Emotionen & Entscheidung

Sentiment & Topic Analyse
der Bundestagsdebatten zur Pandemie

Karlsruher Institut für Technologie

Paula Hofmann
paulahofmann@icloud.com

EINLEITUNG

Sentiment Analysen können eine zentrale Rolle bei der Analyse politischer Reden & Kommunikationen spielen, indem sie versuchen die Einstellung und Reaktion auf politische Themen und Akteure zu quantifizieren.

Im Zusammenhang mit der Covid-Pandemie ermöglicht Sie Trends, Stimmungs-änderungen und politische Reaktionen auf die Krise zu verfolgen und kann dazu beitragen, evidenzbasierte Entscheidungen und politische Maßnahmen im Kampf gegen die Pandemie zu evaluieren.[1,2]

Gibt es einen Zusammenhang zwischen der **7-Tages-Inzidenz** und dem **Sentiment im Bundestag?**

Welche **Themen** sind dabei relevant?

METHODIK

1. **Plenarprotokolle der 19. & 20. Wahlperiode**
   - Filterung der Reden (zeitlich + inhaltlich bezogen auf „Covid-19“)
2. **Datensatz des RKI**
   - Download & filtern
3. **Transformer Modell: „chkla/parlbert-topic-german“[4]**
   - Welche Themen waren an den Peaks besonders relevant?

Transformer Modell: „german-sentiment-bert“[3]

SENTIMENT ANALYSE

auf Satzebene

Dataframes der 19. & 20. Wahlperiode inkl. Sentiment

Dataframes zusammenführen

7-Tages-Inzidenz zu den Fallzahlen in DE

KORRELATIONS ANALYSE

TOPIC ANALYSE

Täglicher Sentimentscore pro PolitikerIn

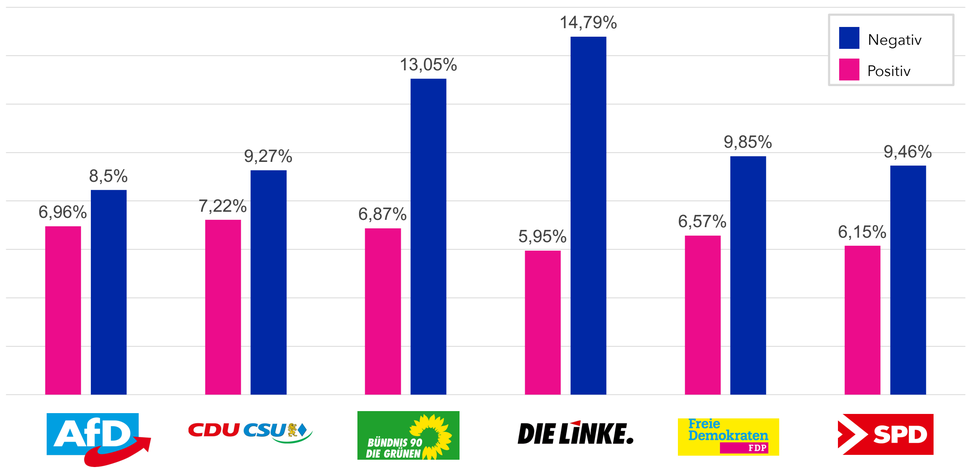
Berechnung durch den Ø Sentiment Wert über alle Sätze eines Politikers/Politikerin an einem bestimmten Tag.

ERGEBNISSE

SENTIMENTANALYSE:

Es zeigt sich innerhalb der Parteien marginale Unterschiede innerhalb der Verteilung des Sentiments über die 19. & 20. Wahlperiode, wobei die Linke und die Grünen am häufigsten über ein negatives Sentiment verfügen.

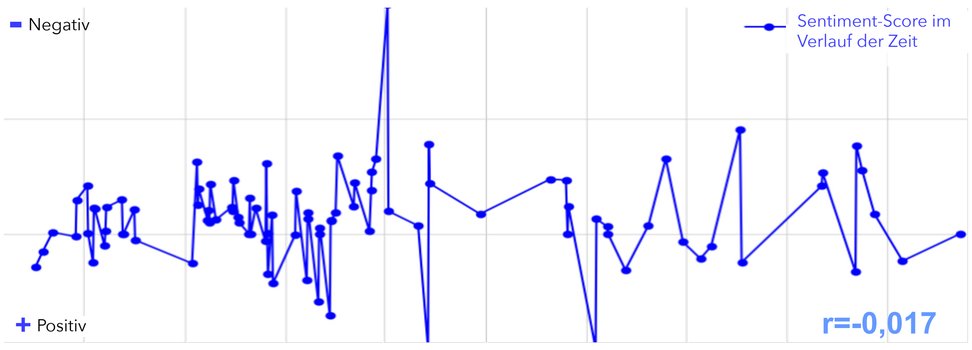
Fig.1: Sentiment Verteilung im Zeitraum Jan.2020-Mai.2023



KORRELATIONSANALYSE:

Es zeigt sich kein linearer Zusammenhang zwischen der 7-Tages Inzidenz und dem Ø Sentiment Score über alle Parteien pro Tag.

