GECO: A Twitter Dataset of COVID-19 Conspiracy Theories Related to the Berlin Parliament and Washington Capitol Riots

Stefan Brenner[1], Daniel Thilo Schroeder[2,3,5], and Johannes Langguth[3,4]

[1] Stuttgart Media University, Stuttgart, Germany
[2] Oslo Metropolitan University, Oslo, Norway
[3] Simula Research Laboratory, Oslo, Norway
[4] BI Norwegian Business School, Oslo, Norway
[5] SINTEF, Oslo, Norway

Abstract. On August 29, 2020, a precursor to the widely known January 6 United States Capitol attack in Washington D.C., USA, occurred in Berlin, Germany, where a group of protesters participating in a demonstration against COVID-19 pandemic measures attempted to storm the German parliament in Berlin. While the event in Berlin was less dramatic than January 6 of 2021 in the US - the protesters were repelled by the police, and no serious damage or injuries were reported - in both cases, mobilization through conspiracy theories on social media is widely considered a significant factor leading both events. Both events were widely reported in the traditional media; moreover, they were often compared with each other and perceived as similar by the public. In this paper, in order to study such social media content, we present an analysis based on a manually labeled dataset of 23,417 German tweets sampled from a large set of COVID-19 related tweets in temporal proximity to the event in Berlin. Moreover, we provide an analysis that is based on a set of tweets following the January 6 event for comparison. The labels distinguish eight different classes of conspiracy theories, as well as other misinformation. This allows for studying the prevalence of different misinformation narratives around events of note. The purpose of this dataset analysis is to allow further study of the phenomena, as well as train machine learning systems capable of detecting conspiracy theory content.

Keywords: Misinformation, Conspiracy Theories, Twitter, Human-annotated data

1 Introduction

The COVID-19 pandemic has had an enormous, worldwide impact that extended far beyond the medical domain. Disagreement on how to deal with the situation carried over into the political sphere, where different approaches to dealing with the pandemic became political positions that were adopted by various parties.

In Germany, a large number of protests against the COVID-19 policies of the government took place in 2020, as well as in the following pandemic years. While the protesters had a multitude of motivations and objectives [14], a vocal group among them was promoting far-fetched conspiracy theories.

On August 29, 2020, during a protest against the COVID-19 restrictions, a crowd of people, spurred by the speech of an alternative practitioner and QAnon supporter who claimed that the then President of the United States of America, Donald J. Trump, and US troops had come to Berlin to replace the German government, tried to force their way into the Reichstag building in Berlin, the seat of the German parliament[6] [15]. In the following, we refer to this event as Berlin Event.

Five months later, on January 6, 2021, a large crowd forced their way into the US Capitol in Washington D.C., ready to use violence against politicians and employees on-site. They were likely incited by the claim uttered by the then US President Donald J. Trump that the 2020 United States presidential elections were manipulated[7]. As a result of the attack on the US Capitol, at least seven people died, moreover 150 officers of the Capitol Police, the Metropolitan Police and local agencies were injured, while numerous workers got traumatized by the violence [9]. In the following, we call this event the Capitol Event.

Some argue that there is a certain societal and democratic value to conspiracy theories because they may force governments to be more transparent about their decisions, expenditures, and actions [32]. On the other hand, the damage conspiracy theories can cause to individuals and societies are manifold, with the most drastic examples leading to physical violence. In the wake of the pandemic, QAnon made the transition from endorsing a conspiracy theory to supporting and carrying out real-world mass violence [18]. Another development that can be observed is a widespread attitude among proponents of such conspiracy theories is that journalists and the media are covering up malicious activities by the government. This has been linked to an increase in attacks on journalists during the COVID-19 pandemic [25]. While conspiracy theories were common among the social circles of the perpetrators of the Capitol Event, their impact on the event has been the topic of a wide discussion [7, 17, 5].

In order to facilitate a similar discussion around the Berlin Event and to better understand how violent political events, in general, may be predicted, we propose to analyze the underlying narratives of conspiracy content spread in online social networks and compare both events.

To this end, we created an annotated dataset of 23,417 messages sampled from Twitter around the Berlin and the Capitol events in which we distinguish different types of conspiracy narratives.

[6] The German federal parliament is called *Bundestag*, but its seat in Berlin is called *Reichstag building*. after the 19th and early 20th century name of the German parliament

[7] The extent to which Donald Trump was responsible for incitement of the riot is subject to a wide discussion and ongoing legal proceedings.

Since the tweets are sampled randomly from COVID-19 related tweets, the annotations also provide insight into the prevalence of such content.

