TECHNISCHE UNIVERSITÄT ILMENAU
Institut für Praktische Informatik und Medieninformatik
Fakultät für Informatik und Automatisierung
Fachgebiet Datenbanken und Informationssysteme

Term Paper Research Seminar

Data Science in Social Media: Graph Analysis and Text Mining in Practice

Submitted by

Bipasha Roy, 63330
Suvarup Ghosh, 63345
Ender-Mark Merkinger, 62993
Hammad Tahir, 63741
Katrin Hess, 62687

Supervisor:

Prof. Emese Domahidi
Dr. Nadine Steinmetz

Ilmenau, 30/09/21

Table of Contents

Table of Figures

Introduction

To understand the impact of COVID-19, it is important to collect data. Data helps to identifying the most vulnerable demographic group and to recognize which treatments are effective and which are not. Therefore, collecting data is a vital process to understand further possibilities regarding COVID-19 (R. Tang,2021). In this study we only consider the citizen perspective of German people. (Lafaro 2020).

To frame the problem many conspiracy theories are spreading in Germany and based on that point, the Querdenker Movement (Lateral Thinking Movement) helps to spread these conspiracy theories with their protests. The Lateral Thinking Movement organizes Corona demonstrations against the Corona measures which are implemented by the German government. In this work we are focusing on the Pro-Querdenkers, a grouping of people who are provoking the citizens against COVID-19 measures and try to determine. We define protest as the number of tweets (Welle 2021).

The COVID-19 disease has attracted the attention on social media. People start to inform and exchange information on these platforms when an event happens. Twitter provides a great possibility to track the dynamics of the Coronavirus (Aguilar-Gallegos et al. 2020).

The aim of the study is to build a good dataset to represent the Querdenker in the German population. We focus on the Pro-Querdenker as representation of the freedom restriction protests to achieve our dataset. The Querdenker protests suggest that there is a link between the government measures and Querdenker protests, that means the Querdenker protests results from the government measures. From this connection a research question can be derived: Is there a relationship between the announcement of Corona measures by the German government (freedom reduction) and the Querdenker protests from March 2020 until March 2021?

The process of building a dataset is restricted to identifying relevant twitter accounts and hashtags, to gather the data and of course to clean the data (Rakshana B. S, 2021; A. Kumar, 2021). The process of analysing the data is not included in our work. Twint is an open-source Python Library and was used to gather the data. It is a tool for Twitter scrapping. This tool allows to extract data from Tweeter profiles and accounts. With Twint you can scrape Tweets from specific users, search Tweets from certain hashtags, topics, and trends. It also allows to sort information from Tweets like a phone number or E-Mail. The reason why we did not used Twitter API is that Twitter API has restrictions on the number of Tweets. It is limited on the last 3200 Tweets (Jose 2021).

Following on from the introduction, the next chapter will give an insight in the related work. The approach will explain our procedure and how we select the Twitter accounts and hashtags. The last chapter deals with the conclusion and whether we have been able to answer our objectives adequately.

Related Work

The research proposal and data retrieval in this paper bases on a literature review. For this purpose, the authors conducted a literature search on *Google Scholar*. The search-term ‘Querdenken AND Corona’ was used to identify relevant literature. No filters were applied. In total, 600 search results appeared. The ten first results were revisited. Four of these papers were available and suit our literature review.

The reviewed papers contain studies about the Querdenken movement itself. Especially, it was analysed who participates in the Corona protests and why. This is important to know for our study because these factors determine if the Querdenken movement is a Corona specific occurrence or if the movement might sustain a longer period. Further, such information can be set into relation to our analysis and might be able to explain the results. First, it should be noted that the heterogeneity of the Querdenken movement is undeniable (Grande et al., 2021; Koos, 2021; Pantenburg et al., 2021). The following paragraphs elaborate on this.

Studies found that the protest participants are older people with slightly more females participating (Koos, 2021). Koos (2021) notes that more than 50% of the participants are older than 50 years. This is in contast to Koos & Binder (2021), who report a stronger affiliation of younger people towards the Querdenken movement than older people as well as a stronger support from men.

Interestingly, the authors of the papers conclude that participants have a higher educational level (Grande et al., 2021; Koos, 2021; Koos & Binder, 2021). Especially, most participants earned a Realschulabschluss or an academic degree. These are interesting findings because in media outlets the Querdenken movement is portrayed as a movement with a lower educational level. However, the educational level changed over time. Later, the general educational level of the Querdenken movement participants was lower than when the movement started (Grande et al., 2021).

Further, it was found that the participants of the Querdenken movement are most likely to be employed (Koos, 2021). Koos (2021) notes that the level of self-employed protest participants is higher than in the German population. Only 2% are unemployed (Koos, 2021). It is widely assumed that the Querdenken movement has its origins in far-right corner of the political spectrum. However, the findings of the studies are a different one. Whereas only 2% identify with the *Alternative für Deutschland (AfD),* a right-wing party, the majority (55%) feel not properly represented by any party (Koos, 2021) but originate from the political middle (Grande et al., 2021). These findings are not shared between all the authors. Another study find that the affiliation with the AfD is overrepresented compared to the AfD supporters in the general population (Grande et al., 2021). But the extremes are still a minority (Grande et al., 2021). Koos (2021) sees the origin of this in the lack of trust in German institutions. However, he also finds that 94% reject a dictatorship. Many Querdenken protestors don’t feel properly represented by the German government. It is important to note that over time a shift of the movement to the right is recognizable (Grande et al., 2021).

Findings show that most participants are more aware of the negative consequences on the macroeconomic level rather than their personal problems (Koos, 2021; Koos & binder, 2021).

