

Coronavirus Perception and Evolution

Term Project - Web Data & Digital Analytics

F. BOEGLIN, A. CHABAUD, K. DONMEZ, R. HUG

- 01 Introduction – Coronavirus under the Loop
- 02 Network Analysis – Worldwide Air Routes
- 03 Text Mining – Countries' Reactions
- 04 Conclusion & Limitations

INTRODUCTION

How can we use network analysis and text mining to analyze the perception and evolution of Coronavirus?

By using available information about the spread of the Coronavirus and the measures taken by different countries, we try to see :

- **Network analysis:** what are the similarities between the spread of the virus and the structure of worldwide air routes?
- **Text Mining:** how did the overall sentiment about this pandemic evolve over time, and what were different countries' reactions?

By combining our findings, we develop conclusions about the perception and evolution of Coronavirus between selected countries.

NETWORK ANALYSIS

THE DATA

- OpenFlights – 2012
- 59,036 routes
- 3,209 airports
- 531 airlines
- Directed network

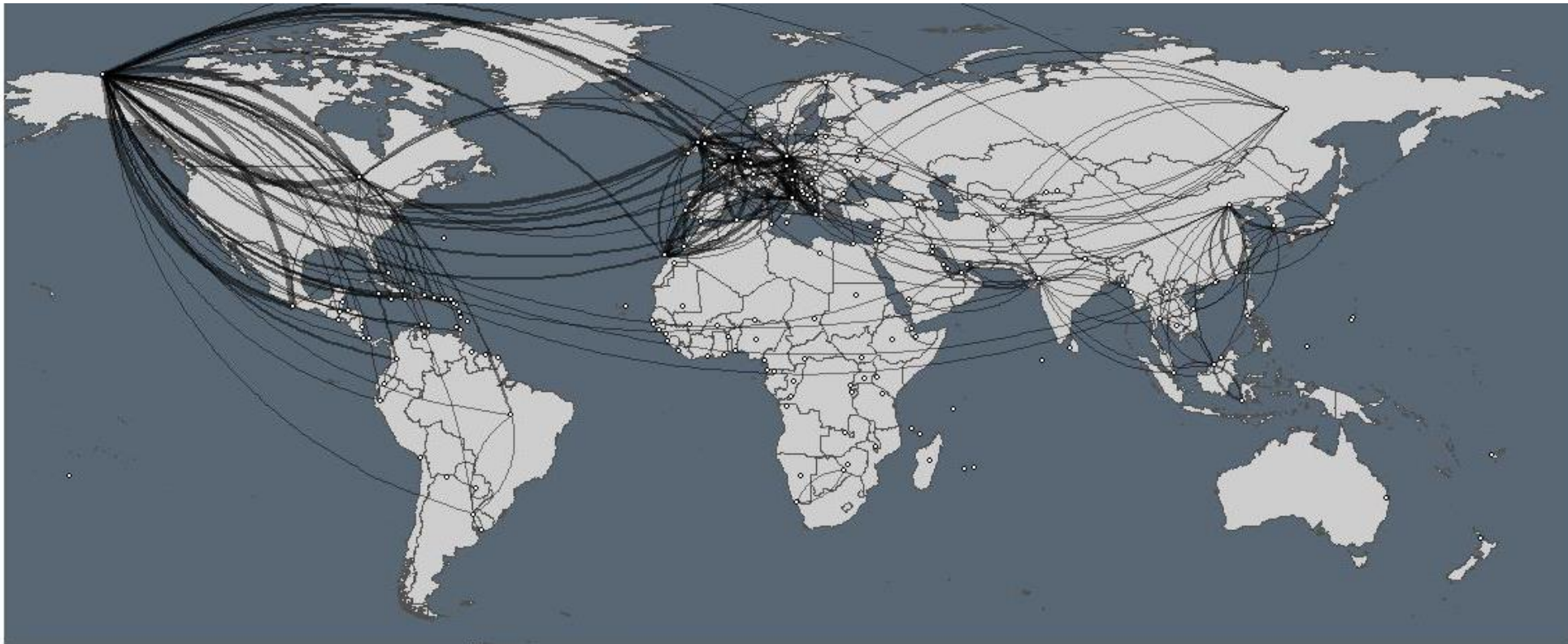
THE ANALYSIS

- Modeling the network
- Hubs
- Comparison with
Coronavirus map

Modeling

Worldwide air routes

- **Very heavy database:** originally listing the connections between airports, by airlines; grouped by country
- **Weight of connection:** same connection from different airports
- **Representation:** only the connections with weight > 5 ; width of the edges represents weight
- **Observations:** Graph density is 0.214, but some areas seem more connected
 - Mediterranean basin
 - Europe-US
 - US-Central America
 - South-East Asia



Hubs

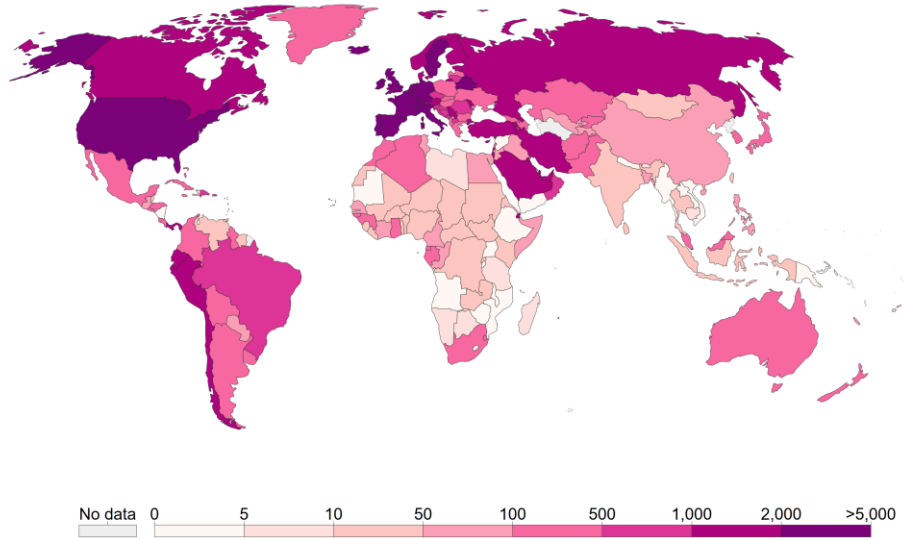
Central countries for worldwide air routes

- **In line with model visualization:** mostly countries in Mediterranean basin + United States & Mexico
- **Apparent similarity with COVID-19 cases:** 8 of the 14 most touched countries are also among the top 14 worldwide hubs

Hubs Ranking	
United States	1
Germany	0,764852
United Kingdom	0,700827
France	0,605118
Canada	0,590904
Italy	0,583538
Mexico	0,498101
Spain	0,465886
Greece	0,264316
Morocco	0,143263
Austria	0,132506
Netherlands	0,112909
Algeria	0,106361
Switzerland	0,10193

COVID-19 Cases per one million population*	
Spain	5661
Ireland	4657
Belgium	4612
United States	4133
Singapore	4072
Italy	3623
Switzerland	3502
United Kingdom	3229
France	2711
Portugal	2705
Sweden	2606
Belarus	2530
Netherlands	2488
Germany	2051

*Of countries with a population over four million; China and Iran excluded
 As of 11 May 2020 10:39 GMT
 Source: worldometer



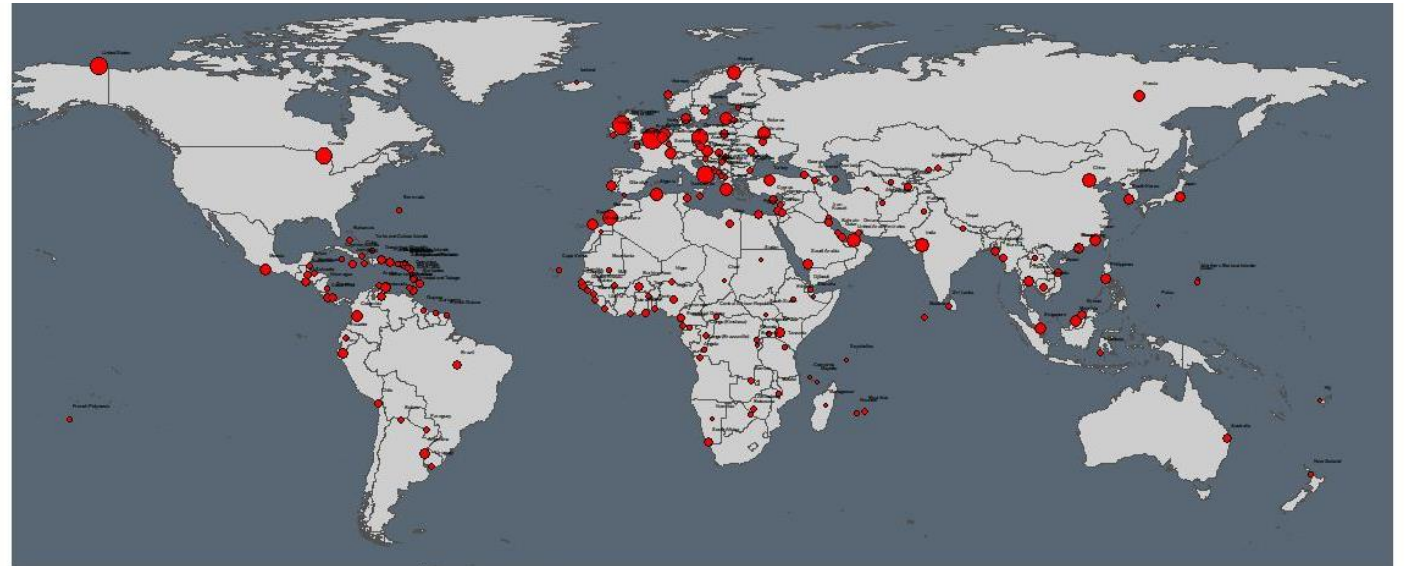
Source: European CDC – Situation Update Worldwide – Last updated 10th May, 11:00 (London time) OurWorldInData.org/coronavirus • CC BY

