	...	...	...
533	Schussattentat	NOUN	n	-0.844	-0.871001
534	Exekution	NOUN	n	-0.848	-0.875129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

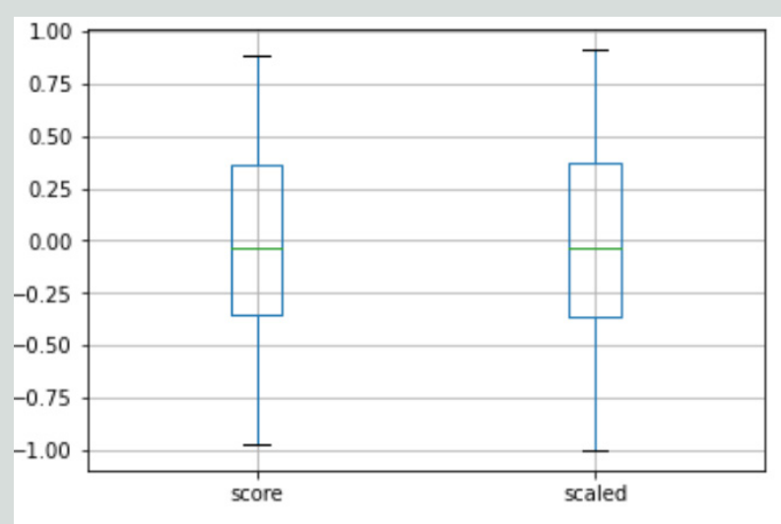
538 rows × 5 columns

Post Processing

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



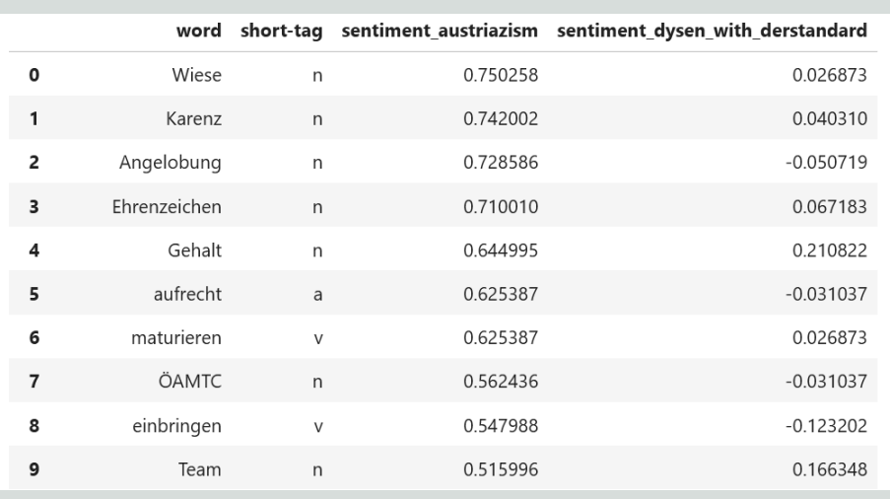
amc with STP after applying SPLM



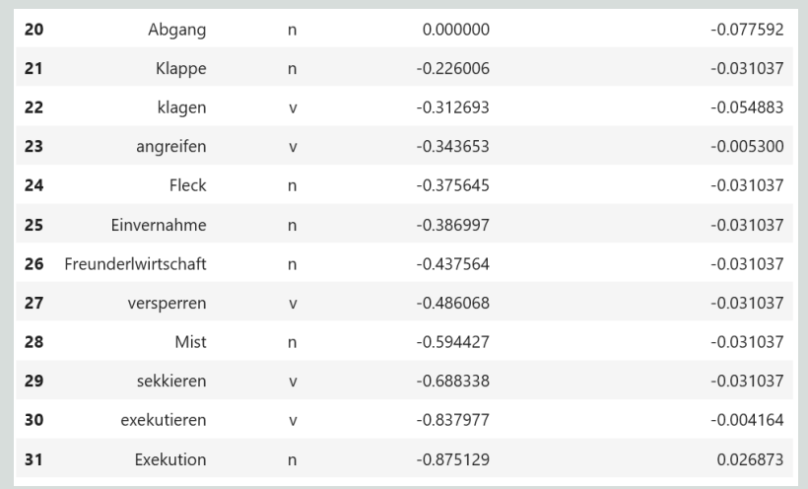
Austriacisms after applying BWS

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