We then present a preliminary data analysis of the data, along with insights we gained during the annotation process, and open the possibility for other researchers to perform deeper or comparative analyses.

2 Dataset Creation

In order to create the dataset, we collected a large amount of COVID-19 related Twitter data during the pandemic via the Twitter search API. The data collection started shortly after the initial phase of the pandemic. Thus, we obtained relevant data from the days leading up to the key events. Consequently, this also implies that the initial data collection was not specifically geared towards both events under investigation.

Among the statuses collected as described above, we removed all retweets, replies, and quoted replies, leaving only tweets. We used the language field of the tweet objects to select German-language tweets only. For the Berlin Event, we used tweets from August 22 to September 6, 2020, and for the Capitol Event, we used data from January 6 to January 12, 2021. To prepare manual labeling, from all the available German-language tweets, we selected 5%[8] uniformly at random for each day under study. In this manner, we get a reliable estimate of the prevalence of misinformation in COVID-related tweets.

The events surrounding the attack on the US Capitol in Washington D.C. show similarities to the Berlin Event in the parliament building on August 29, 2020, which is why the latter is used as a reference point. However, we did expect to find German Twitter messages that are related to the event in the days leading up to the event, and thus we do not include tweets in the week before January 6. On the other hand, the event was widely discussed in German-speaking countries in the period after January 6. Consequently, we use the week after the event, from January 6 to January 12, 2021 as a point of reference. These tweets are part of the dataset.

Tweets that contained only hashtags or URLs and no actual text were removed, as were tweets that had been classified incorrectly as German. While URLs can provide valuable insight into the context of a tweet, and using them is a common technique [28, 23], our analysis focuses on the text contained in the tweets, and it is impossible to understand the intention of tweets that only consist of a URL without following the URL and evaluating the contents found there. In place of the discarded tweets, additional ones were selected uniformly at random to guarantee at least 5% coverage for each individual day. For each day randomized samples between 11,599 to 28,384 tweets were incorporated into the dataset.

[8] In practice, this amounts to slightly more than 5% due to processing batches of tweets.

Table 1. Number of Tweets captured and evaluated for each day under study.

Berlin Event	Total tweets	5%	Annotated
Saturday, 22 August 2020	15,251	763	780
Sunday, 23 August 2020	14,881	744	750
Monday, 24 August 2020	17,533	877	878
Tuesday, 25 August 2020	19,772	989	989
Wednesday, 26 August 2020	25,576	1,279	1,286
Thursday, 27 August 2020	22,308	1,115	1,218
Friday, 28 August 2020	22,572	1,129	1,151
Saturday, 29 August 2020	21,110	1,056	1,129
Sunday, 30 August 2020	16,656	833	912
Monday, 31 August 2020	19,140	957	988
Tuesday, 1 September 2020	18,930	947	992
Wednesday, 2 September 2020	18,696	935	963
Thursday, 3 September 2020	16,110	806	874
Friday, 4 September 2020	15,558	778	796
Saturday, 5 September 2020	11,725	586	599
Sunday, 6 September 2020	11,599	580	603

Capitol Event	Total tweets	5%	Annotated
Wednesday, 6 January 2021	28,577	1,429	1,434
Thursday, 7 January 2021	20,537	1,027	1,027
Friday, 8 January 2021	25,830	1,292	1,293
Saturday, 9 January 2021	19,487	974	994
Sunday, 10 January 2021	21,127	1,056	1,069
Monday, 11 January 2021	25,023	1,251	1,262
Tuesday, 12 January 2021	28,384	1,419	1,430

2.1 Class Labeling

The main labeling was performed by the lead author of this study in a two-stage process. First, all tweets were read, evaluated, and assigned to one of three classes. In addition, a second annotator with substantial experience in the field who is also an author of this paper read all tweets that were labeled as positive or at least one type of conspiracy for quality assurance for quality assurance. For all tweets with disagreement, labels were discussed until agreement was reached. As a result, 61 out of 751 labels were changed to their final value. Labeling did not use any software except for Microsoft Excel.

All tweets that support a known conspiracy theory or promote belief in some clandestine plot were assigned to the first class, *conspiracy*. We define conspiracy theories as narratives and beliefs that are scientifically impossible or highly implausible or that consist of disproven or unproven allegations against individuals or groups perceived as powerful in providing an explanation for disruptive or perceived disruptive economic, cultural, social, political, violent, or other events in the past, present, or future by spreading claims of secret, malevolent agendas. Common conspiracy theories in the context of COVID-19 posit that the virus either does not exist, was released intentionally, or that vaccines serve highly improbable goals such as mind-control.

