The German-Speaking Twitter Community Reference Data Set

Johannes Pflugmacher and Stephan Escher
TU Dresden, Germany
<firstname.lastname>@tu-dresden.de

Jan Reubold and Thorsten Strufe
Karlsruhe Institute of Technology (KIT), Germany
<firstname.lastname>@kit.edu

***Abstract*—News providers and politicians increasingly publish and disseminate their content on online social media to reach broader audiences effectively. Directed by ubiquitous mobile use, the majority of individuals reportedly consume daily news directly on these platforms, mainly in an incidental manner. This bears many risks of misconceptions and misinformation: Social media users tend to extend unwarranted trust in posts that are distributed by contacts on the platform and therefore have difficulties evaluating the credibility and trustworthiness of information and its sources. Reduced political proficiency and social understanding have been reported as directed results, as well as the risk of succumbing to partisan echo chambers, user groups amplify and reinforce their own beliefs due to almost exclusive exposition.**

Measuring and understanding these phenomena requires analysis of the user behavior on these platforms, and a virtually complete data set of one representative community. We focus on Twitter and present collection techniques to obtain a complete data set of specified sub-groups of its users, with the example of the German-tweeting community, in this paper. We show how to collect a representative snapshot of all tweets pertaining to this community over the period of two months. The resulting sample includes 77 million tweets and 6.9 million users. We validate the sample with exhaustive evaluations, and identify the notable impact of political events, such as the 2019 European Parliament election.

***Index Terms*—Twitter, OSN, Behavior Analysis, Community**

I. Introduction

The rapid development and mass adoption of social media, platforms such as Facebook, Twitter, and YouTube, have amassed remarkable influence over the last decades. Several studies reveal the trend of consuming news exclusively through these channels [1], [2]. The reasons are manifold and range from facilitating access (commonly free of charge) to democratizing the publishing industry, since the costs for production and publication are essentially disappearing. However, democratization entails new challenges.

Consumers had to learn and judge some, to a few dozen newspapers, TV stations, and other media channels in the past. The plethora of news sites, blogs, and video platforms that have emerged in the meantime complicate this endeavour [3]. The number of potential sources of information fed into social media [4] and the resulting flood of information, not least due to casual news consumption in social networks (OSNs) [2], mean that information and its sources are increasingly either placed in the wrong context or judged wrongly in terms of their trustworthiness. Profit-oriented false contributions, targeted propaganda or simple satire contributions thus gain relevant attention [5].

This work in part was funded by DFG EXC 2050/1 - ID 390696704 'CeTI'.

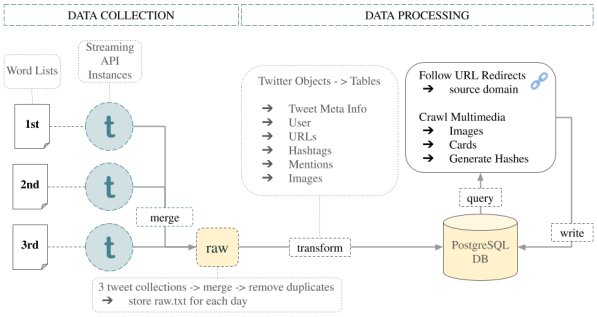

Fig. 1. Data collection pipeline.

Analysing such impact on well-defined subgroups, or communities requires a virtually complete data set of its corresponding activities on the platform [6]. Data sets covering international users who tweet in English are simple to obtain. It is questionable if conclusions on this set are transferable to other subgroups, but sampling other communities has not convincingly been done, so far [7].

Therefore, in this work, we propose

- a well-documented method that is fully reproducible and defines the requirements for uniform data collection of tweets published in specified target languages (Fig. 1),
- a strategy to categorize domains from hyperlinks in an automated manner, classifying the majority of domains from hyperlinks in our data set.

For the data collection process, we chose a time-frame based on the 2019 European Parliament election (23-26 May 2019), collecting data between the 2nd of April and the 2nd of June. We focused on the German-speaking Twitter community, with the goal to collect a virtually complete data set. Based on this collection, we report

- what type of external content (e.g., news, multimedia streaming, lifestyle, marketing, spam, etc.) is prominent in the entire network,
- behavior patterns of users in the German Twitter community,

- how different types of content are received by the German Twitter-sphere.