Especially, the participants are concerned about job security, financial security, critical situations of families and the restrictions of basic laws in Germany. He notes that participants protest as deputies of the affected above-mentioned social groups. Further, 73% of the participants are socially engaged and help people in need e. g. helping with the grocery shopping. These findings contrast with Grande et al. (2021). Grande et al. (2021) conclude that people who are stronger affected by the restrictions have a higher understanding for protests. It suggests that conspiracy theories are widely accepted in the Querdenken movement e. g. that influential businessmen want to force the public to get vaccinated or that scientists hide prove to manipulate the public (Grande, 2021; Koos, 2021; Koos & Binder, 2021). 46% of the participants of the Querdenken movement are sceptical of the validity of the public voices which claim that Corona is a dangerous virus (Koos & Binder, 2021; Koos, 2021). Based on these findings, Grande et al. (2021) suggest a potential of radicalization of the Querdenker protestants.

The media consumption of Querdenker also plays an important role. 90% of participants use the Internet, 52% use Telegram, Whatsapp or family and friends (Koos 2021). In general, protestants are mistrustful of media, politics, and science (Koos & Binder, 2021; Pantenburg et al., 2021). Specifically, participants see a relation of scientific publications and political and economic interests and educate themselves about new findings (Pantenburg et al., 2021). Most important to the participants of Corona protests are the abolishment of Corona measures in Germany (Grande et al., 2021; Koos 2021; Koos & Binder, 2021; Pantenburg et al., 2021). These measures are perceived as too strong and not legitimate (Grande et al., 2021; Koos 2021; Koos & Binder, 2021). Protestants recognize public censorship (Pantenburg et al., 2021). Based on this, they demand basic rights and open public discussions without being criticised (Koos, 2021). 75% don’t agree with further radicalisation (Koos 2021). Critics argue that the Querdenken movement isn’t a rational one but bases on emotions (Pantenburg et al., 2021).

To conclude, the literature suggests that Covid restrictions in Germany are the most important reason why participants protest. Therefore, it can be assumed that the Querdenken protests are stronger when restrictions are implemented. In turn, it can be assumed that lower restrictions are associated with less protest by the Querdenken movement.

Approach

In our approach, we have used an open-source Python Library named *Twint* for gathering data from selected Twitter accounts and we also have used *nest_asyncio* for loop patching. Twint is an advance tool for Twitter scrapping and this allows data scraping from Tweeter accounts and profiles in relatively easier way. Twint utilizes Twitter's search operators to let us scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets.

Algorithm

- STEP 1: Import necessary packages (twint, nest_asyncio etc.)
  - STEP 1.1: Search data with respect to the pro-Querdenken usernames and make a loop so that we can collect the data from all the usernames in one shot
  - STEP 1.2: From our collected hashtags, define the searchlist
  - STEP 1.3: Select the language in which it will collect data
  - STEP 1.4: Mention the timeline and the format of storing the data (.csv, .xlsx etc.)
- STEP 2: Classification of Data
  - STEP 2.1: Based on Features
  - STEP 2.2: Based on Pro-Querdenken twitter accounts
- STEP 3: Data Cleaning
  - STEP 3.1: Removal of punctuations
  - STEP 3.2: Removal of retweets 'RT'
  - STEP 3.3: Removal of stop words
  - STEP 3.4: Removal of tweet repetitions
  - STEP 3.5: Removal of non-related tweets and based on that, further classification
  - STEP 3.6: Identify Hashtags which are associated with anti-Querdenker (e. g. #covidiots,…)
  - STEP 3.7: Create new columns which count the words supportive of pro-Querdenker (positive words) and words which are not supportive of pro-Querdenker (negative words)
  - STEP 3.8: Manually delete Tweets which are not supportive of Querdenken and couldn‘t be identified through 3.6 and 3.7.
- STEP 4: Data Pre-processing
  - STEP 4.1: Simplification of the data
  - STEP 4.2: Extract fact-checkable tweets
  - STEP 4.3: Inclusion of new columns (where required)
  - STEP 4.4: Inclusion of class and ranks to prioritize which tweets are relevant to our research question
  - STEP 4.5: Ranking of tweets based on: (a) Numbers and Quantity, (b) Location, (c) Organiation Names, (d) Resources Present, (e) Contact Information

Data Collection

Hashtags are a combination of keywords or phrases preceded by the # symbol, excluding any spaces or punctuations. As an example, if we put the # symbol ahead the words "social media," it becomes a hashtag #socialmedia (K. Rahul, 2021). Hashtags help group Tweets and conversations round the same topic so people can easily find and follow what interests them. So, when someone clicks on or searches a particular hashtag, they'll be able to find all the profiles and public posts that use that hashtag. As an example, if we look for #homecooking, we will see many posts of home-cooked meals. Hashtags essentially make your content searchable and grouped into relevant topics. By using hashtags to hitch in on subjects like industry discussion or Twitter trends, it'll appear when people search that hashtag. One can get sight to a wider audience outside our following and boost the discoverability and reach of the content.

We have only considered the tweets of "*Pro-Querdenkers*" in this study. We visited the official twitter page of *"Querdenken Movement"* where they listed all regional movements including their Twitter accounts. From there, we gathered significant Hashtags and we made a Hashtag set from it. Also, from the tweets of Querdenken page, we analysed the tweets based on three things: place, activists, and keywords (*Querdenken, lockdown, COVID-19, Coronavirus* etc.). And based on these three conditions, we were able to identify the accounts of *"Pro-Querdenkers".* Now, for the classification (G. M. Raza, 2021), we could use the keywords for segregation but as we already identified the accounts of Pro-Querdenkers, therefore the data that we collected is already classified.

