

Sentiment Analysis

ÖAW AI Winter School 2023

Thomas E. Kolb

PhD Student / Research Assistant
CDL - RecSys @ TU Wien

 thomas.kolb@tuwien.ac.at

Introduction

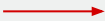


Nehammer
(politician)

Introduction (cont.)



Nehammer
(politician)



©Twitter/@NoeWehrtSich
polarizing action

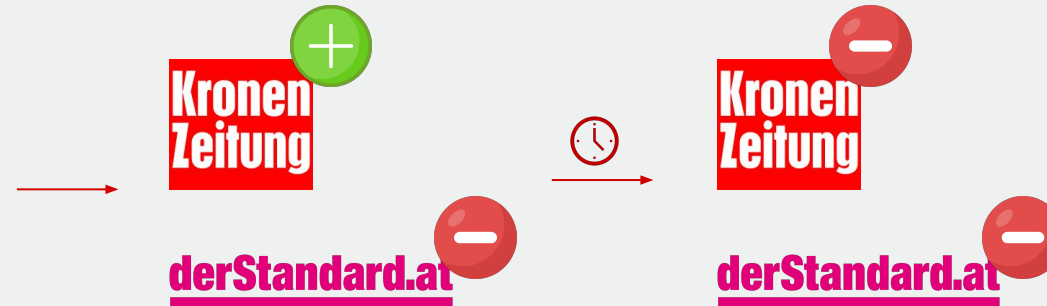
Introduction (cont.)



Nehammer
(politician)



©Twitter/@NoeWehrtSich
polarizing action



possible connection between polarizing
action and change in media reporting

Introduction (cont.)



Example of a research question (RQ) requiring sentiment analysis:

RQ: To what extent is it possible to predict the polarization* of politicians over time in different news media outlets?

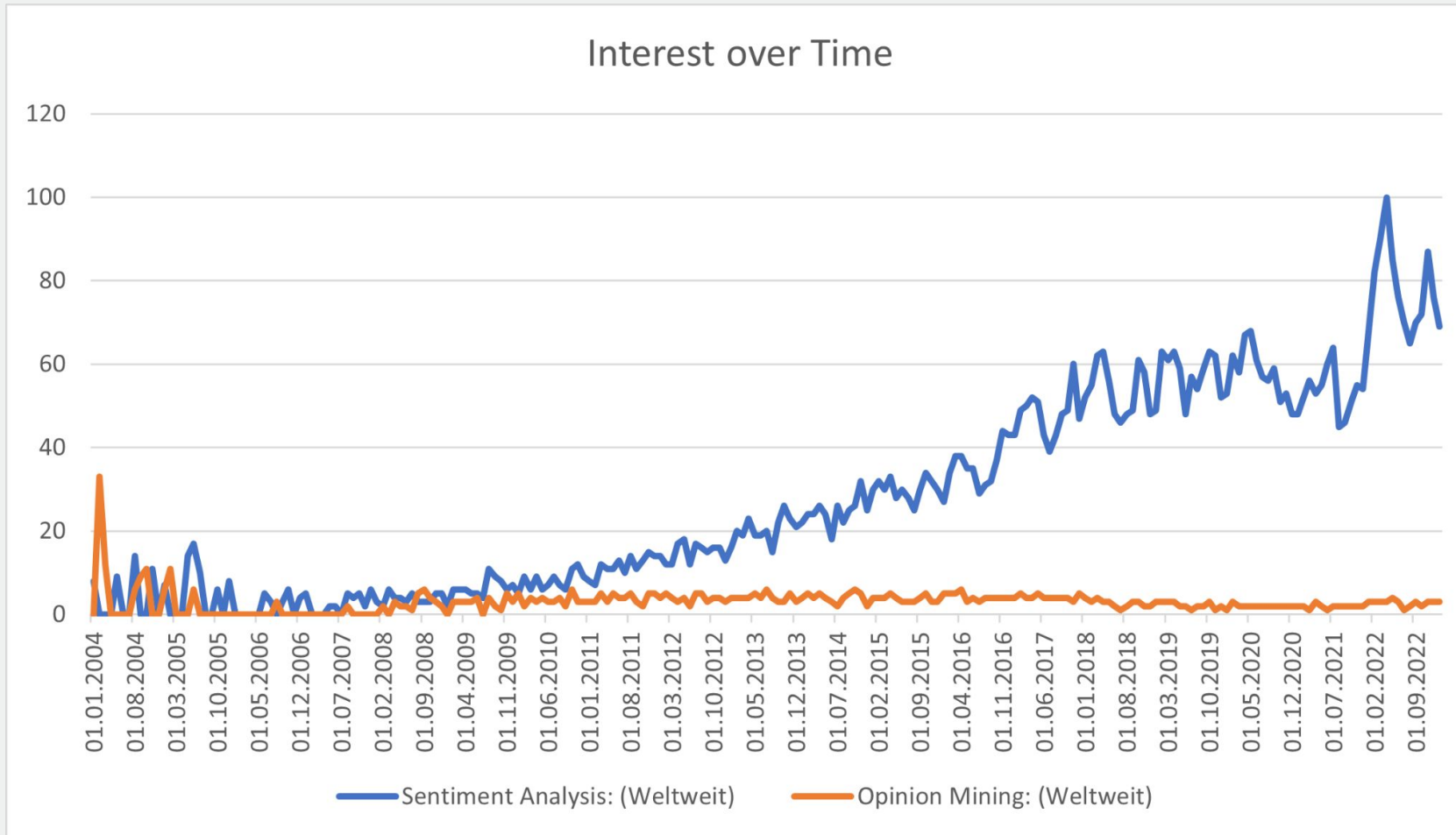
*by analysing their corresponding sentiment

Introduction to Sentiment Analysis

“Sentiment analysis (SA), also called Opinion Mining (OM) is the task of extracting and analyzing people’s opinions, sentiments, attitudes, perceptions, etc., toward different entities such as topics, products, and services.” [1]

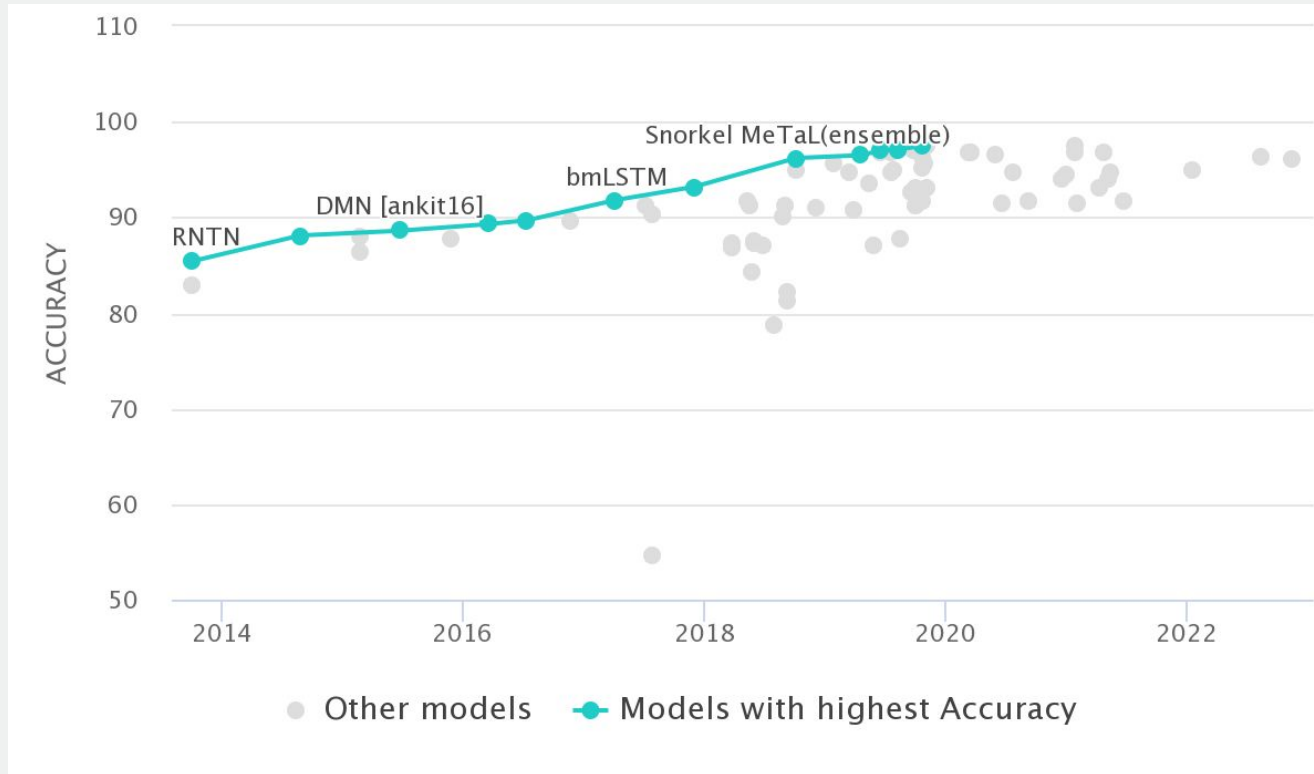
[1] Marouane Birjali, Mohammed Kasri, & Abderrahim Beni-Hssane (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.