In the rest of the paper we first explore the related work in §2, then present our data collection and a description of the sample in §3, analyse the sample in depth in §4, and finally conclude our work.

II. Related Work

The goal of this paper is to capture a virtually complete snapshot of the German Twitter traffic. Therefore, in the following, we take a closer look at state-of-the-art data collection techniques for OSNs.

From a data mining perspective, a complete data set on the usage of an OSN is comprised of all user-generated data within a specific time frame. In the past, the academic community leveraged methods towards collaborative data collection [8]. However, according to Twitter's policies, the public sharing of its particular contents is prohibited [9]. Therefore, researchers started to develop customized data crawling techniques that fit their particular research scope. A reliable data collection process should also be transparent and reproducible for evaluation through future research.

In 2010, Kwak and others [10] crawled the entire Twitter platform. Utilizing 20 machines operating with different IPs that crawled tweets via the Twitter Search API over several weeks to bypass Twitter's rate-limit, they obtained 41.7 million user profiles, 1.47 billion social relations, $4,262$ trending topics, and 106 million tweets. Following the growth of Twitter over recent years, this approach is prohibitively costly and time-intensive, and it can be considered infeasible to collect a complete data set.

Analysing a specific subgroup of Twitter users (German users in our example) in a complete data set requires thorough pre-processing, as there are 500+ million new tweets generated every day. Recently published studies avoid this overhead by using Twitter's Streaming API, which allows researchers to obtain a limited number of real-time tweets that match a specific word-filter. Accordingly, the size of the acquired data set depends on the prevalence of the determined search terms (e.g. event-related hashtags) [11]–[13]. A downside of the Streaming API is that Twitter restricts the total number of tweets that can be crawled per day to 1% of all data. If the number of tweets matching a word-filter exceeds the limit, the stream will return a random sample of all matching tweets. For research purposes, this is an undesired outcome, as studies have revealed that the Streaming API provides a non-representative sample tweets [14]. Furthermore, it is not sufficient to fix the sample discrepancy by using multiple machines to combine simultaneous samples from the Streaming API [15].

Scheffler captures a representative snapshot of the German Twitter traffic despite the limitations [7]. She configured Twitter's Streaming API based on an exclusively German word-filter list. Since the number of German tweets within the whole Twitter-sphere is considerably small, the number of captured tweets only slightly exceeded the 1% limit. Thereby, she minimized the effects of Twitter's downsampling. By collecting every tweet that matched at least one word from the word list, Scheffler also collected a great number of tweets that were not German. She used a language detection algorithm to filter for German tweets in consequence. However, due to insufficient labeled data, the algorithm had to be evaluated manually on a small subset of the captured tweets. Regarding the effects of Twitter's downsampling, Scheffler concluded that these were negligible, as they accounted for under 3% of missing data. Nonetheless, it must be noted that the data collection process was conducted in 2013. Therefore, a future-proof data collection method should consider the possibility of a rising number of German tweets.

III. Data Acquisition and Sample Description

In the following we describe the sampling process and provide information on our data set. After giving a short overview of the type of data that is generated on Twitter, we detail the different elements of our algorithms. Finally, we describe our sample, providing information on the quantities of different types of data contained.

A. Twitter OSN and Functionalities

Twitter offers its users different types of *Tweet-objects* to generate content on the platform. *Original Tweets* are the the basic way of posting. The user can write a message also known as status update to his Timeline. The Timeline of a user represents a roster of posts to record activities and make them visible to followers. Furthermore, the Timeline displays activities of *followed* others, to whom the user has subscribed. Following many others, such as news providers, celebrities, and friends, produces a news-feed like overview of current events and activities. *Retweets* are another type of post, which allows a user to copy a tweet from another user to his own Timeline. Therefore, it is visible to his respective followers and visitors. Users can also create a *Reply*, to comment on any given tweet, except Retweets. There finally are *Quotes*, which are tweets that include the tweet from another user (except retweets). Thereby, users can display the original message of another user and directly comment on it.

It is also possible to tweet multimedia and interactive content. This may be multimedia content, like (*photos*, *videos*, *animated GIFs*), interactive content (*hashtags*, *user mentions*), places (*geolocation*), and (*URLs* linking to external sources, which then commonly are visualized as *Twitter Cards*). Besides manually embedded user mentions (@username), Twitter automatically adds *mentions* in front of content that implies an interaction between users (retweets, replies and quotes).