Fig.2: Ø Sentiment-Score im Zeitraum Jan.2020-Mai.2023

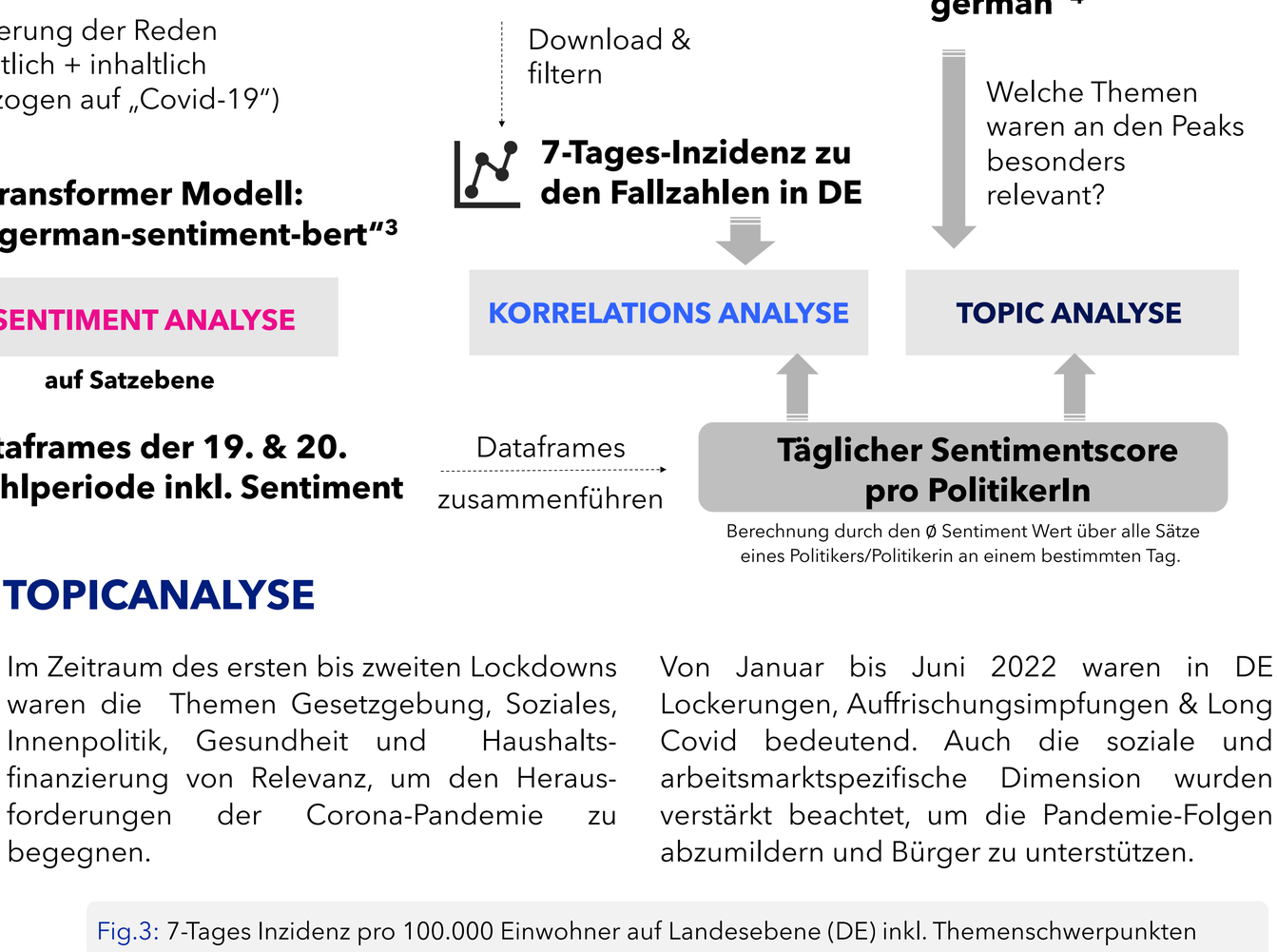


TOPICANALYSE

Im Zeitraum des ersten bis zweiten Lockdowns waren die Themen Gesetzgebung, Soziales, Innenpolitik, Gesundheit und Haushalts-finanzierung von Relevanz, um den Heraus-forderungen der Corona-Pandemie zu begegnen.

Von Januar bis Juni 2022 waren in DE Lockerungen, Auffrischungsimpfungen & Long Covid bedeutend. Auch die soziale und arbeitsmarktspezifische Dimension wurden verstärkt beachtet, um die Pandemie-Folgen abzumildern und Bürger zu unterstützen.

Fig.3: 7-Tages Inzidenz pro 100.000 Einwohner auf Landesebene (DE) inkl. Themenschwerpunkten



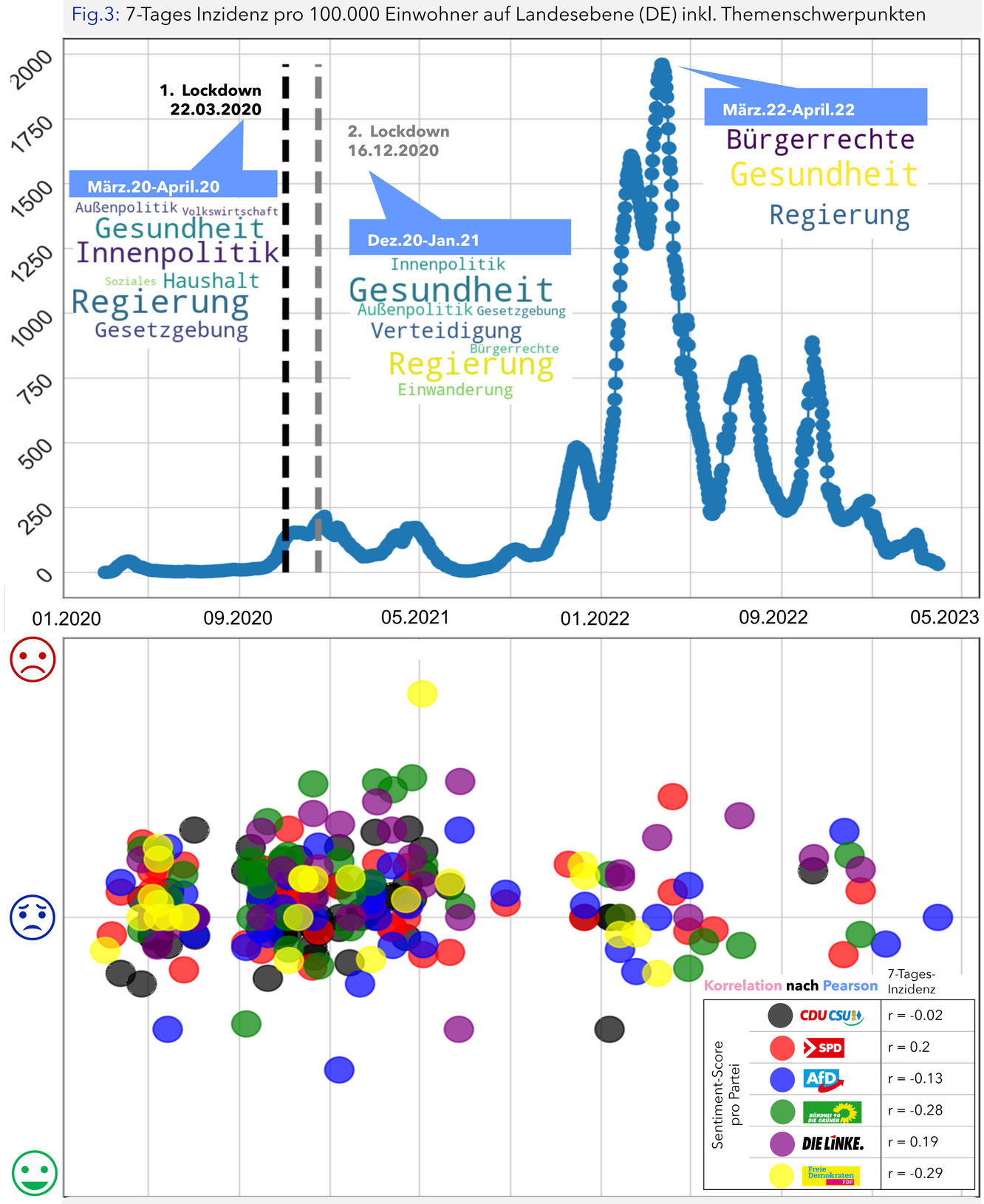


Fig.4: Ø Sentiment der jeweiligen Parteien inkl. den Korrelationswerten zwischen der 7 Tages-Inzidenz pro Tag und dem Sentiment-Score der jeweiligen Parteien

FAZIT

Die Sentiment-Analyse im Bundestag während der COVID-19-Pandemie ergab, dass kein linearer Zusammenhang existiert und unterschiedliche Themen von Bedeutung sind.

Das Sentiment schwankt im Zeitverlauf und innerhalb der Parteien aufgrund wechselnder Schwerpunktthemen und Debatten. Das Fehlen eines klaren Zusammenhangs zwischen der 7-Tages-Inzidenz und der Bundestagsstimmung betont die Komplexität der Situation.