- **Data Collection Using Twint**

  We used an open-source Python Library named Twint for gathering data from selected Twitter accounts. Twint is an advance tool for Twitter scrapping. It allows data scraping from Tweeter accounts and profiles. Twint utilizes Twitter's search operators to allow us for scraping Tweets from specific users, scrape Tweets referring to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers.

  Above all, Twint has these major benefits:

  1. Twitter API has restrictions to scrape only the last 3200 Tweets. But Twint can fetch most Tweets.
  2. Setting up is quick as there is no hassle of setting up Twitter API.
  3. It can be used anonymously without Twitter sign-up.
  4. It's free! So, no pricing limitations.
  5. Provides easy to use options to store scraped tweets into different formats — CSV, JSON, SQLite, and Elasticsearch.

- **Data Collection Using Specific Search Strings**

  We also used some specific search strings to scrape the tweets. We can specify different search strings to filter tweets from the web. At first, we have used *'Querdenken'*, *'Covidioten'*, *'Querdenken ist am Ende'*, *'Querdenken Demo'*, *'coronazis'* and *'Covid-Aktivisten'* and then all the datasets are merged to induce an overall dataset.

- **Using Specific Timeline**

  We mentioned a specific timeline in case of collecting the data. So, since and Until feature is employed to line the date range from March 2020 to March 2021. Next, we gathered some important people and their Tweeter Usernames of the ***'Querdenken'*** group. The dataset that we have collected from using the keyword ***'Covid-Aktivisten'***, gives us a list of all user id who is currently involved with Querdenken.

- **Scrape the Tweets of Users**
  In the next step of work, we used the user id's and collect data from their twitter account by using the identical hashtag set. So, after collecting all the data, we merged all these data to get the final dataset.

- **Using Places**
  On the official internet page of "Querdenken Movement", all regional querdenken accounts were listed. Therefore, we visited the official page as well as all the regional querdenken accounts pages and found out about regional movements including their respective Twitter accounts.

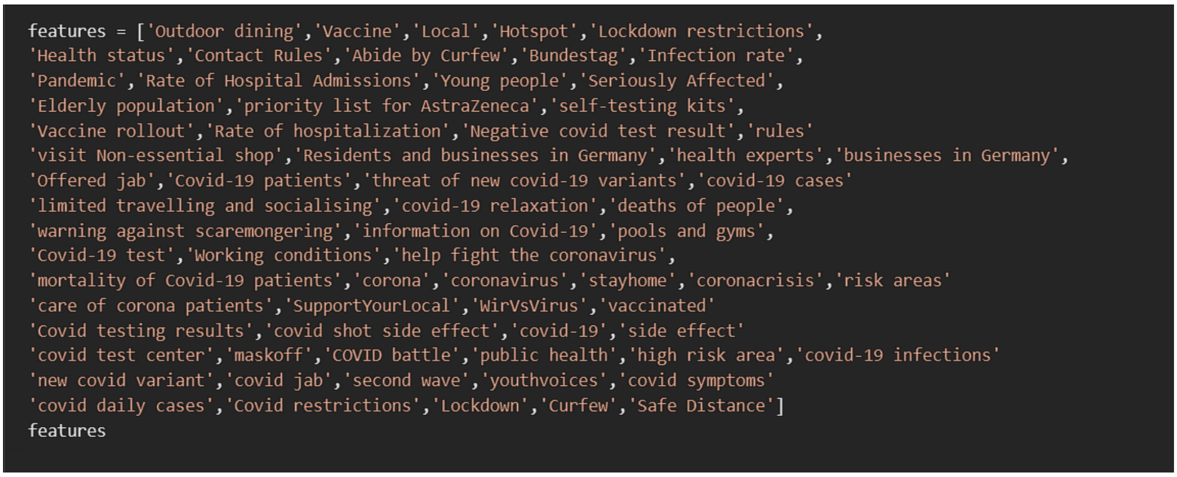
Data Classification

Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently (Jelodar H, 2020). On a basic level, the classification process makes data easier to locate and retrieve. Data classification is of particular importance when it comes to risk management, compliance, and data security.

We have followed two major procedures for data classification:

(1) Feature selection:

Feature selection is one of the most important steps in data mining and knowledge discovery (S. Roy, 2020). The idea is to select the best features that improve the classification accuracy. So, we already have our feature set and based on that, we are segregating and then performing classification. (Figure 1)