Map Comparison

Inferences by comparing the two maps

- **Similarities:** in general, it seems that there is some correlation between the worldwide air routes and the spread of the Coronavirus
- **Multitude of factors:** there are many factors that explain the spread, and the worldwide air routes are only one aspect
- **Limitations:** our model presents limitations, as our data is static & older vs. a dynamic & constant evolution of the pandemic

- **Countries for text mining:** choice of countries based on this network analysis
 - US: main hub and strongly touched
 - Germany: second biggest hub, but not too strongly touched
 - UK: third hub and interesting change of strategy



TEXT MINING

THE DATA

- Opinion leaders' tweets from selected countries
- US, UK, Germany
- A politician, a national influencer, a news media

THE ANALYSIS

- Measures of the three countries
- Sentiment analysis over time
- Comparison between opinion leaders in each country
- Comparison across countries

negative

- [illegible]

positive

- [illegible]

Figure 1 displays three horizontal bar charts showing the top 10 terms in each of the three LDA topics. The x-axis represents the probability of a term belonging to a topic, ranging from 0.00 to 0.03.

Topic 1 (Red):

Term	Probability (approx.)
thank	0.005
presidenttrump	0.003
presid	0.002
begin	0.002
greatstat	0.002
great	0.002
work	0.002
realdonaldtrump	0.002
republican	0.002

Topic 2 (Green):

Term	Probability (approx.)
countin	0.004
work	0.003
democrat	0.003
make	0.003
coronavirustaskforc	0.003
realdonaldtrump	0.003
presid	0.003
see	0.003
r	0.003

Topic 3 (Blue):

Term	Probability (approx.)
thank	0.012
peopl	0.008
get	0.006
rtuhitahouselivepressbrief	0.004
coronavirus	0.004
greatstat	0.003
know	0.002
done	0.002
greatjob	0.002

-

- All the graphs are in the appendix

United States

Opinion Leaders' Level

OBSERVED PERSONS

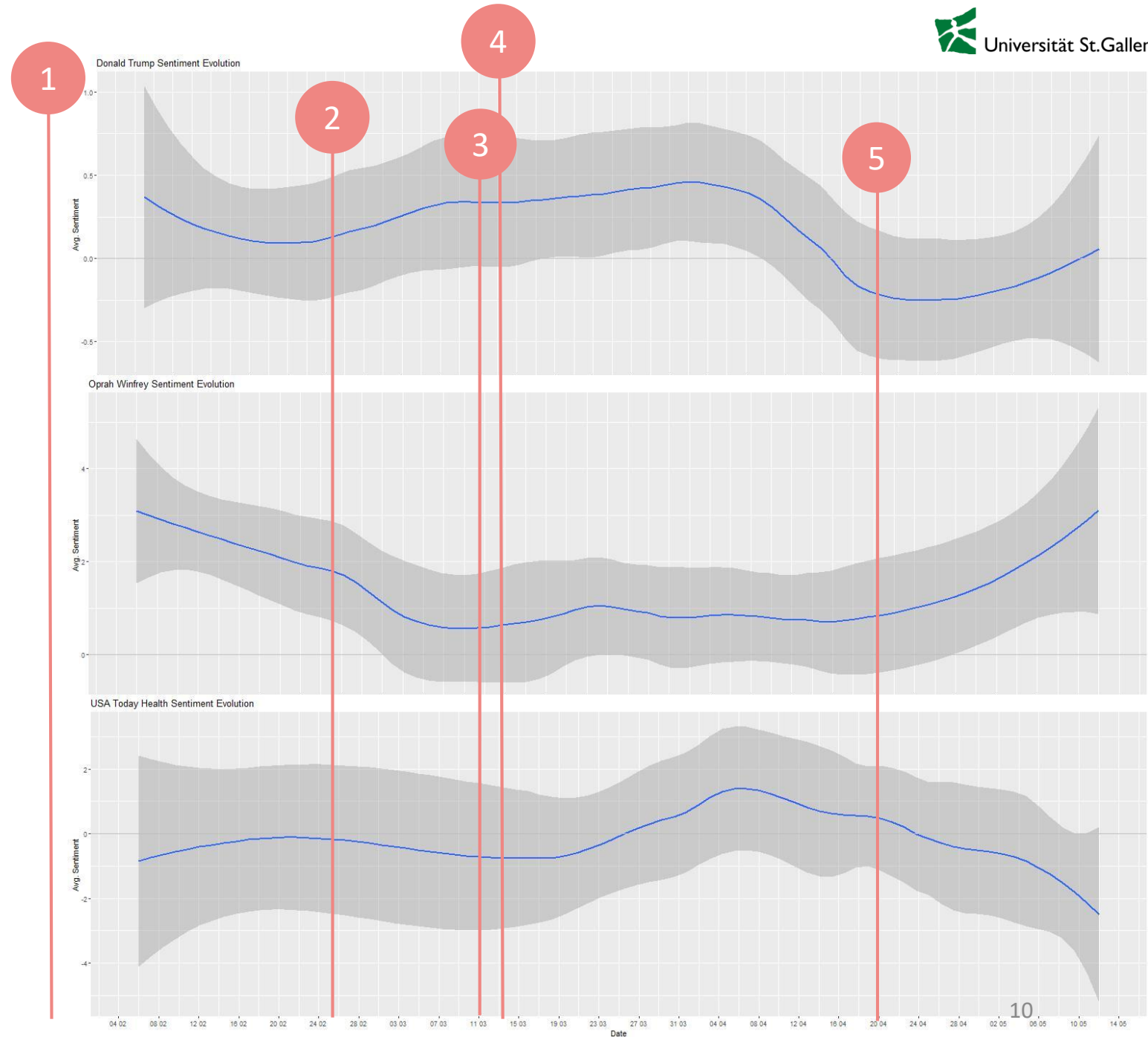
- **Trump** (with retweets) – sentiment values are close to 0
- **Oprah Winfrey** (with retweets) – highest sentiment values
- **USA Today Health** – sentiment is most of the times below 0

SENTIMENT EVOLUTION

- Sentiment evolution of Donald Trump continuously increases after end of February while Oprah Winfrey and USA Today Health sentiment evolution goes down
- Sentiment of Donald Trump and Oprah Winfrey goes up after denial of immigration; sentiment of USA Today health goes down

IMPORTANT POINTS IN TIME/POSSIBLE TURNING POINTS

1. First COVID-19 case confirmed
2. US CDC warns that the spread to the United States is likely and that people should prepare
3. WHO declares COVID-19 as a pandemic
4. States start to cancel gatherings of more than 250 people
5. Donald Trump announces he will temporarily suspend immigration to the US for 60 days by executive order.



- New daily cases seem to have an inverse relationship with daily sentiment
- The sentiment decrease can be seen starting before the virus started growing
- Wordcloud shows “fake” (comes from Donald Trump) and “hard”, as well as “strong”, “protect” and “support”

The figure consists of two vertically stacked line charts sharing a common x-axis representing time from January 27, 2020, to May 14, 2020.

The top chart, titled "U.S. new cases", plots the number of new cases on the y-axis (0 to 30,000). A blue line shows the daily new cases, which remain near zero until mid-March, then rise sharply to a peak of approximately 30,000 in early April, followed by a decline. A gray shaded area represents the confidence interval around the blue line.

The bottom chart, titled "US", plots the average sentiment on the y-axis (-0.5 to 1.0). A blue line shows the sentiment index, which starts around 0.25, peaks near 0.65 in late February, then generally declines to a low of about -0.2 in mid-April, before rising back towards 0.3 by mid-May. A gray shaded area represents the confidence interval around the blue line.