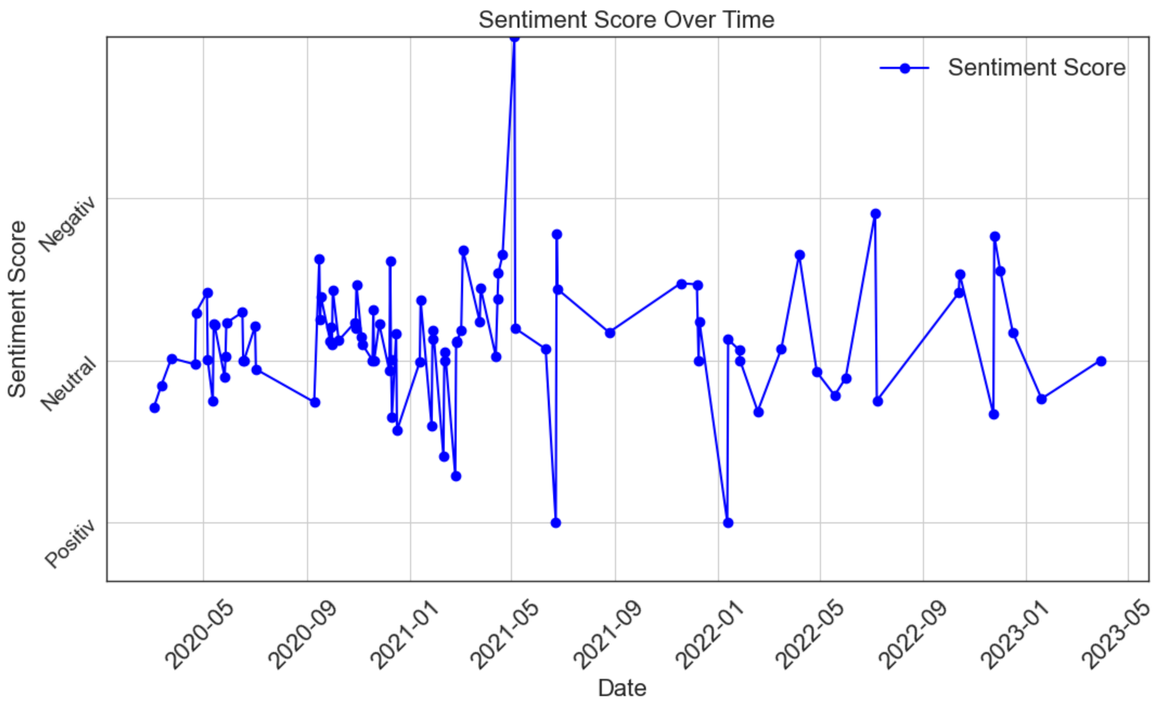
Referenzen:
1. Ragothaman, A., & Huang, C. Y. (2021). Sentiment analysis on Covid-19 twitter data. International Journal of Computer Theory and Engineering, 13(4), 100-107.
2. Khan, Rijwan, et al. "Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data." J. Crit. Rev 7.9 (2020): 2761-2774.
3. Guhr, Oliver, et al. "Training a broad-coverage German sentiment classification model for dialog systems." Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.
4. Klamm, Christopher, Ines Rehbein, and Simone Paolo Ponzetto. "FrameASt: A framework for second-level agenda setting in parliamentary debates through the lens of comparative agenda topics." (2022): 92-100.

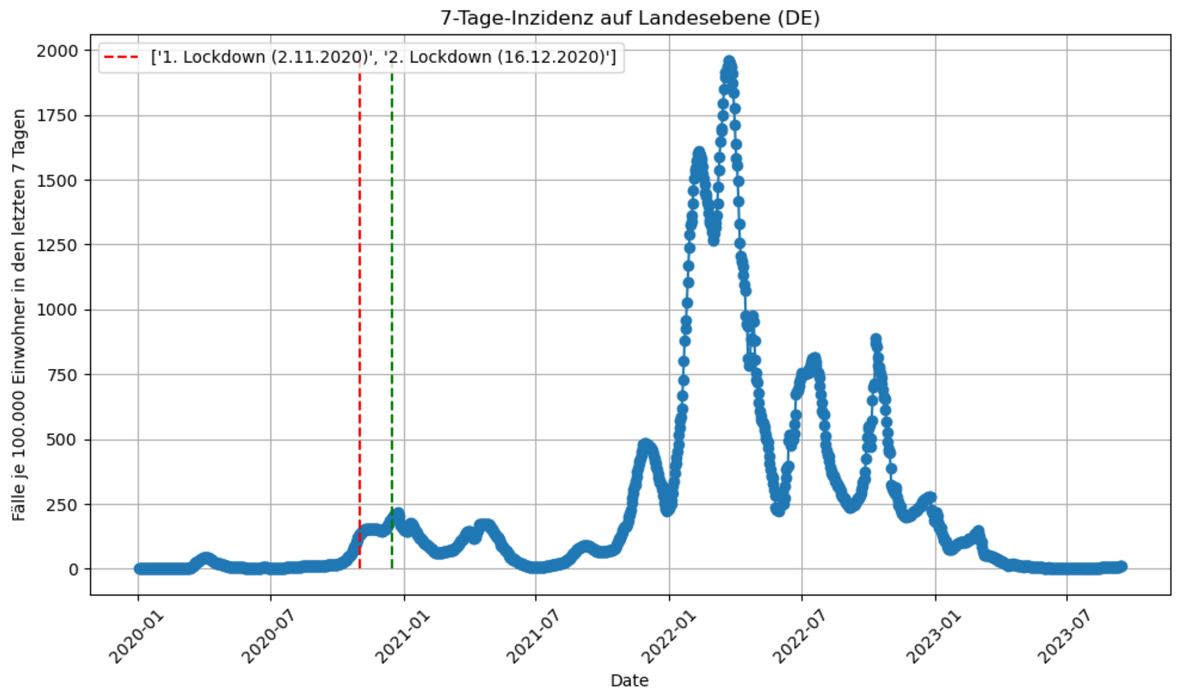
Seminar: Politische Debatten und Polarisierung im Bundestag | Sven Banisch

