gesis Leibniz Institute for the Social Sciences

Don‘t Look Away! Studying the Tweets that Disappeared.

Leon Fröhling, Dennis Assenmacher, Katrin Weller

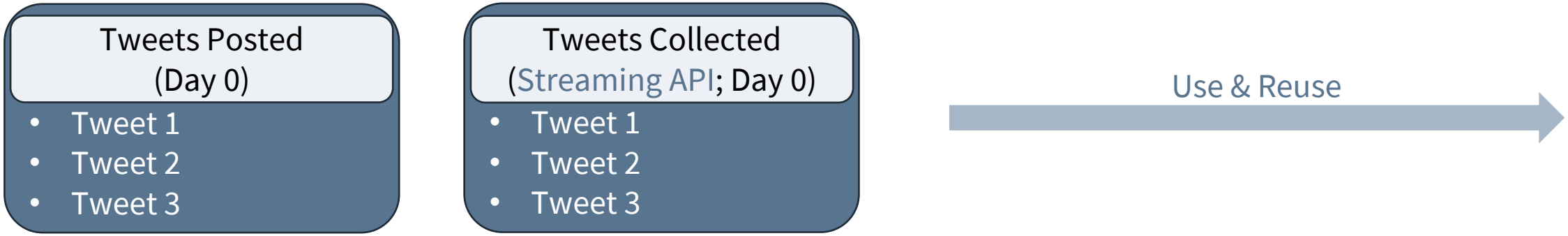
Nachhaltige Archivierung Sozialer Medien – Twitter Und Danach

Deutsche Nationalbibliothek – 20.03.2024

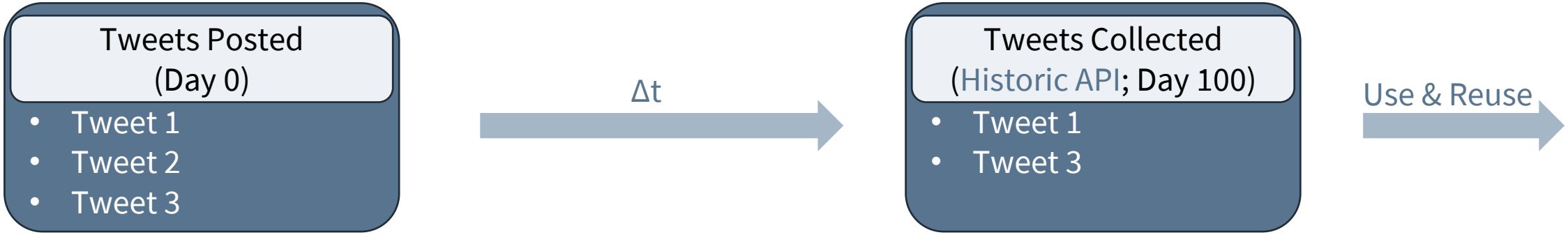
Leibniz Association

Once upon a time …

Collecting Tweets via the Twitter Streaming API

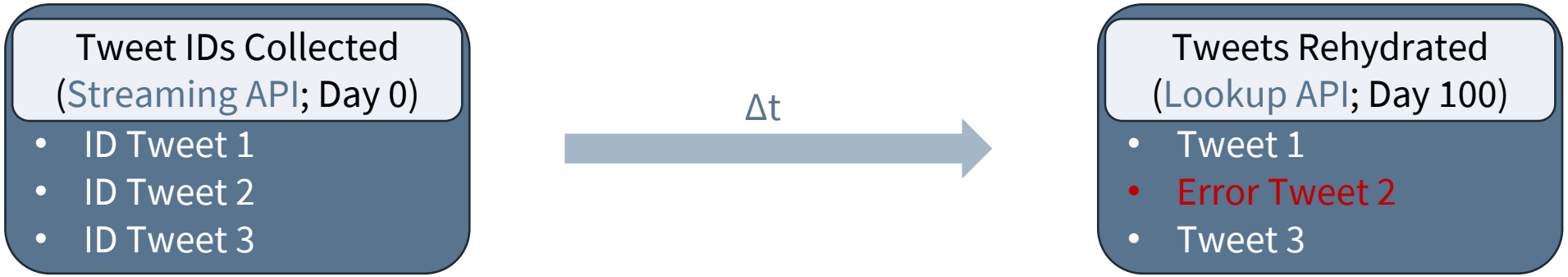


Collecting Tweets via the Twitter Historic API



Once upon a time …

Rehydrating Tweets via the Twitter Tweet Lookup API



API error mystery


```
{'resource_id': '1[Tweet ID]7',
 'parameter': 'ids',
 'resource_type': 'tweet',
 'section': 'data',
 'title': 'Authorization Error',
 'value': '1[Tweet ID]7',
 'detail': 'Sorry, you are not authorized to see the Tweet with ids: [1[Tweet ID]7].',
 'type': 'https://api.twitter.com/2/problems/not-authorized-for-resource'},

{'value': '1[Tweet ID]3',
 'detail': 'Could not find tweet with ids: [1[Tweet ID]3].',
 'title': 'Not Found Error',
 'resource_type': 'tweet',
 'parameter': 'ids',
 'resource_id': '1[Tweet ID]3',
 'type': 'https://api.twitter.com/2/problems/resource-not-found'},
```

API error mystery

Resource Unauthorized Problem	https://api.twitter.com/2/problems/not-authorized-for-resource	A problem that indicates you are not allowed to see a particular Post, User, etc.

```
{'value': '1[Tweet ID]3',
 'detail': 'Could not find tweet with ids: [1[Tweet ID]3].',
 'title': 'Not Found Error',
 'resource_type': 'tweet',
 'parameter': 'ids',
 'resource_id': '1[Tweet ID]3',
 'type': 'https://api.twitter.com/2/problems/resource-not-found'},
```

API error mystery

Resource Unauthorized Problem	https://api.twitter.com/2/problems/not-authorized-for-resource	A problem that indicates you are not allowed to see a particular Post, User, etc.

Resource Not Found Problem	https://api.twitter.com/2/problems/resource-not-found	A problem that indicates that a given Post, User, etc. does not exist.

gesis Leibniz Institute for the Social Sciences

API error mystery

Did the author change the visibility settings for the Tweet?

Did the author change their account to private?

Is Twitter withholding the Tweet from being displayed in my region?

Resource Unauthorized Problem	https://api.twitter.com/2/problems/not-authorized-for-resource	A problem that indicates you are not allowed to see a particular Post, User, etc.

Resource Not Found Problem	https://api.twitter.com/2/problems/resource-not-found	A problem that indicates that a given Post, User, etc. does not exist.

Did the author delete the Tweet?

Did Twitter delete the Tweet?

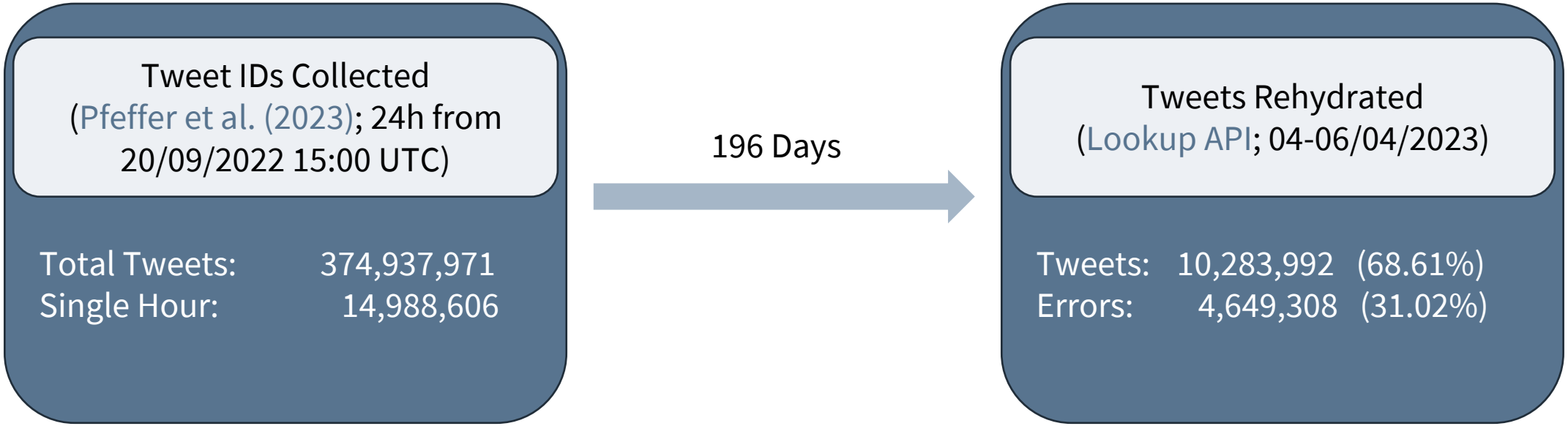
Was the previous Tweet in a conversation deleted?

Did Twitter ban the author‘s account?

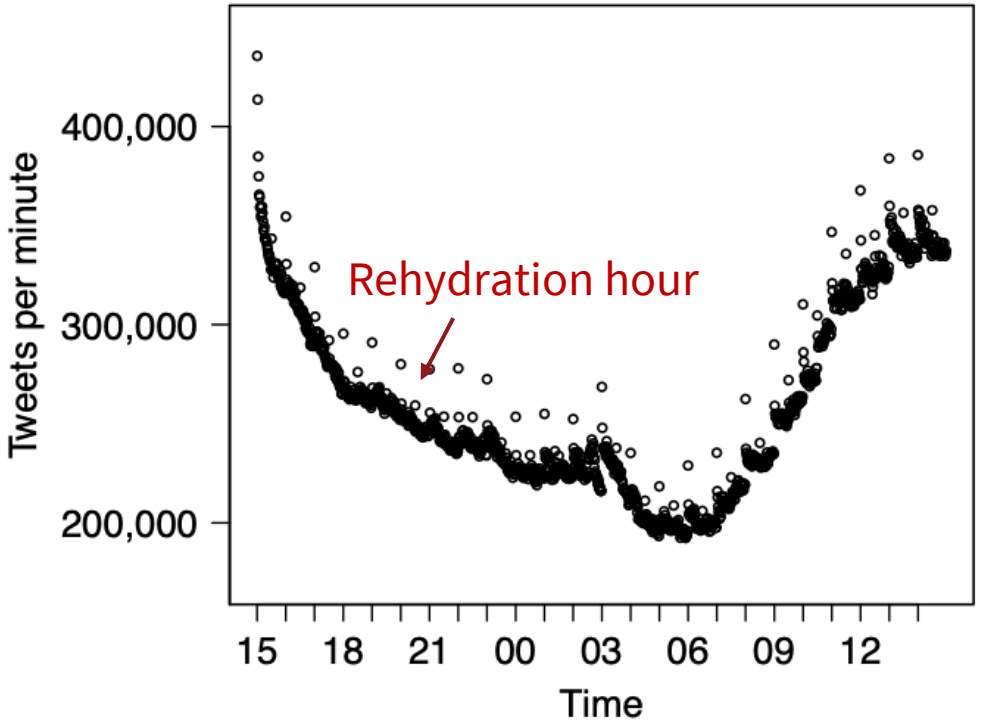
Did the author delete their account?