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



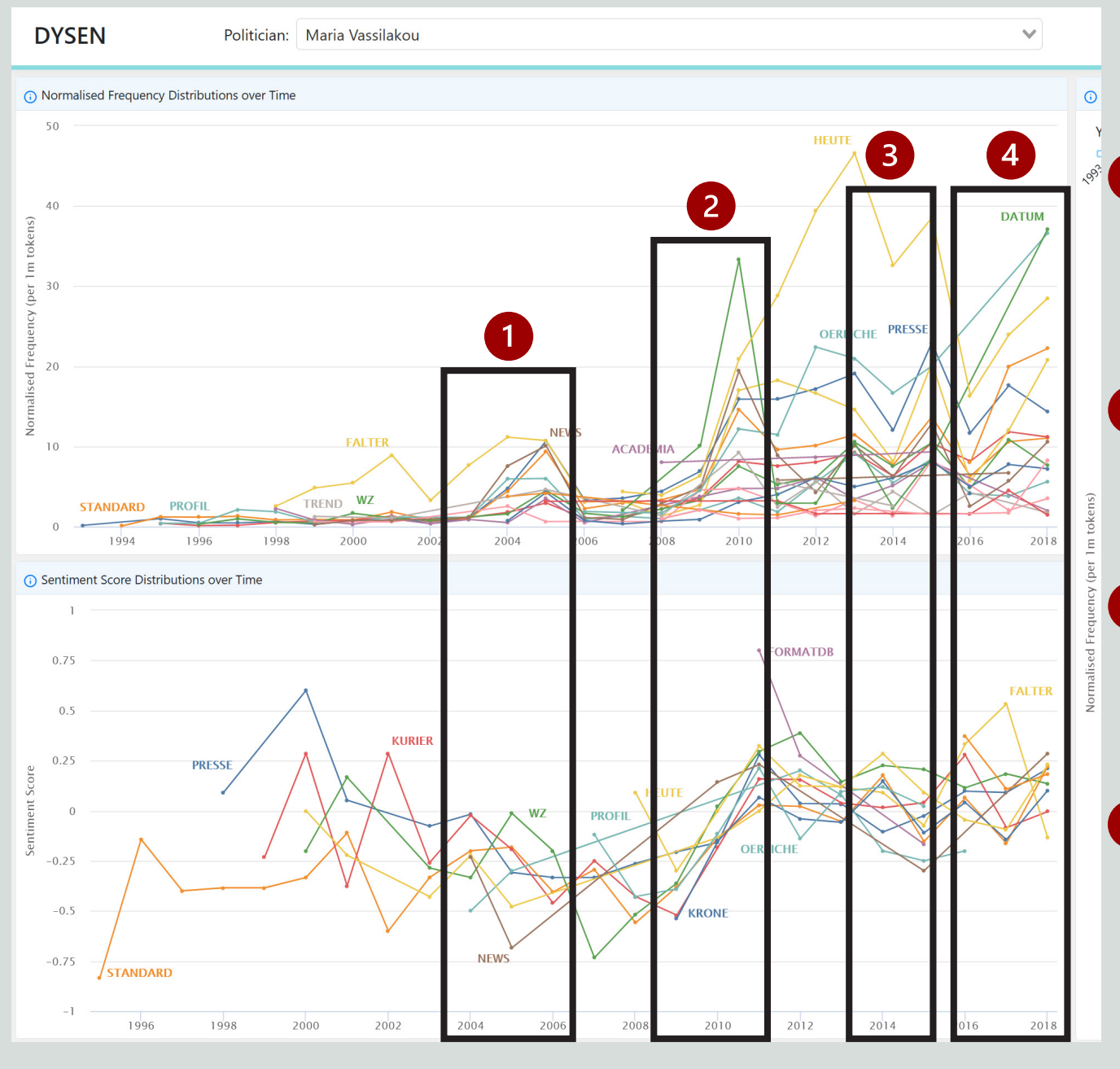
	word	short-tag	sentiment_austriazism	sentiment_dysen_with_derstandard
0	Wiese	n	0.750258	0.026873
1	Karenz	n	0.742002	0.040310
2	Angelobung	n	0.728586	-0.050719
3	Ehrenzeichen	n	0.710010	0.067183
4	Gehalt	n	0.644995	0.210822
5	aufrecht	a	0.625387	-0.031037
6	maturieren	v	0.625387	0.026873
7	ÖAMTC	n	0.562436	-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.