The second class, *other misinformation*, is assigned to all tweets that do not support or promote a conspiracy theory but contain other known misinformation, such as incorrect statements about COVID-19 that are not connected to any perpetrator or purpose. This includes tweets that misrepresent otherwise correct statistics or medical information. Note that this constitutes a flagging of rather obvious misinformation rather than a fine-grained fact-checking of every single statement, which would be beyond the scope of this paper.

The third class consists of all tweets that were included in *neither* the first nor the second class.

During labeling, we did not use any additional information, such as other tweets by the same author, or images or URLs embedded in tweets. Doing so makes it possible to train simple NLP systems using the tweet texts and labels

we provided based on the available text and labels alone. Incorporating all the information available would require highly complex, custom designed multimodal models.

Consequently, due to the limited amount of text provided in a tweet, the distinction between the three classes is often difficult. Here, we label relatively aggressively, i.e., we opted for a low threshold for labeling a tweet as conspiracy-promoting. We only required that tweets insinuate conspiracy theories, point towards known conspiracy theories, or name a typical culprit to be considered a conspiracy tweet. A reason for this is that due to the random sampling, very few tweets will be evaluated that clearly spell out entire conspiracy theory narratives. Our goal here is to determine the frequency of tweets related to or supporting conspiracy theories. As a result of these labeling rules, tweets making much weaker statements are labeled as conspiracy related. This procedure differs from that used for many other datasets, which make use of keyword-based selection of potential conspiracy tweets [21, 1].

2.2 Categories

In the second stage, we revisited all tweets in the *conspiracy* class to determine which categories of conspiracy theories they support or promote.

In order to distinguish between different conspiracy theories, we first analyzed a limited number of tweets. We studied 400 randomly selected tweets from September 6, 2020, which is the last day of the first period of observation. The last day was chosen in case any new conspiracies arose during the period of observation.

This was complemented by building a list of well-known conspiracy theories and commonly mentioned conspiracy theories that appeared during the COVID-19 pandemic and comparing them with categories proposed in other studies and literature [3, 8, 21]. In this manner, we found a substantial number of narratives that were eventually merged into seven categories, plus an additional category for other conspiracy theories not covered by the first seven categories.

The fundamental ideas of many conspiracy theories discovered in this manner existed prior to the COVID-19 pandemic. For example, *New World Order* has been a topic among conspiracy theorists for a long time [34], but here it is being discussed in context of COVID-19. Similarly, opposition to 5G wireless technology predates the COVID-19 pandemic, but it became widely known only after conspiracy theorists linked it to COVID [20].

Each tweet was then classified as positive or negative for each of these eight categories, where positive means that the tweet supports or promotes that conspiracy topic. Simply mentioning a conspiracy theory without supporting it is classified as negative unless the tweet makes a direct reference to known conspiracy theories, e.g., #GreatReset, Great Replacement, or New World Order. Here, the reference to a conspiracy theory provides the framing to understand the tweet in the context of the conspiracy theory. Thus, such tweets are labeled with the relevant category.

The categorization can be illustrated with the following example based on a tweet from our dataset: "Covid was precisely planned and is the first big step to change our society into a Marxist state, e.g. New World Order of Soros, Gates, Bilderberg, Clintons and others from this criminal lying group"

This tweet contains an extensive accumulation of different antagonists – George Soros, Bill Gates and the Clintons. They are linked to the New World Order (NWO) conspiracy theory which stands for an alleged Marxist world government in this case. The term Bilderberg refers to the Bilderberg conference, which is linked to a wide variety of conspiracy theories [2, 16]. Based on this, the tweet was classified for containing and spreading conspiracy theories of category number seven (Secret Societies).

The largest number of positive labels for a single tweet was three. The categories are defined as follows:

1. **Suppressed Cures and Treatments**: This category collects narratives proposing that effective medications and treatments for COVID-19 were available but whose existence or effectiveness has been denied by authorities, either for financial gain by the vaccine producers or some other harmful intent, including ideas from other conspiracy categories listed below.
2. **Autocracy and Control**: In this category, we collected narratives containing the idea that the pandemic is being exploited to control the behavior of individuals through fear, through laws that are only accepted because of fear, or through state controlled media and propaganda. While annotating tweets to the Autocracy and Control category, we distinguished between concerns about democratic conditions and fears of abolishing the freedom to demonstrate and conspiracy-ideological statements in which a fictional authoritarian state or dictatorship already exists or is in immediate preparation.
3. **Antivax and Harmful Medicine**: In this category, we collect all statements that suggest that the COVID-19 vaccines serve some hidden nefarious purpose. This includes narratives about vaccinations as the cause of a disease or that vaccines are actually used for euthanasia. This category does not include concerns about vaccine safety or efficacy, or concerns about the trustworthiness of the producers, since these are not conspiracies, even though they may contain misinformation. Furthermore, we do not consider *forced vaccination* a conspiracy narrative since many Western countries introduced vaccine mandates for some professions or tried to introduce a mandatory vaccination [12, 22].