Our data basis

Rehydrating Tweets via the Twitter Tweet Lookup API



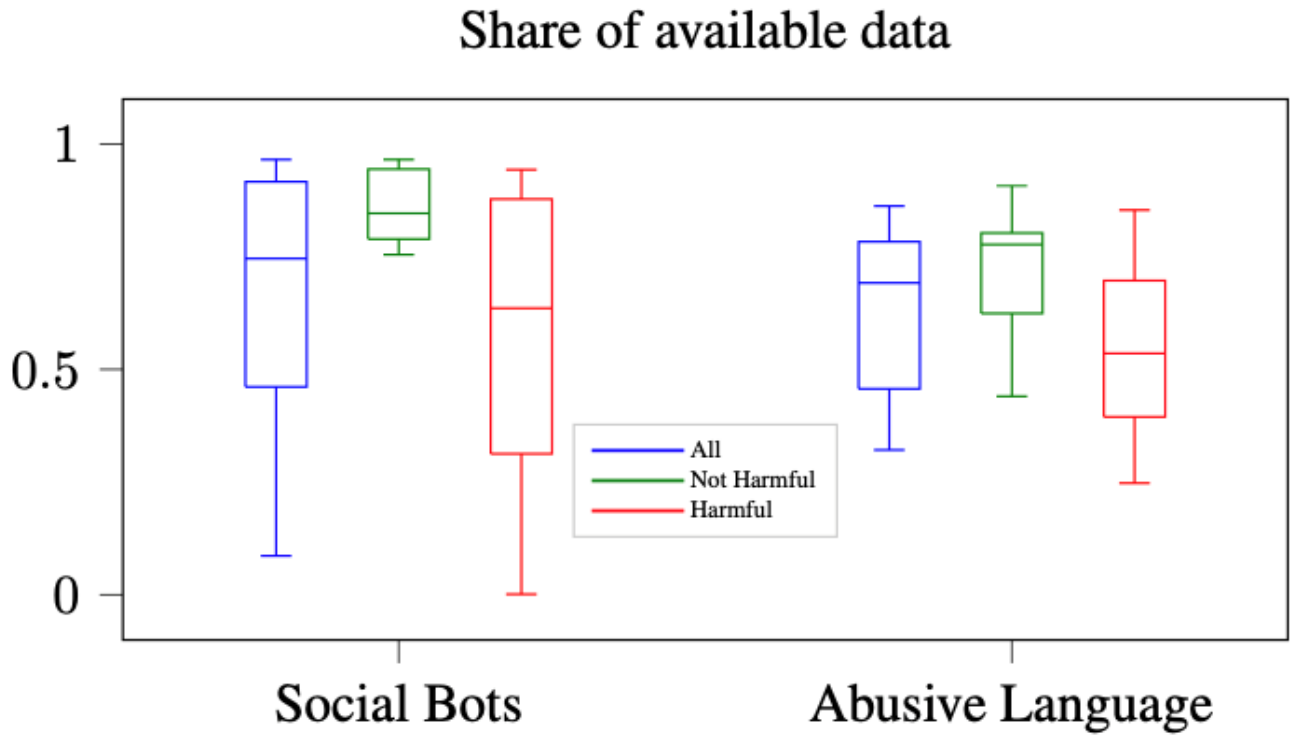
Just another day on Twitter



Category	Hashtags	Occurrence	
Celebrities	159	20,809,742	25.5%
Sex	104	20,529,196	25.2%
Iranian Protests	15	13,488,295	16.6%
Entertainment	45	4,392,227	5.4%
Advertisement	32	4,644,540	5.7%
Politics	38	3,858,550	4.7%
Finance	30	3,549,107	4.4%
Games	21	3,348,128	4.1%
Other	31	2,672,291	3.3%
Unknown	25	4,176,432	5.1%
Sum	500	81,468,508	100.0%

Pfeffer et al. (2023) - Just another day on twitter: A complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 17, pp. 1073-1081). https://doi.org/10.1609/icwsm.v17i1.22215

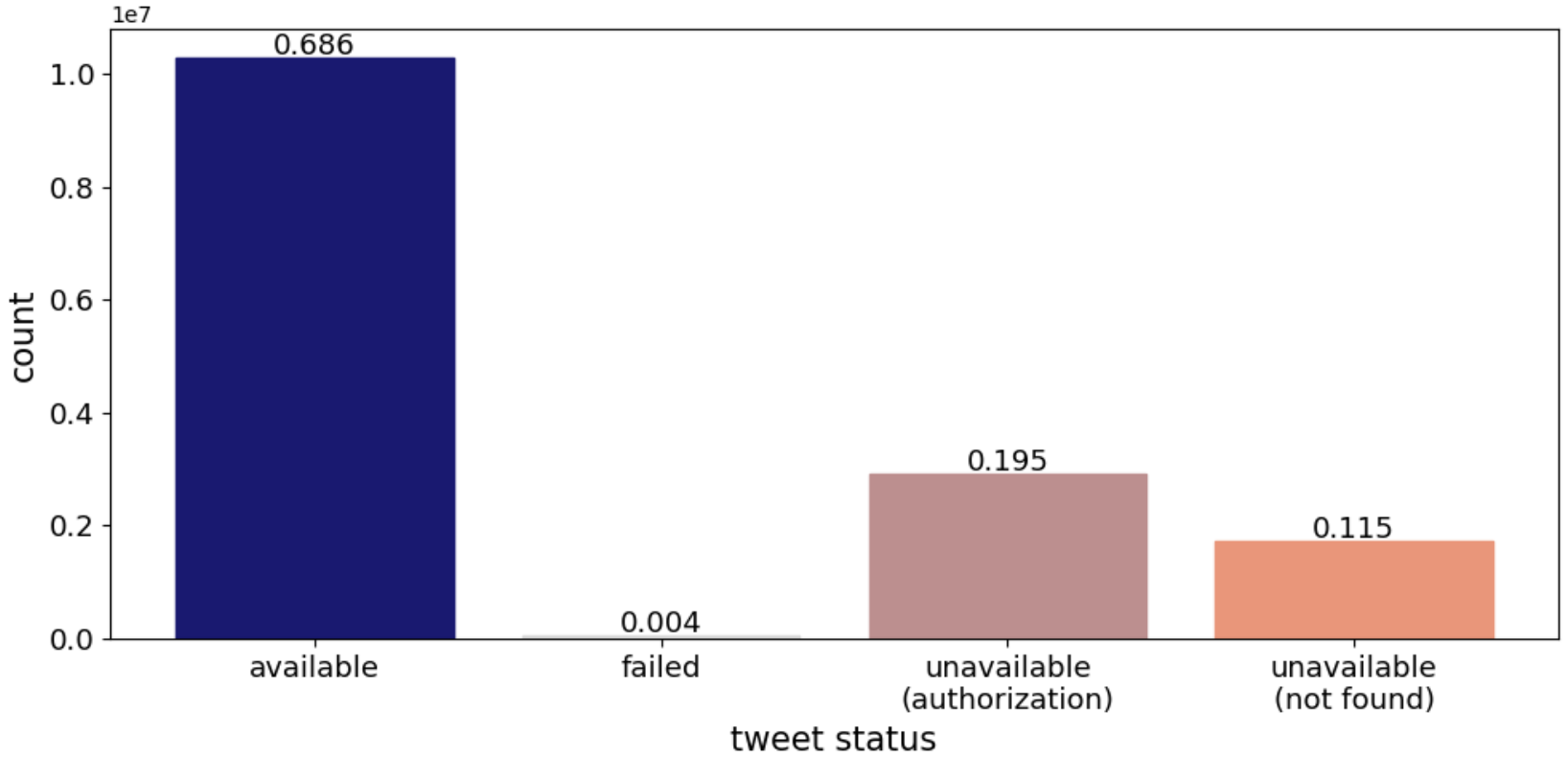
Implications for machine learning



➡ Different types of Tweets decay at different rates!