User-Objects provide a variety of meta-data including multiple free-text fields (e.g., name, description, URL), statistics about the user's social-links, such as follower and friends count, and also statistics about the users' activities, such as favorites and tweet count (statuses).

Interactions between users manifest themselves in the form of direct User Mentions within a tweet or indirectly by using connected tweet types, such as Retweets, Quotes, and Replies.

TABLE I
TWITTER-OBJECTS CAPTURED DURING THE DATA COLLECTION PROCESS.

Object-Type	Count	Tweets with (%)	Users using (%)
Tweet	77,390,122	-	-
User	6,919,206	-	-
Mention	85,155,158	72	80
URL	18,358,074	23	25
Hashtag	39,197,019	22	29
Multimedia	19,702,261	19	56
Place	1,189,696	1.5	2.2

TABLE II
DISTRIBUTION OF TWEET VARIANTS WHEN PERFORMING ACTIONS

Action	Tweet Variant		
	Original Tweet (%)	Reply (%)	Quote (%)
Retweeting	66.8	21.7	11.5
Replying to	24.7	71.7	3.6
Quoting	76.5	14.5	9.0

In contrast to static following-links, these interactions can be used to learn more about relationship dynamics over time.

B. Data Acquisition

We propose to extend the approach by Scheffler [7] (cmp. Fig. 1). Our evaluation of different collection methods confirmed Scheffler's findings. We hence decided to leverage word-lists for our purpose. In contrast to Scheffler we do not collect-then-filter to remove tweets in other languages, but we leverage the built-in language identification of Twitter. We hence created word filters, encompassing the $1,200$ most frequent German words. Our choice is based on multiple text corpora, provided by the Leipzig Corpora Collection [16] and one corpus of frequently used words from OpenSubtitles.org[1]. The latter encompasses terms that are more prevalent in informal conversations. Twitter enforces a maximum of 400 keywords per instance, so we divided our word-filter into 3 different lists and utilized three individual data streams in parallel. We ranked these by the amount of captured tweets during a test run. Based on the ranking, the 400 most frequent words from the first corpus were used as a filter for the first Streaming API instance. The remaining 800 words are a combination of the remaining top words from each corpus. All streams obtained a high number of tweets from 600k to 1.2M on average. Thus our approach does not exceed Twitter's rate limitations of 1% ($\approx$ 5M tweets). During the collection process, we drop duplicated entries and merge the stream outputs.

We enrich the sample of tweets by additional data. We extracted all attributes and child objects from the collected Tweet-Objects. This may entail collecting additional (non-German) Tweet-/ User-Objects. We argue that users who do not tweet in German but interact with German tweets have to be included into the sample in consequence.

Further, we developed an algorithm that resolves shortened URLs to reveal their source domains. Leveraging McAfee's domain categorization tool TrustedSource[2] we obtained the category of each domain (e.g. News, Lifestyle, Political Opinion, Spam, etc).

Additionally, we analyzed the most prominent OSNs, measured by the amount of shared external content originated from these platforms (see Fig. 4). To enable measurements on the influence of specific personalities we then developed web crawlers for the most prominent platforms, i.e. YouTube, Facebook, and Instagram. By utilizing the YouTube Data API v3[3] and the HTML and JavaScript sources from Facebook and Instagram, we identified YouTube Channels, Facebook Pages, and Instagram profiles that were shared by the users in our corpus.

C. Collected Data

We only capture German tweets, so our approach does not exceed Twitter's rate limitations, and the data is not subject to downsampling. Instead of collecting and subsequently filtering for German tweets, as Scheffler, we rely on Twitter's language detection and the completeness of our data depends on it. Twitter's algorithm is missing a thorough documentation. However, research showed that Twitter's language identification can outperform established alternatives, such as Google's Compact Language Detector [17], and we hence are optimistic that our sample covers the entirety of tweets in German.