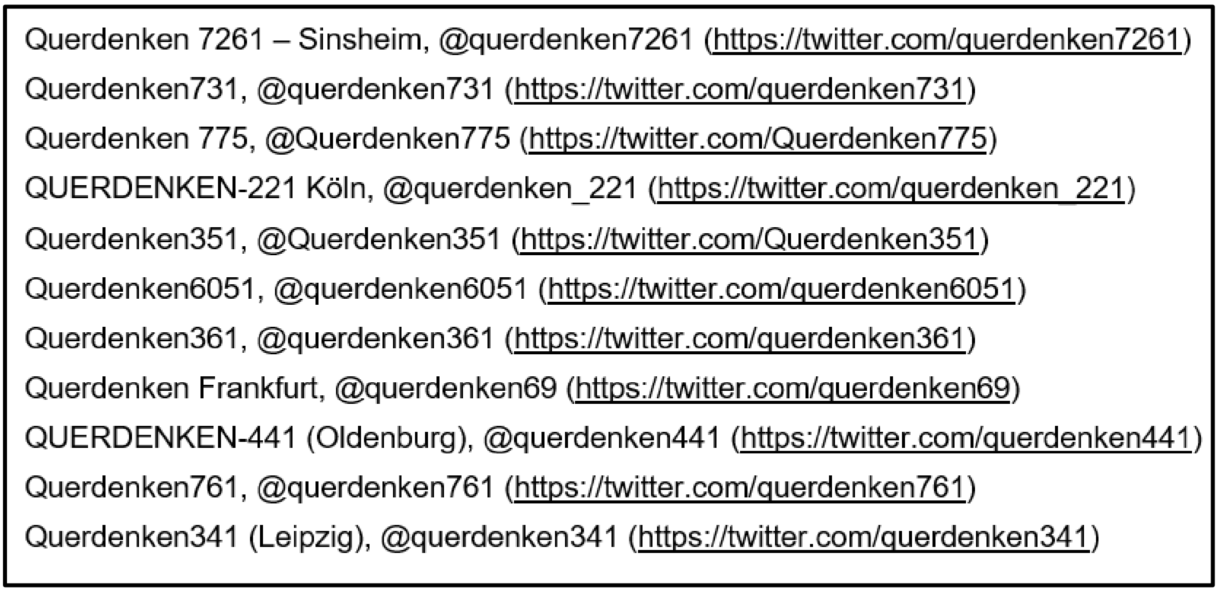


```
features = ['Outdoor dining','Vaccine','Local','Hotspot','Lockdown restrictions',
'Health status','Contact Rules','Abide by Curfew','Bundestag','Infection rate',
'Pandemic','Rate of Hospital Admissions','Young people','Seriously Affected',
'Elderly population','priority list for AstraZeneca','self-testing kits',
'Vaccine rollout','Rate of hospitalization','Negative covid test result','rules'
'visit Non-essential shop','Residents and businesses in Germany','health experts','businesses in Germany',
'Offered jab','Covid-19 patients','threat of new covid-19 variants','covid-19 cases'
'limited travelling and socialising','covid-19 relaxation','deaths of people',
'warning against scaremongering','information on Covid-19','pools and gyms',
'Covid-19 test','Working conditions','help fight the coronavirus',
'mortality of Covid-19 patients','corona','coronavirus','stayhome','coronacrisis','risk areas'
'care of corona patients','SupportYourLocal','WirVsVirus','vaccinated'
'Covid testing results','covid shot side effect','covid-19','side effect'
'covid test center','maskoff','COVID battle','public health','high risk area','covid-19 infections'
'new covid variant','covid jab','second wave','youthvoices','covid symptoms'
'covid daily cases','Covid restrictions','Lockdown','Curfew','Safe Distance']
features
```

Figure 1: Selected Features for data classification

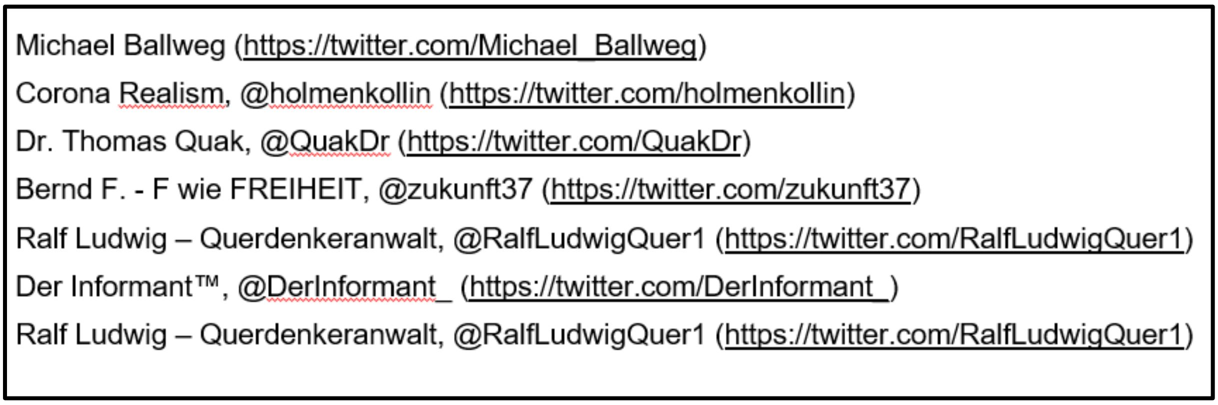
(2) Pro-Querdenken twitter accounts:

"Pro-Querdenkers" are only considered in this study. We visited the official twitter page of "***Querdenken Movement***" where they listed all regional movements including their Twitter accounts (Figure 2 & Figure 3). From there, we gathered significant Hashtags and we made a Hashtag set from it. Also, from the tweets of Querdenken page, we analysed the tweets based on three things: place, activists, and keywords (Querdenken, lockdown, COVID-19, Coronavirus etc.). And based on these three conditions, we were able to identify the accounts of "***Pro-Querdenkers***". Now, for the classification, we could use the keywords for segregation but as we already identified the accounts of Pro-Querdenkers, therefore the data that we collected is already classified.



Querdenken 7261 – Sinsheim, @querdenken7261 (https://twitter.com/querdenken7261)
Querdenken731, @querdenken731 (https://twitter.com/querdenken731)
Querdenken 775, @Querdenken775 (https://twitter.com/Querdenken775)
QUERDENKEN-221 Köln, @querdenken_221 (https://twitter.com/querdenken_221)
Querdenken351, @Querdenken351 (https://twitter.com/Querdenken351)
Querdenken6051, @querdenken6051 (https://twitter.com/querdenken6051)
Querdenken361, @querdenken361 (https://twitter.com/querdenken361)
Querdenken Frankfurt, @querdenken69 (https://twitter.com/querdenken69)
QUERDENKEN-441 (Oldenburg), @querdenken441 (https://twitter.com/querdenken441)
Querdenken761, @querdenken761 (https://twitter.com/querdenken761)
Querdenken341 (Leipzig), @querdenken341 (https://twitter.com/querdenken341)