Emotionen & Entscheidung

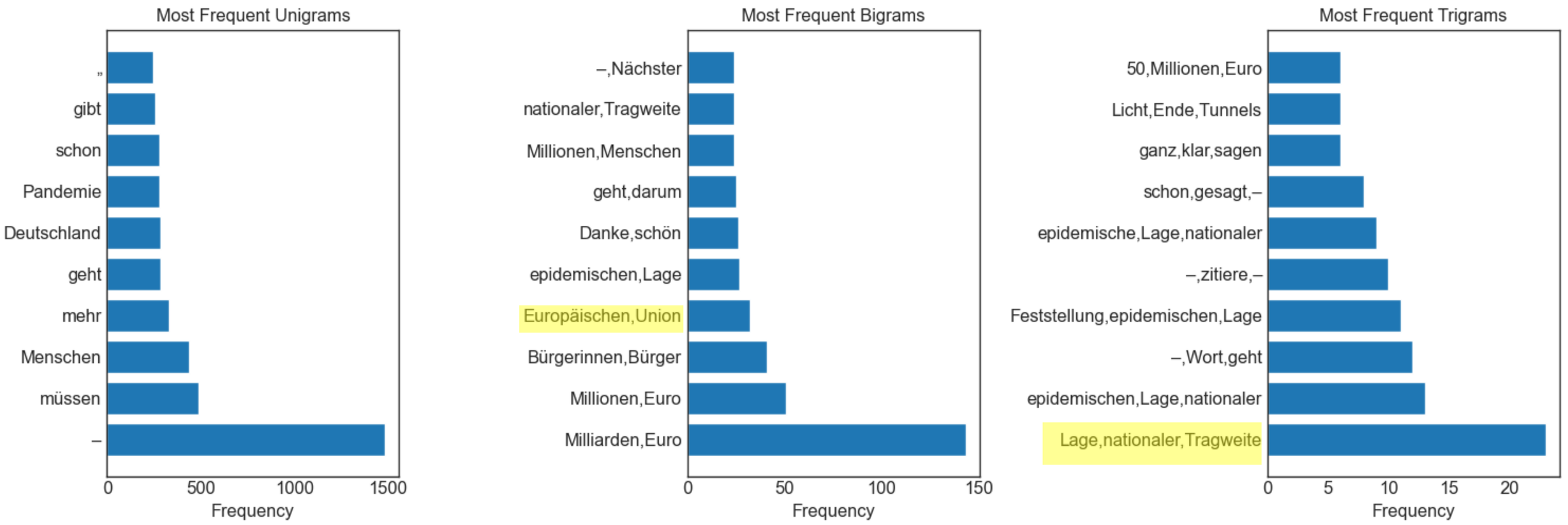
1. Sentiment Analyse

Sentiment-Score im Zeitverlauf vs. 7-Tages-Inzidenz

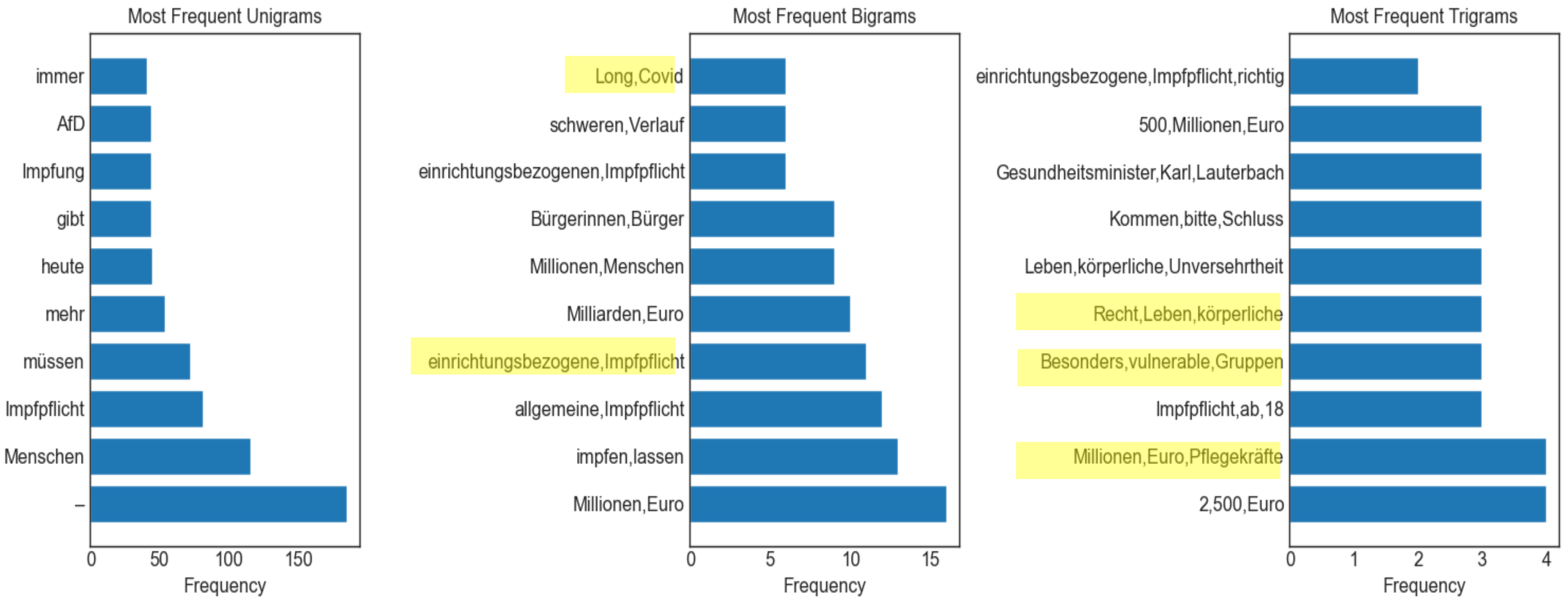




Häufigste n-gramme in der Wahlperiode 19.

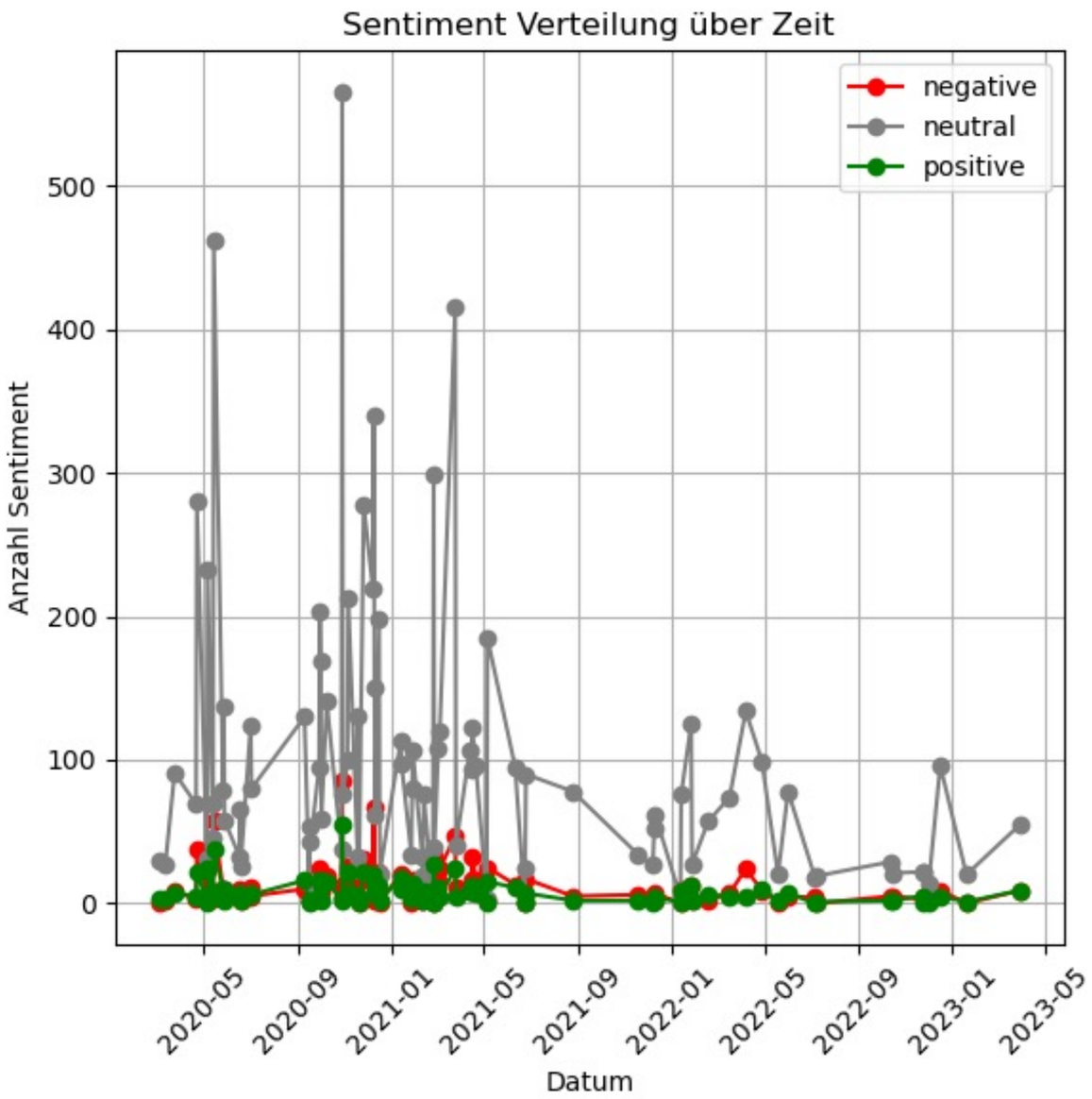


Häufigste n-gramme in der Wahlperiode 20.