Still Interesting?



According to Google Trends (<https://trends.google.com/trends/>)

Towards the State of the Art



Sentiment Analysis on SST-2 Binary classification
(<https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary>)

Dataset: SST ([Stanford Sentiment Treebank](#)) = benchmark dataset

- 8 out of the top 10 performing algorithms are transformer based approaches (e.g. BERT)
- Previous approaches often based on CNN / LSTM
- Early approaches were often based on dictionaries

Applications of Sentiment Analysis

- User reviews (products, movies, music, ...)
- News domain (comments section, forum, ...)
- Analysis of user generated content (social media e.g. Twitter)
- ...

The results can be used to create recommendations for users or to analyse public opinion on a specific event (COVID-19, elections, ...).

Pre-Processing Based on the Example of *fastText*

Pre-processing always depends on the planned application and the method used!

“FastText is an open-source, free, lightweight library that allows users to learn text representations and text classifiers. It works on standard, generic hardware. Models can later be reduced in size to even fit on mobile devices.” [2]

[2] <https://fasttext.cc/>

Pre-Processing Based on the Example of *fastText* (cont.)

- *Data Cleanup (punctuation, upper to lowercase, emoticons, stopwords, spell checking, ...)*
- *Splitting up the data set into train-, test- and validation data set*
- *epochs, learning rate*
- *word n-grams*

<https://fasttext.cc/docs/en/supervised-tutorial.html#making-the-model-better>

Pre-Processing Based on the Example of *fastText* (cont.)

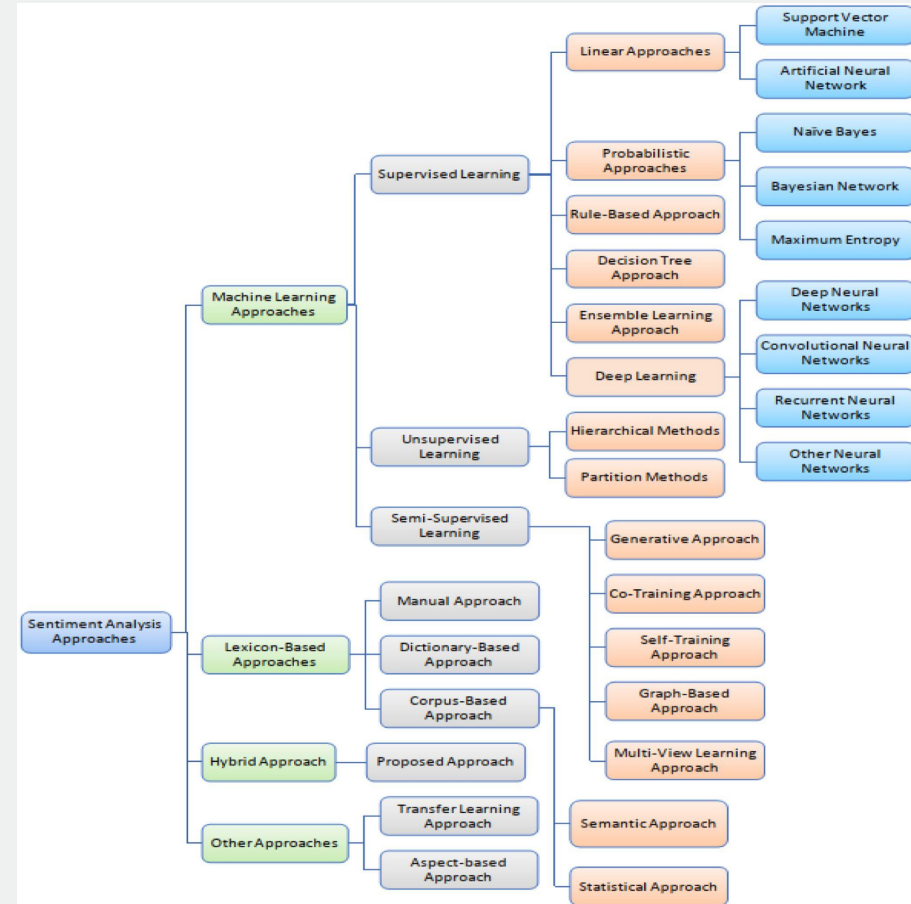
Frameworks

- Natural Language Toolkit (NLTK) (<https://www.nltk.org/>)
- spaCy (<https://spacy.io/>)

But often basic approaches like “sed”, “awk”, “sort” (= Linux packages) can help a lot if the data set is very big.

Sentiment Analysis Techniques

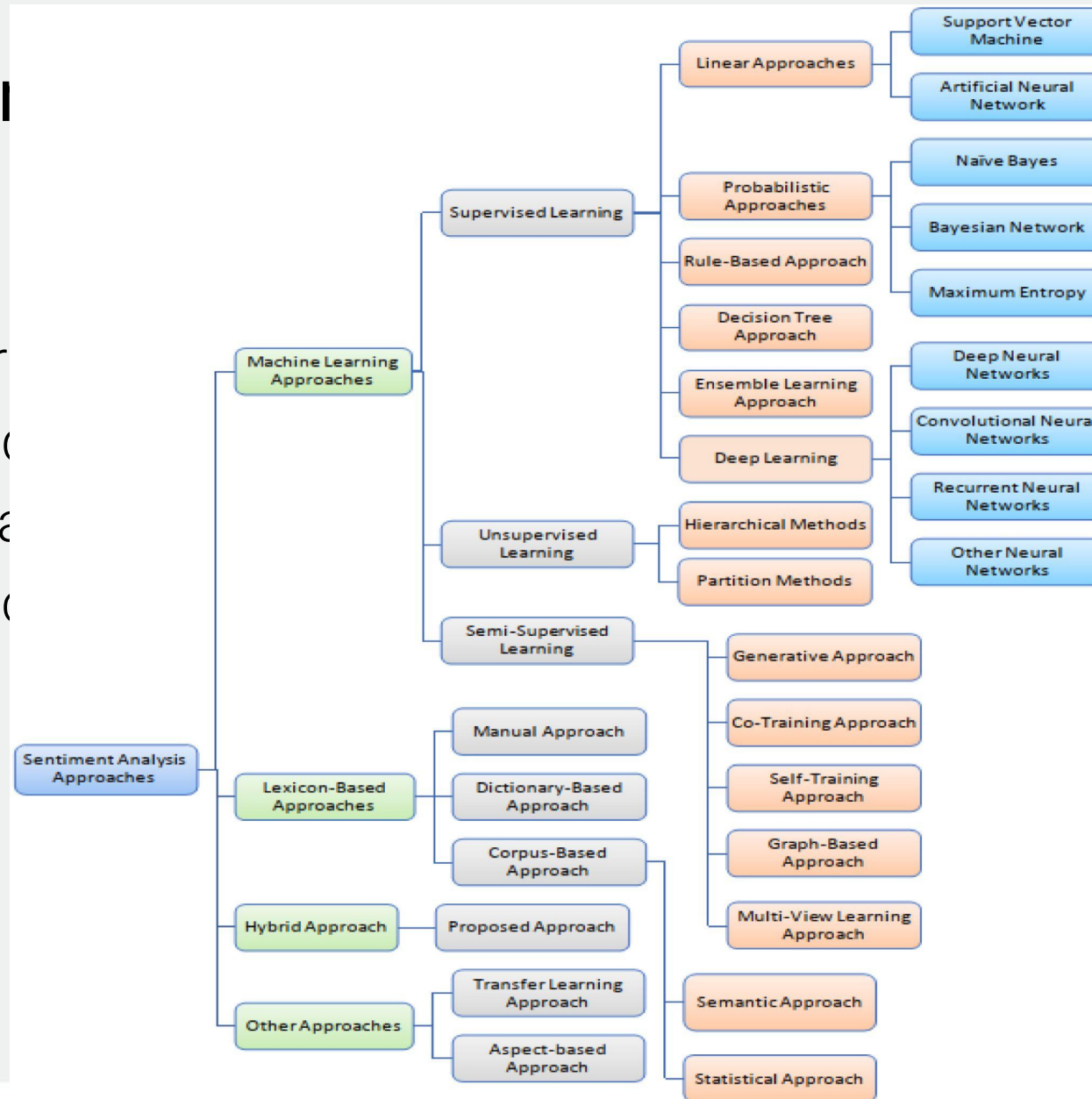
- Machine Learning Approaches
- Lexicon Based Approaches
- Hybrid Approaches
- Other Approaches



Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.
<https://doi.org/https://doi.org/10.1016/j.knosys.2021.107134>

Sentiment Analysis

- Machine Learning
- Lexicon Based
- Hybrid Approaches
- Other Approaches



Comprehensive survey on sentiment analysis:
Systems, 226, 107134.
107134

Levels of Sentiment Analysis

- Document level
- Sentence level
- Phrase level (e.g. aspect based sentiment analysis)

Phrase level extraction often requires named entity recognition to get a specific phrase around a target e.g. a politician name.

Which level to use is always domain and task dependent!

P. Balaji, O. Nagaraju and D. Haritha, "Levels of sentiment analysis and its challenges: A literature review," *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDACI)*, Chirala, Andhra Pradesh, India, 2017, pp. 436-439, doi: 10.1109/ICBDACI.2017.8070879.

Evaluation Metrics

Accuracy

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

Precision

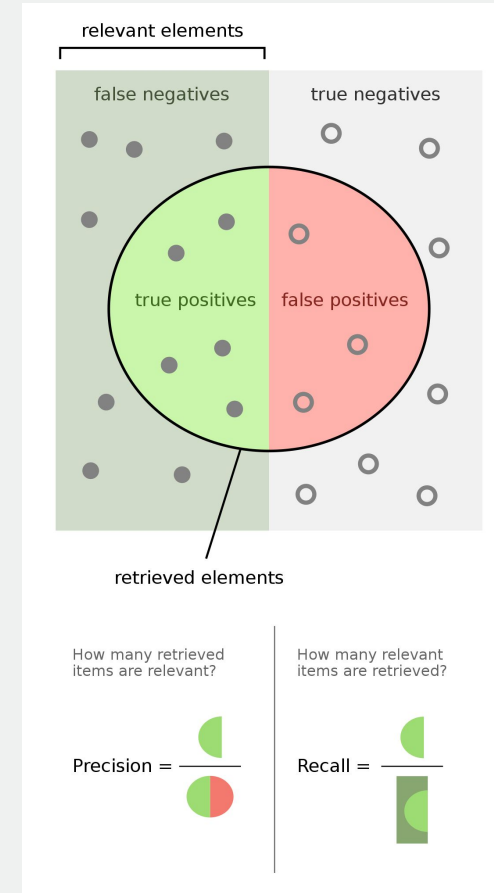
$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$



Walber, CC BY-SA 4.0, via Wikimedia Commons

Saxena, A., Reddy, H., Saxena, P. (2022). Introduction to Sentiment Analysis Covering Basics, Tools, Evaluation Metrics, Challenges, and Applications. In: Biswas, A., Patgiri, R., Biswas, B. (eds) Principles of Social Networking. Smart Innovation, Systems and Technologies, vol 246. Springer, Singapore.
https://doi.org/10.1007/978-981-16-3398-0_12

Evaluation Metrics (cont.)

There are many more relevant metrics in this area e.g.:

- Cohen's Kappa: measure inter-rater reliability (two raters)
- Fleiss Kappa: measure inter-rater reliability (any number of raters)
- ...

https://en.wikipedia.org/wiki/Fleiss%27_kappa

https://en.wikipedia.org/wiki/Cohen%27s_kappa

Application in Research

DYSEN Project

Dynamic Sentiment Analysis as Emotional Compass for the Digital Media Landscape



RQ: 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



TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

ÖAW

AUSTRIAN
ACADEMY OF
SCIENCES



universität
wien

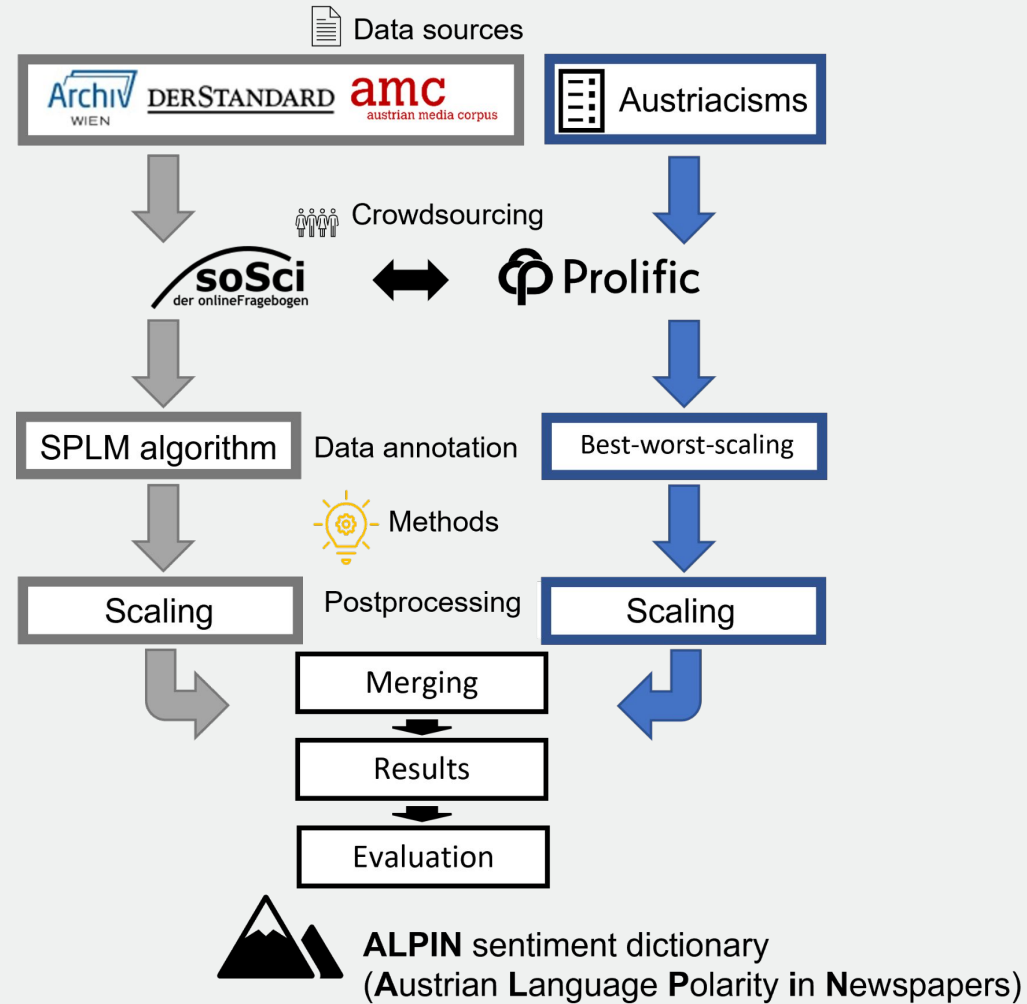
Funded by (DigHum Call):



Stadt
Wien

Grant number:
MA7-737909/19

DYSEN Project (cont.)