4. **Fake Pandemic**: Prominent narratives that surfaced early on in the pandemic were that there is no COVID-19 pandemic, or that the pandemic is just an over-dramatization of the annual flu season, or a statistical error produced by the detection methods in use. Typically, such narratives claim that the intent of the authorities is to deceive the population in order to hide deaths from other causes or to create irrational fear in order to control the

behavior of the population.

5. **Intentional Pandemic**: This straightforward narrative posits that the cause of the pandemic is purposeful human action pursuing some illicit goal. It thus produces a culprit for the situation. Note that this is distinct from asserting that COVID-19 is a *bioweapon* or discussing whether it was created in a laboratory [29] since this does not preclude the possibility that it was released accidentally, which would not produce a culprit and thus not qualify as a conspiracy theory.

6. **Harmful Technology**: This class of conspiracy theories bundles all ideas that connect COVID-19 to harmful technologies like wireless transmissions, especially from 5G equipment.

7. **Secret Societies**: This category collects narratives in which the perpetrators are alleged to be part of some secret society like the Illuminati, New World Order (NWO), Rothschild family, Deep State, or a satanic cult, who perform objectionable rituals, or make use of occult ideas or symbols.

8. **Other Conspiracy Theory**: We added a catchall category for tweets that promote other known conspiracy theories in the light of COVID-19 or connect some of the above categories to preexisting conspiracy theories, for example, claiming that the moon landing was faked, or that the earth is hollow or flat.

3 Technical Dataset Description

We make the dataset available to interested academic researchers on request. In order to preserve user privacy, we remove *tweet ids* and *user ids*, as well as mentions of user names in the text. We also remove the exact posting time of the tweets but keep the dates. Furthermore, we do not provide the original text, only an English translation of the tweets. This limitation is necessary for privacy reasons since most tweets and thus their authors can easily be found based on the tweet text using search engines.

Naturally, the quality of the translation may vary due to typographic errors in the German tweets. Their value as training data may also vary due to cultural differences between German and English-speaking countries. For example, the Great Reset conspiracy theory in German tweets is often perceived as a fascist conspiracy, while in English tweets from the US it is commonly seen as a communist or socialist plot [21].

For each tweet, the dataset contains the publication date, the number of *likes* and *retweets* at the time the tweet was received from the Twitter API, as well as the *friends* and *follower* numbers of the tweet author at that time. Since the tweets were collected on a daily basis, tweets were received at most one day after being published.

4 Quantitative Dataset Description

In this section, we provide quantitative information about the dataset. We start with the classes and categories, i.e., the numbers of assigned labels. In addition, we show which labels occur together frequently.

Table 2 list the number of times a tweet was evaluated as positive for each category label on the main diagonal, as well as the number of co-mentions in the rest of the table. Suppressed Cures and Treatments, for example, has little overlap with other categories, while Secret Societies has a relatively strong overlap.

In total, we assigned 751 positive labels for the conspiracy categories, and 564 tweets (i.e., about 2.5%) were assigned at least one conspiracy label. In addition, 1,545 tweets (i.e., about 6.5%) were considered substantial misinformation other than conspiracy theories, which means that about 9% of all sampled tweets contain misinformation. Thus, the remaining 21,308 fall in the *neither* class, which means that they promote no conspiracy theories and do not contain obvious misinformation.

The tweets were authored by 15,878 different users, 3,211 of which have at least two tweets in the dataset, and 124 users have ten tweets or more. The maximum number of tweets among all users was 49, which amounts to more than two tweets per day on average. The users had a median of 244 followers and an average of 5,032. The large difference between the median and average is due to a large number of popular accounts belonging to media, influencers, public institutions, and football clubs. There are 133 accounts with more than 100,000 followers and 713 accounts with more than 10,000 followers in the dataset. The maximum belongs to the *Borussia Dortmund* football club with more than 3.6 million followers.

The top 133 accounts make up 63%, and the top 713 accounts make up 84% of all follower relationships. Thus, the situation is somewhat different from, e.g., US Twitter, with more followers concentrated on large public institution accounts but fewer followers concentrated at the very top (e.g., in the US, Barack Obama and Elon Musk have more than 130 million followers each)[31].

Fake Pandemic is by far the largest group, and it has a strong overlap with Autocracy and Control as well as with Antivax and Harmful Medicine, even though the latter two only have a small overlap with each other.