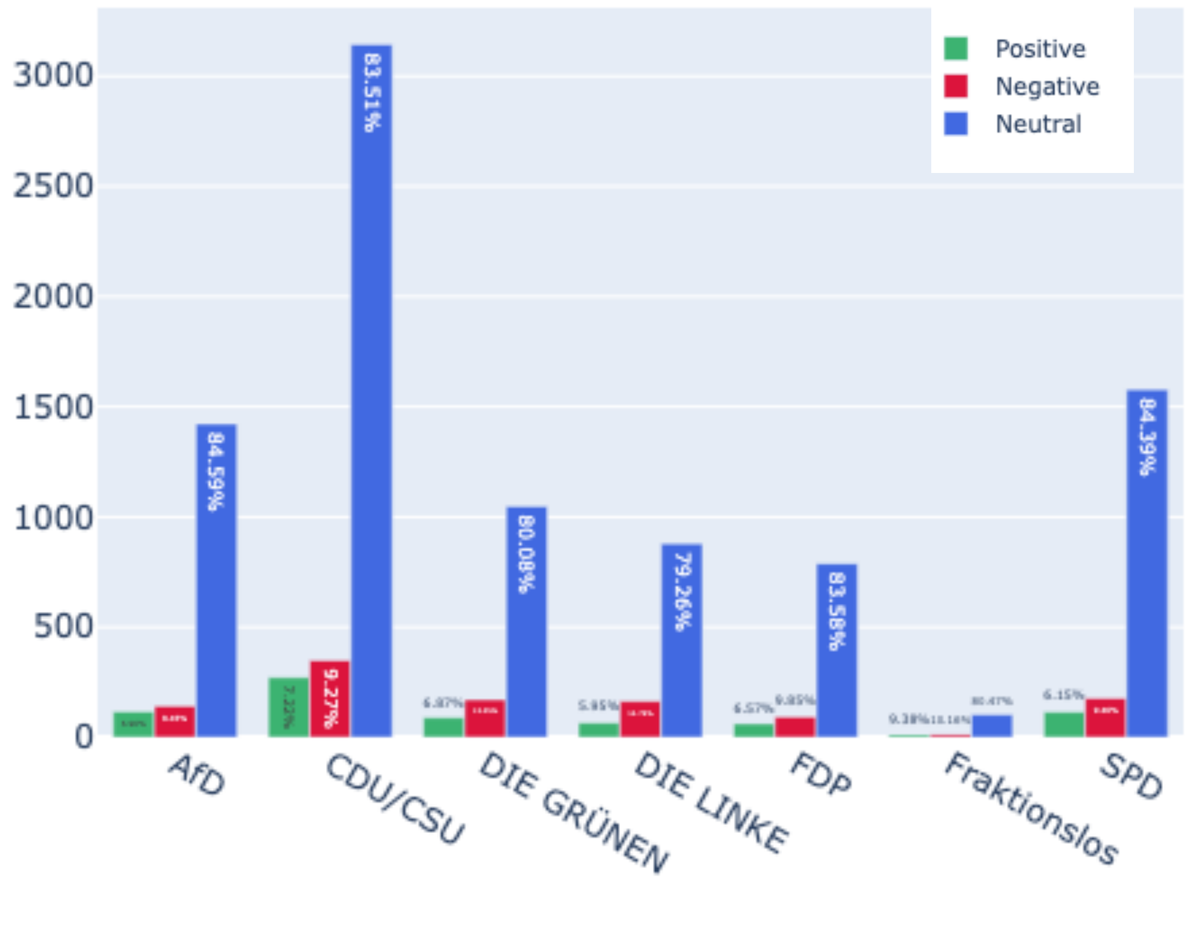


Emotionen & Entscheidung

1. Sentiment Analyse

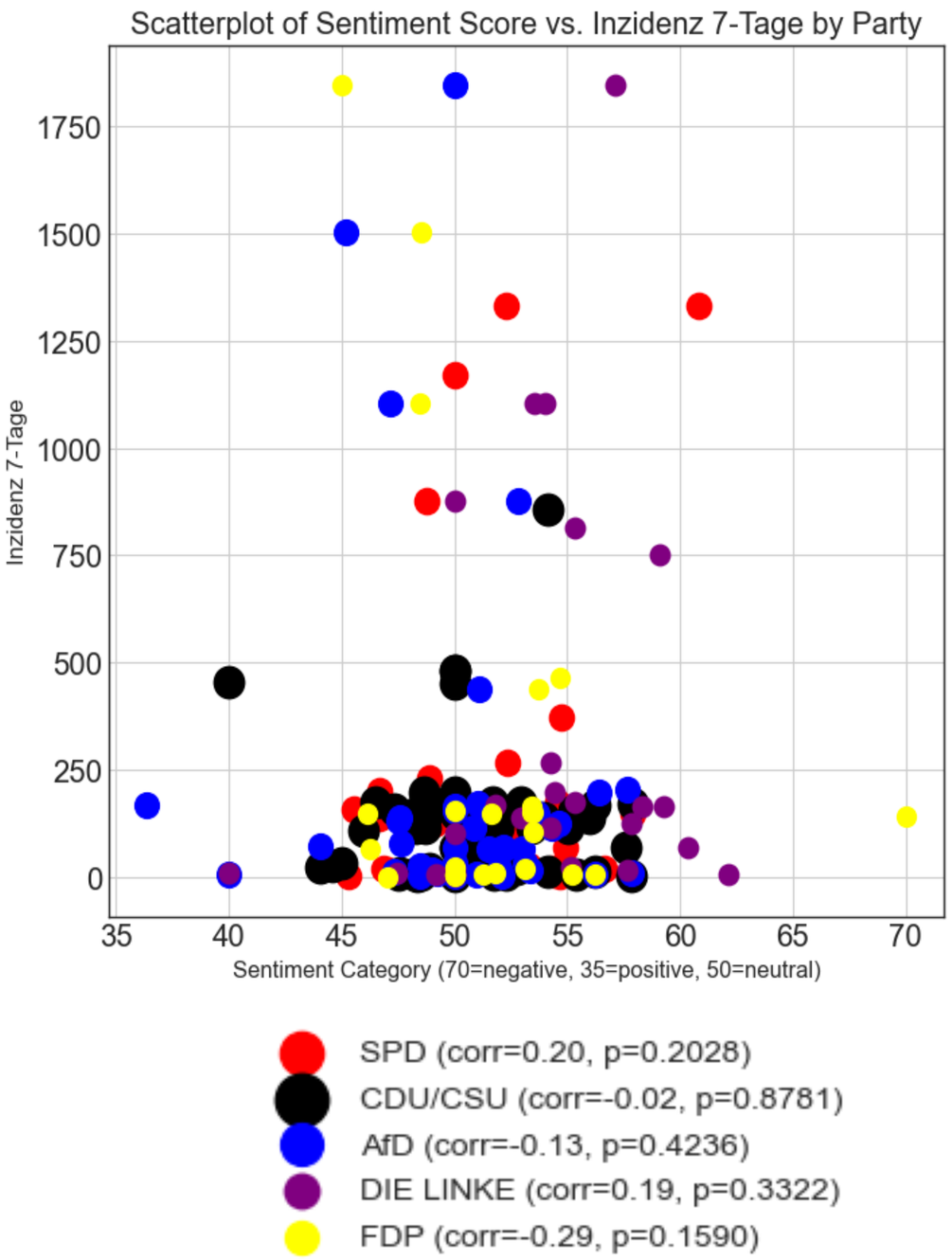


Häufigkeitsverteilung der jeweiligen Sentiment Kategorien über die Wahlperiode 19 & 20



2. Korrelationsanalyse

<u>Korrelationsergebnisse:</u>
7-Tages-Inzidenz vs. Sentiment Score der jeweiligen Politiker in Abhängigkeit der Parteizugehörigkeit