Data Collection & Preprocessing



Linguistic annotated Austrian Media Corpus (Ransmayr et al., 2017); contains around 45 million articles by covering the print media landscape of Austria

Extraction criteria:

- National & regional print media related to Vienna between 1996 and 2017
- Text areas limited to area extraction around politician names with around 60 tokens (legal limitation / copyright of the amc corpora)



Politician archive of Vienna (*POLAR*) of the Vienna City and State Archives¹

- Politicians which were active between the 13th and 20th parliamentary term (1983 to 2020)

AMC and POLAR are combined to extract text areas around Viennese politicians

¹<https://polar.wien.at>

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 ($\geq 75\%$ correct)
- Two annotation runs (1st 70 annotators; Fleiss-Kappa: 0,295,
2nd 88 annotators; Fleiss-Kappa: 0,283)

Restricted annotators by:

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

Data Collection & Preprocessing

DERSTANDARD

1 Million Posts Corpus (Schabus et al., 2017)

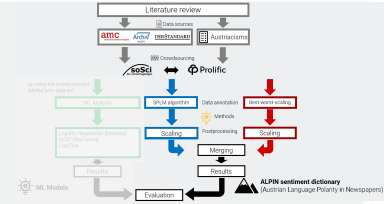
- Posts from 2015 to 2016
- 3599 posts with sentiment annotated by employed forum annotators

Austriacisms

Based on:

- „Variantenwörterbuch des Deutschen“ (VWB; words specific to Austria) (Ammon et al., 2016)
- Austriacism list of Wikipedia¹
- 1600 words checked by the whole project team

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



Crowd Sourcing: Austriacisms

Preselection survey

- 1600 words
- Quality control ($\geq 75\%$ correct)
- Four options (positive, neutral, negative, unknown)

Main survey

- Best-worst-scaling (BWS) method
- 1074 tuples
- Quality control ($\geq 75\%$ correct)
- 34 annotators

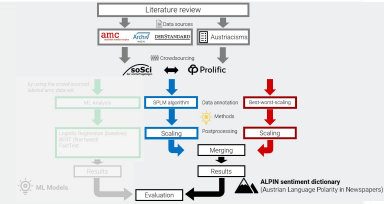
Restricted annotators by:

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

	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	Halbpreisspass	versumpfern	Ermäßigungsausweis	versumpfern
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

Labeled dataset after main survey



Methods (Dictionary Based)

...for generating sentiment scores

SPLM method

(Almatarneh & Gamallo, 2018)

Used for: **amc** austrian media corpus & **DERSTANDARD** datasets

... based on a labelled (positive, neutral, negative) dataset

Best-Worst-Scaling (BWS) method

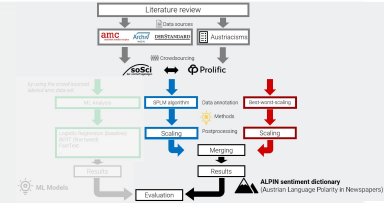
(Kiritchenko & Mohammad, 2017)

Used for: Austriacism list

... based on tuple pairs (best, worst) of words

Spearman correlation:

0.9159 (+/- 0,0051) by applying split-half reliability



Sentiment score word lists based on...

... AMC & standard posts (**SPLM**)

	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

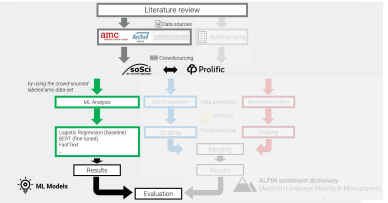
8929 rows × 4 columns

... Austriacisms (**BWS**)

	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

Evaluation



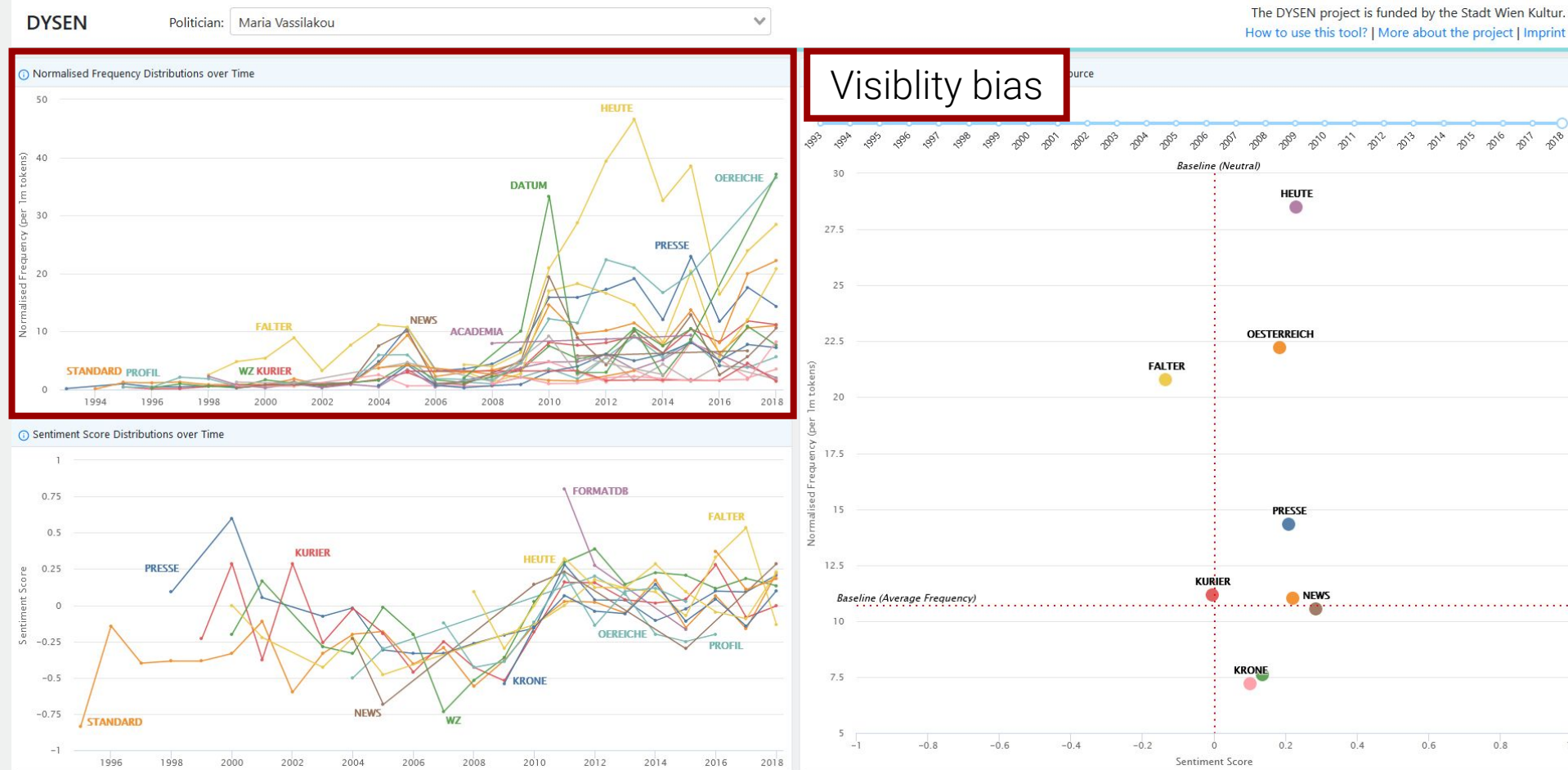
Model	Accuracy	Precision	Recall	F1
DummyClassifier v1 (stratified)	0,52	0,54	0,51	0,52
DummyClassifier v2 (uniformly gen. pred.)	0,52	0,54	0,57	0,56
BERT (dbmdz/bert-base-german-cased) Finetuned with the AMC dataset	0,78	0,82	0,76	0,79
ALPIN (dictionary based approach)	0.70	0.74	0.70	0.72

preprocessing:

- Model specific (stratified k-fold, train/test/validation)
- Different tokenization requirements
- Encoding
- ...