We have seen that among the conspiracies the Fake Pandemic narrative, along with Autocracy and Control in second place with less than half the number of tweets. Fake Pandemic also has a strong overlap with Autocracy and Control and with Antivax and Harmful Medicine, even though the latter two only have small overlap with each other. Note that Antivax and Harmful Medicine is quite small since the both key events and thus the tweets occurred before the COVID vaccine rollout.

The Fake Pandemic and the Intentional Pandemic narratives are in competition since the former claims that the virus does not exist or is not particularly dangerous, while the latter implies that the virus is real and dangerous. We considered this during the labeling, consciously deciding which of the alternatives

Table 2. Left: Labels and common occurrences of labels. The number of tweets by category is given on the main diagonal. Note that some tweets have more than one category. Right: Average number of tweets per day by conspiracy category for both time periods.

Conspiracy co-mentions	Suppressed Cures	Autocracy	Antivax	Fake Pandemic	Intentional	Harmful Tech	Secret Societies	Other Conspiracy
Suppressed Cures and Treatments	12	0	0	2	0	0	1	0
Autocracy and Control	0	136	5	57	0	0	19	9
Antivax and Harmful Medicine	0	5	43	20	1	1	3	2
Fake Pandemic	2	57	20	317	0	1	35	12
Intentional Pandemic	0	0	1	0	45	3	12	2
Harmful Technology	0	0	1	1	3	11	4	2
Secret Societies	1	19	3	35	12	4	95	6
Other Conspiracy Theory	0	9	2	12	2	2	6	63

Category/class	August 2020	January 2021
Suppressed Cures and Treatments	0.63	0.29
Autocracy and Control	7.19	3.00
Antivax and Harmful Medicine	1.81	2.00
Fake Pandemic	15.75	9.29
Intentional Pandemic	2.19	1.43
Harmful Technology	0.50	0.43
Secret Societies	4.31	3.71
Other Conspiracy Theory	3.19	1.71
Conspiracy class	35.56	21.86
Misinformationclass	69.06	63.86

a tweet author believed in. On the other hand, Secret Societies has substantial overlap with other categories. The reason for this is that it is defined by a perpetrator rather than a means and is thus compatible with other categories.

The large overlap between Fake Pandemic and Autocracy and Control can be explained based on the epidemiological and political situation. In August 2020, the COVID situation in Germany was calm, with a few hundred daily cases and about four deaths per day. In hospitals, about 1% of the ICU beds were occupied by COVID-19 patients. At the same time, significant COVID-19 restrictions were still in effect.

Thus, the narrative that the government and media were exaggerating the COVID-19 threat and artificially inflating the number of cases in order to restrict civil liberties was relatively prominent. Many such tweets also fall into the Autocracy and Control category. By the same token, there is a large overlap between Fake Pandemic and Antivax and Harmful Medicine. Such tweets typically claim that medical measures are not appropriate because the threat is exaggerated.

The demonstration on August 29, 2020, in Berlin, Germany, was banned by the Senate Department for the Interior on August 26, 2020 [26]. However, the ban was overturned by the Berlin Administrative Court on August 28, 2020 [27]. Thus, tweets about Autocracy and Control peaked between August 26 and August 28, 2020, and got rare after.

In August 2020, the announcement of the first COVID-19 vaccines was still several months away. Thus, Antivax and Harmful Medicine was a topic that few people cared about, with most tweets in that category relating to masks which were a more prominent topic at that time. Still, the number of such tweets was lower than expected.

The Harmful Technology and Suppressed Cures and Treatments categories play no significant role here. Suppressed Cures and Treatments had prominent proponents in the US (Donald Trump) and France (Didier Raoult) [6] but not in Germany. The Harmful Technology narratives were very prominent in the UK

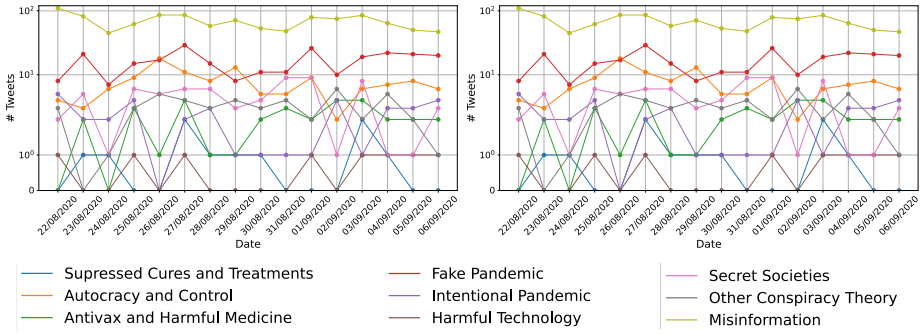

Fig. 1. Distribution of tweets by category over time around the Berlin Event (left) and the Capitol Event (right)

earlier in 2020, resulting in attacks on wireless telecommunication infrastructure, but had little impact in Germany [20].