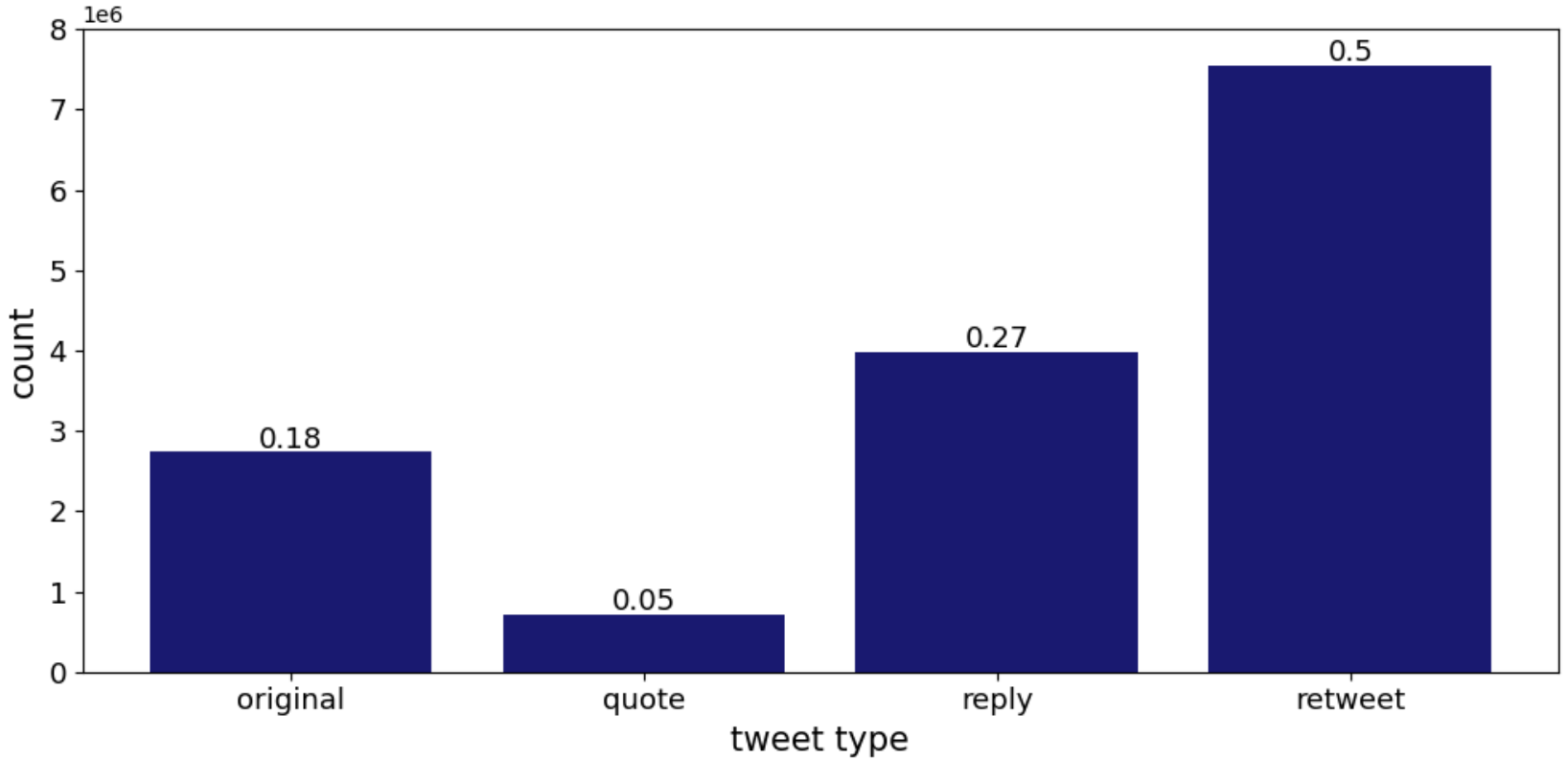
Assenmacher et al. (2023) – The End of the Rehydration Era – The Problem of Sharing Harmful Twitter Research Data. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*. https://doi.org/10.36190/2023.56

How much is lost?

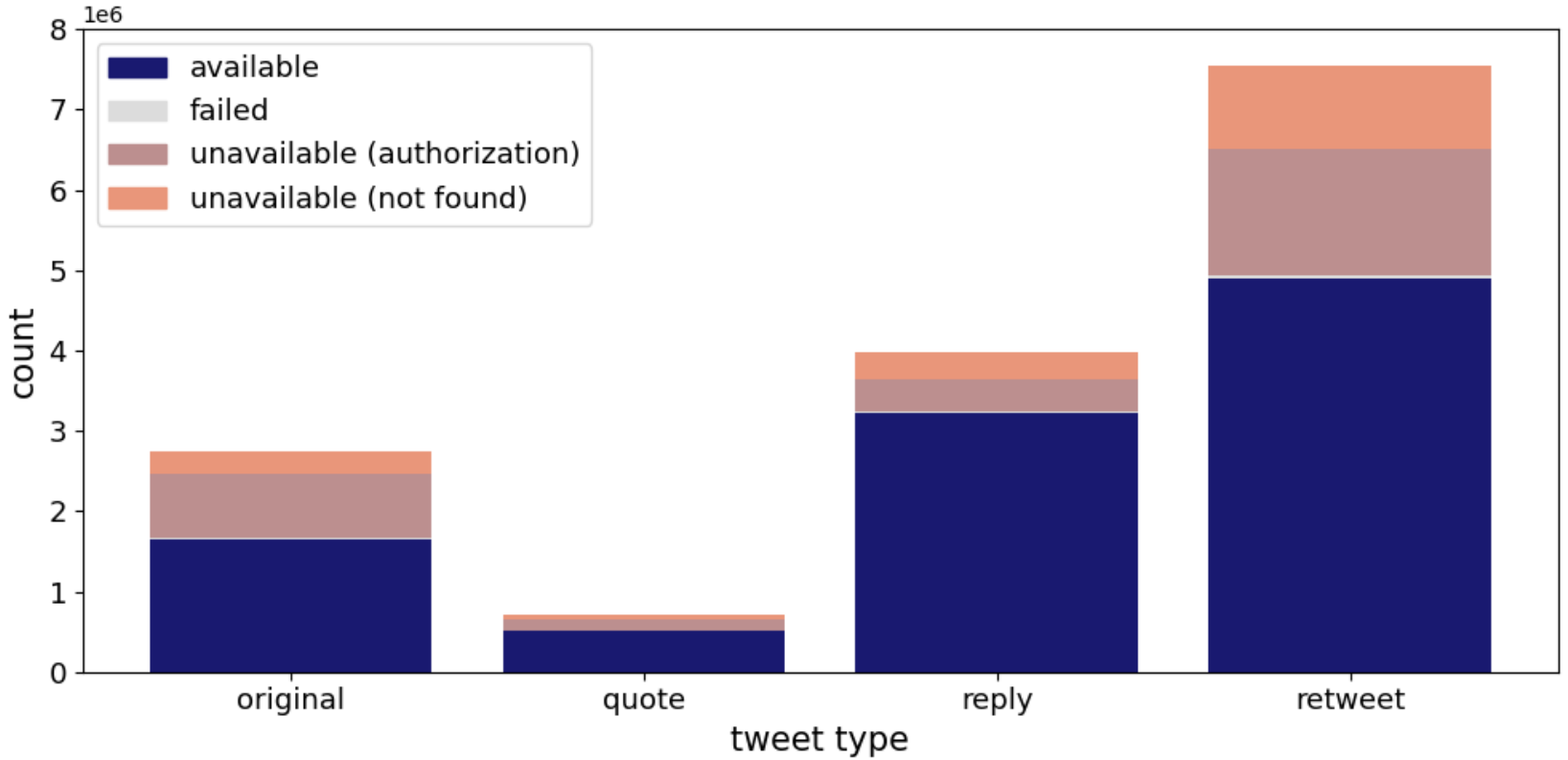


➡ After only half a year, more than 30% of Tweets have disappeared

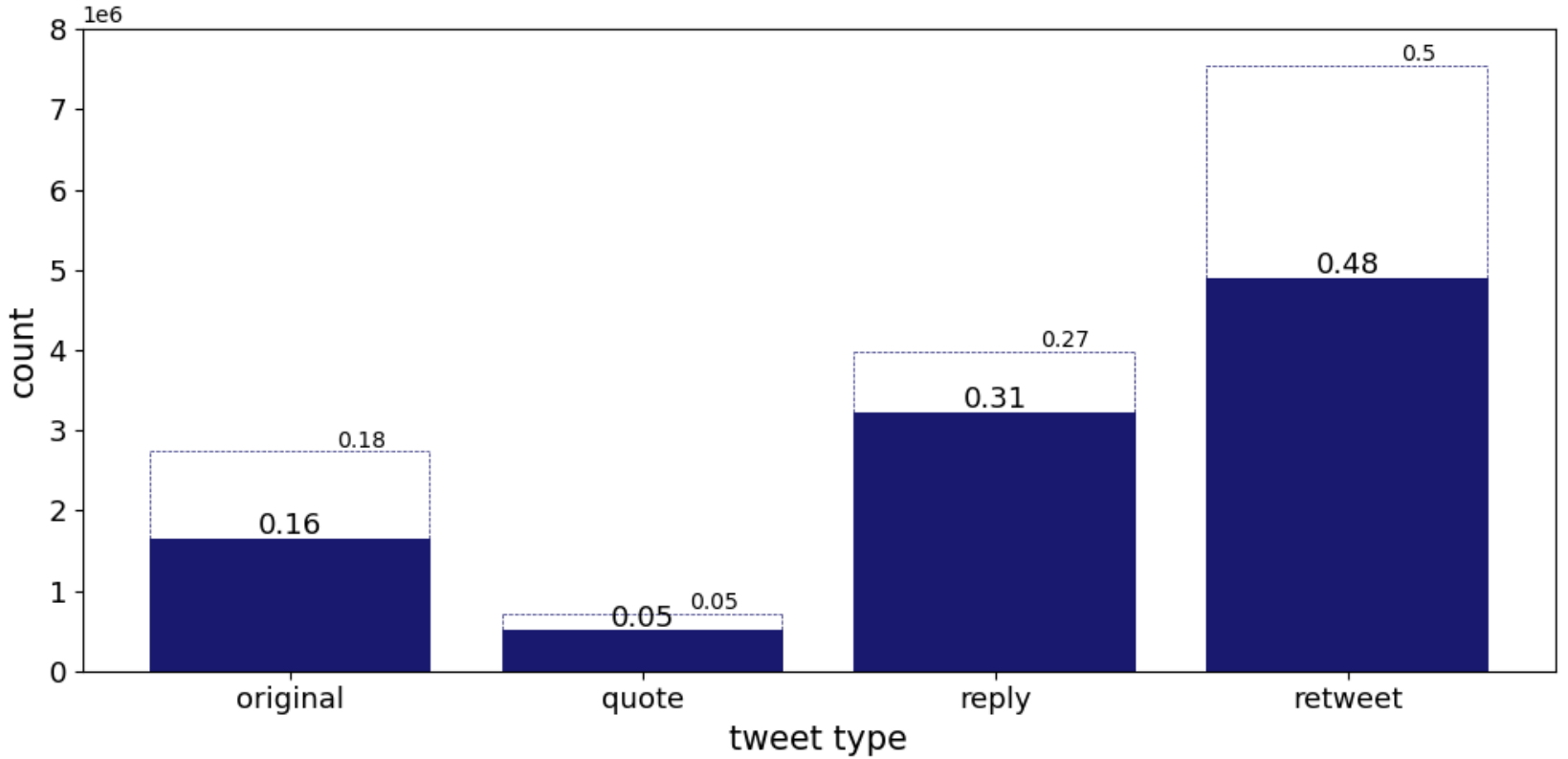
What tweet types **were** in the dataset?