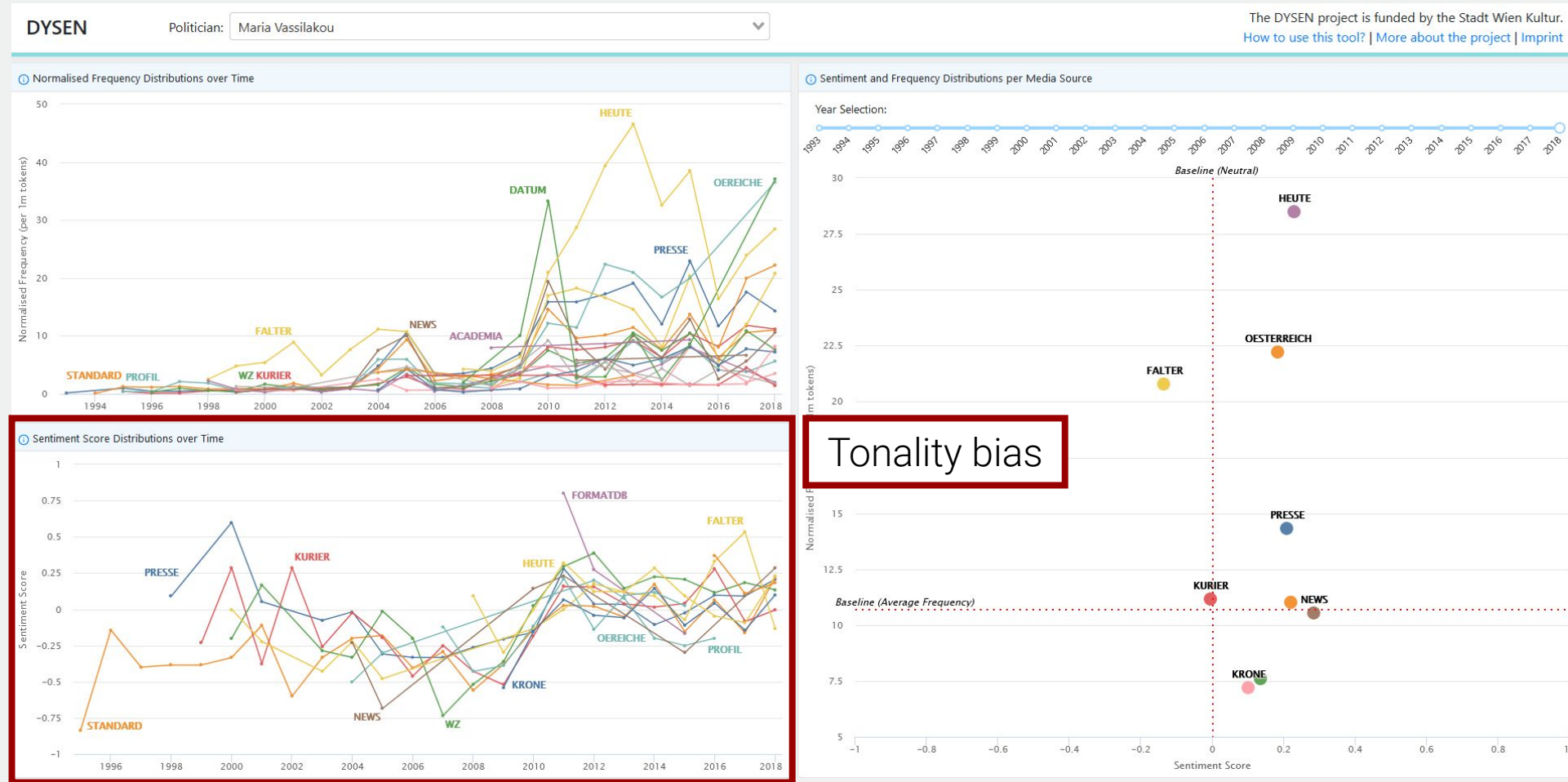
Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



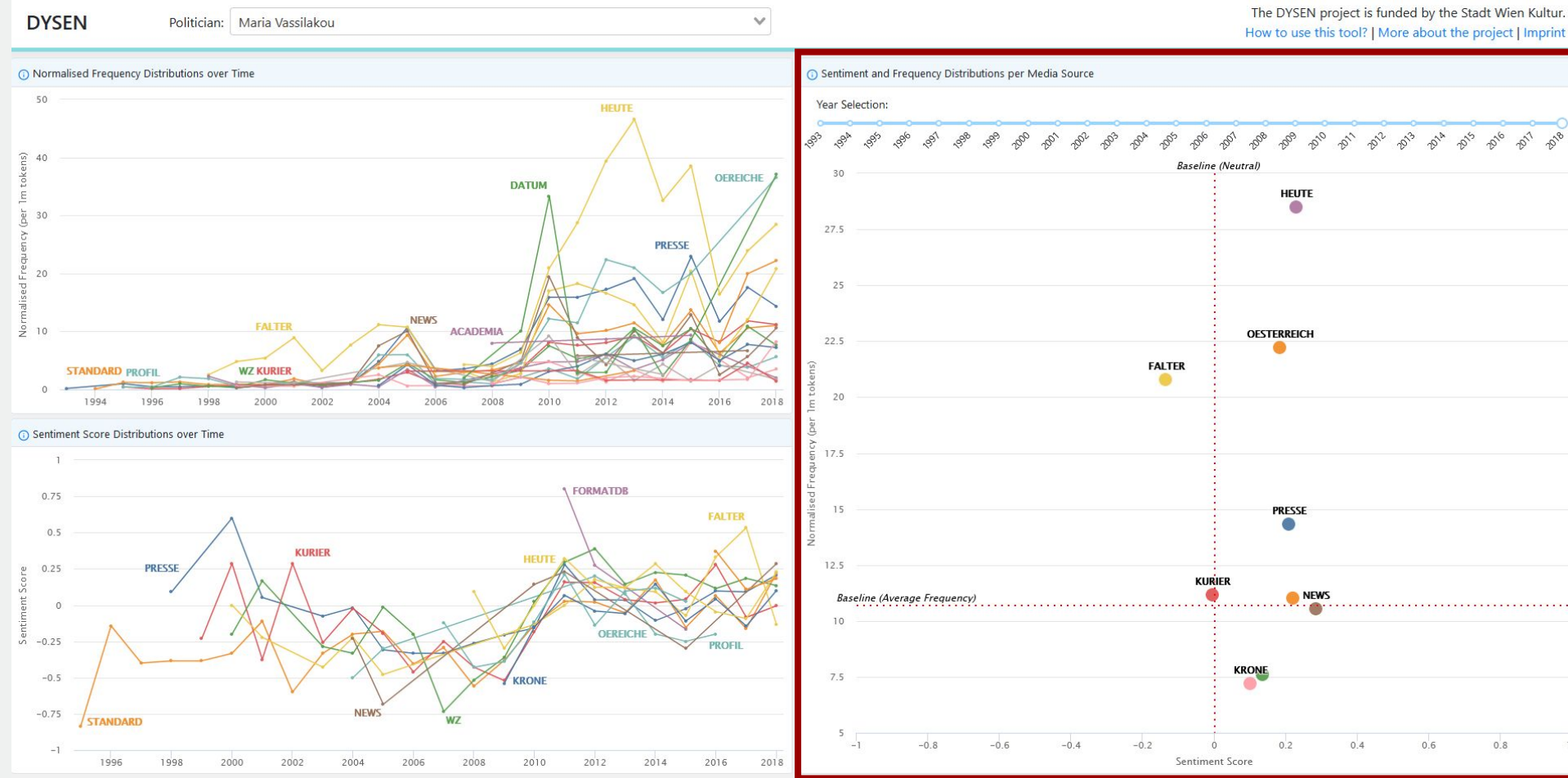
Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