Figure 2: Regional Querdenken Twitter handles



Michael Ballweg (https://twitter.com/Michael_Ballweg)
Corona Realism, @holmenkollin (https://twitter.com/holmenkollin)
Dr. Thomas Quak, @QuakDr (https://twitter.com/QuakDr)
Bernd F. - F wie FREIHEIT, @zukunft37 (https://twitter.com/zukunft37)
Ralf Ludwig – Querdenkeranwalt, @RalfLudwigQuer1 (https://twitter.com/RalfLudwigQuer1)
Der Informant™, @DerInformant_ (https://twitter.com/DerInformant_)
Ralf Ludwig – Querdenkeranwalt, @RalfLudwigQuer1 (https://twitter.com/RalfLudwigQuer1)

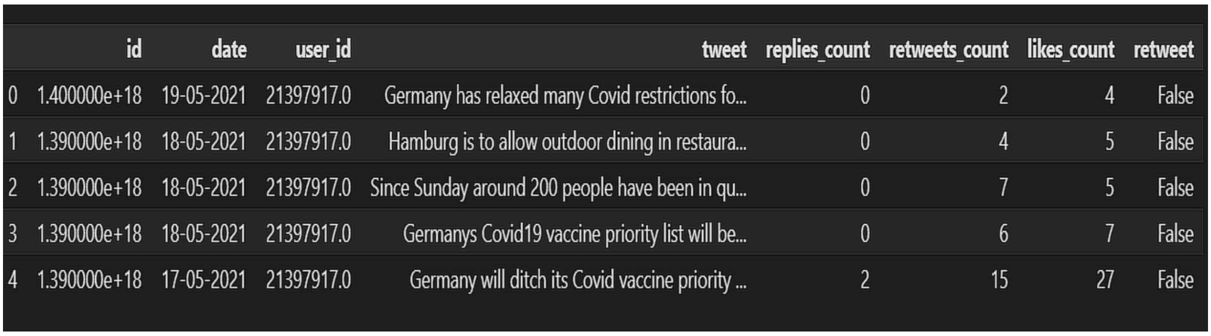
Figure 3: Querdenken Usernames

Data Pre-Processing Procedure

Data pre-processing is a crucial step that helps enhance the standard of data to help the extraction of meaningful insights from the data. Data pre-processing refers to the technique of preparing (cleaning and organizing) the raw information to make it suitable for further tasks. In simple words, data pre-processing is a data mining technique that transforms raw information into a noticeable and readable format.

There are many ways to perform data pre-processing. Our approach is explained in three simple steps:

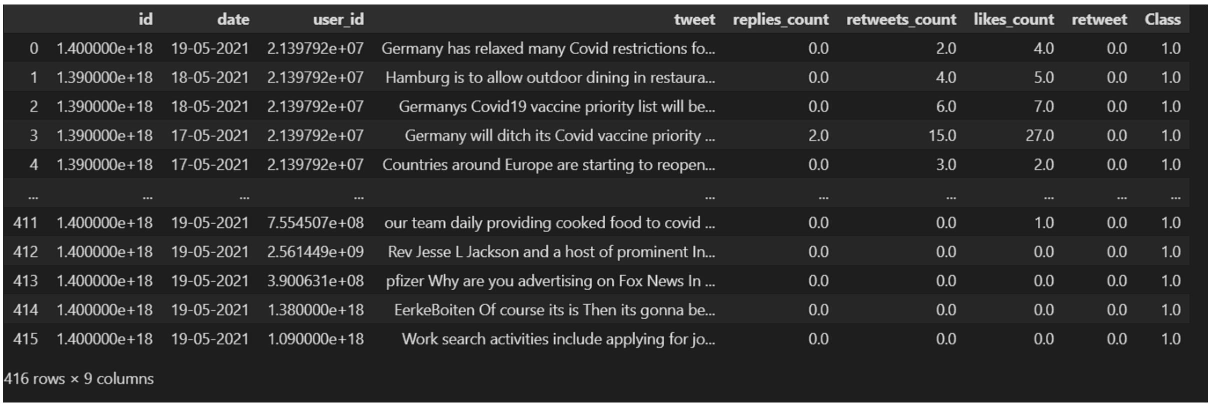
(1) Simplification of data:
In our dataset (as shown in Figure 4), primarily there are a lot of data which were irrelevant or not necessary for the further tasks. So, from the dataset, we have selected precise data columns which provides us valid information required to answer our question. Only these columns are selected for the further processing (Figure. 4).



	id	date	user_id	tweet	replies_count	retweets_count	likes_count	retweet
0	1.400000e+18	19-05-2021	21397917.0	Germany has relaxed many Covid restrictions fo...	0	2	4	False
1	1.390000e+18	18-05-2021	21397917.0	Hamburg is to allow outdoor dining in restaura...	0	4	5	False
2	1.390000e+18	18-05-2021	21397917.0	Since Sunday around 200 people have been in qu...	0	7	5	False
3	1.390000e+18	18-05-2021	21397917.0	Germanys Covid19 vaccine priority list will be...	0	6	7	False
4	1.390000e+18	17-05-2021	21397917.0	Germany will ditch its Covid vaccine priority ...	2	15	27	False

Figure 4: Simplified data for further processing

(1) Extract Fact-checkable Tweets:
In our study, we have used tweets from Twitter as it is very popular nowadays for microblogging. We have addressed this problem by separating all the fact-checkable tweets from the non-fact-checkable once based on some features. To generate more accurate result, we have initially classified all the tweets then rank them. However, this is a very challenging task to handle as because no pure data can be easily obtained from Twitter. It consists of various noises such as url, punctuation, re-tweets, stopwords, etc. which is irrelevant to us (Figure 5).