United Kingdom

Opinion Leaders' Level

OBSERVED PERSONS

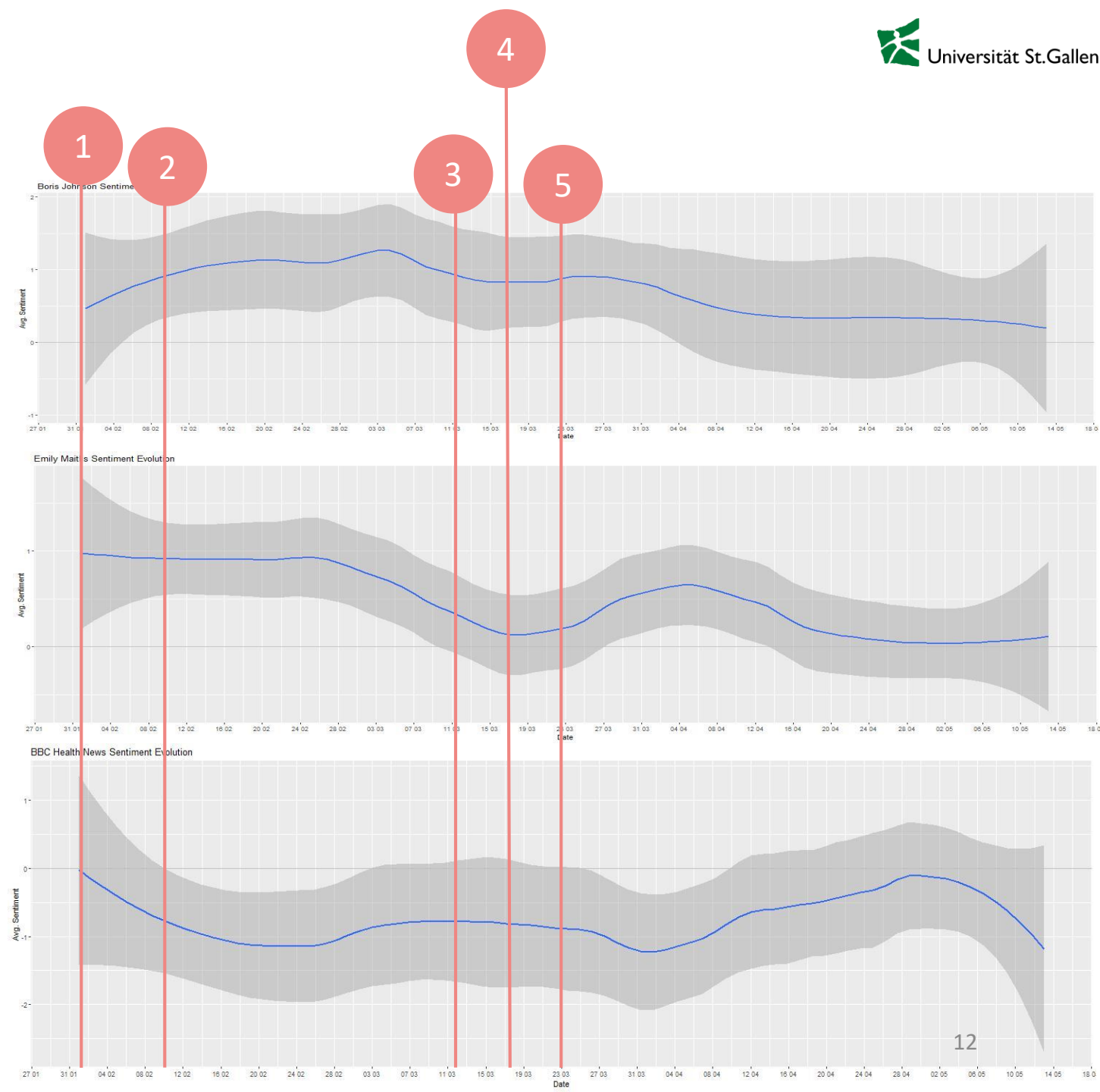
- **Boris Johnson** – sentiment goes up and then gradually decreases
- **Emily Maitlis** – compared to the others, high starting point and with time close to 0
- **BBC Health News** – always below 0

SENTIMENT EVOLUTION

- Emily Maitlis' sentiment evolution develops against the direction of the others (beginning of April)
- Beginning of April, sentiment of Boris Johnson goes gradually down
- BBC health decreases and then increases again mid March; it sinks again after May
- BBC's sentiment increases when Boris' and Emily's gradually goes down

IMPORTANT POINTS IN TIME/POSSIBLE TURNING POINTS

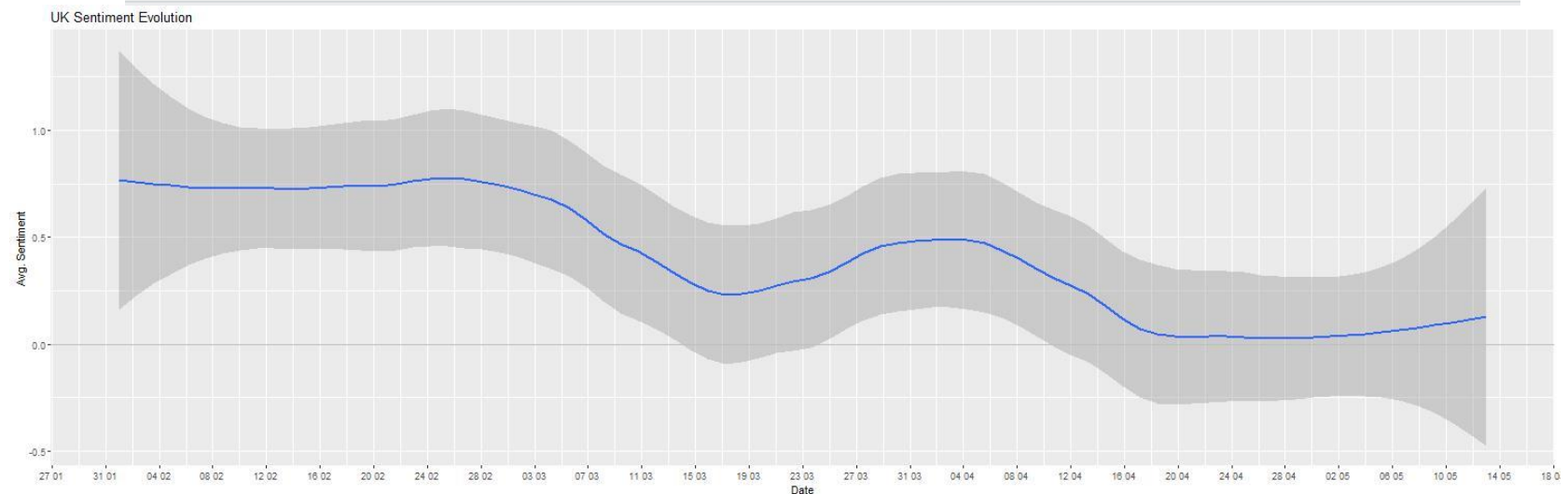
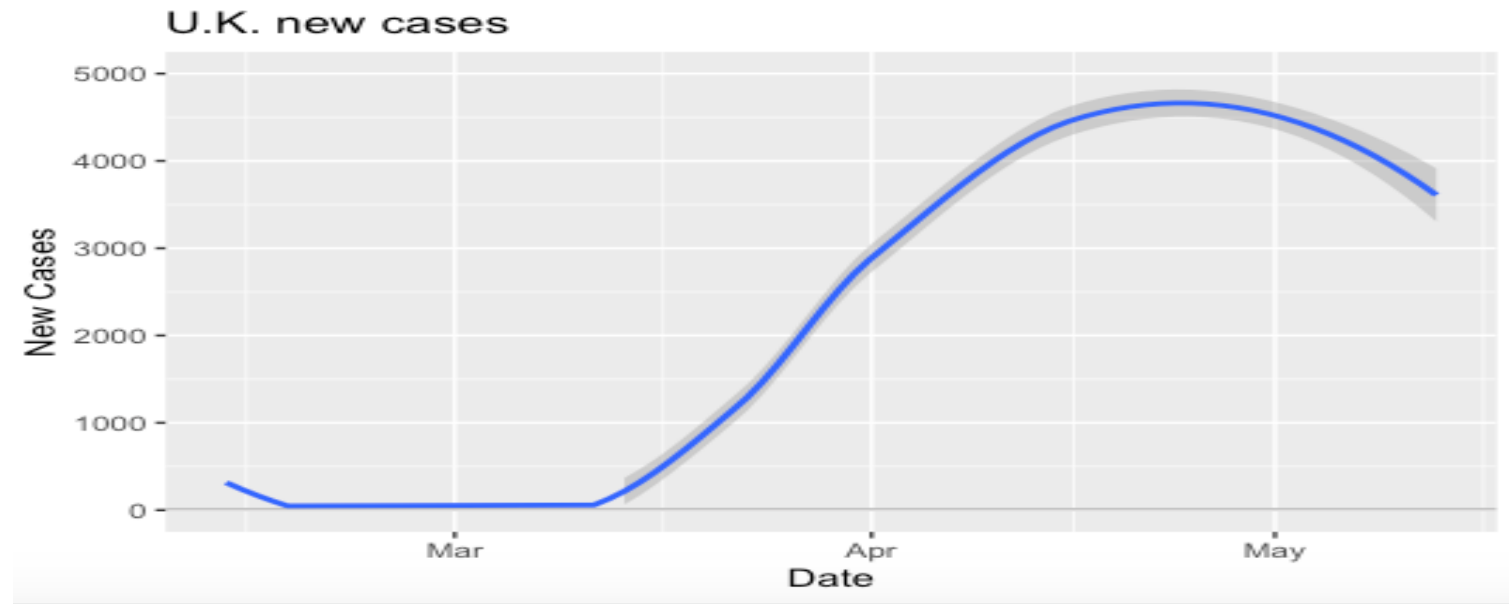
1. First COVID-19 case confirmed
2. WHO declares COVID-19 as a pandemic
3. Declaration that the coronavirus constitutes a serious and imminent threat to public health
4. UK advises citizens against all but essential international travel and strongly starts advising social distancing
5. Lockdown