We sampled tweets throughout a period of two months – between the 2nd of April and the 2nd of June 2019. The sample contains 77 million tweets and 6.9 million user profiles (ref. Table I for an overview). In the following we take a closer look at the data types we collected (e.g. Tweet-Types, Tweet-Content, etc).

a) Tweet Types: By categorizing these Tweet-Objects based on their tweet type (i.e., Original Tweet, Retweet, Reply, Quote), we found that the most frequent action was Retweeting. The majority of activity in our sample is not innovative, but reactive. Retweets account for 38% of all tweets in our corpus and are used to distribute content from other users, Replies for 31%, and Original Tweets, creating novel content or initiating conversations, account for only 27%. Quotes are rarely used at all (3.7% of the sample). The high number of replies shows that users are willing to discuss or comment on others' content. Interestingly, we found that fewer users in our sample use Replies (23%) than Retweets (64%). As there are more replies in total than retweets, this shows that replying users are more active than those who retweet.

Besides investigating tweet types, we also analyzed their interaction. Table II shows that the reaction to Tweets commonly was to be retweeted (66%), followed by Replies (21.7%) and only rare Quotes (11.5%). Looking at Quotes, the distribution of reactions is very similar. For Reply tweets, however, the majority of reactions are other Replies (71.7%).

[1]https://github.com/hermitdave/FrequencyWords/

[2]https://trustedsource.org/

[3]http://youtube.googleapis.com

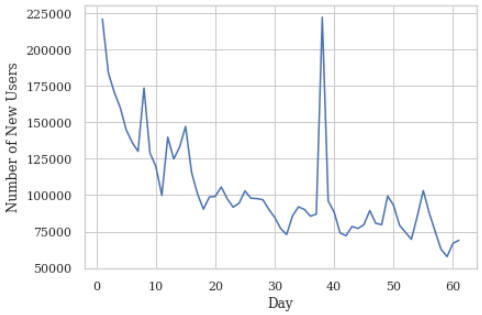

Fig. 2. Daily number of new users captured during the data collection process (an international soccer match explains the peak at day 38).

b) Tweet Content: The content of each tweet can consist of text and/or additional, interactive content (Entity-Objects). Table I shows statistics on the usage of different content types.

The most prominent Entity Objects were *Mentions* (85 million). Since every retweet, reply and quote contains at least one Mention to the originator of the tweet, we conclude that 33% of the 85 million mention-objects are User Mentions, which are added manually into a Tweet in form of @username. *URLs* (18 million) are the second most prominent objects which are found in 23% of all tweets. There were 6,667,962 distinct URLs shared that originated from 275,078 different domains.

Beside these external sources we extracted 19.7 million (5,874,013 distinct) multimedia-objects. The majority of the multimedia contents shared are photos (82%), followed by videos (12%) and animated GIFs (4%), shared by a total of 56% of the users. It must be noted that we can only obtain multimedia content from tweets that also contain text, as at least a single word is needed to identify a tweet to be German. Further, 29% of the users in our data set shared 39 million hashtags in 22% of all tweets. However, while there are more Tweets with Hashtags in our corpus than multimedia-objects, a larger group of users share multimedia contents (56%) than Hashtags (22%). We also found that users using Hashtags are about two times more active on Twitter than users sharing multimedia content, which explains this effect to some extent. A feature that is almost entirely neglected by the majority of users in our data set is the submission of geolocation data (*Places*). Only 2% of the users share their location when tweeting.

c) Users: Regarding users, we can see a steadily decreasing volume of new Twitter users each day (see Fig. 2). By observing the hashtag usage of new users on the day with the unexpected peak, we could identify many soccer fans among them. The Hashtag #CHESGE was used by many users, following the UEFA Europa League semi-final between the Chelsea F.C. and Eintracht Frankfurt.

Based on the users captured during the data collection, we calculate the mean and median values of a variety of meta-data of the User-Objects (see Table III). The mean values are heavily influenced by users with unusual high activities.