How much is lost – per Tweet type?



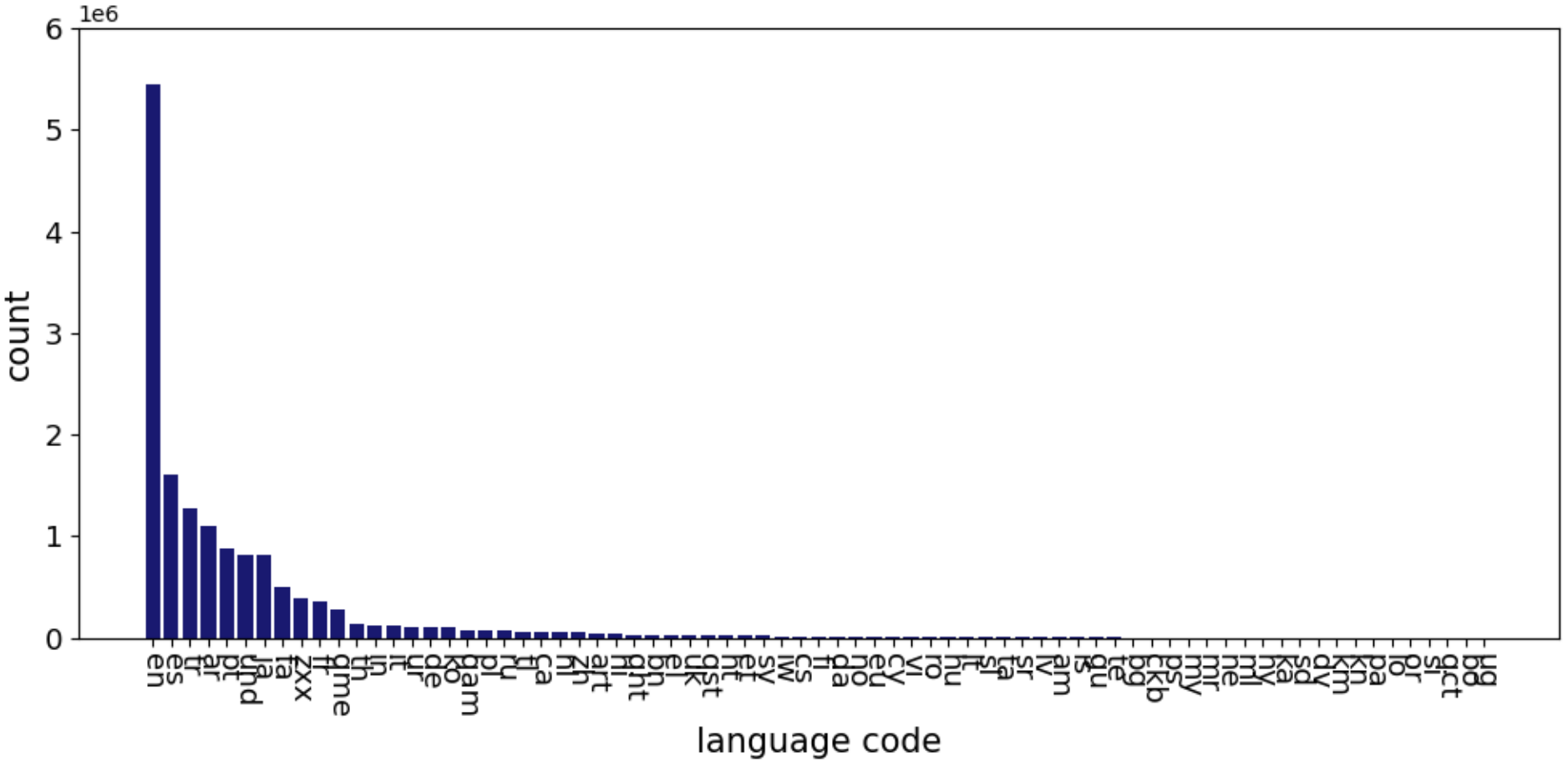
And what remains?



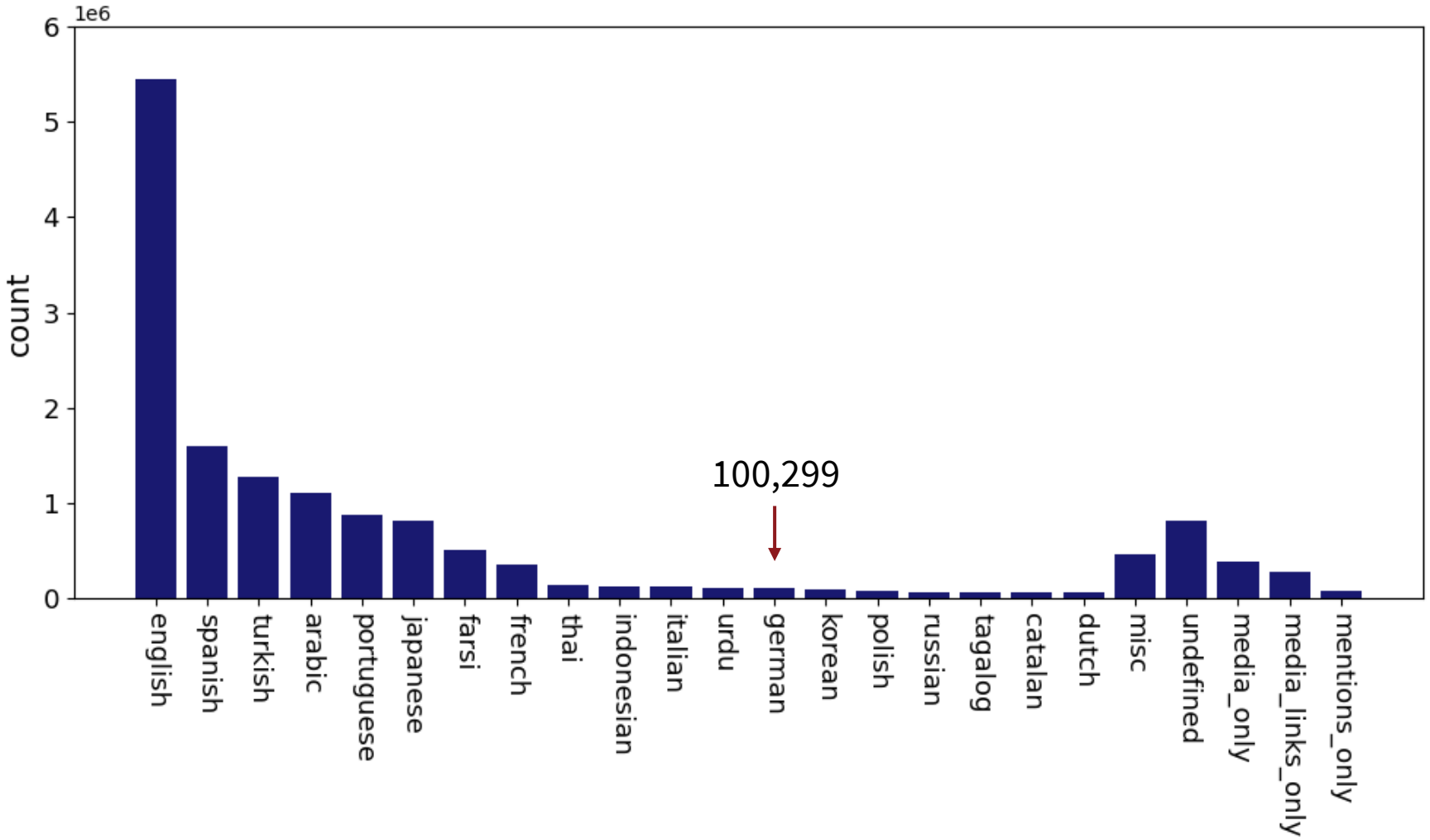
Not just did the total number of Tweets (drastically) decrease, but also the distribution across Tweet types is (slightly) shifted.