United Kingdom

Country's Level

- Interesting case due to complete change of strategy
- Interestingly, sentiment started turning positive as cases started growing; it only started going down again after change of strategy and lockdown implementation
- Wordcloud points to heavy focus on protection but also points towards current crisis and death



Germany

Opinion Leaders' Level

OBSERVED PERSONS

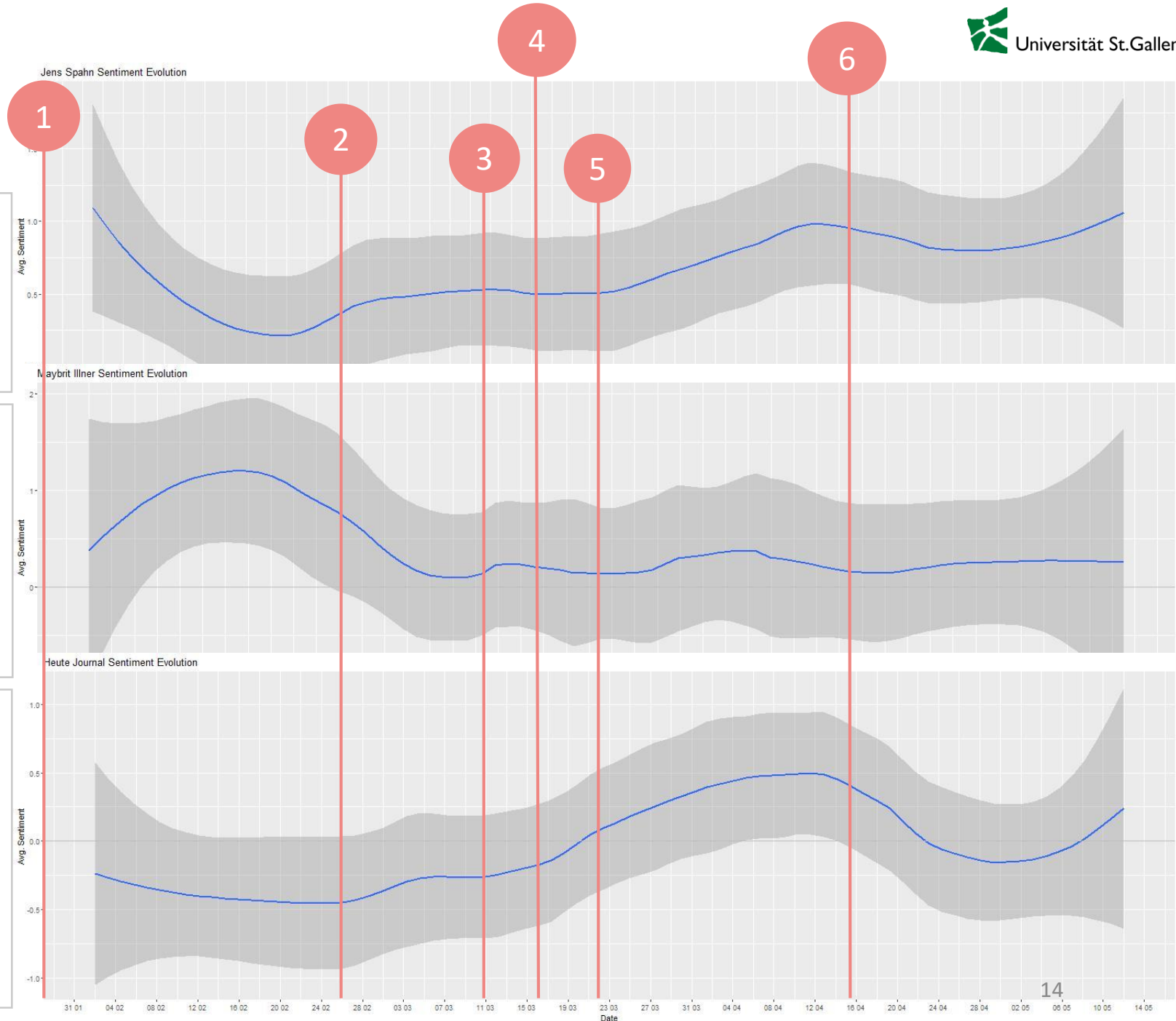
- **Jens Spahn** – low sentiment at the beginning and then gradually up
- **Maybrit Illner** – up and down at the beginning and then stays close to 0
- **Heute Journal** – stays close to 0, increase and decrease

SENTIMENT EVOLUTION

- Jens Spahn's sentiment is contrasting with Illner's and starts rising at the end of February; aligns with Heute Journal's sentiment evolution
- Maybrit Illner's sentiment peaks midst of February and then stays close to zero, without big changes
- Heute Journal's sentiment goes gradually up, unlike Maybrit's; it goes down after mid April and then rises again