TABLE III
USER ACTIVITY BASED META-DATA

Meta-Data	Count		Growth over time	
	Mean	Median	Mean	Median
time elapsed	-	-	39 days	28 days
followers	4375	178	+124	+1
friends	705	251	+23	+3
favorites	11465	2755	+1012	+131
tweets	15715	3513	+953	+178

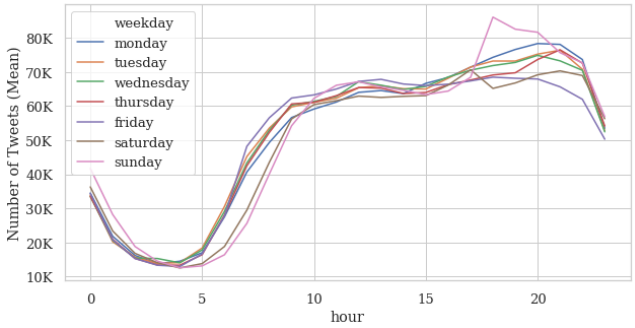

Fig. 3. Twitter activities over the course of every weekday.

Therefore, the median is better suited to derive the central tendency of the average user activities in our German tweet corpus. Based on these values, we conclude that the majority of users in our data set have already established their profiles and maintained their activity over a longer period of time.

Since users can generate an arbitrary number of tweets, we also captured multiple User-Objects from the same users. Whenever users tweeted something, we obtained an updated version of their profile. This allowed us to record the evolution of the meta-data over time. We followed the changes of 2,841,529 user profiles within our Twitter corpus. Table III shows the evolution of the statistical values over time.

IV. Experiments for User Behavior

In the following we present analyses and report on findings regarding the user behavior of the German-speaking Twitter community.

A. Tweets over Time

The volume of daily captured tweets varies from 1M to 1.6M messages with an average of 1.2M. By examining the average collection of tweets by weekdays, we observed that German-speaking Twitter users were more active from Sunday to Tuesday and had a decreasing interest in Twitter from Wednesday to Saturday, with the lowest activity on Saturdays. The overall daily usage (see Fig. 3) is moderate in the morning, increases during after-work hours, and drops to its lowest point at night between 1 am and 5 am. At the weekend, Twitter usage naturally starts a few hours later in the morning. The oddly shaped peak on Sunday evenings is the result of the high volume of tweets during the election night of the 2019 European Parliament election. The daily Twitter activities

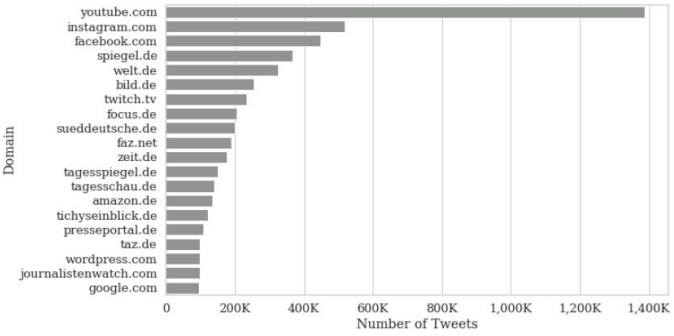

Fig. 4. Distribution of domains of shared external sources.

Category	Tweets %	Users %	URLs %	OT %	RT %	RP %	QT %	Third Party %
General News	32	21	23	40	57	2	1	23
Blogs/Wiki	10	16	10	52	44	3	1	36
Streaming Media	10	36	8	42	52	6	1	17
Media Sharing	8	33	6	39	54	7	1	14
Social Networking	7	18	12	69	30	1	0.36	64
Entertainment	4	10	4	56	41	2	1	38
Business	4	8	5	65	31	3	2	42
Politics/Opinion	3	5	1	27	68	4	2	14
Internet Services	3	7	4	73	24	2	1	60
Marketing/Merchandising	3	5	4	73	24	3	1	59
Sports	3	4	3	66	32	1	0.50	49
Games	3	5	2	57	42	1	1	37
Online Shopping	3	4	3	71	25	3	0.49	52
Public Information	2	3	2	68	29	3	1	59
Pornography	2	2	2	49	51	0.45	0.06	63

Fig. 5. The 15 most distributed categories within the German Twitter-sphere.

matches Central European Time and the working schedule of people from Germany and Austria.

B. Occurred Events

Looking at the statistics of daily usage we observe unusual peaks of tweet traffic. By examining the most popular Hashtags during these high peaks, we identified the corresponding influential events. Overall, there is a high amount of politically motivated Hashtags every time we observed an excessive increase in Twitter usage (e.g. #Strache, #Rezo). This indicates a high interest of German-speaking Twitter users in politics. We observed the highest activity at the end of the election campaigns of the 2019 European Parliament election (26th of May). The top Hashtags shared on Twitter during this period corresponds to the election and discussions on election results. When comparing the hashtag popularity of parties and their election results we come to the conclusion that the popularity of Hashtags is rather the result of lively discussions than a reflection of political-party affiliation. The far-right party Alternative für Deutschland (AfD), for instance, is close to leading the Hashtag ranking (#AfD), even though it only came in fourth place in the elections.