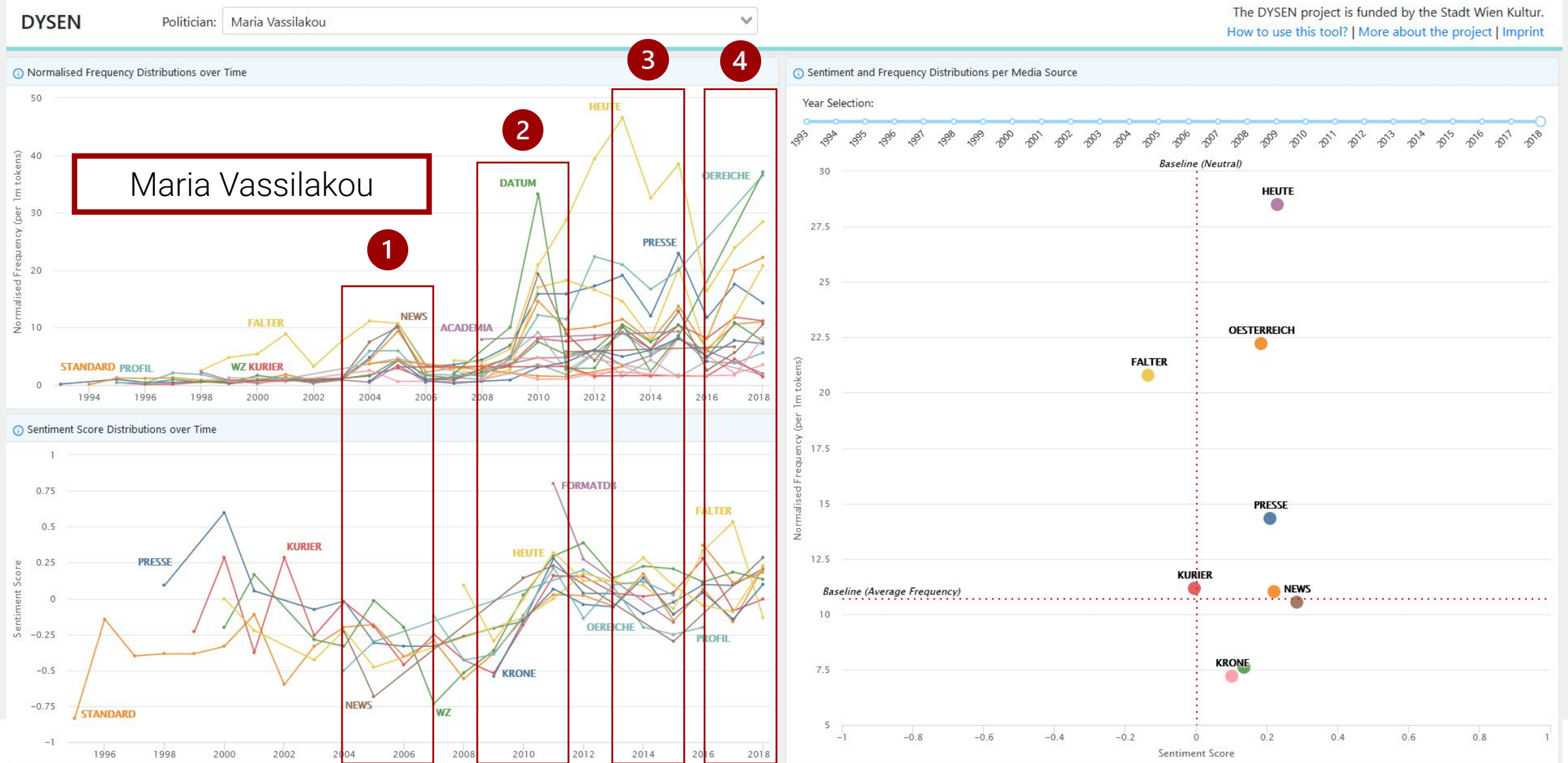
Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



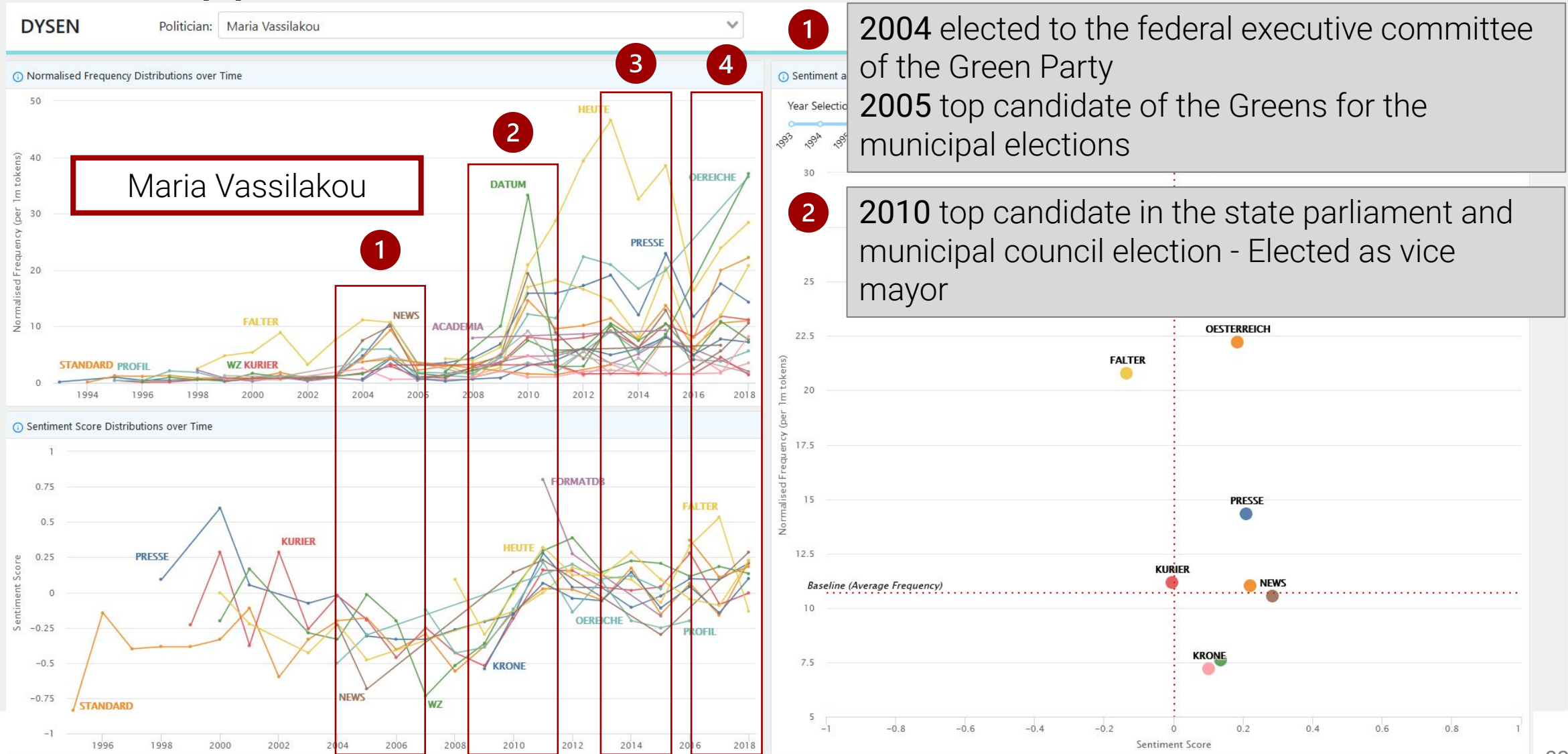
Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



