Humboldt-Universität zu Berlin

Department of Social Sciences

Master Studies in Social Sciences

Summer Term 2017

Course 53112: Web Data Collection and Social Media Mining

Teacher: Dr. Simon Munzert

Research Paper

Twitter Bots in the German Election

Employing OSoMe's *Botometer* in the German Twittersphere

By: Tim König

Enrolment number 584174

tim.koenig.2@student.hu-berlin.de

0176/80180070

Weserstr. 178, 12045 Berlin

Berlin, the 5th of January 2018

Content

Introduction 1

1. Methodology 3

1.1 The data sample 3

1.2 Employing OSoMe's *Botometer* API 4

2. Data analysis 6

3. Limitations of the approach 10

Conclusion 12

Bibliography 13

Appendix: hashtags used for generating the sample 14

Introduction

The emergence of social media stirs politics and social sciences alike. Conceptions of democratic discourse and campaigning need to be reapplied and reimagined in light of new forms of conversing, discussing and information diffusion (Murthy, 2012; Owen, 2014). At the same time, social media is becoming a fixed part of political campaigning in modern democracies (Lilleker et al. 2015). While social media has played a major role in U.S. campaigning since Obama's successful hybrid media approach during the 2008 election (Chadwick, 2013), it became a medium for campaigning only in the 2017 German federal elections. With all major parties and 40% of the party candidates active on Facebook and Twitter (Scholz & Friends et al. 2017), social media was used extensively for advertising, informing and discussing about political events and the parties' programmes.

From being highlighted for its democratic potential during the Arab Spring (Howard & Hussain, 2011) to becoming a cited source for international news agencies through president Trump's infamous tweets, Twitter has emerged as the archetype of "new" social media that has come to stay. Twitter's potential for open, unfiltered online discussion, however, bears not only potential for true democratic discourse and the display of grassroot movements, but also for manipulation. Automated programs simulating human behaviour, so-called bots, can be used to manipulate online discourse by artificially inflating support for or smearing certain political candidates or persons, simulating grass-roots support or spreading misinformation and fake news throughout the social network (Ferrara et al., 2016). For the 2016 U.S. presidential election, Bessi and Ferrara found that as much as 400,000 bots were involved in the online discussion on Twitter, voicing support for both candidates Clinton and Trump and being responsible for roughly 20% of the total tweet volume (Bessi & Ferrara, 2016). With Twitter gaining importance for the German elections and a new populist party, the AfD *(Alternative für Deutschland),* on the rise, it seems high time to explore the possibility of social bots distorting the democratic discourse not only in the U.S., but also in European countries.

This paper aims to estimate the prevalence of Twitter bots active during the 2017 German federal election by applying *Botometer*, the same classification technique as Bessi and Ferrara in their 2016 paper on Twitter bots in the U.S. presidential election (Bessi & Ferrara, 2016). It will shed light on the process of bot detection by explaining its methodology (1.), proceed with the analysis of the data (2.) and discuss the limits of the approach for non-English tweets (3.). In analysing the data (2.), it will discuss probability scores for the bot classification in the sample, give an overview of bots, humans, suspended and deleted users in the sample and their respective tweet volume in the observation period. Furthermore, it will visualise and discuss differences in the spatio-temporal dynamics of content generation for bots and humans as well as tweet/retweet ratios for both bots and humans.

1. Methodology

1.1 The data sample

By setting up a dedicated server via Amazon Webservices and running automated scripts with the programming language *R*, I was able to stream tweets related to the German federal election for 27 days from the Twitter Streaming API, starting on September 3rd and ending on September 30th with the election being held on September 24th. By discriminating via hashtags related to the major parties, their top candidates and slogans as well as the elections in general,[1] I was able to distinguish the tweets related to the election and generate a sample of a total 2,066,488 unique tweets, published by 200,223 users. As expected, this sample for the German election is much smaller than the 20,772,153 tweets, 2,782,418 user sample generated by Ferrara and Bessi for the U.S. presidential election (Bessi & Ferrara, 2016). This, however, is not only related to a more limited amount of resources resulting in a shorter observation period (35 vs. 27 days), but mainly to a smaller user base in Germany, the smaller overall population, the language barrier for potential international participants in the online discourse and the lower geopolitical interest in the topic. Even with limited resources, however, I was able to continuously stream the election-related tweets, with only one actual data loss on September 16tth, with most of the tweets for that day, unfortunately, getting lost (for all other minor errors in the streaming process, I was able to retrieve the lost tweets via Twitter REST API). The overall lower tweet volume for the German election also meant, however, that the Streaming API could be used without running the risk of losing data. For the Streaming API limits its output at a maximum of 1% of the global tweet volume produced at any given time, a continuous use of the REST API (that is, twitter searches) can yield better, more exhaustive results (Morstatter et al., 2013). With estimates as high as 400 Million Tweets a day already in 2013 (Morstatter et al., 2013), it is highly unlikely that the 2 Million Tweet-sample generated here hit the 1% threshhold at any given point.