	id	date	user_id	tweet	replies_count	retweets_count	likes_count	retweet	Class
0	1.400000e+18	19-05-2021	2.139792e+07	Germany has relaxed many Covid restrictions fo...	0.0	2.0	4.0	0.0	1.0
1	1.390000e+18	18-05-2021	2.139792e+07	Hamburg is to allow outdoor dining in restaura...	0.0	4.0	5.0	0.0	1.0
2	1.390000e+18	18-05-2021	2.139792e+07	Germanys Covid19 vaccine priority list will be...	0.0	6.0	7.0	0.0	1.0
3	1.390000e+18	17-05-2021	2.139792e+07	Germany will ditch its Covid vaccine priority ...	2.0	15.0	27.0	0.0	1.0
4	1.390000e+18	17-05-2021	2.139792e+07	Countries around Europe are starting to reopen...	0.0	3.0	2.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...
411	1.400000e+18	19-05-2021	7.554507e+08	our team daily providing cooked food to covid ...	0.0	0.0	1.0	0.0	1.0
412	1.400000e+18	19-05-2021	2.561449e+09	Rev Jesse L Jackson and a host of prominent In...	0.0	0.0	0.0	0.0	1.0
413	1.400000e+18	19-05-2021	3.900631e+08	pfizer Why are you advertising on Fox News In ...	0.0	0.0	0.0	0.0	1.0
414	1.400000e+18	19-05-2021	1.380000e+18	EerkeBoiten Of course its is Then its gonna be...	0.0	0.0	0.0	0.0	1.0
415	1.400000e+18	19-05-2021	1.090000e+18	Work search activities include applying for jo...	0.0	0.0	0.0	0.0	1.0

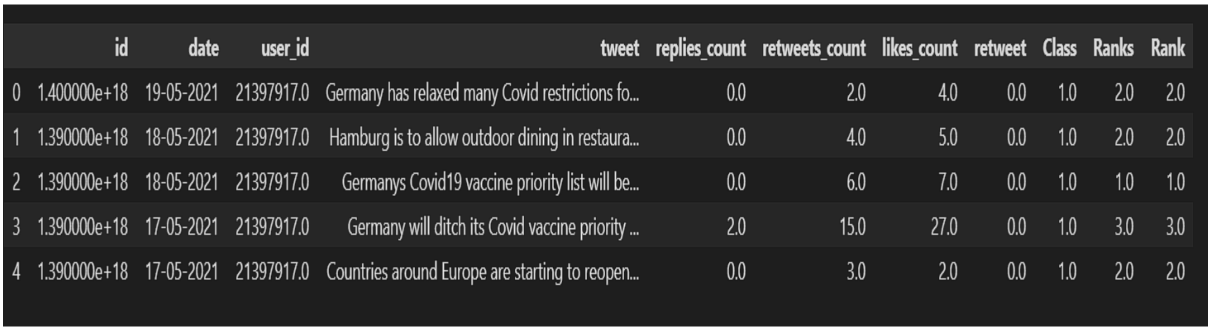
416 rows × 9 columns

Figure 5: Fact-checkable Tweets

(2) Inclusion of Class and Ranks:
In the rank-based approach, we've categorized the fact-checkable tweets based on the ranks (Figure 6). The ranks of the tweets are given supported the data content within the tweets (F. Es-Sabery et al, 2021). The content is checked using NER Tagger and RE (Regular Expression) to seek out the presence of: (a) Numbers and Resources, (b) Contact Information, (c) Organization Names, (d) Location.

As RE (Regular Expression) is majorly used for pattern identification like feature extraction, OCR identification, etc. Similarly, we've got used this process for matching of numbers and phone number in tweets to spot and verify the importance of the tweets (S. Roy, 2020).

NER Tagger (Name Entity Recognizer) is JAVA implemented to extract and label a sequence of text from datasets with any extensions. Names particularly consists of 3 main classes *viz.* PERSON, LOCATION, ORGANIZATION.



	id	date	user_id	tweet	replies_count	retweets_count	likes_count	retweet	Class	Ranks	Rank
0	1.400000e+18	19-05-2021	21397917.0	Germany has relaxed many Covid restrictions fo...	0.0	2.0	4.0	0.0	1.0	2.0	2.0
1	1.390000e+18	18-05-2021	21397917.0	Hamburg is to allow outdoor dining in restaura...	0.0	4.0	5.0	0.0	1.0	2.0	2.0
2	1.390000e+18	18-05-2021	21397917.0	Germanys Covid19 vaccine priority list will be...	0.0	6.0	7.0	0.0	1.0	1.0	1.0
3	1.390000e+18	17-05-2021	21397917.0	Germany will ditch its Covid vaccine priority ...	2.0	15.0	27.0	0.0	1.0	3.0	3.0
4	1.390000e+18	17-05-2021	21397917.0	Countries around Europe are starting to reopen...	0.0	3.0	2.0	0.0	1.0	2.0	2.0

Figure 6: Filtered dataset with Class and Rank

Data Filtering

Data filtering or data cleaning is that the method of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or dataset and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

After cleansing, an information set should be to keep with other similar data sets within the system. The inconsistencies detected or removed may are originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of comparable entities in numerous stores. Data cleaning differs from data validation and during this validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of information.

Our approach of data cleaning can be explained within the following:

(1) **Removal of Punctuation:**
Text cleaning or Text pre-processing is a crucial step when we are working with text in Natural Language Processing (NLP). In real-life human writable text data contain various words with the incorrect spelling, short words, special symbols, emojis, etc. and that we must clean this sort of noisy text data before going further.