gesis Leibniz Institute for the Social Sciences

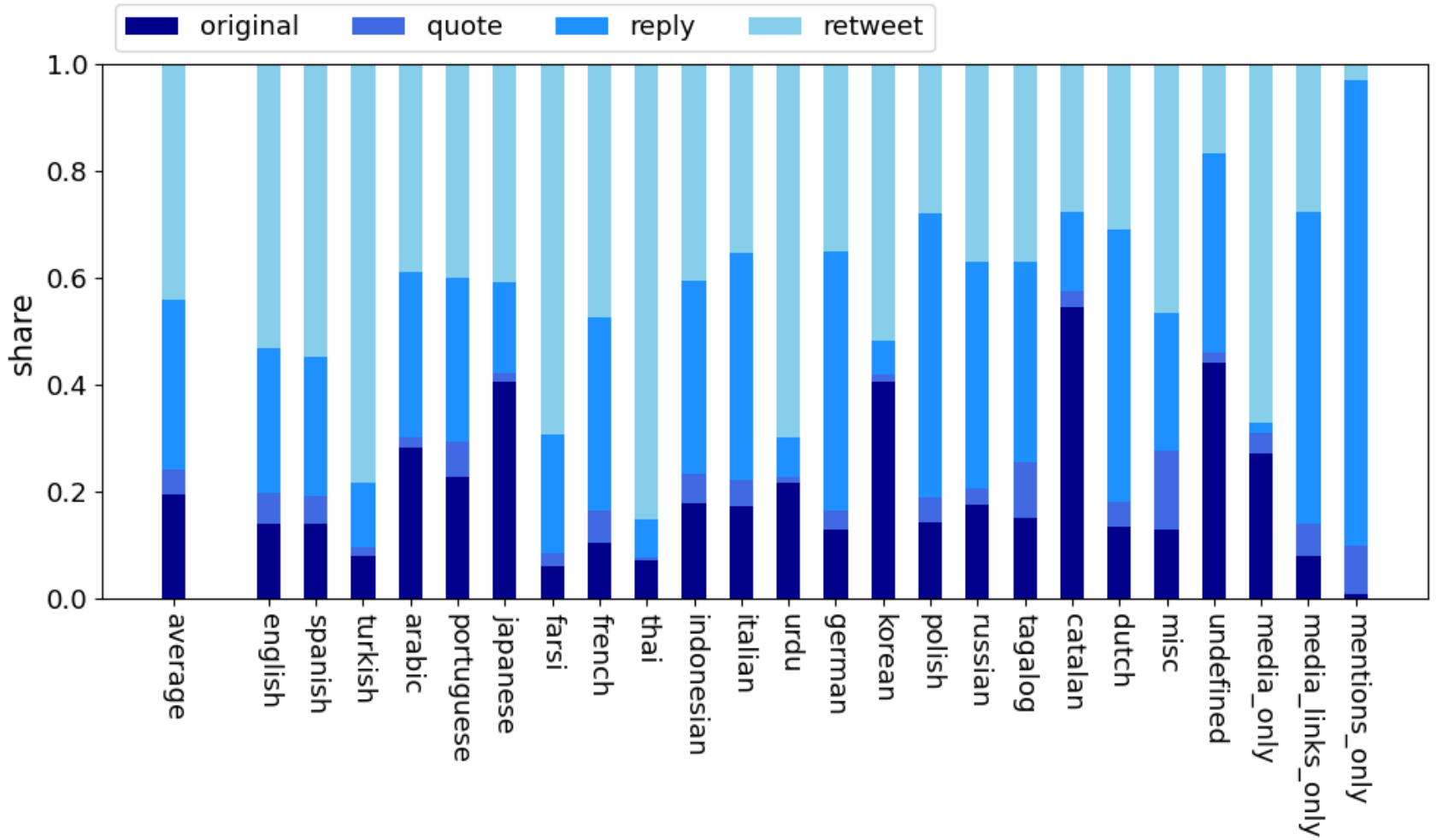
What languages are (were) in the dataset?



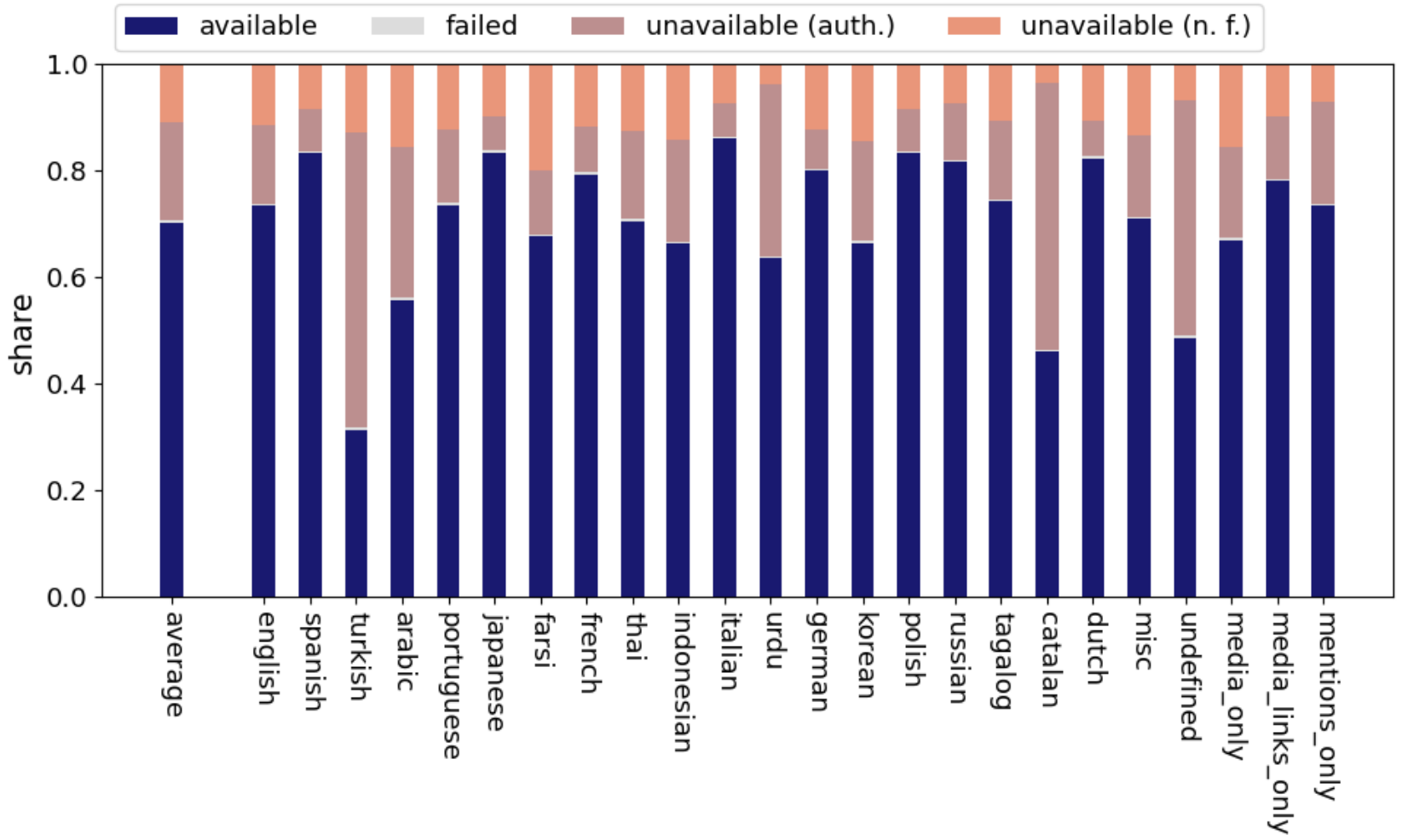
What „big“ languages are (were) in the dataset?



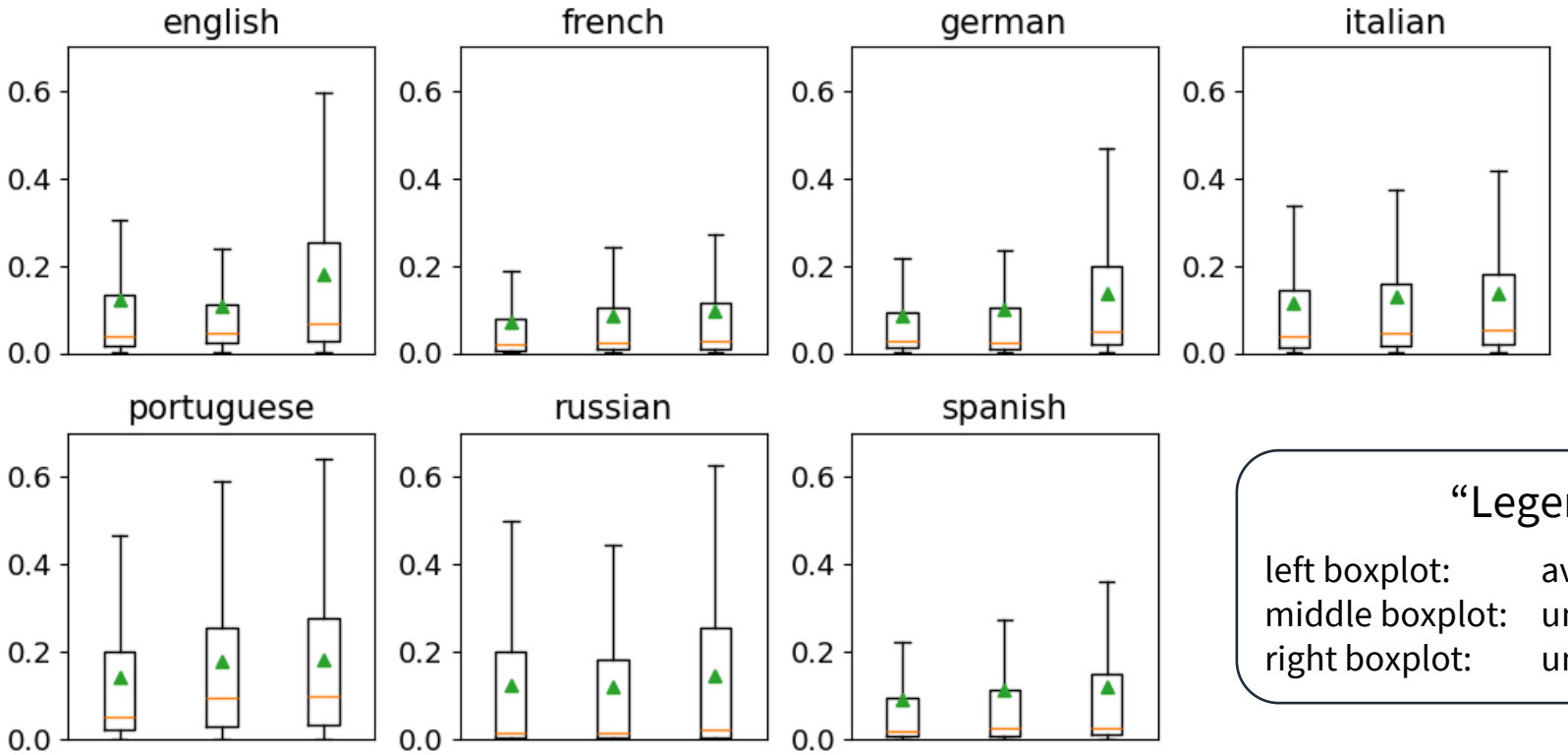
Tweeting habits across languages



Disappearance patterns across languages



Are Tweets that disappeared more problematic?