Web Application



DYSEN
Politician: Maria Vassilakou
Normalised Frequency Distributions over Time
Normalised Frequency (per 1m tokens)
Sentiment Score Distributions over Time
Sentiment Score
STANDARD
PROFIL
TREND
WZ
FALTER
NEWS
ACADEMIA
HEUTE
OERICHE
PRESSE
DATUM
KURIER
FORMATDB
KRONE
1
2
3
4

Politician: Maria Vassilakou

1. **2004** elected to the federal executive committee of the Green Party
2005 top candidate of the Greens for the municipal elections

2. **2010** top candidate in the state parliament and municipal council election - Elected as vice mayor

3. **2015** state parliament and municipal council election – Controversy surrounding her declaration to resign if Green Party loses vote share

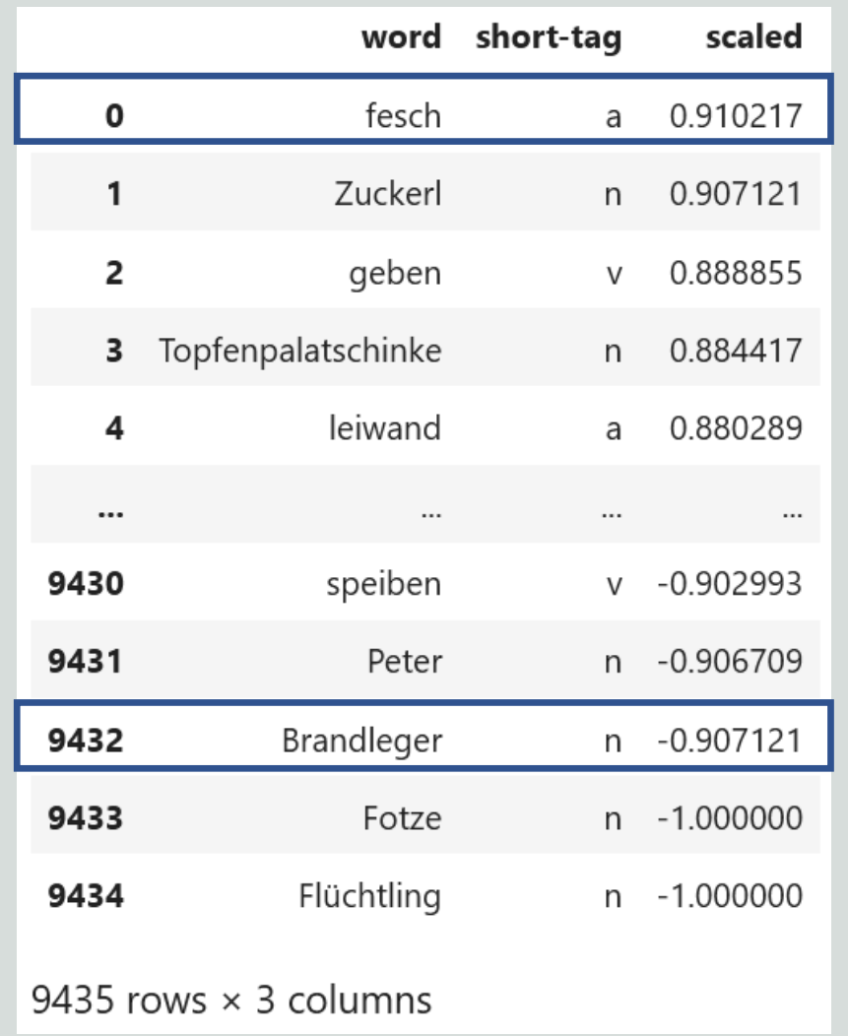
4. **2017** controversial high-rise project at the Heumarkt in Vienna; UNESCO sets the City of Vienna onto the Red List of World Heritage in Danger
2018 Announcement that she will not run in the next state parliament and municipal council election

DYSEN Website based on the ALPIN dictionary. Link: https://dysen-tool.acdh-dev.oeaw.ac.at/

Results

amc + derStandard + 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

Evaluation

Evaluated the dictionary which is based on amc, derStandard and the austriacism list against "derStandard" 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

Discussion

- 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

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). Onemillion posts: A data set of german online discussions. In Proceedings of the 40th International ACMSIGIR 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.
Almatarneh, S., & Gamallo, P. (2018). Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification. In F. De la Prieta, Z. Vale, L. Antunes, T. Pinto, A. T. Campbell, V. Julián, A. J. R. Neves, & M. N. Moreno (Eds.), Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017 (pp. 175–182). Springer International Publishing.
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

Contact

1) **Faculty of Informatics, TU Wien**
thomas.kolb@tuwien.ac.at, julia.neidhardt@tuwien.ac.at
2) **Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences**
tanja.wissik@oeaw.ac.at
3) **Department of European and Comparative Literature and Language Studies and Department of English and American Studies, University of Vienna**
bettina2.kern@univie.ac.at, andreas.baumann@univie.ac.at

Data

https://doi.org/10.5281/zenodo.5857150

TECHNISCHE UNIVERSITÄT WIEN
universität wien
ÖAW AUSTRIAN ACADEMY OF SCIENCES
funded by:
Stadt Wien
GO!DIGITAL NEXT GENERATION
Grant number: MA7-737909/19
https://dylen.acdh.oeaw.ac.at/dysen/