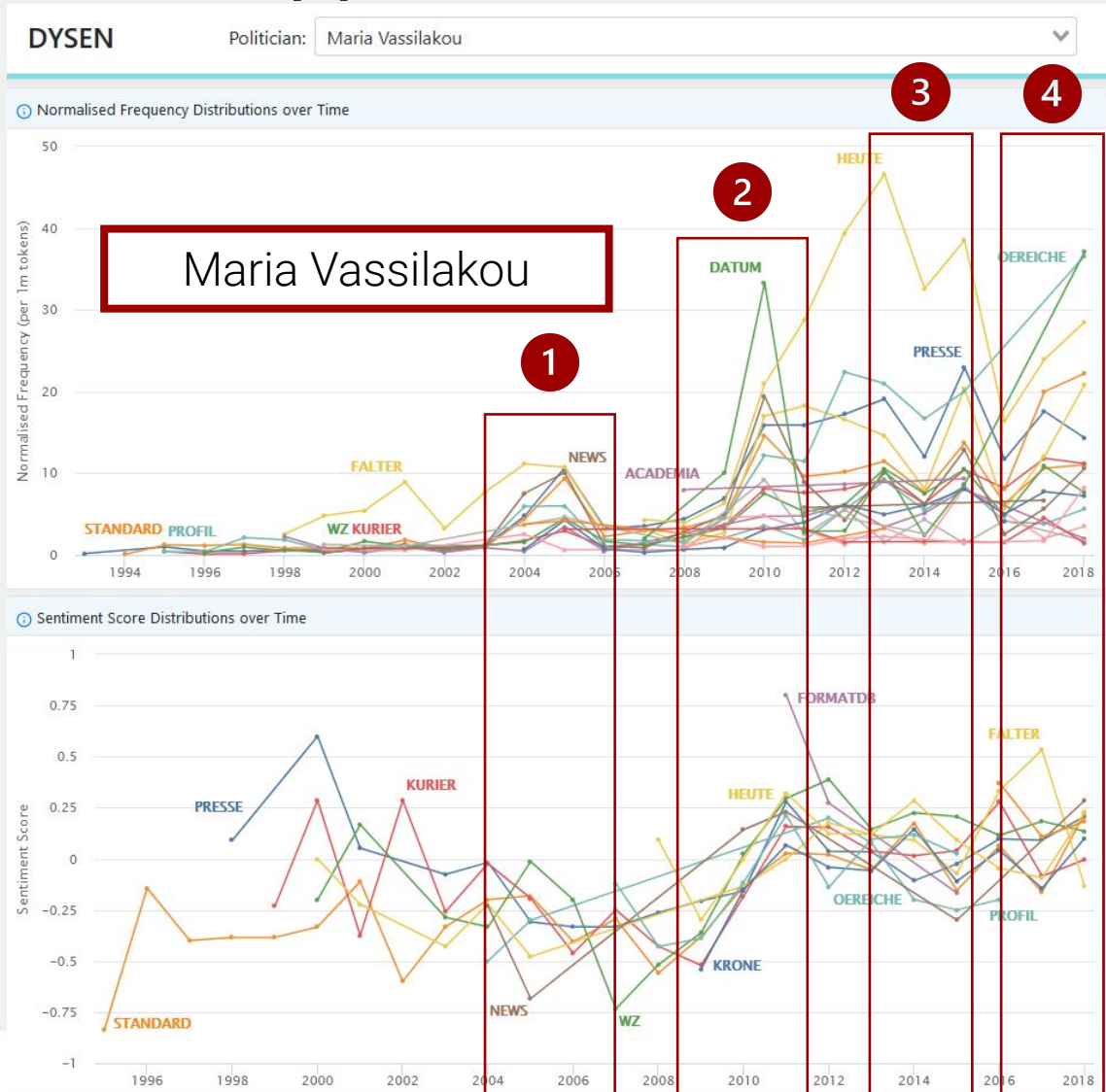
Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