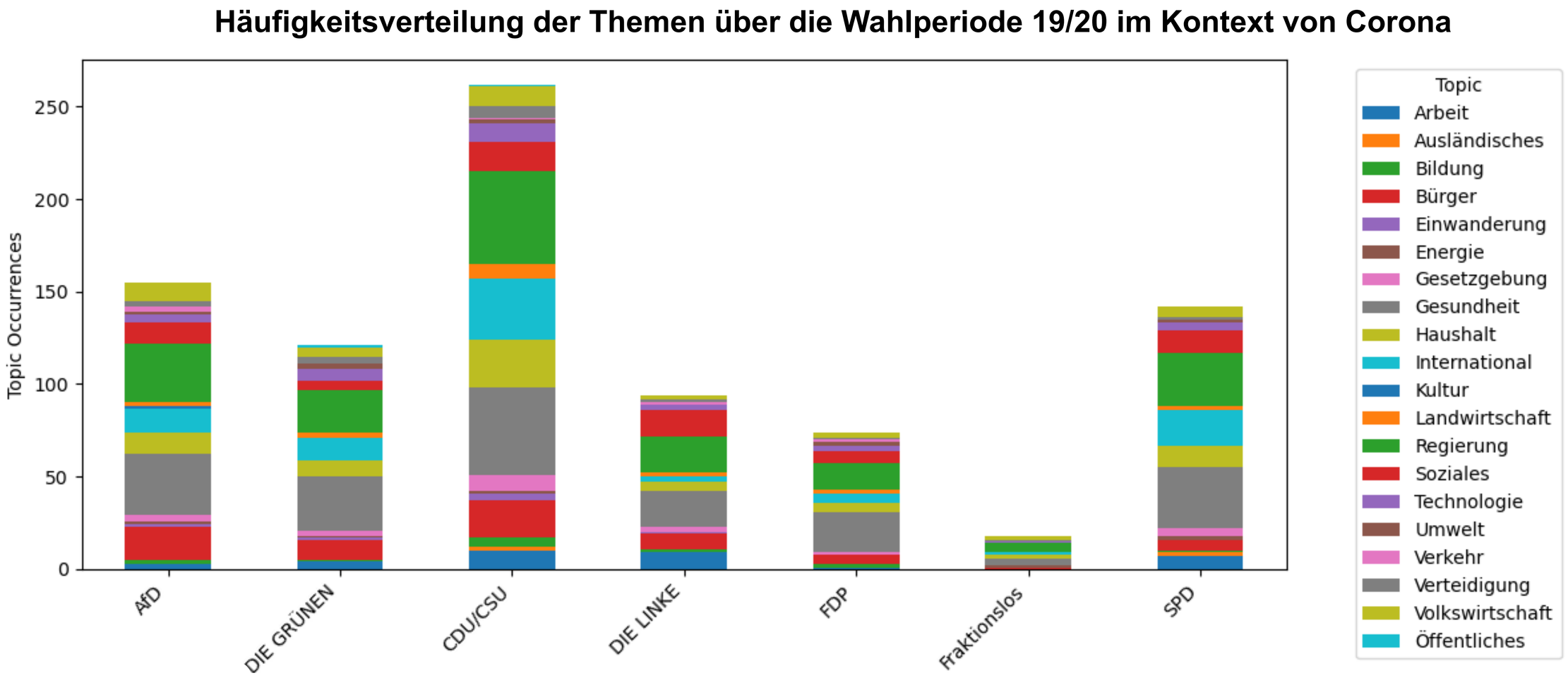
→ Bei der Korrelationsanalyse ergeben sich keine signifikanten Zusammenhänge zwischen der 7-Tages Inzidenz und den Sentiment Scores der einzelnen Parteien

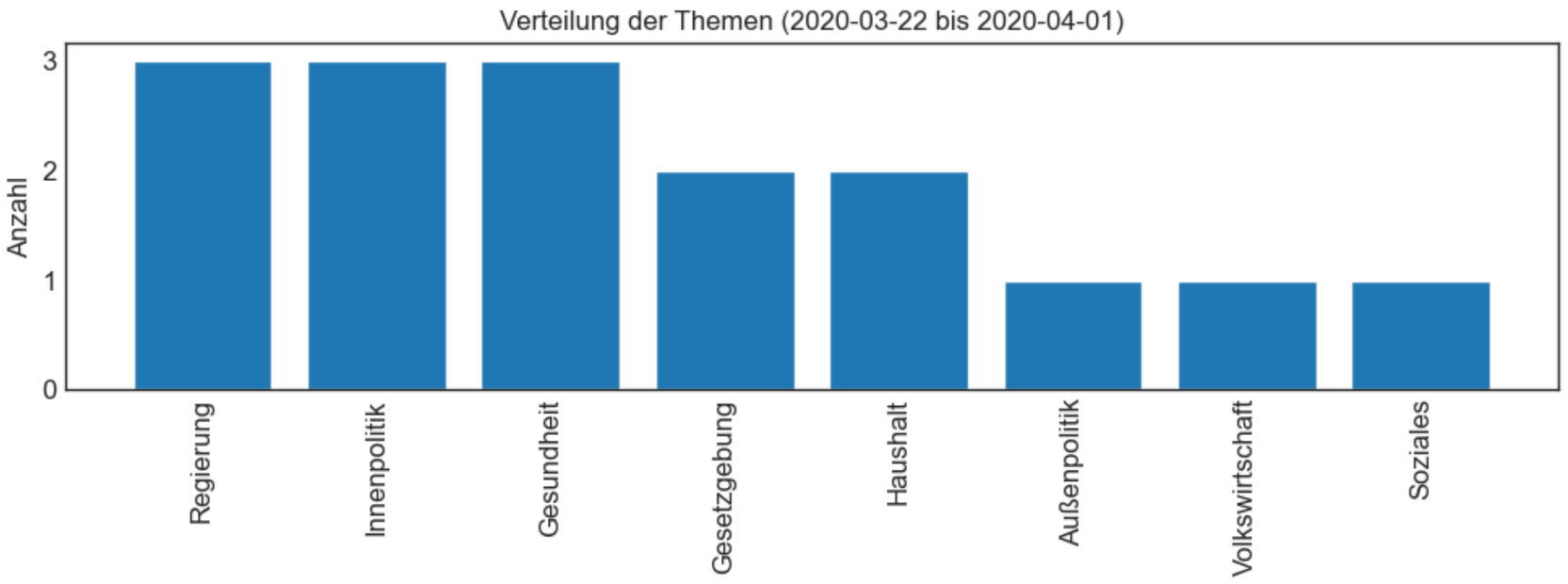
Auch wenn wir den Sentiment Score unabhängig von der Partei betrachten ergibt sich keine signifikante Korrelation

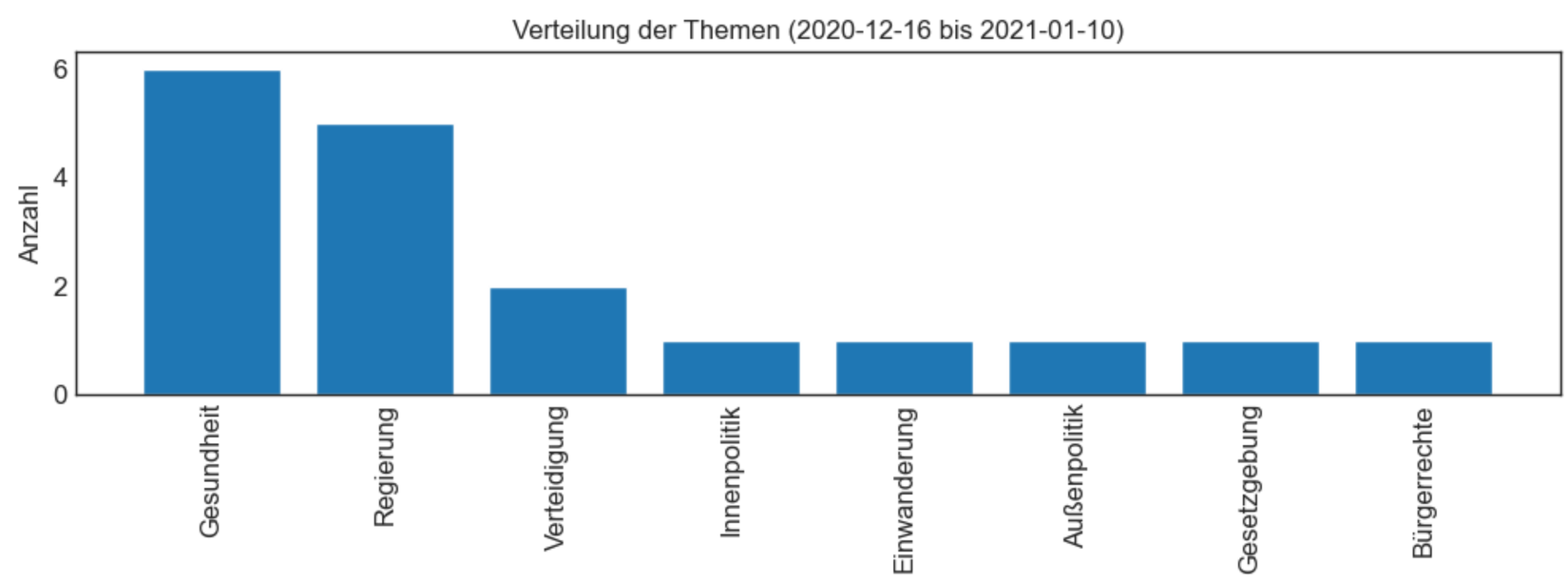
<u>Tabelle:</u>
7-Tages-Inzidenz vs. Durchschnittlicher Sentiment Score, Fälle Gesamt, Fälle-neu und Fälle-7-Tage