IMPORTANT POINTS IN TIME/POSSIBLE TURNING POINTS

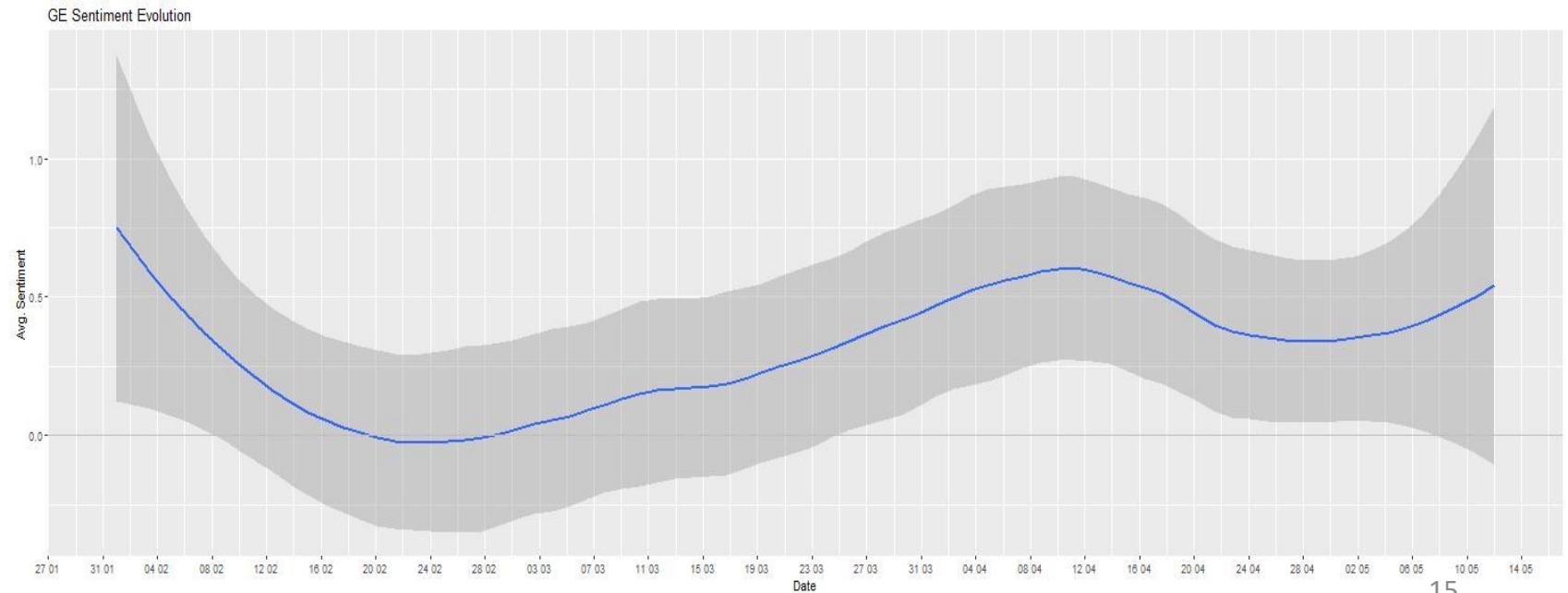
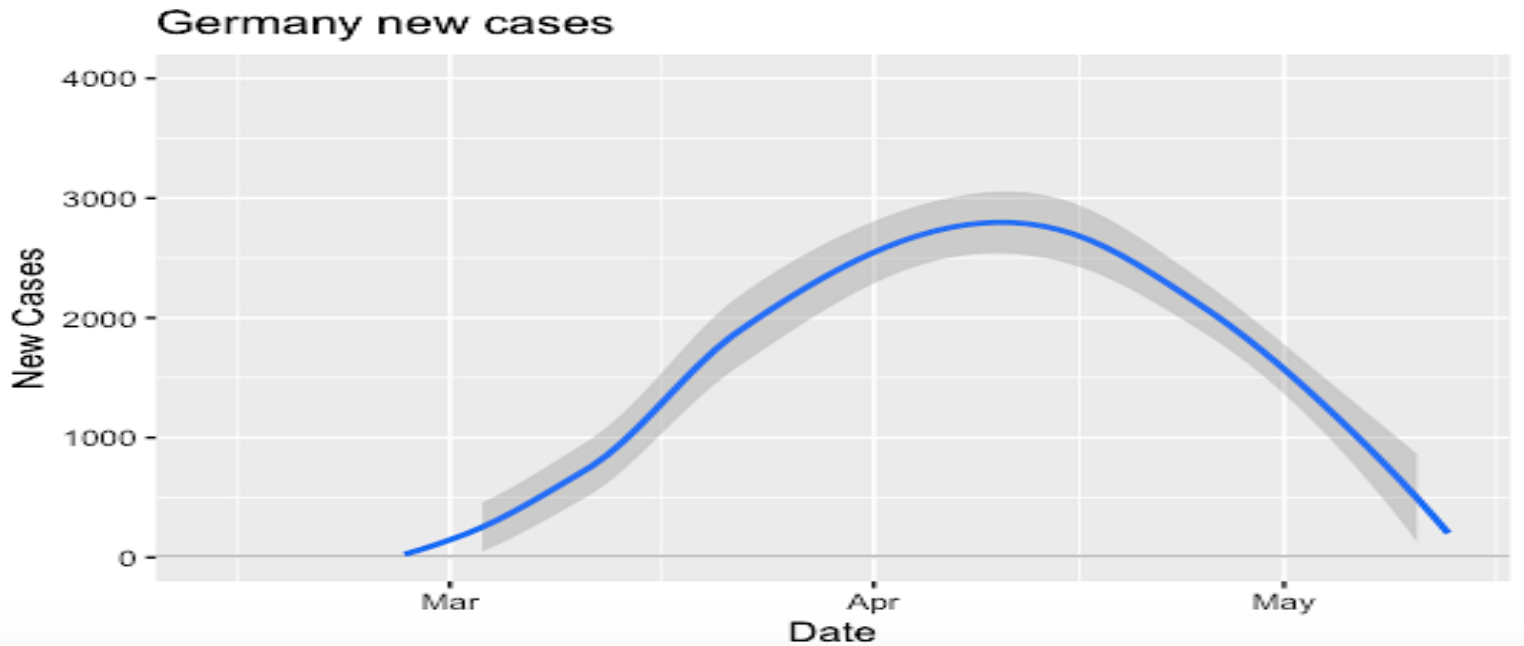
1. First COVID-19 case confirmed
2. Germany's minister of health announces that the country is at the beginning of an epidemic
3. WHO declares COVID-19 as a pandemic
4. Germany closes borders
5. "Contact ban" of more than two people
6. Germany announces plans to reopen their economy starting April 20th



Germany

Country's Level

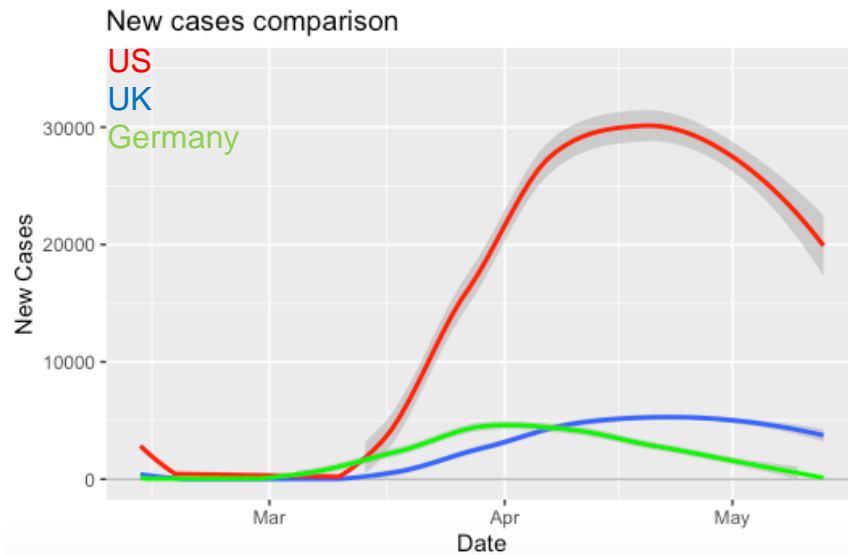
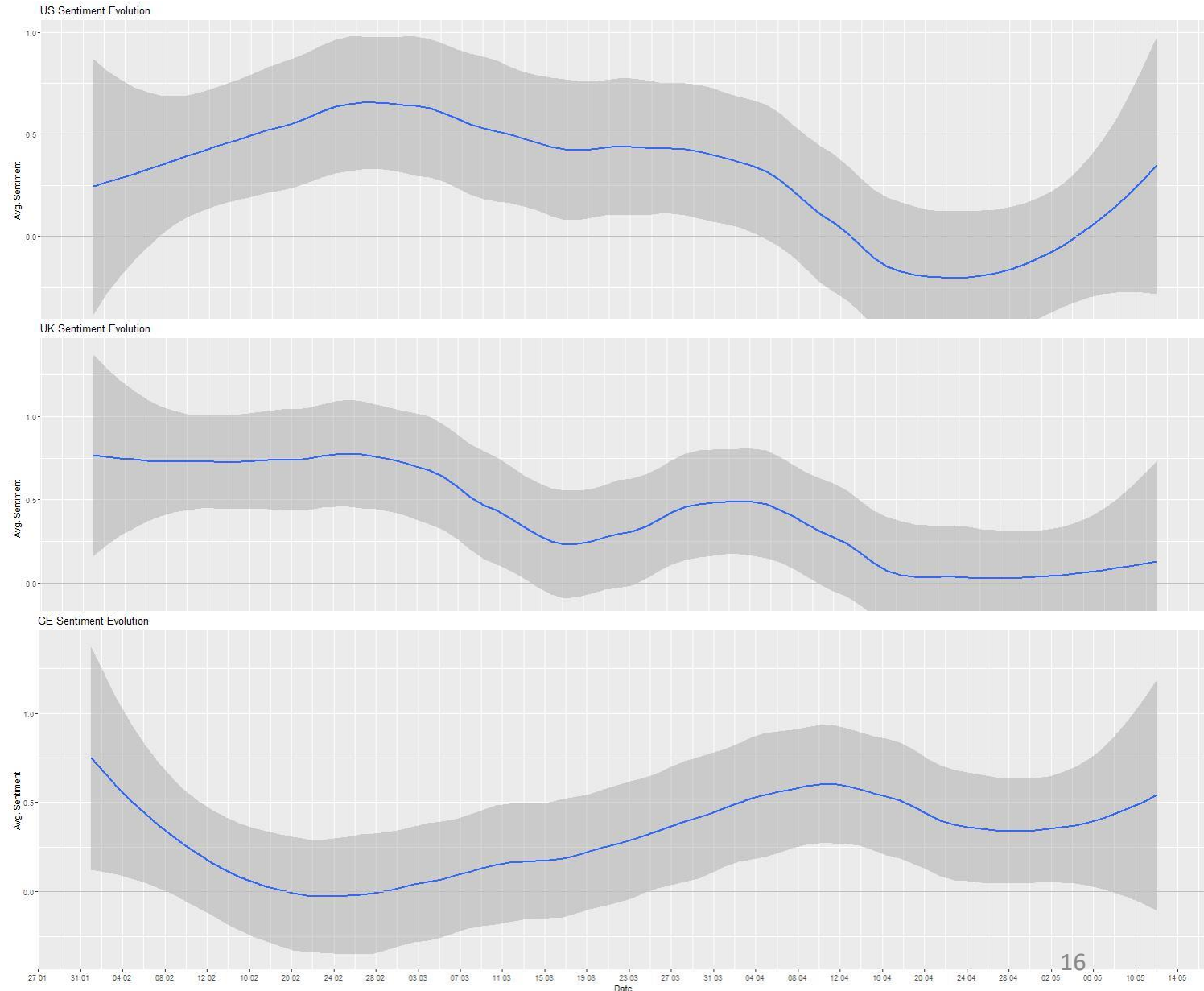
- There seems to be a lag between sentiment evolution and virus spread: sentiment declines before the virus starts growing, then it increases and starts declining again mid-April, until it goes back up
- The sentiment probably correlates with the measures, as people feel safer when they know that something is put in place to protect them
- Negative words are about the difficult time and the measures that are taken, positive words are about going through the situation together