In order to study the development over time, Figure 1 shows the frequency of conspiracy tweets for each day under observation in both time periods. There are some visible trends, such as a reduction of the Autocracy and Control tweets in September 2020 and a reduction in Fake Pandemic tweets later in January 2021. However, for most categories, the number of conspiracy tweets is too low to reliably observe trends.

To get a better overview over the development, Table 2 on the right shows the average number of tweets in each category per day. Clearly, the daily number of conspiracy tweets declined despite an increase in the total number of daily tweets (see Table 1).

Especially Suppressed Cures and Treatments, Autocracy and Control, and Other Conspiracy Theory declined by almost half, while Fake Pandemic and Intentional Pandemic declined by about one third. Only Antivax and Harmful Medicine increased, which is not surprising since the topic of vaccinations became much more relevant in 2021. On the other hand, Harmful Technology declined very little, but due to the small number of tweets in this category, the difference is not significant. In contrast to the conspiracy categories, the Misinformation class remained fairly constant.

Since our dataset contains *retweet* and *like* counts, we present these numbers for each conspiracy category. They come with the caveat that they are numbers at the time of recording, not the final number of *retweets* and *likes* that a tweet attains. However, since there is no indication that any category is more affected by this than other category, we can analyze the comparative differences.

As shown in Table 2, the Autocracy and Control and Fake Pandemic categories are by far the most frequent. Here, we observe that they are also the most popular. Most other COVID-19 conspiracy categories receive far fewer likes and retweets. This indicates that the popularity of conspiracy theories is influenced strongly by current events. In comparison, other narratives, e.g., Antivax and

Table 3. Average number of retweets and likes per tweet.

Category/class	Avg. Retweets	Avg. Likes
Suppressed Cures and Treatments	0.92	1.25
Autocracy and Control	4.05	10.83
Antivax and Harmful Medicine	0.40	1.44
Fake Pandemic	1.31	4.26
Intentional Pandemic	0.87	2.13
Harmful Technology	0.18	0.64
Secret Societies	0.63	2.24
Other Conspiracy Theory	1.30	5.44
Conspiracy class	1.63	5.60
Misinformation class	2.01	5.70
All tweets in dataset	1.66	7.65

Harmful Medicine, became far more frequent later in the pandemic [21]. Interestingly, the Other Conspiracy Theory class is almost as popular as Autocracy and Control and Fake Pandemic and comparable to the non-conspiracy tweets.

While the averages are driven by very small numbers of highly influential tweets and thus are not statistically significant, the overall trend remains stable when removing the top tweets. The median values in most cases are 0 or 1 since the majority of tweets do not get any likes or retweets. Interestingly, the Misinformation class has the highest number of retweets. The higher number of likes in the other tweets is expected though, since the median number of followers in the dataset (252) is significantly higher than that in the *conspiracy* (153) and *other misinformation* classes (167).

5 Qualitative Dataset Description

Conspiracy theories are not consistent, as it is possible to create a wide variety of different narratives by replacing the alleged culprit or changing small details within a conspiracy storyline. In some cases, even a conspiracy ideological umbrella term can have opposed meanings. For example, the term *Plandemic*, a portmanteau of the terms plan and pandemic, has two different meanings, and it depends on the context whether the term is used to describe an Intentional Pandemic and the deliberate infection of the world population with a virus or a Fake Pandemic in the sense of a feigned crisis [19].

By far, the most common type of conspiracy theory is that of the Fake Pandemic category. This observation is hardly surprising since this conspiracy theory was established early on as a unifying core element in the conspiracy-ideological protest movement against pandemic protective measures. In Germany, this has led to the protest movement often being referred to as Corona Deniers in the media. Moreover, the Fake Pandemic narratives were observed to be exceedingly connectable to other categories of conspiracy theories, but especially to those that can be associated with perceived elitist or power structures (governments, media, or some clandestine masterminds). Consequently Fake Pandemic tweets were most frequently observed in combination with conspiracy theories from the Autocracy and Control category and the Secret Societies category. Furthermore, it was possible to confirm an observation that has already been documented in

other research. Conspiracy theories can reach a peak abruptly in a limited period of time, only to disappear again afterwards [4, 20]. This is the case for the Autocracy and Control category, which became very prominent for a few days, before suddenly disappearing in the background noise.