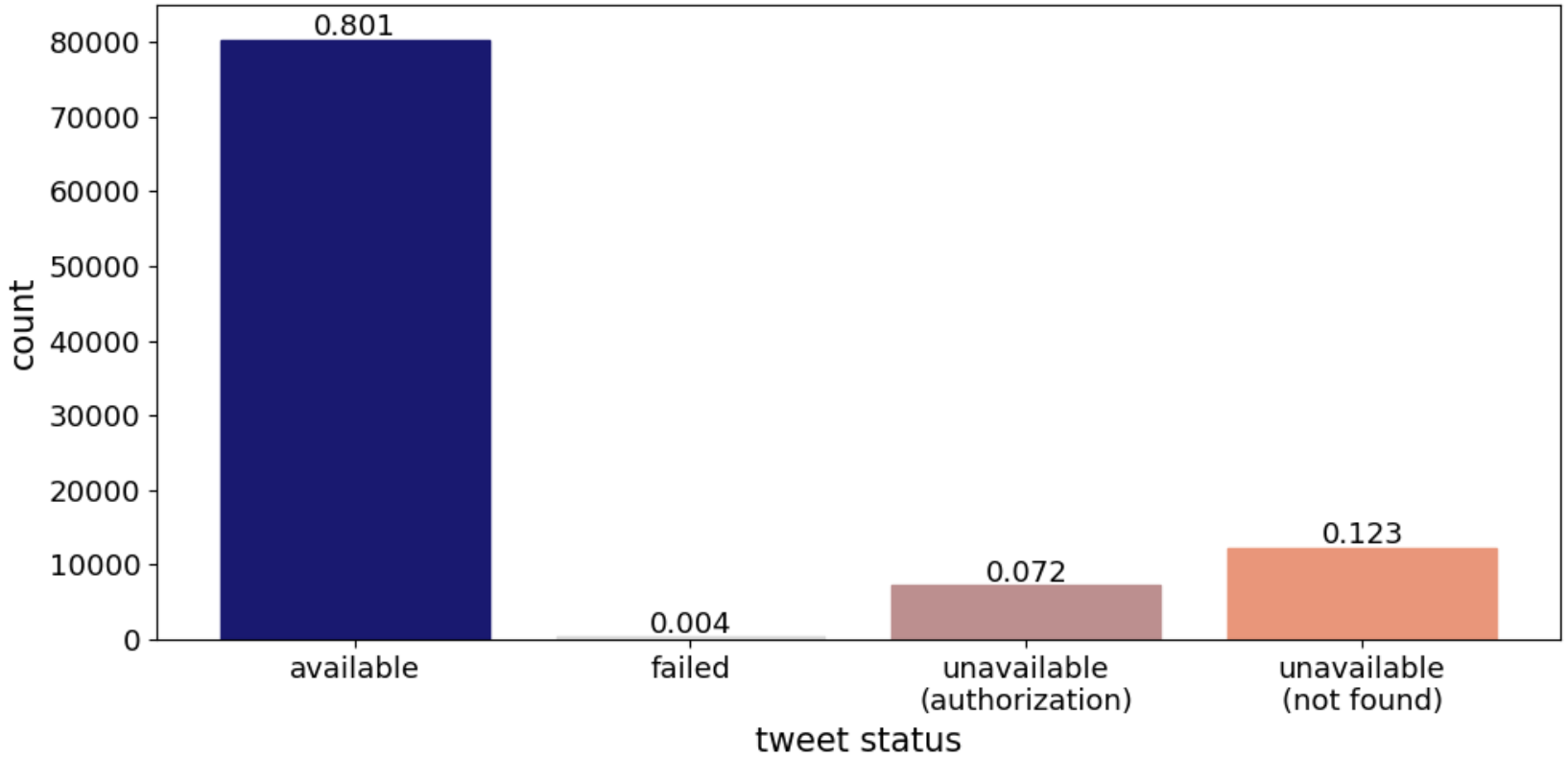


"Legend"

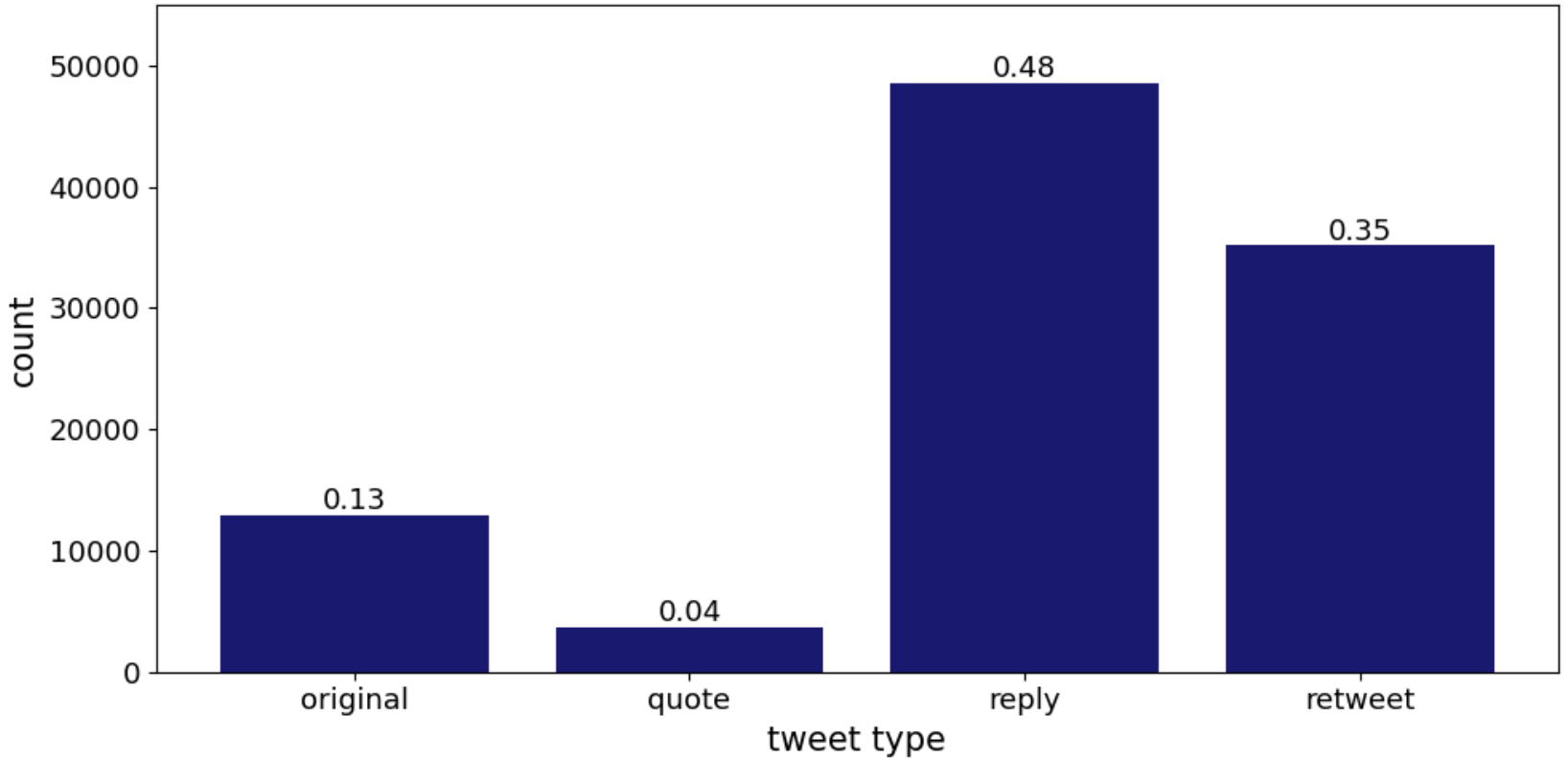
left boxplot:	available
middle boxplot:	unavailable (auth.)
right boxplot:	unavailable (n. f.)

➡ Tweets that disappeared tend to be „more toxic". Particularly, the most extreme cases seem to have disappeared.