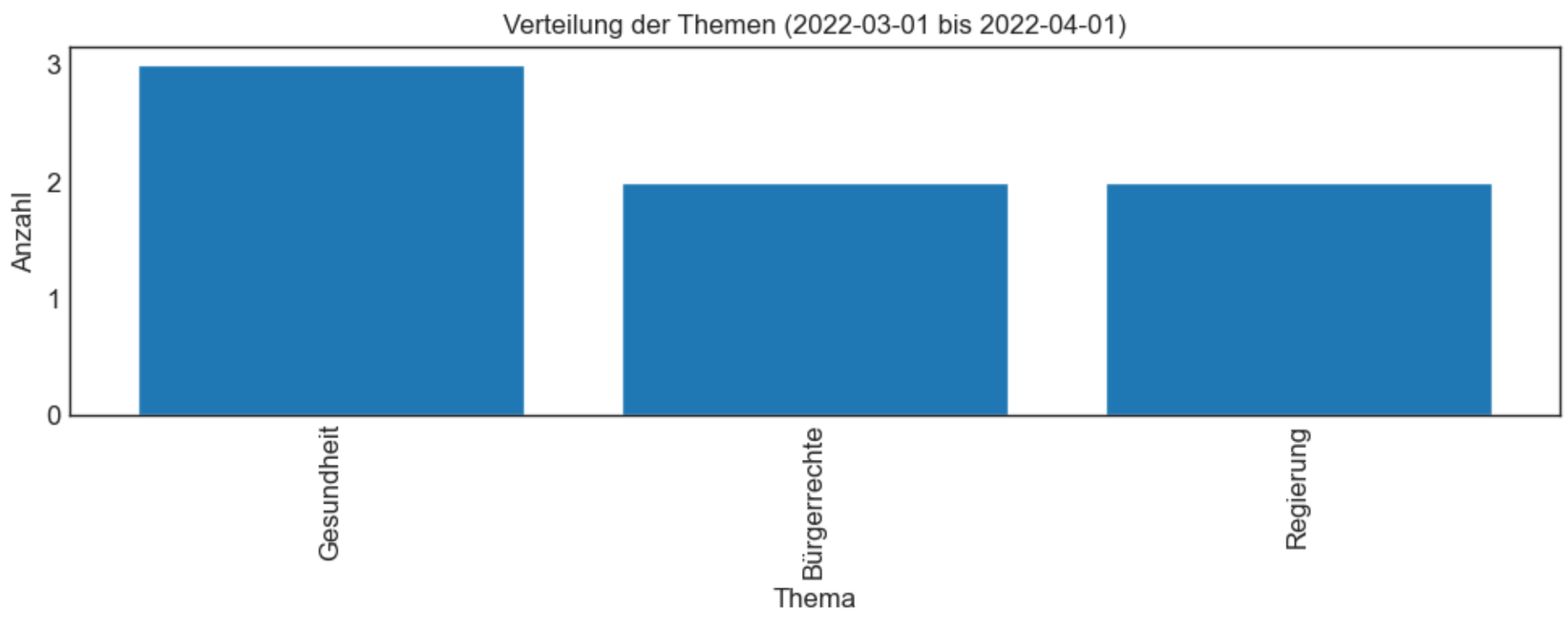
Variable 1	Variable 2	r-Wert	P-Value
Inzidenz_7-Tage	period	-0.76932	8.4326583
	Sentiment_Score	-0.01939	0.7635793
	Faelle_gesamt	0.533644	2.7485562
	Faelle_neu	0.992647	5.0085470
	Faelle_7-Tage	0.999999	0.0

3. Topic Analyse







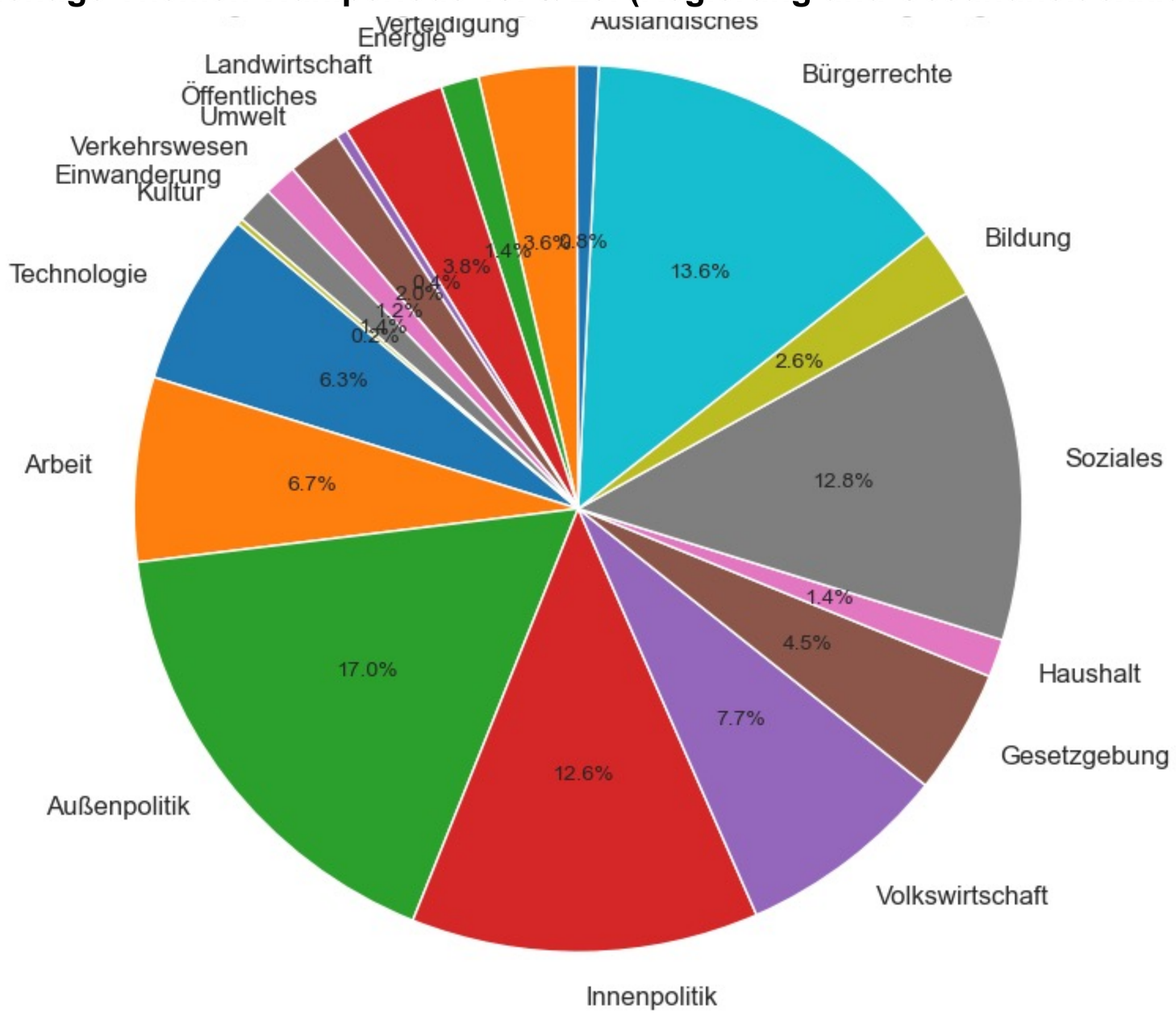


3. Topic Analyse

Ausschnitt aus dem topic_df zum Verständnis der Analyse:

index	id	date	party	name	text	period	Sentiment_Score	probabi...	*topic*
76	ID1918009...	2020-10-0...	CDU/CSU	Josef Rief	Sehr geehrte Frau Präs...	19	48.9130434783	0.998749...	Arbeit,Gesundheit
1	ID1915300...	2020-03-1...	SPD	Olaf Scholz	Sehr geehrter Herr Prä...	19	48.4375	0.999096...	Arbeit,Gesundheit,Regierung,Außenpolitik