(2) **Removal of Retweets:**
In the twitter datasets, there is also other information as retweets. So, all of this should be ignored and removed from the dataset because that information makes no sense in our work.

(3) **Classification (i.e., removal of non-related data/tweets):**
The data that we gathered from Twitter also consists of some non-related data tweets i.e., the information is irrelevant to our research question. Therefore, we will classify those data as class 0 and then we are able to remove those as well.

(4) **Removal of stop-words:**
Stop words are generally thought to be a "single set of words". We would not want these words taking over space in our database. For this using NLTK and applying a "Stop Word Dictionary". So, the stop words are removed as they are not useful.

(5) **Removal of tweet repetitions:**
We should also remove the tweet repetitions because we have already got the information once. So, similar data but more than once; just make the general procedure more inefficient and time consuming. Therefore, tweet repetitions are removed.

Evaluation

Our topic of research is to find Citizen's Perspective in Germany regarding COVID 19. To analyze the people's perspective, we compare two famous platforms for gathering the user's data which are Twitter and Reddit.
Tools used for respective platforms.

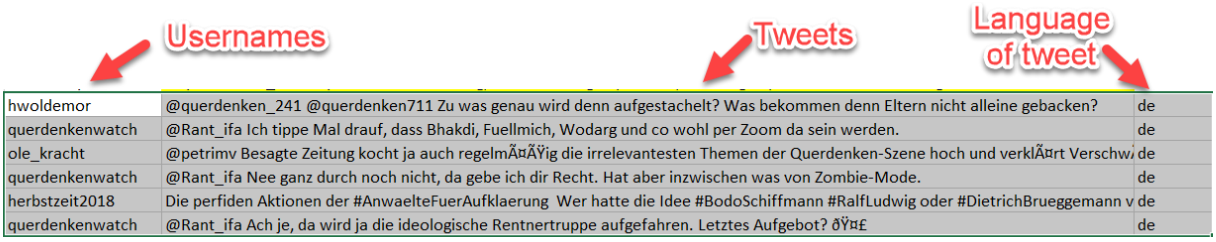
- For Scrapping Reddit, we use PRAW.
- For Scrapping Twitter data, we use Twint.

Limitations with Reddit (PRAWN)

PRAW is a python wrapper for Reddit API, which we can use to scrap posts subreddits. But there are some limitations, we cannot define the number of 'hashtags, usernames and keywords' to retrieve the related data which is the biggest hurdle to obtain clean, related and clear data on which analysis can be done for our studies.

Twint

On the other hand, Twint is a useful open-source python library in which we can define number of keywords, usernames, and dates to get precise and related data. Below is a selected small sample of the extracted dataset sample obtained from using Twint (Figure 7).



hwoldemor	@querdenken_241 @querdenken711 Zu was genau wird denn aufgestachelt? Was bekommen denn Eltern nicht alleine gebacken?	de
querdenkenwatch	@Rant_ifa Ich tippe Mal drauf, dass Bhakdi, Fuellmich, Wodarg und co wohl per Zoom da sein werden.	de
ole_kracht	@petrimv Besagte Zeitung kocht ja auch regelmÃ¤ÃŸig die irrelevantesten Themen der Querdenken-Szene hoch und verklÃ¤rt Verschwi	de
querdenkenwatch	@Rant_ifa Nee ganz durch noch nicht, da gebe ich dir Recht. Hat aber inzwischen was von Zombie-Mode.	de
herbstzeit2018	Die perfiden Aktionen der #AnwaelteFuerAufklaerung Wer hatte die Idee #BodoSchiffmann #RalfLudwig oder #DietrichBrueggemann v	de
querdenkenwatch	@Rant_ifa Ach je, da wird ja die ideologische Rentnertruppe aufgefahren. Letztes Aufgebot? ðŸ¤£	de

Figure 7: Tweeter data sample

Limitations with Twitter

Even Twitter is having some limitations, which is difficult to omit. We only used Twitter, because they offer comments of pro-Querdenker on official accounts (more representative than retrieving data from several sources, e. g. the same person comments on Twitter and Facebook similar content on the same day). Not all sub-movements are found on Twitter (e. g. Ärzte für Aufklärung)

Conclusion and Outlook

This paper describes a step-by-step guide how we retrieved data from Twitter and how we cleaned it. We suggest that the Querdenken movement is a phenomenon which has a stronger support when more restrictions are implemented. On the other hand, we assume that the support for Querdenken protests is lower when less or no restrictions are in place. Upon this, the current literature on the Querdenken movement was revised. The literature review suggests that, indeed, Covid restrictions by the government are the main cause for protests. This correlation is expected to be seen when the analysis of the data is performed. To analyse our dataset, we assume a correlation between the announcement of Corona restrictions by the German government and the amounts of tweets on Twitter. We define a higher number of tweets as a stronger protest against the Corona measures.