1 2004 elected to the federal executive committee of the Green Party

Sentiment a

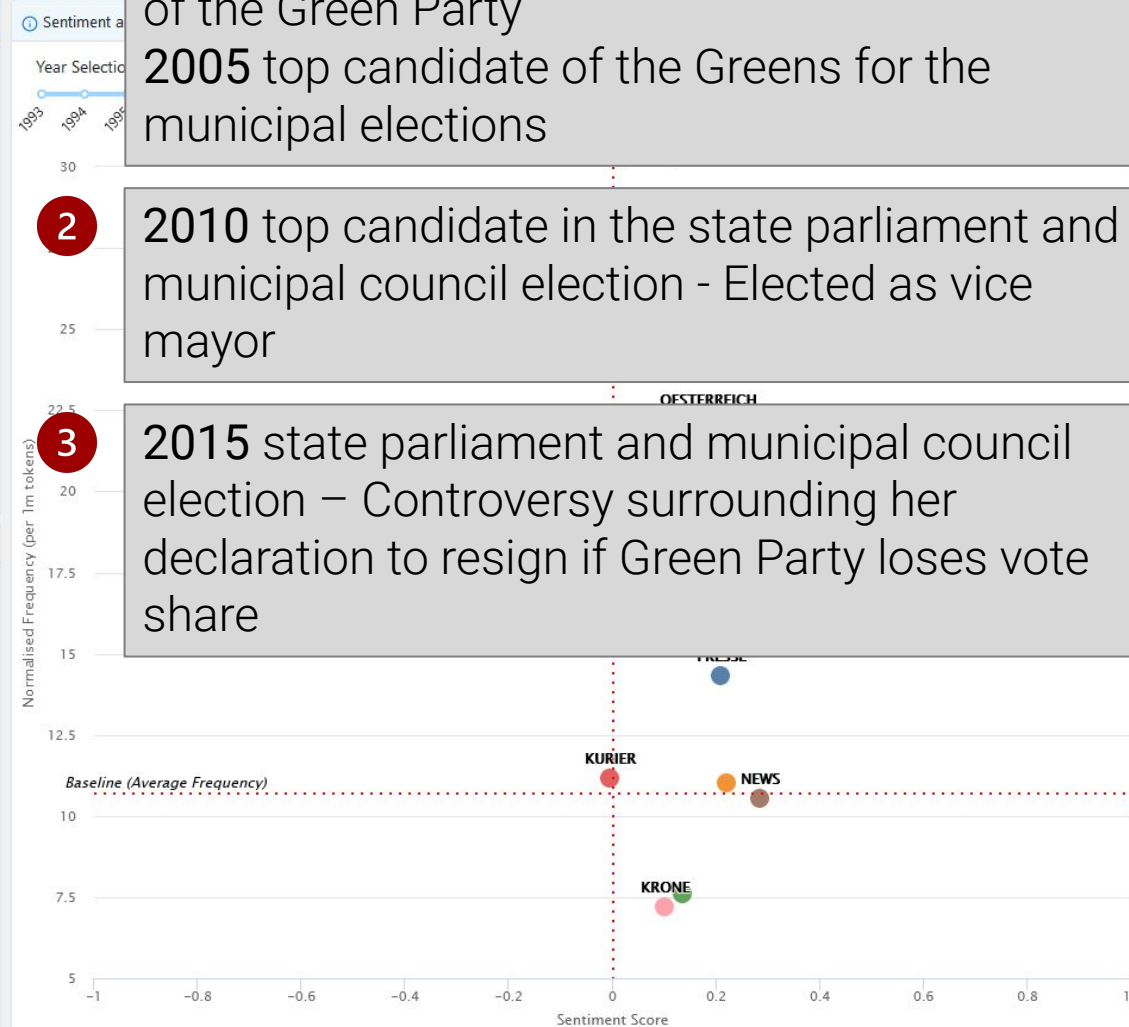
Year Selection

1993 1994 1995

2005 top candidate of the Greens for the municipal elections

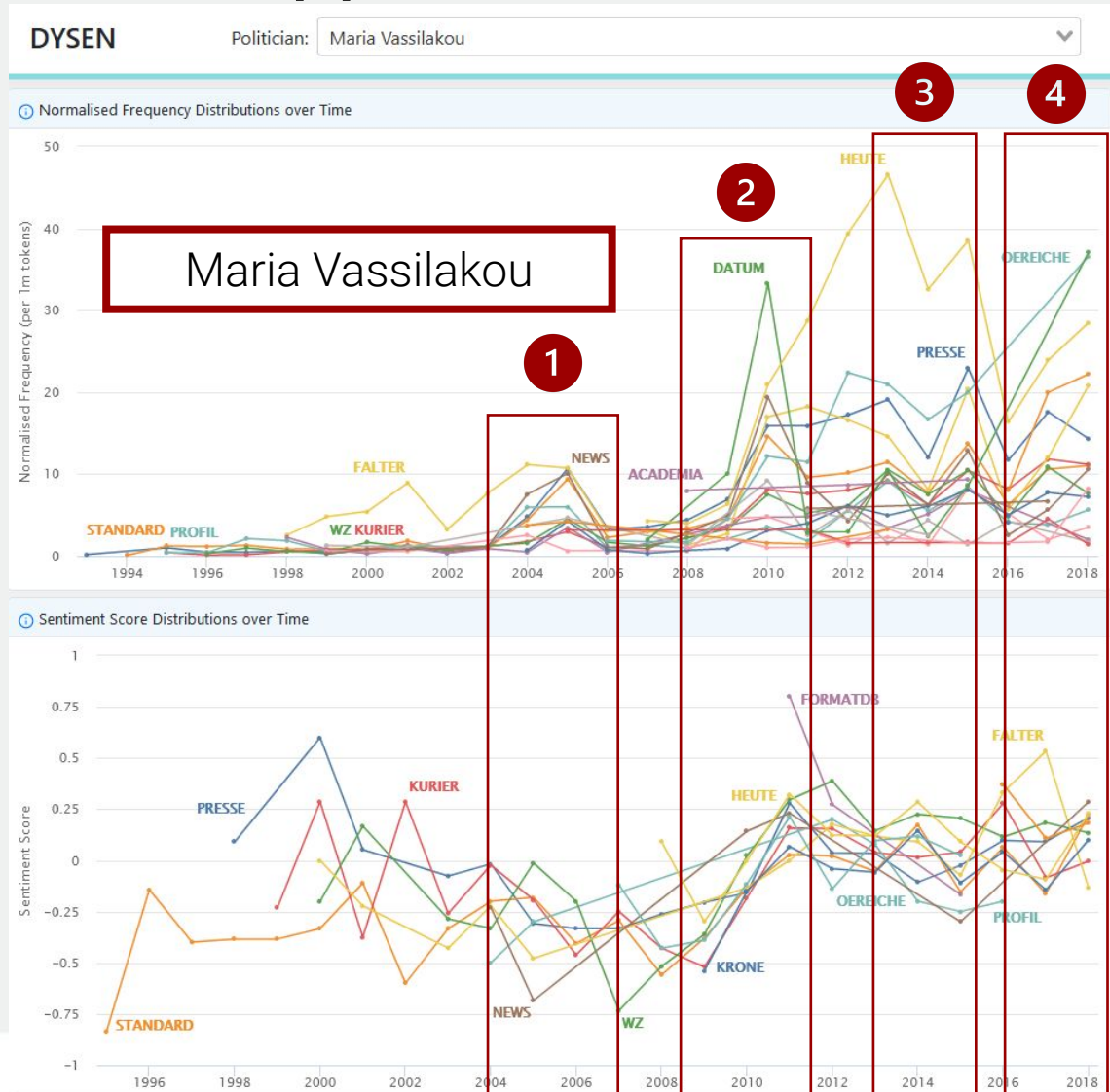
2 2010 top candidate in the state parliament and municipal council election - Elected as vice mayor

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



Web Application¹

¹data based on a master-thesis (Kolb, 2022), tool developed by the team of ÖAW; available at: <https://dysen-tool.acdh.oeaw.ac.at/>



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

The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers

Kolb, T. E., Kern, B. M., Sekanina, K., Wissik, T., Neidhardt, J., Baumann, A.,
(2022) The ALPIN Sentiment Dictionary: Austrian Language Polarity in
Newspapers.

<https://zenodo.org/record/5857151>

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

The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers

Thomas E. Kolb¹, Katharina Sekanina², Bettina M. J. Kern³, Julia Neidhardt¹, Tanja Wissik³, Andreas Baumann³

What Is This All About?
This publication is part of the **DYEN Project** which stands for *Dynamic Sentiment Analysis as Emotional Compass for the Digital Media Landscape*.
Research Question: How do print media report about Viennese Politicians?
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:
List of all Viennese politicians in the last 100 years of the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 22nd, 23rd, 24th, 25th, 26th, 27th, 28th, 29th, 30th, 31st, 32nd, 33rd, 34th, 35th, 36th, 37th, 38th, 39th, 40th, 41st, 42nd, 43rd, 44th, 45th, 46th, 47th, 48th, 49th, 50th, 51st, 52nd, 53rd, 54th, 55th, 56th, 57th, 58th, 59th, 60th, 61st, 62nd, 63rd, 64th, 65th, 66th, 67th, 68th, 69th, 70th, 71st, 72nd, 73rd, 74th, 75th, 76th, 77th, 78th, 79th, 80th, 81st, 82nd, 83rd, 84th, 85th, 86th, 87th, 88th, 89th, 90th, 91st, 92nd, 93rd, 94th, 95th, 96th, 97th, 98th, 99th, 100th.
Standard Posts (STP):
List of all standard posts in the last 100 years of the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 22nd, 23rd, 24th, 25th, 26th, 27th, 28th, 29th, 30th, 31st, 32nd, 33rd, 34th, 35th, 36th, 37th, 38th, 39th, 40th, 41st, 42nd, 43rd, 44th, 45th, 46th, 47th, 48th, 49th, 50th, 51st, 52nd, 53rd, 54th, 55th, 56th, 57th, 58th, 59th, 60th, 61st, 62nd, 63rd, 64th, 65th, 66th, 67th, 68th, 69th, 70th, 71st, 72nd, 73rd, 74th, 75th, 76th, 77th, 78th, 79th, 80th, 81st, 82nd, 83rd, 84th, 85th, 86th, 87th, 88th, 89th, 90th, 91st, 92nd, 93rd, 94th, 95th, 96th, 97th, 98th, 99th, 100th.
Austracisms:
List of all austracisms in the last 100 years of the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 22nd, 23rd, 24th, 25th, 26th, 27th, 28th, 29th, 30th, 31st, 32nd, 33rd, 34th, 35th, 36th, 37th, 38th, 39th, 40th, 41st, 42nd, 43rd, 44th, 45th, 46th, 47th, 48th, 49th, 50th, 51st, 52nd, 53rd, 54th, 55th, 56th, 57th, 58th, 59th, 60th, 61st, 62nd, 63rd, 64th, 65th, 66th, 67th, 68th, 69th, 70th, 71st, 72nd, 73rd, 74th, 75th, 76th, 77th, 78th, 79th, 80th, 81st, 82nd, 83rd, 84th, 85th, 86th, 87th, 88th, 89th, 90th, 91st, 92nd, 93rd, 94th, 95th, 96th, 97th, 98th, 99th, 100th.

Methodology:
The methodology section describes the process of creating the ALPIN sentiment dictionary. It starts with the data sources and the standard posts (STP). The data is then processed through a series of steps: data cleaning, data normalization, data annotation, and data evaluation. The final step is the creation of the ALPIN sentiment dictionary, which is a tool that can detect change of emotional polarization of politicians in Austrian Newspapers.

Post-Processing:
The post-processing section describes the steps taken to refine the dictionary. This includes removing words that are not relevant to the domain of news media and politics, and adding words that are missing. The final result is a refined dictionary that is ready for use.

Web Application:
The web application section describes the tool that was developed to make the dictionary accessible to users. It allows users to search for words and see their sentiment scores. It also provides a visual representation of the data, showing the distribution of sentiment scores for different words.

Results:
The results section presents the findings of the study. It shows that the dictionary is able to detect change of emotional polarization of politicians in Austrian Newspapers. It also shows that the dictionary is able to identify words that are relevant to the domain of news media and politics.

Evaluation:
The evaluation section describes the steps taken to assess the quality of the dictionary. This includes comparing the dictionary to other sentiment dictionaries and to human judgments. The results show that the dictionary performs well in terms of accuracy and reliability.

Discussion:
The discussion section discusses the implications of the study. It highlights the importance of having a sentiment dictionary for the domain of news media and politics. It also discusses the limitations of the study and suggests areas for future research.

Future Work:
The future work section outlines the plans for extending the dictionary. This includes adding more words and improving the accuracy of the sentiment scores.

References:
The references section lists the sources used in the study. These include academic papers, books, and online resources.

Contact:
The contact section provides information about the authors and how to reach them. It includes email addresses and a website.

Data:
The data section provides a link to the ALPIN sentiment dictionary. It also includes a QR code that can be used to access the dictionary.

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

Logos:
The logos of the funding organizations and the institutions involved in the project are displayed at the bottom of the page. These include the Technische Universität Wien, the Österreichische Akademie der Wissenschaften, and the Austrian Research Promotion Agency.

Ongoing Work

Detection of sentiment in user comments of a large Austrian news company to ...

... identify the change of positivity or negativity within the commentary section over time.

... investigation potential polarization over time towards a certain topic (e.g. the topic "Coronavirus")

Beyond Sentiment Analysis: What's next?

Detection of ...

... emotions

... hate speech

... toxicity

... populism

Many very recent research topics in this field can be found here:

<https://semeval.github.io/>

Summary?

- Introduction to sentiment analysis
- Applications and challenges of sentiment analysis
- Sentiment analysis techniques
- Evaluation metrics
- Applications in research
- Future topics



Dipl.-Ing. Thomas E. Kolb
Research Unit of Data Science
thomas.kolb@tuwien.ac.at
<https://recsys-lab.at>