Countries' Comparison

Sentiment Analysis Across Countries

- In all cases, it seems that a decrease of sentiment started before the virus actually started growing
- The sentiment curves of the US and UK seem to be similar when compared to Germany, which could be explained by their similarities in the “new cases” curves
- Germany seems to have dealt with the crisis faster and more effectively than the US and UK
- Hypothesis – Germany realized the dangers of the virus early, compared to UK and the US



FINAL NOTES

CONCLUSION & POTENTIAL EXTENSION

- Our results are mostly hypotheses, but they allow for some better understanding of the situation
- Big potential for many hypotheses testing, with a bigger and more representative sample:
 - How the sentiment evolves during a crisis (lag or not),
 - How efficient are different countries' reactions,
 - Who really influences the global sentiment during a crisis
 - ...
- Comparison with past disease outbreaks could give us a better understanding of how to use sentiment as a prediction for crisis development

LIMITATIONS

- Old and static data for network analysis
- Limited data for tweets (maximum of 3,200 tweets)
- Non-representativeness of our sample and risk of bias
- Different languages, so difficulty of analysis
- Dynamic data, as the situation is not over



THANK YOU!

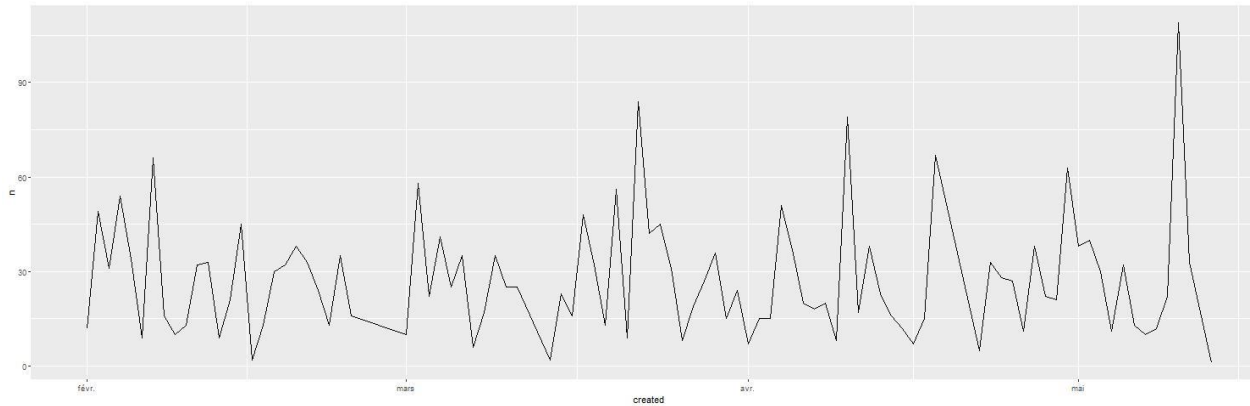
ANY QUESTION?

APPENDIX

Additional data

Donal Trump

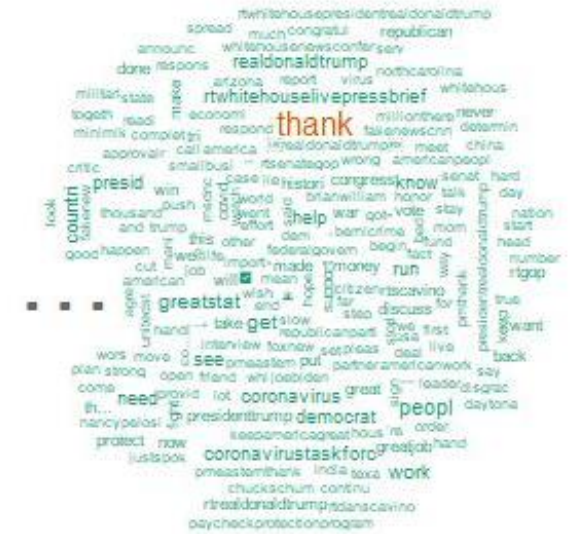
Wordclouds & Topics



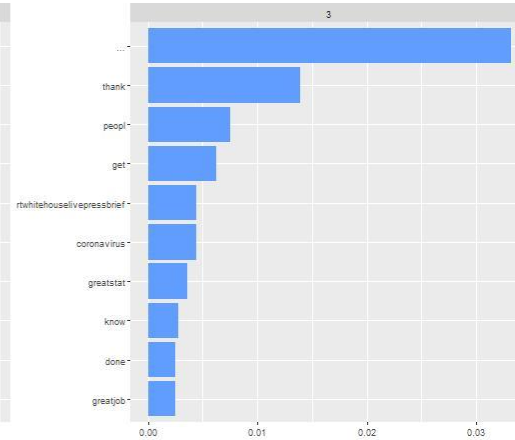
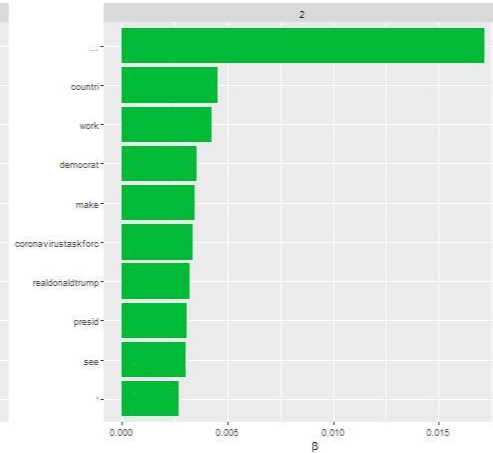
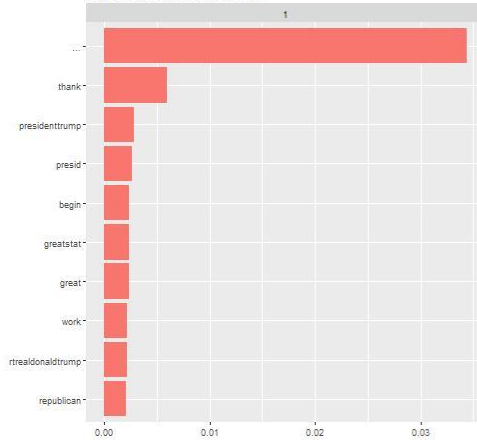
negative



positive



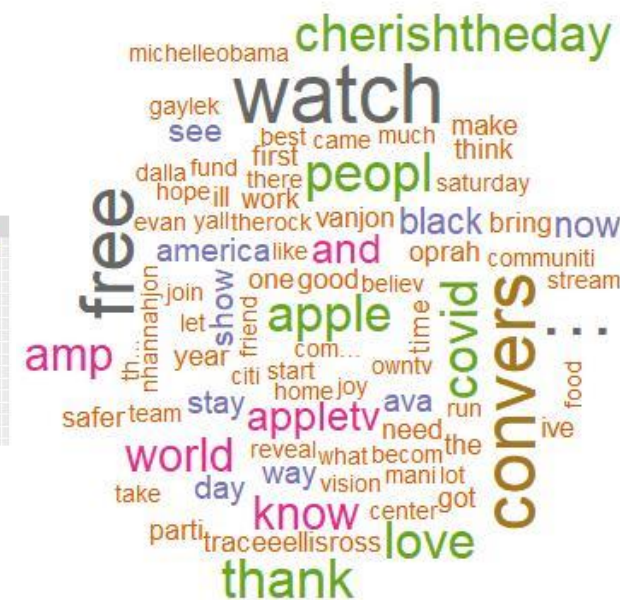
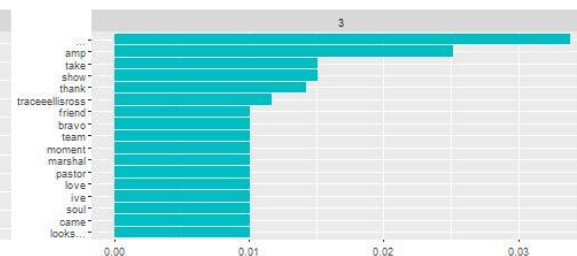
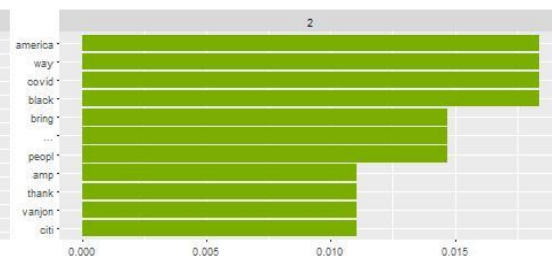
Top 10 terms in each LDA topic