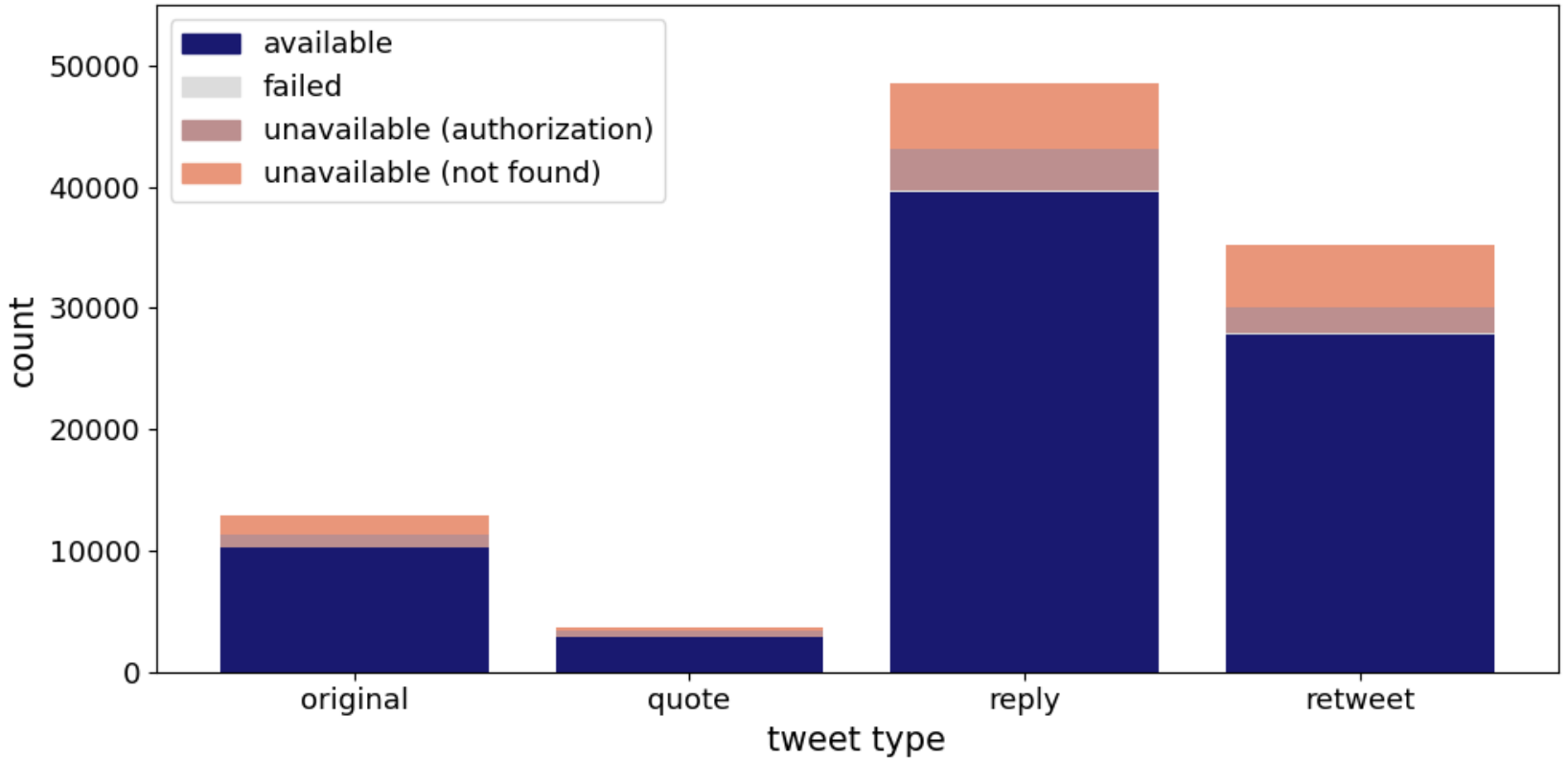
How much is lost? – *German edition*



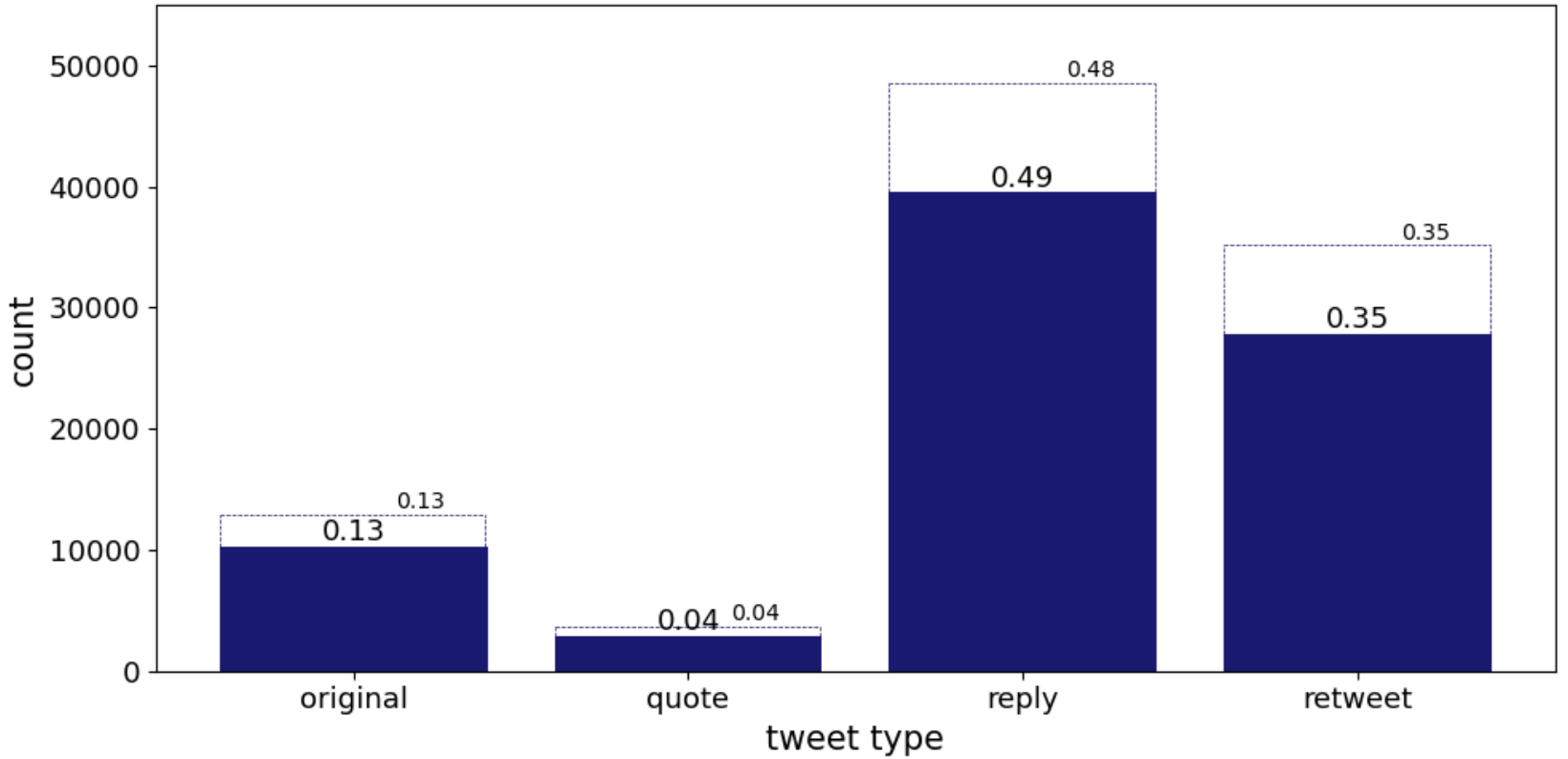
German Tweet types



German Tweet types getting lost



German Tweet types remaining



➡ The number of Tweets decreased rather uniformely across types, leaving the distribution of Tweets across types largely unaltered.

Top tokens per availability status

status	available		unavailable (auth.)		unavailable (not found)	
rank	**token**	**c-tf-idf**	**token**	**c-tf-idf**	**token**	**c-tf-idf**
1	mal	0.003816	mal	0.002435	einschalten	0.005324
2	mehr	0.002855	heute	0.002340	stream	0.005154
3	ja	0.002672	mehr	0.002293	the	0.003056
4	heute	0.002531	schon	0.002053	edition	0.002591
5	immer	0.002497	nacht	0.001808	german	0.002341
6	einfach	0.002406	via	0.001745	for	0.002335
7	gut	0.002150	gute	0.001731	nft	0.002312
8	macht	0.002123	immer	0.001671	and	0.002170
9	wer	0.001999	ja	0.001667	dance	0.002020
10	geht	0.001908	gut	0.001544	kit	0.001988

c-tf-idf idea and implementation from Maarten Grootendorst (https://github.com/MaartenGr/cTFIDF).

gesis Leibniz Institute for the Social Sciences

Most „productive“ Users

	author (rank)	Tweets	available	unavailable (auth.)	unavailable (not found)
retweets animal rescue page →	1	87	1	0	86
retweets anti-government content →	2	85	80	4	1
retweets anti-government content →	3	83	70	1	12
retweets anti-government content →	4	80	75	2	3
retweets pornographic content →	5	78	1	0	77
retweets anti-government content →	6	71	0	0	71
retweets @NVIDIAGeForceDE →	7	70	70	0	0
retweets anti-government content →	8	66	55	2	9
retweets anti-government content →	9	64	60	2	2
tweets songs currently “on-air” →	10	59	0	59	0

gesis Leibniz Institute for the Social Sciences

Tweet topics – available

Topic	Count	Name	Representation	Representative_Docs
-1	4150	-1_mal_heute_mehr_immer	[mal, heute, mehr, immer, ja, schon, via, waru...	[uhr live wer dabei , heute schon jahr...
0	284	0_ukraine_putin_russland_referenden	[ukraine, putin, russland, referenden, putins,...	[sperrt russland sogar gaslieferungen china ...
1	192	1_latenightberlin_klaas_jakob_cringe	[latenightberlin, klaas, jakob, cringe, joko, ...	[latenightberlin , antwortmöglichkeiten bl...
2	186	2_tag_tagesthemen_wickert_tage	[tag, tagesthemen, wickert, tage, miosga, ulri...	[tagesthemen caren miosga löst ulrich wickert...
3	163	3_maischberger_weidel_kalkofe_lanz	[maischberger, weidel, kalkofe, lanz, wagenkne...	[bitte los ard weidel ukraine maischber...
4	130	4_twitter_tweets_tweet_instagram	[twitter, tweets, tweet, instagram, rabattcode...	[gute nacht twitter , gute nacht twitter, ...
5	100	5_musik_band_songs_music	[musik, band, songs, music, album, musikzeit, ...	[musik schnürt herz zusammen, musik datent...
6	98	6_valentine_busch_schmitt_schlft	[valentine, busch, schmitt, schlft, schn, schl...	[schlafen könnt , nowplaying robin schul...
7	97	7_michelle_michael_and_morgan	[michelle, michael, and, morgan, drake, ludwig...	[kaleb michael jackson federline kevin federl...
8	92	8_iran_mahsaamini_proteste_tod	[iran, mahsaamini, proteste, tod, protesten, r...	[tote protesten iran , iran erleben anf...
9	83	9_hartzundherzlich_elvis_christine_hrtesten	[hartzundherzlich, elvis, christine, hrtesten,...	[hartzundherzlich geht tür , tschüssikow...
10	77	10_gute_nacht_enough_erwacht	[gute, nacht, enough, erwacht, sinne, trumt, a...	[gute nacht, gute nacht , gute nacht]
11	73	11_gas_gasumlage_gasspeicher_winter	[gas, gasumlage, gasspeicher, winter, strom, h...	[würdest mal spd wählen atomkrieg habeck h...
12	72	12_news_published_been_has	[news, published, been, has, headline, gendert...	[news städtische versorger könnten erste opfe...
13	70	13_datingohnegrenzen_timeline_dating_gino	[datingohnegrenzen, timeline, dating, gino, mi...	[sagt datingohnegrenzen, sehen datingohn...
14	70	14_buyer_seller_magiceden_tx	[buyer, seller, magiceden, tx, bmo, sol, sold,...	[bmo sold sol magiceden v tx sell...
15	69	15_brechen_afdpolitiker_effzehmv_mitgliederver...	[brechen, afdpolitiker, effzehmv, mitgliederve...	[scharfe kritik afdpolitiker brechen reise d...
16	69	16_deutschland_deutschlands_durchschnittspreis...	[deutschland, deutschlands, durchschnittspreis...	[deutschland today wurde gerade veröffentlicht...
17	68	17_film_netflix_filme_filmen	[film, netflix, filme, filmen, hellraiser, tra...	[schauspielerin anne zander spielt hauptroll...
18	67	18_derschiffsarzt_eindruck_omi_asexuellen	[derschiffsarzt, eindruck, omi, asexuellen, fr...	[frau ärztin eifersüchtig derschiffsarzt, ...
19	66	19_uhrzeit_aktuelle_remain_chelsea	[uhrzeit, aktuelle, remain, chelsea, salzburg,...	[aktuelle uhrzeit , aktuelle uhrzeit ...
20	66	20_bayern_bundesliga_fc_dortmund	[bayern, bundesliga, fc, dortmund, krise, fifa...	[woran gearbeitet bayern krise eintrac...