6 Related Work

Conspiracy theories on Twitter have been investigated in the past. A study by Wood [33] on the prevalence and attempts to contain conspiracy theories around the Zika virus in 2015 combines approaches from folk narrative research with statistical network analysis methods. A data collection of 88,523 tweets published over a seven-month period was collected using search terms observed in the context of Zika virus conspiracy theories. The tweets were then assigned to categories according to their relationship to these conspiracy theories. For example, whether they were actively spreading a Zika conspiracy theory, referencing a conspiracy theory, or expressing disbelief in a conspiracy theory. The different positions were then mapped and contrasted across a network. Wood was able to show that Twitter accounts that spread conspiracy theories about the Zika virus often referenced centralized sources of information.

A substantial number of COVID-19 misinformation datasets have been released in the wake of the pandemic [10]. Darius and Urquhart [11] study conspiracy theories related to COVID-19. However, unlike our dataset, they rely on hashtag analysis rather than human annotation.

Twitter datasets with human annotations dealing with conspiracy theories have been released recently, dealing either specifically with 5G-related COVID-19 conspiracy misinformation [24, 30, 20] or a larger variety of conspiracy theories [21, 13].

7 Conclusion

We have studied COVID-19 related German conspiracy theories on Twitter around the time of significant political motivated events. Unlike most previous studies, we selected tweets randomly from a large set of COVID-19 related tweets, which means that the composition of the labeled data resembles the composition on Twitter. We found that about 2.5% of the tweets contain conspiracy content, and 6,5% contain other forms of misinformation. When using the dataset for training machine learning systems, this class imbalance presents a challenge, but since it is present in real-world data, it needs to be taken into account when designing systems capable of dealing with real-world data.

References

1. Ahmed, W., López Seguí, F., Vidal-Alaball, J., Katz, M.S.: Covid-19 and the "film your hospital" conspiracy theory: Social network analysis of twitter data. Journal of medical Internet research **22**(10), e22374 (2020)

2. Aupers, S.: Decoding mass media/encoding conspiracy theory, pp. 469–482. Routledge, Abingdon, Oxon; New York, NY (2020)
3. Bartoschek, S.: Bekanntheit von und Zustimmung zu Verschwörungstheorien-eine empirische Grundlagenarbeit. jmb Verlag, Hannover (2020)
4. Batzdorfer, V., Steinmetz, H., Biella, M., Alizadeh, M.: Conspiracy theories on twitter: emerging motifs and temporal dynamics during the covid-19 pandemic. International journal of data science and analytics **13**(4), 315–333 (2022)
5. Baum, M., Druckman, J., Simonson, M.D., Lin, J., Perlis, R.: What i saw on the road to insurrection: Internal political efficacy, conspiracy beliefs and the effect of depression on support for the january 6th storming of the capitol (2021)
6. Berlivet, L., Löwy, I.: Hydroxychloroquine controversies: clinical trials, epistemology, and the democratization of science. Medical anthropology quarterly **34**(4), 525–541 (2020)
7. Bond, B.E., Neville-Shepard, R.: The rise of presidential eschatology: conspiracy theories, religion, and the january 6th insurrection. American Behavioral Scientist **67**(5), 681–696 (2023)
8. Butter, M., Knight, P.: Routledge handbook of conspiracy theories. Routledge, London (2020)
9. Cameron, C.: These are the people who died in connection with the capitol riot (2022), \url{https://www.nytimes.com/2022/01/05/us/politics/jan-6-capitol-deaths.html}, last accessed 12 September 2022
10. Cui, L., Lee, D.: Coaid: COVID-19 healthcare misinformation dataset. CoRR **abs/2006.00885** (2020), https://arxiv.org/abs/2006.00885
11. Darius, P., Urquhart, M.: Disinformed social movements: A large-scale mapping of conspiracy narratives as online harms during the covid-19 pandemic. Online Social Networks and Media **26**, 100174 (2021)
12. Deutscher Bundestag: Gesetzentwurf für allgemeine impfpflicht ab 18 jahren (2022), https://www.bundestag.de/presse/hib/kurzmeldungen-883000
13. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1163–1168. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1138
14. Frei, N., Nachtwey, O.: Quellen des "Querdenkertums": eine politische Soziologie der Corona-Proteste in Baden-Württemberg. Universität Basel, Fachbereich Soziologie, Basel (2021). https://doi.org/10.31235/osf.io/8f4pb
15. Fröhlich, A., Ismar, G.: Steinmeier dankt Beamten für Einsatz: Diese Polizisten stoppten die Reichstags-Demonstranten. Der Tagesspiegel Online (Aug 2020), https://www.tagesspiegel.de/berlin/diese-polizisten-stoppten-die-reichstags-demonstranten-6863287.html
16. Girard, P.: Conspiracy theories in europe during the twentieth century, pp. 569–581. Routledge, Abingdon, Oxon; New York, NY (2020)
17. Heck, K.B.: The January 6 th Capitol Attack: Exposure to Election Conspiracies and Support for Political Violence. Ph.D. thesis, Saint Louis University (2023)
18. Hodwitz, O., King, S., Thompson, J.: QAnon: The Calm Before the Storm. Society (Mar 2022). https://doi.org/10.1007/s12115-022-00688-x, https://doi.org/10.1007/s12115-022-00688-x
19. Kearney, M.D., Chiang, S.C., Massey, P.M.: The Twitter origins and evolution of the COVID-19 "plandemic" conspiracy theory. Harvard Kennedy School Misinformation Review (Oct 2020). https://doi.org/10.37016/mr-2020-42, https://misinforeview.hks.harvard.edu/?p=3397