negative



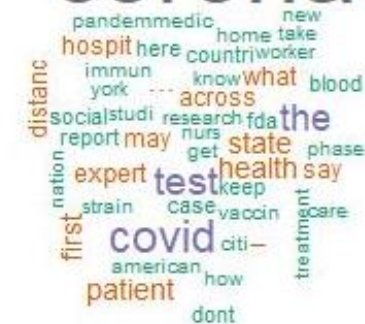
positive



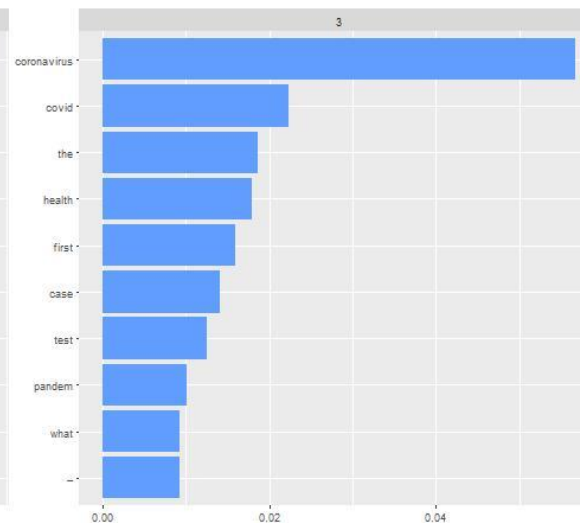
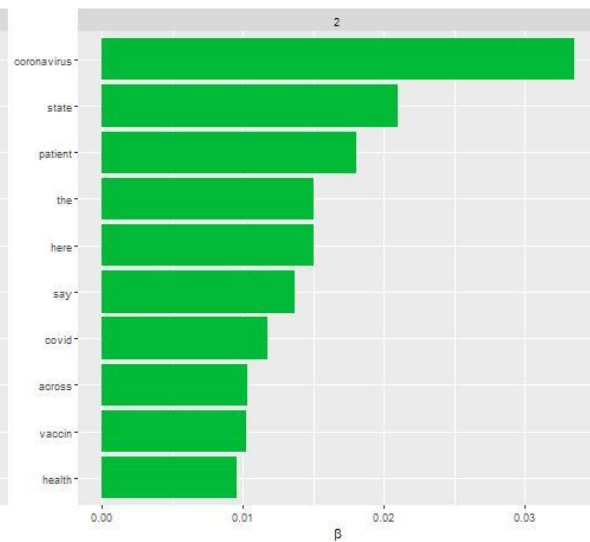
Month	Count
Jan	2
Feb	1
Mar	1
Apr	5
May	1

A word cloud featuring various words in shades of red, orange, and teal. The words include: worried, vulnerable, shortage, missed, frustrated, touted, fear, damaged, kill, limits, worst, horrific, alarm, illness, sore, ease, calm, safe, healthy, clarity, and effective. The words 'calm', 'healthy', and 'safe' are the largest and most prominent, rendered in a teal color, while the others are in various sizes and shades of red and orange.

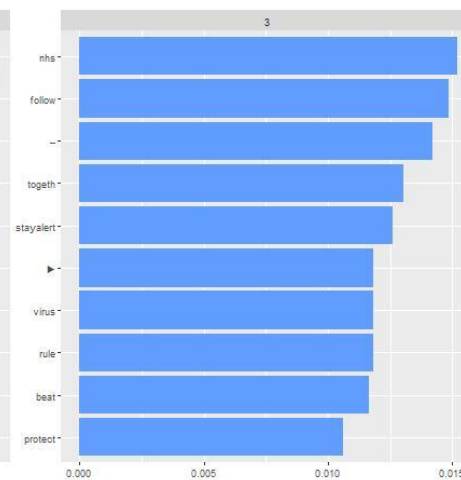
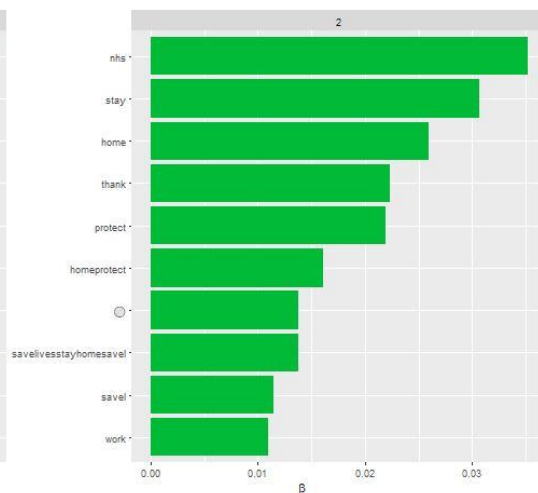
coronavirus



Word	Proportion
coronavirus	0.045
test	0.025
covid	0.020
distanc	0.015
may	0.012
countri	0.010
patient	0.008
expert	0.007
keep	0.005

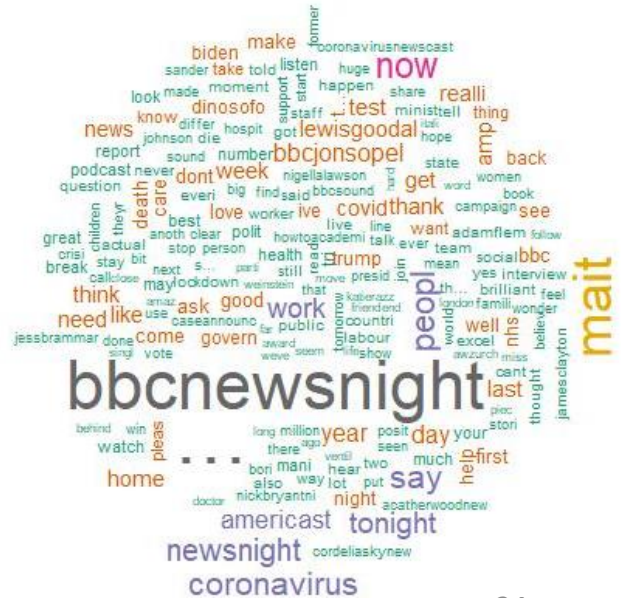


The graph displays the frequency of tweets containing the hashtag #COVID19. The x-axis represents the time of creation, with labels for February (fevr.), March (mars), April (avr.), and May (mai). The y-axis represents the count of tweets (n), ranging from 0 to 20. The data shows a relatively stable but fluctuating number of tweets (mostly between 1 and 5) from February through early April. A sharp increase occurs in late April, with a peak of approximately 15 tweets. Following this, the count drops but remains elevated, with another significant spike reaching over 20 tweets in early May, before settling back around 5 tweets by the end of the period shown.

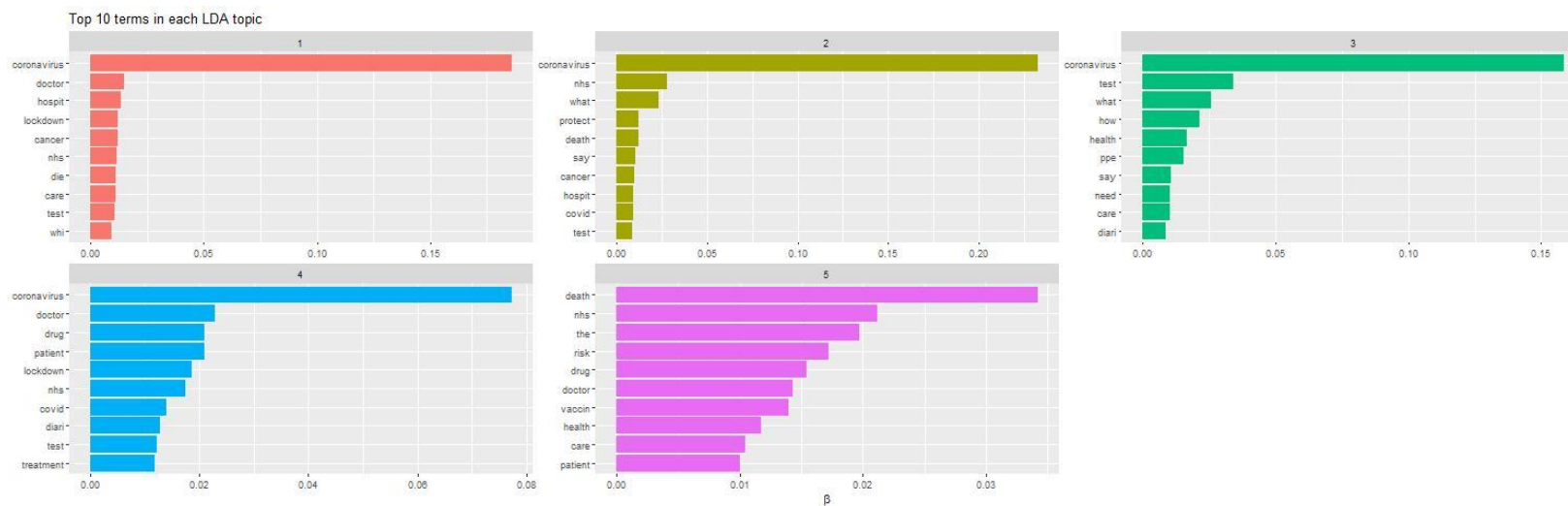
[illegible]

St.Gallen

The graph displays the number of tweets (n) on the y-axis (0 to 50) against the date created on the x-axis (fevr., mars, avril, mai). The data shows a highly volatile trend with several peaks. The highest peak occurs in late February, reaching nearly 55 tweets. Other notable peaks are in early April (around 48 tweets) and mid-March (around 35 tweets). The number of tweets generally decreases after the initial peak in February, with a secondary rise in April followed by a decline in May.