Tweet topics – unavailable (not found)

Topic	Count	Name	Representation	Representative_Docs
-1	706	-1_ich_und_die_ist	[ich, und, die, ist, das, der, nicht, zu, mit,...	[Vergiss nicht, es ist dein Leben! Mach was du...
0	137	0_dunsparce_top_my_on	[dunsparce, top, my, on, nft, friends, pc, no,...	[Dunsparce\nIV58 0/14/12 CP1072 LVL25\nGL Rank...
1	134	1_ich_nicht_so_mich	[ich, nicht, so, mich, das, soll, aber, mehr, ...	[ICH WÜRDE MIT DIR ALLES TEILEN AUCH WENN ICH ...
2	104	2_und_ein_einfach_ich	[und, ein, einfach, ich, die, ist, so, fr, hat...	[Holt euch den Artikel Romantisches, den ich ...
3	69	3_die_der_und_deutschland	[die, der, und, deutschland, ist, auch, wird, ...	[Genau so wie ich es immer gesagt habe - auch ...
4	61	4_russland_ukraine_der_putin	[russland, ukraine, der, putin, die, in, und, ...	[Über die Ereignisse in der Ukraine und Russla...
5	56	5_nacht_gute_euch_gut	[nacht, gute, euch, gut, ich, guten, dir, alle...	[Gute Nacht https://t.co/458usWX5ZG, Gute Nach...
6	52	6_twitter_tweet_auf_und	[twitter, tweet, auf, und, hier, ihr, tweets, ...	[Kennt ihr auch diese Tage, an denen ihr super...
7	35	7_maischberger_weidel_bei_lanz	[maischberger, weidel, bei, lanz, rr, alice, h...	[#Achduscheiße\n\n#Wagenknecht bei #Lanz\nund ...
8	30	8_tag_tagesthemen_wickert_in	[tag, tagesthemen, wickert, in, tage, die, ulr...	[Ulrich Wickert überraschend noch mal in den #...
9	27	9_twitch_ich_live_bin	[twitch, ich, live, bin, video, stream, gegen,...	[Hey wir sind jetzt Live auf Twitch. Wer mich ...
10	23	10_latenightberlin_hat_ich_witzig	[latenightberlin, hat, ich, witzig, schadeners...	[Es wird nie wieder wmjk oder dudg geben, weil...
11	21	11_frauen_frau_das_namen	[frauen, frau, das, namen, power, judentum, ch...	[Frauen besingen das Blut von #Jesus!\nDas hat...
12	19	12_zib2_vdb_orf_ist	[zib2, vdb, orf, ist, entlassen, zimmermann, k...	[#ZiB2 heute und VdB nicht bei Armin Wolf sond...
13	19	13_musik_song_playlist_rap	[musik, song, playlist, rap, deutschrap, you, ...	[Wer diesen You thought I was feeling you munc...
14	18	14_entdeckt_queen_lego_einen	[entdeckt, queen, lego, einen, mir, ich, den, ...	[Mir isr so kalt ich raste aus ich werde mir 2...
15	18	15_winter_habeck_wird_robert	[winter, habeck, wird, robert, gerne, man, auf...	[Ja wen haben wir denn da?\n\nRobert Habeck ge...
16	16	16_schwarz_kurzhaar_katze_getigert	[schwarz, kurzhaar, katze, getigert, weiblich,...	[#Katze #gefunden https://t.co/doJP3GPX8Z 8730...
17	16	17_internet_auf_kann_cyberpunk	[internet, auf, kann, cyberpunk, sehr, mitmach...	[kann mir jemand helfen mein internet geht nur...
18	14	18_herrin_geldherrin_goddess_findom	[herrin, geldherrin, goddess, findom, yootalk,...	[Will ich erstattet haben, sowie ein neu gefül...
19	12	19_temperatur_00kmh_luftdruck_luftfeuchtigkeit	[temperatur, 00kmh, luftdruck, luftfeuchtigkei...	[WSHoH 2\n🕒 22h | 🌫 Klar mit Nebel\n\nWind: Wi...
20	11	20_schlafen_schlaf_gehen_zu	[schlafen, schlaf, gehen, zu, kann, abgrundtie...	[Ich kriege beim schlafen immer Panikattacken ...

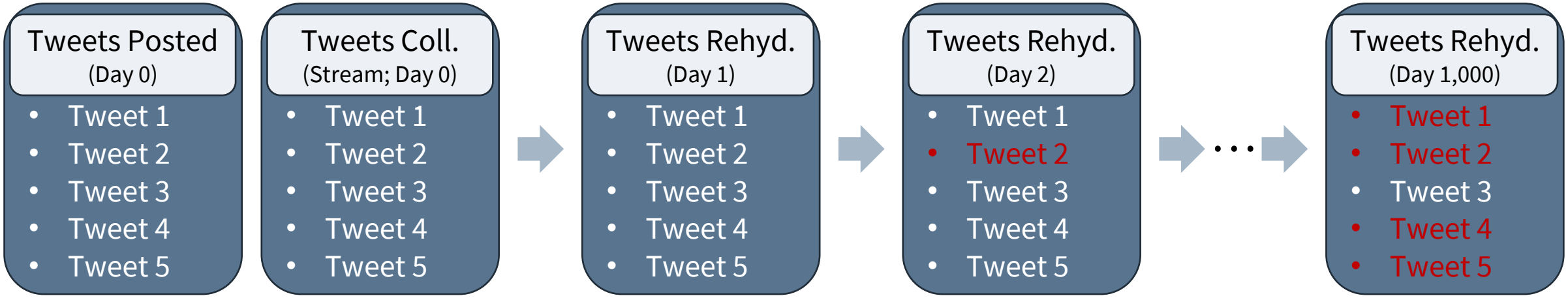
Tweet topics – unavailable (authorization)

Topic	Count	Name	Representation	Representative_Docs
-1	31	-1_massage_dubai_in_planet	[massage, dubai, in, planet, garfield, vip, lu...	[VIP massage luxury in DUBAI https://t.co/pHV3...
0	606	0_die_ich_und_auf	[die, ich, und, auf, ist, von, der, das, jetzt...	["und du" ist die ineffizienteste 'frage' die ...
1	193	1_kit_for_pack_12	[kit, for, pack, 12, amp, glass, steel, and, b...	[ROG Strix GL10 Gaming Desktop PC, AMD Ryzen™ ...
2	89	2_the_blowjob_book_and	[the, blowjob, book, and, auburn, michelle, tr...	[Blowjob Cosplay Cum Cute TikTok\n\n#Blowjob #...
3	67	3_daum_094psszhsj_xrp1337_wah79j1fvg	[daum, 094psszhsj, xrp1337, wah79j1fvg, wo6xzu...	[gebte KSHSSGSVXVVX X XVXDHSHE, tchuhiraum dau...
4	23	4_nft_ethereum_degen_btc	[nft, ethereum, degen, btc, nfts, crypto, bitc...	[NFT news.\n\nNFT Marketplace OpenSea to Suppo...
5	22	5_shgesoig_httpstcovihhvvwufw_togehter_4tyk4jh52	[shgesoig, httpstcovihhvvwufw, togehter, 4tyk4...	[To spark, often burst in hard stone. -- Willi...
6	18	6_nftcoin23_yogapetz_keung_busterdustland	[nftcoin23, yogapetz, keung, busterdustland, r...	[the_dustland Yogapetz rumjahn keung buster_du...
7	11	7_immenseharmony_boreunwieldy_underwearnebulou...	[immenseharmony, boreunwieldy, underwearnebulo...	[Stepmom👄 https://t.co/fxk9x6slPZ, ❤️❤️❤️Gege❤️...
8	11	8_wl_mint_master_jackquaid92	[wl, mint, master, jackquaid92, plagucidaskora...	[WL airdrop @tobyjfynnsen @Renlaoshi_eth @ys00...

All Topic Models shown here were done using BERTopic (https://maartengr.github.io/BERTopic/).
Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Dreaming of a perfect API world …

„Half-life of Tweets“



„Tweet Decay Probability“ $P(\text{Tweet has disappeared}|\Delta t)$

Results & Observations

- Tweets decay at a (surprisingly) high rate and pace – ~30% of tweets disappeared after only 6 months
- Tweets decay at different rates for different types, languages, contents, …, with immediate implications for, e.g., the development of ML methods
- Disappeared Tweets tend to be slightly more toxic, as well as slightly more likely to show signs of spam
- There is no way of knowing the actual reason for a Tweet's disappearance, but many alternative candidates
  - No way of knowing the "mechanism", e.g., did only the Tweet or did the whole User disappear?
  - *Content analysis* of available vs. unavailable Tweets much more ambiguous than expected

Considerations & Questions for the future

- What should future data access ideally look like? Can we ever be content with not having access to the live feed of a platform's activity?
- Knowing that 30% of Tweets have disappeared after only 6 months – how should we reflect upon data quality of existing collections? When is the decay problematic / not problematic?
- (How) Do we need to monitor the disappearance of Tweets from our datasets?
- Should we prioritize the collection of certain types or contents of Tweets to avoid their untimely loss?

References

Pfeffer et al. (2023) - Just another day on twitter: A complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 17, pp. 1073-1081). https://doi.org/10.1609/icwsm.v17i1.22215

Assenmacher et al. (2023) – The End of the Rehydration Era – The Problem of Sharing Harmful Twitter Research Data. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*. https://doi.org/10.36190/2023.56

c-tf-idf idea and implementation from Maarten Grootendorst (https://github.com/MaartenGr/cTFIDF).

All Topic Models shown here were done using BERTopic (https://maartengr.github.io/BERTopic/).
Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.