Wichtige Themen Wahlperiode 19. & 20. (Regierung und Gesundheit exkludiert)



Beide Wahlperioden

	Topic	Count
0	Technologie	32
1	Gesundheit	187
2	Arbeit	34
3	Außenpolitik	86
4	Regierung	173
5	Innenpolitik	64
6	Volkswirtschaft	39
7	Gesetzgebung	23
8	Haushalt	7
9	Soziales	65
10	Bildung	13
11	Bürgerrechte	69
12	Ausländisches	4
13	Verteidigung	18
14	Energie	7
15	Landwirtschaft	19
16	Öffentliches	2
17	Umwelt	10
18	Verkehrswesen	6
19	Einwanderung	7
20	Kultur	1

19. Wahlperiode

	Topic	Count
0	Technologie	30
1	Gesundheit	148
2	Arbeit	33
3	Außenpolitik	82
4	Regierung	149
5	Innenpolitik	60
6	Volkswirtschaft	39
7	Gesetzgebung	23
8	Haushalt	7
9	Soziales	48
10	Bildung	13
11	Bürgerrechte	57
12	Ausländisches	4
13	Verteidigung	18
14	Energie	7
15	Landwirtschaft	17
16	Öffentliches	2
17	Umwelt	9
18	Verkehrswesen	6
19	Einwanderung	6
20	Kultur	1

20. Wahlperiode

	Topic	Count
0	Gesundheit	39
1	Bürgerrechte	12
2	Umwelt	1
3	Regierung	24
4	Soziales	17
5	Technologie	2
6	Außenpolitik	4
7	Arbeit	1
8	Innenpolitik	4
9	Einwanderung	1
10	Landwirtschaft	2

Emotionen & Entscheidung

4. Datenvorbereitung

		Anzahl Daten		Anzahl Daten
Start	Reden aus der 19. Wahlperiode	n= 25.187	Reden aus der 20. Wahlperiode	n=10.791
1. Schritt	Reden filtern nach dem Stichwort „Covid 19“. Es werden nur Reden enthalten sein, die das Stichwort Covid-19 enthalten, davor wurden verschiedene Suchstrategien entwickelt unter anderem unter dem Stichwort “Corona“, welches zu schwachen Ergebnissen & Reden unabhängig vom Kontext Corona geführt hat.			
Zwischenschritt	df 19	n=200	df 20	n = 43
2. Schritt	Reden transformieren, sodass in jeder Zeile ein Satz vorhanden ist. Dadurch wird jede Rede als einzelner Satz in dem Dataframe eingespeichert. Da das Transformer Modell die Sentiment Analyse nur auf Satzebene durchführen kann.			
Zwischenschritt	df19_cleaned	n=9397	df20_cleaned	n=1411
3. Schritt	Anwendung des Transformer Modells: **german-sentiment-bert** zur automatischen Bestimmung des Sentiments auf Satz-ebene. Dieses vortrainierte Modell gibt eine Klassifikation auf Basis der Label „positiv“, „negativ“ und „neutral“ aus und ein Wahrscheinlichkeit Score der wiedergibt, wie wahrscheinlich die Klassifikation erreicht wird, siehe dazu auch https://huggingface.co/oliverguhr/german-sentiment-bert.			
Ende	df_sentiment19	n=9397	df_sentiment20	n=9397

5. Datenverarbeitung

		Anzahl Daten		Anzahl Daten
Start	df_sentiment19	n=9397	df_sentiment20	n=9397
1. Schritt	Nach einer kurzen deskriptiven Analyse werden beide Dataframes zum merged_df zusammengeführt. Ziel ist es zu untersuchen, ob zwischen dem Sentiment Verlauf innerhalb des Bundestages und der 7-Tages-Inzidenz ein Zusammenhang besteht. Dazu werden von dem Github Resp des RKI die Zahlen zu den Coronadaten gedownloadet, hier interessiert vor allem die Variable Inzidenz_7-Tage.			
Zwischenschritt	merged_df	n=10808	Corona_Fallzahlen RKI	n =9506
2. Schritt	In einem nächsten Schritt werden die Corona Zahlen gefiltert, wir wollen die 7-Tage-Inzidenz unabhängig von der Altersgruppe erhalten → Ausprägung Altersgruppe = 00+.			
Zwischenschritt	merged_df	n=10808	corona_fallzahlen_gesamt	n=1358
3. Schritt	Derzeit haben wir für jeden Tag eine Vielzahl an Sätzen und den zugehörigen Sentiments in unserem merged_df, um die Vergleichbarkeit mit dem Verlauf der 7-Tages-Inzidenz zu gewährleisten wollen wir einen Durchschnitts Sentiment-Score für jeden Tag pro Politiker berechnen. Dazu müssen wir zuerst die Label „Positiv“, “neutral“ und “negativ“ in einen Score umwandeln, wir nutzen die Skala: 'positive’ = 0.0, 'neutral': 50.0 & 'negative': 100.0. Dieses ist zur grafischen Visualisierung einfacher. Danach wird auf Basis der eindeutigen Redner ID für jeden Politiker der durchschnittliche Sentiment Score pro Tag gebildet. In einem abschließenden Schritt filtern wir den corona_fallzahlen_gesamt Dataframe, um für die vorgegebenen Daten die passende 7-Tages-Inzidenz zu erhalten, da wir nur für ausgewählte Daten Reden vorhanden haben.			
Ende	Result_Df	n=243	corona_fallzahlen_gesamt	n=243

- Auf Basis dieser Daten wurden dann die Korrelationsanalyse und Topicanalyse durchgeführt

5. Limitation der Ausarbeitung

I. Abhängigkeit von den Trainingsdaten der Transformer-Modelle:

- Die Qualität der Sentimentbewertung und Topicmodellen hängt stark von den Trainingsdaten der Transformer-Modelle ab.
- Das German-Sentiment Modell ist nicht speziell auf politische Reden trainiert, was zu häufigen Fehlklassifikationen führen kann.

II. Beschränkung auf einen kleinen Datensatz:

- Die Untersuchung wurde auf einen kleinen Datensatz von Reden beschränkt.
- Man könnte die Reden genauer nach anderen Stichworten im Kontext von Corona filtern, um die Genauigkeit zu erhöhen.
- Die p-Werte der Korrelationsanalysen waren nicht signifikant, dieses kann auch an einem zu kleinen Datensatz liegen.
- Insgesamt war es schwierig signifikante Zusammenhänge zwischen der Inzidenz und dem Sentiment zu identifizieren.
- Sentiment und Inzidenz müssten genauer untersucht werden auch im Kontext von Zeitreihenanalyse, ggf. andere Parameter nutzen.

II. NLP-Verarbeitung & Transformer-Modelle sind sehr ressourcenintensiv

- NLP-Methoden nutzen große Datensätze, was zu langen Verarbeitungszeiten und einem hohen Bedarf an Rechenleistung führt.
- Die hohen Rechenanforderungen von Transformer-Modellen haben zu einer Begrenzung auf einen kleinen Datensatz geführt.