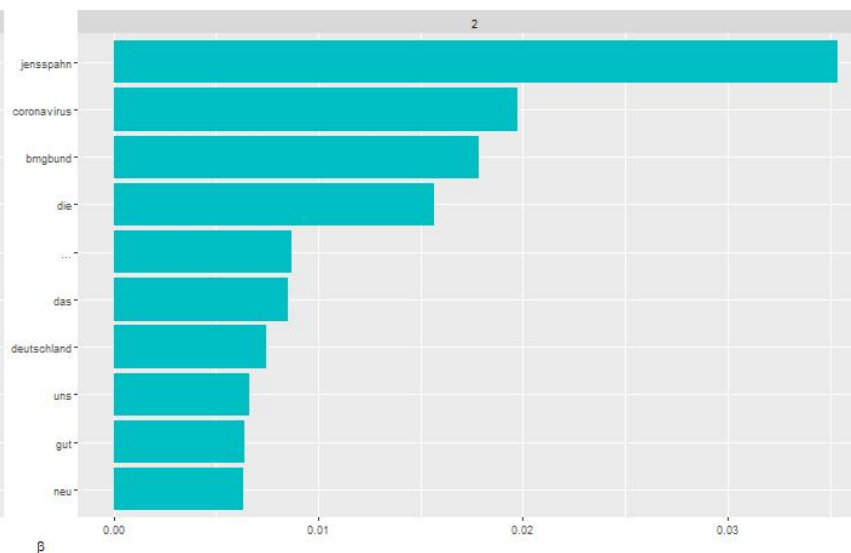
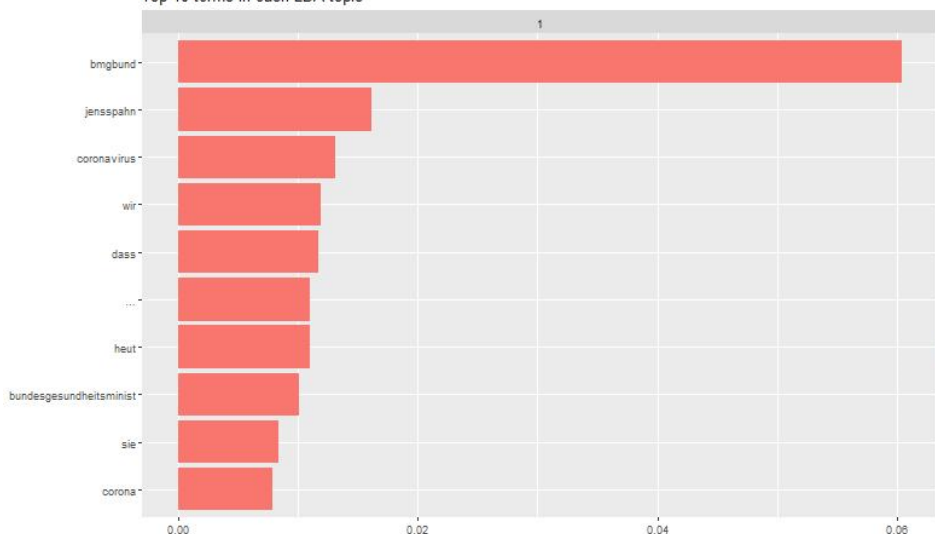
The line plot displays the number of tweets (n) on the y-axis against the creation date (created) on the x-axis. The y-axis has major ticks at 0, 2, 4, 6, and 8. The x-axis is labeled with months: fevr., mars, avr., and mai. The data shows a highly volatile trend. It starts in February with values between 1 and 3. In March, it fluctuates between 1 and 6. A significant peak of 8 occurs in late April. Following this, the values drop and then fluctuate between 2 and 5 through May.

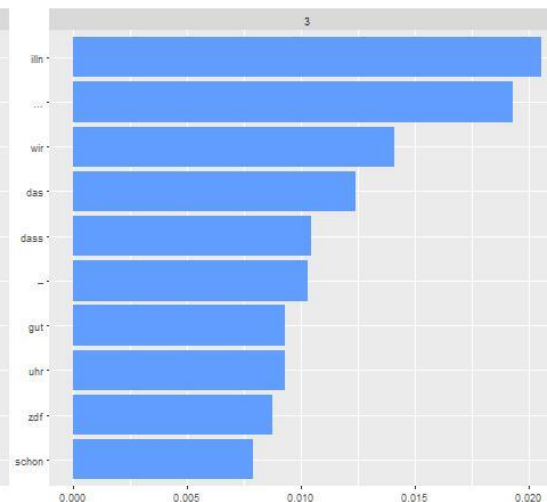
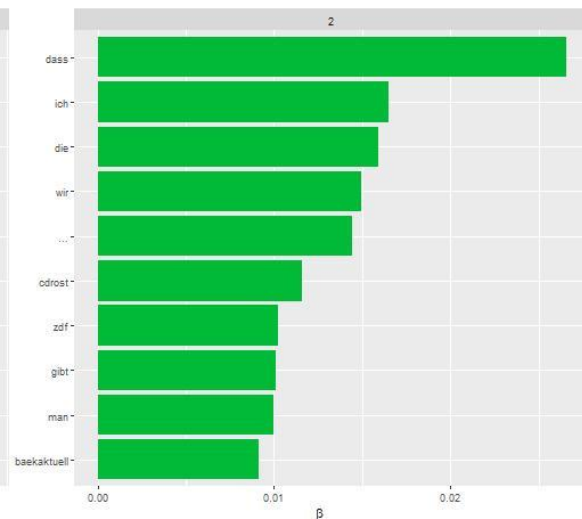
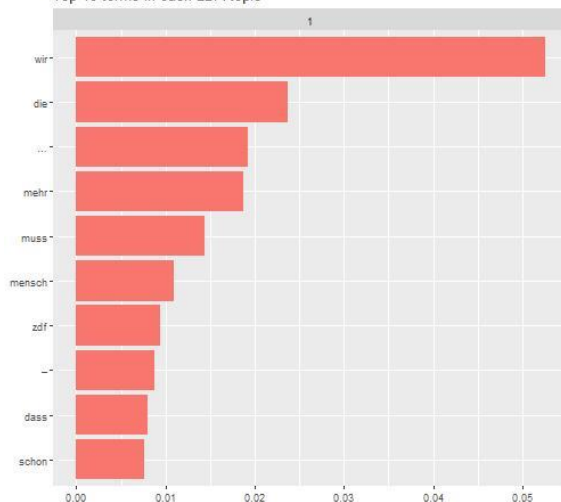


The line plot displays the frequency of tweets for the hashtag #BlackLivesMatter from February 1, 2020, to May 1, 2020. The y-axis, labeled 'n', represents the count of tweets, ranging from 0 to 35. The x-axis, labeled 'created', shows the timeline with major ticks for febr., mars, avr., and mai. The data shows a period of low activity in February, followed by a sharp increase in early March, peaking at approximately 25 tweets. After a decline, there is a second, even higher peak of nearly 35 tweets in early April. This is followed by a period of high volatility with multiple peaks between 10 and 20 tweets, and a final significant peak of about 20 tweets in late May.



pos



[illegible][illegible]

The line plot displays the number of tweets (n) over time (created). The y-axis represents the count of tweets, ranging from 0 to 20. The x-axis represents time, with labels for February (févr.), March (mars), April (avr.), and May (mai). The plot shows a highly volatile trend with several sharp peaks. Notable peaks occur in late February (reaching approximately 20), early April (reaching 20), and mid-April (reaching approximately 17). The number of tweets generally fluctuates between 5 and 15 for most of the period, with a slight downward trend towards the end of May.



Term	Proportion (approx.)
wir	0.042
du	0.020
die	0.018
ein	0.016
mehr	0.014
muss	0.012
zu	0.011
schon	0.010
zu	0.009
das	0.008