Besides political events, the increase in daily Twitter volume is caused by pop-cultural events (e.g., #GNTM, #ESC2019). There are also non-German Hashtags referencing the Korean pop band BTS, which achieved high music chart rankings over several weeks in Germany,released a single, generating several trending hashtags. Nevertheless, the majority of the top Hashtags correspond to events within German-speaking countries. These events also dominated the news in Germany during the data collection period. Therefore, we can conclude that our data collection strategy results in a corpus that correctly captures German tweets.

C. External Media usage

When analyzing the usage of external media sources we examined the 20 most shared domains (see Fig. 4). Overall 13 of the 20 domains link to popular German news providers/political blogs. However the most shared URLs link to other OSNs. In order to give a complete picture of shared external media sources we resolved links to YouTube, Facebook, and Instagram to identify popular YouTube Channels, Facebook Pages, and Instagram profiles.

a) Online Social Networks: We discovered 1.4M tweets that shared 374k distinct YouTube-URLs. While the number of shared Instagram-URLs (520k) is only a third of the distributed YouTube-URLs, they contain a similar number of distinct URLs (370k). The same finding is true when looking at content from Facebook. Regarding the type of media shared via these platforms YouTube links seldom contained other content then video links (97%). These videos originated from 97k YouTube channels. Via Instagram the most common shared media types are images (71%) followed by PostPages (12%), which also contain multimedia content and profile pages (10%). The content from Facebook-links is mainly textual (post: 53%; story: 13%) and less multimedia-based (photo: 10%; video: 8%). There are only a few events and groups shared within our corpus.

b) Functional Groups: Given the collected and complemented URLs, our approach automatically categorized 98.3% of all shared URLs. Besides the 100 hierarchical categories, McAfee provides 12 semantic subsets of categories called Functional Groups (FGs). Based on this preliminary work combined with the FGs, we report on the media-consuming behavior of a quarter of the entire German Twitter population.

c) Largest user base: The group with the largest user base is *Entertainment/Culture*. The majority of the users are interested in multimedia content, such as videos and photos (Streaming Media: 36%; Media Sharing: 33%). Users often add these contents to their Replies to communicate with each other via memes and videos. Based on our YouTube investigation, we know that most of the Streaming Media content belongs to music and political events regarding the 2019 European elections.

d) Most traffic: The German user base generates most of its traffic in the *Information/Communication* group (47% tweets). Most of this content is related to news (General News: 32%) and personal blogs (Blogs/Wiki: 10%), which both mainly consist of content from online news media and personalized political websites (see Fig. 5). Based on the high number of Retweets in this group (52%), news and blog content seems to be well-received by the German user base. We observed the same popularity of political domains

in the Functional Group *Society/Education/Religion*, which is comprised of even more elaborate political content. Most of the URLs captured during our data collection are from domains within the Information/Communication, which means a high number of distinct news articles are generated and distributed on Twitter. Despite this variety of articles, the Twitter community still reacts to these links by spreading them via Retweets and Replies. In contrast to news and political content, lifestyle-related content (Functional Group: Lifestyle) results in fewer Retweets (34%), which indicates a lower acceptance by the German community. An exception in this group is the category of Controversial Opinions, which includes domains that share highly opinionated political content (e.g., journalistenwatch.com, philosophia-perennis.de, pi-news.net). The number of Retweets in this category is 70%, which further supports the assumption that political content on Twitter is widely distributed and acknowledged.

e) Most original tweets: The Functional Groups with the most Original Tweets are related to marketing campaigns (*Purchasing*: 75% OT), business advertising (*Business/Service*: 71% OT), and online technologies (*Information Technologies*: 69%). The majority of tweets in these groups are generated by third-party services. We assume that most of these domains conduct an automated distribution of their products as a marketing strategy. The lack of Retweets within the respective categories indicates that this distribution approach is not overly effective in the German Twitter community.