20. Langguth, J., Filkuková, P., Brenner, S., Schroeder, D.T., Pogorelov, K.: Covid-19 and 5g conspiracy theories: long term observation of a digital wildfire. International Journal of Data Science and Analytics (May 2022). https://doi.org/10.1007/s41060-022-00322-3
21. Langguth, J., Schroeder, D.T., Filkuková, P., Brenner, S., Phillips, J., Pogorelov, K.: Coco: an annotated twitter dataset of covid-19 conspiracy theories. Journal of Computational Social Science pp. 1–42 (2023)
22. Österreich Parlament, R.: Covid-19-impfpflichtgesetz (2022), https://www.parlament.gv.at/PAKT/VHG/XXVII/A/A_02173/
23. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., Rand, D.G.: Shifting attention to accuracy can reduce misinformation online. Nature **592**(7855), 590–595 (Apr 2021). https://doi.org/10.1038/s41586-021-03344-2
24. Pogorelov, K., Schroeder, D.T., Filkukova, P., Brenner, S., Langguth, J.: WICO text: A labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In: Guidi, B., Michienzi, A., Ricci, L. (eds.) OASIS@HT 2021: Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks, Virtual Event, Ireland, 30 August 2021. pp. 21–25. ACM, New York, USA (2021). https://doi.org/10.1145/3472720.3483617
25. Reporter ohne Grenzen: Nahaufnahme 2022: Die Lage der Pressefreiheit in Deutschland (2022), https://www.reporter-ohne-grenzen.de/nahaufnahme/2022
26. Reuters: Berlin bans protest against coronavirus curbs. Reuters (Aug 2020), https://www.reuters.com/article/us-health-coronavirus-germany-protest-idUSKBN25M1GF
27. Reuters: German court permits Berlin protests against coronavirus curbs. Reuters (Aug 2020), https://www.reuters.com/article/health-coronavirus-germany-protest-idINL8N2FV03L
28. Ribeiro, M.H., Calais, P.H., Almeida, V.A.F., Jr., W.M.: "everything I disagree with is #fakenews": Correlating political polarization and spread of misinformation. CoRR **abs/1706.05924** (2017), http://arxiv.org/abs/1706.05924
29. Ridley, M., Chan, A.: Viral: The Search for the Origin of COVID-19. HarperCollins, New York, USA (2021), https://books.google.no/books?id=o2ozEAAAQBAJ
30. Schroeder, D.T., Schaal, F., Filkukova, P., Pogorelov, K., Langguth, J.: Wico graph: A labeled dataset of twitter subgraphs based on conspiracy theory and 5g-corona misinformation tweets. In: ICAART (2). pp. 257–266 (2021)
31. Statista: Twitter accounts with the most followers worldwide as of January 2023 (2023), https://www.statista.com/statistics/273172/twitter-accounts-with-the-most-followers-worldwide/
32. Swami, V., Coles, R.: The truth is out there. The Psychologist **23**, 560–563 (2010), place: United Kingdom Publisher: British Psychological Society
33. Wood, M.J.: Propagating and Debunking Conspiracy Theories on Twitter During the 2015–2016 Zika Virus Outbreak. Cyberpsychology, Behavior, and Social Networking **21**(8), 485–490 (Aug 2018). https://doi.org/10.1089/cyber.2017.0669, http://www.liebertpub.com/doi/10.1089/cyber.2017.0669
34. Yablokov, I.: Conspiracy Theories in Putin's Russia: The case of the 'New World Order'. In: Routledge Handbook of Conspiracy Theories, pp. 582–595. Routledge, London (2020)