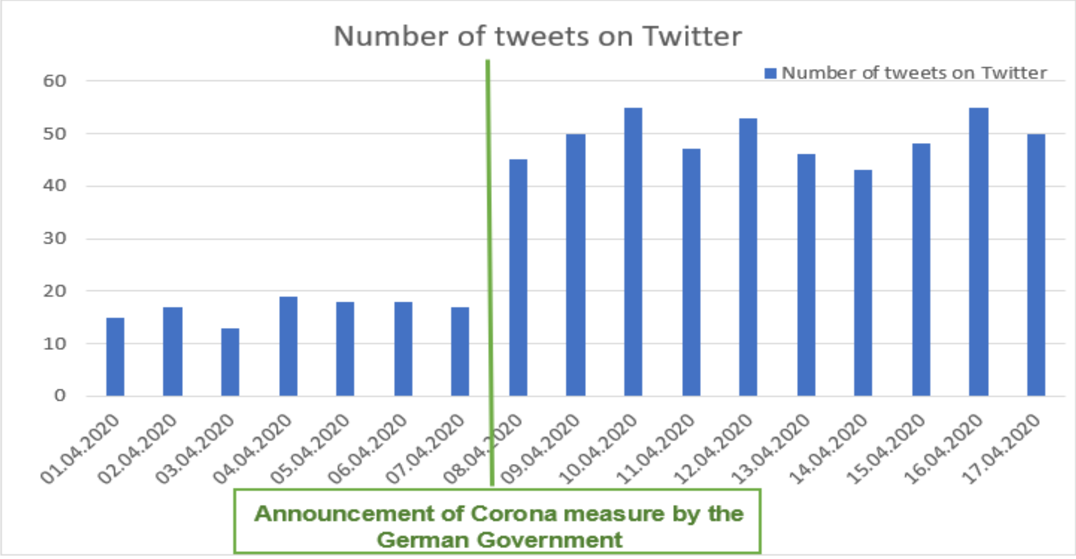


Figure 8: Example of a possible analysis solution (note: fictional). Timeframe snippet from April 1st, 2020, to April 17th 2020.

Figure 8 shows a possible solution of the future analysis. On the x-axis, the dates are presented. In this specific and fictional example, it shows a timeframe from April 1st 2020 to April 17th 2020. As mentioned, our timeframe from the retrieved data comprehends from March 2020 to March 2021. The y-axis represents the number of tweets per day. Here, the number of tweets from April 1st until April 7th 2020 is quite low. Then, a Corona measure is implemented by the German government between April 7th and April 8th 2020. On April 8th 2020, the number of tweets on Twitter is significantly higher than before. This trend is seen in the following days after April 8th 2020. Based on this analysis, it can be confirmed that the number of tweets on Twitter per day – the protest against Corona measures – is stronger when the freedom of the citizens gets restricted by the government. This would confirm that the protest bases on the fear of freedom restriction. Therefore, the protest might be temporary until all restrictions are lifted again.

To perform the analysis with our dataset as described above, researchers are supposed to group the tweets per day. Following on this, the number of tweets per day need to be counted and visualized. Then, the data should be set into relation with the dates of the announcement of Corona measures by the German government and an interpretation can be retrieved from the contextual data.

References

Aguilar-Gallegos, N., Romero-García, L. E., Martínez-González, E. G., García-Sánchez, E. I., & Aguilar-Ávila, J. (2020). Dataset on dynamics of Coronavirus on Twitter. *Data in Brief*, *30*, 105684.

Jose, B. K. (2021, February 24). Twint: Twitter Scraping Without Twitter's API - Basil K Jose - Medium. *Medium*.

Lafaro, A. (2020, October 1). The Importance of COVID-19 Data Collection and Transmission. *UNC Research*.

Welle, D. (2021, September 12). Meet Germany's 'Querdenker' COVID protest movement | DW | 03.04.2021.

Grande, E., Hutter, S., Hunger, S. & Kanol, E. (2021). *Alles Covidioten? Politische Potenziale des Corona-Protests in Deutschland*. WZB Discussion Paper, No. ZZ 2021-601, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin, P. 1-33.

Koos, S. (2021). Die „Querdenker". Wer nimmt an Corona-Protesten teil und warum? : Ergebnisse einer Befragung während der „Corona- Proteste" am 4.10.2020 in Konstanz.

Koos, S. & Binder, N. (2021). Wer unterstützt die »Querdenker«?. In Reichardt, S. (Eds.), *Die Misstrauensgemeinschaft der »Querdenker«* (p. 295-320). Campus Verlag.

Pantenburg, J., Reichardt, S. & Sepp, B. (2021). *Corona-Proteste und das (Gegen-) Wissen sozialer Bewegungen*. In: Politik und Zeitgeschichte (APuZ) ; 71 (2021), 3-4. - S. 22-27.

Jelodar H, Wang Y, Orji R, Huang S. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. IEEE J Biomed Health Inform. 2020 Oct;24(10):2733-2742. doi: 10.1109/JBHI.2020.3001216. Epub 2020 Jun 9. PMID: 32750931.

R. Tang, L. Zhang, G. Zhang and J. Wang, "Analysis of COVID-19 Rebound Based on Natural Language Processing," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 333-336, doi: 10.1109/ICSP51882.2021.9408930.

R. B. S, A. Ezhilan, D. R, A. R and S. R, "Sentiment Analysis and Classification of COVID-19 Tweets," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 821-828, doi: 10.1109/ICOEI51242.2021.9453062.

A. Kumar, K. Yun, T. Gebregzabiher, B. Y. Tesfay and S. G. Adane, "COVID19 Tweeter Dataset Sentiment Analysis," 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2021, pp. 110-115, doi: 10.1109/CCICT53244.2021.00032.

G. M. Raza, Z. S. Butt, S. Latif and A. Wahid, "Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models," 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2021, pp. 1-6, doi: 10.1109/ICoDT252288.2021.9441508.

K. Rahul, B. R. Jindal, K. Singh and P. Meel, "Analysing Public Sentiments Regarding COVID-19 Vaccine on Twitter," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 488-493, doi: 10.1109/ICACCS51430.2021.9441693.

F. Es-Sabery et al., "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier," in IEEE Access, vol. 9, pp. 58706-58739, 2021, doi: 10.1109/ACCESS.2021.3073215.

S. Roy, A. Bhowmik, S. Ghosh and B. Roy, "Statistical Analysis & Categorization for Tweets during Natural Disaster using Classification and Ranking Approach," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154122.