f) Malicious Content: The share of spam and inappropriate content is relatively small (overall 4 to 7%). Moreover, there is just a small margin of users involved in the distribution process. Spam URLs found in our data were mainly shared via Original Tweets (97%) and distributed via third-party services (92%), which further confirms the use of automated distribution in terms of scam or marketing. Based on the low number of retweets, users recognize spam content and do not distribute these any further in the network.

g) Shortened URLs: 14% of URLs shared on Twitter are disguised with link shortening services. By resolving these links, we discovered that news providers and bloggers use marketing services to distribute their content in an automated manner. Another big share of disguised URLs is related to inappropriate content, such as pornography.

V. Conclusion

Analysing a well specified subgroup of all Twitter users can shed insight into regional and cultural differences in social media use. Large data sets of English tweets, or international communities have been published, but it was difficult to extract a corpus corresponding to a specific group, so far. In this paper we proposed and documented an algorithm to collect of exhaustive set of behavioral data of Twitter users defined by the language of their tweets. We adjust the Streaming API parameters to filter for German tweets to significantly decrease the volume of returned tweets. This allows us to stay free of Twitter's 1% threshold. After sampling German tweets for two months, we introduced the collected data set and provided detailed evaluations. These analyses exposed several peculiarities in the behavior of German Twitter users. Sharing and consuming of external media sources is common in the German-speaking community, for instance, and hashtag popularity is the result of lively discussions rather than a reflection of political-party affiliation. Additionally, peaks of discovered users during events such as the football game (Chelsea F.C. vs Eintracht Frankfurt) suggests that there exists a sizable group of Twitter users that only act on special occasions. Further, based on the small median growth of followers and friends, we conclude that the users within our data set are not eager to extend their social-links to other users, rather, they use tweets and favorites to actively participate on the platform. For future work, time-series of user-objects are a great opportunity to observe the evolution of specific users over a longer period of time, i.e. to measure the popularity of politicians. So far we have not found studies that made use of historical data provided by Twitter's Streaming API. The evaluations in general suggest that the quality of the collected data enables research on and in-depth analyses of the inner workings of the German Twitter community.

References

[1] Sascha Hölig and Uwe Hasebrink. Reuters digital news report - ergebnisse für deutschland. Technical report, Reuters, 2018.

[2] Pablo J Boczkowski, Eugenia Mitchelstein, and Mora Matassi. "news comes across when i'm in a moment of leisure": Understanding the practices of incidental news consumption on social media. *New Media & Society*, 20(10):3523–3539, 2018.

[3] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating information: The cornerstone of civic online reasoning. Technical report, 2016.

[4] Nicole Ernst, Sven Engesser, Florin Büchel, Sina Blassnig, and Frank Esser. Extreme parties and populism: an analysis of facebook and twitter across six countries. *iCS*, 20(9):1347–1364, 2017.

[5] Caroline Meyer and S Reiter. Vaccine opponents and sceptics. history, background, arguments, interaction. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 47(12):1182–1188, 2004.

[6] Thorsten Strufe. Profile Popularity in a Business-oriented Online Social Network. In *EuroSys/SNS*, 2010.

[7] Tatjana Scheffler. A german twitter snapshot. In *LREC*, 2014.

[8] Cong Ding, Yang Chen, and Xiaoming Fu. Crowd crawling: Towards collaborative data collection for large-scale online social networks. In *Proceedings of ACM COSN*, 2013.

[9] Audrey Watters. How recent changes to twitter's terms of service might hurt academic research. *Read Write*, 2011.

[10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *The WebConf*, 2010.

[11] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *SSCORE*, 31(6):649–679, 2013.

[12] James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139, 2013.

[13] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[14] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of AAAI ICWSM*, 2013.

[15] Kenneth Joseph, Peter M. Landwehr, and Kathleen M. Carley. Two 1%s don't make a whole: Comparing simultaneous samples from twitter's streaming api. In *Proceedings of SBP*, 2014.

[16] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of LREC*, volume 29, 2012.

[17] Bogdan Pavliy and Jonathan Lewis. The performance of twitter's language detection algorithm and google's compact language detector on language detection in ukrainian and russian tweets. *Bulletin of Toyama University of International Studies*, 8:99–106, 2016.