To avoid checking an unreasonable amount of inactive or semi-active users for their probability of being bots, the sample of 200,233 users was split into “active” and “non-active” users, with the active users being defined as users publishing at least an average of one tweet per day. This resulted in a subsample of 11,897 users classified as “active” to be used for

[1] For a complete overview of the hashtags used, see appendix 1

further analysis. As figure 1 shows, however, the active users were responsible for 1,533,311 of the total 2,066,488 tweets (roughly 3/4) during the observation period.

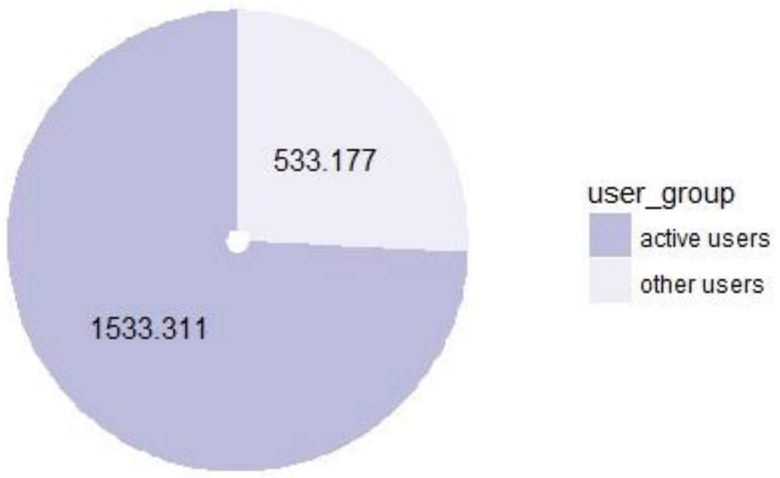

Figure 1*: Tweets by active users and other users in the sample (total of 2,066,488 tweets)*

1.2 Employing OSoMe's *Botometer* API

In order to check the 11,897 users classified as "active" for their probability of being bots, OSoMe's[2] *Botometer* API (formerly known as *BotOrNot*) was employed.[3] First released in 2014 for the purpose of identifying bots and sybil accounts on Twitter, the API outputs a given user's probability of being a bot on a scale between 0.0 and 1.0. If provided with a user's *screen name*, *timeline* and *mentions* of their screen name from Twitter REST API, *Botometer* classifies accounts based on six categories: **Network** features, such as retweets, mentions and co-occurrence of hashtags; **User** meta-data including language, geographic locations (if available), and account creation time; **Friends** features, that is, the account's social contacts in followers, followees, post, re-posts and so on; **Temporal** features based on timing patterns of generated content; features of the generated **Content** itself, based on linguistic cues; and **Sentiment** features that utilise Twitter's built-in sentiment analysis algorithms (Davis et al., 2016). Even though social bots exhibit more and more complex, human-like behaviour such

[2] https://osome.iuni.iu.edu/

[3] See https://*Botometer*.iuni.iu.edu/#!/ for a browser-friendly, easy-to-use implementation of the API

as emulating human-like temporal behaviour in their content generation and consumption, commenting on posts and entertaining simple conversations, and purposeful expanding their social circles (Ferrara et al., 2016), the *Botometer* application was reported to yield an accuracy of above 95% in its predictions and was successfully employed for an analysis of the 2016 U.S. Presidential election Twitter discussion (Davis et al., 2016; Bessi & Ferrara, 2016). While its authors remain vague on the exact algorithms of bot detection, the application employs a machine learning framework initially trained with a sample of 15,000 manually verified social bots and 16,000 legitimate human accounts and a ever growing number of further classifications (Davis et al., 2016). Evaluating more than a thousand features over the aforementioned categories, it analyses statistical features, generating a bot score for each category as well as a english-specific and an universal, language-independent total bot score.[4]

Since the tweets and users analysed in the sample for the German federal election are almost exclusively in German, only the universal bot score could be used, dropping Content and Sentiment features since they do not work reliable for non-English content (https://*Botometer*.iuni.iu.edu/#!/faq).[5] To apply *Botometer*'s bot detection algorithms to the 11,897 active users in the sample, code was written in *R* to automatically access the *Botometer* API endpoint[6] and retrieve the returned bot scores.[7]

---

[4] See Davis et al., 2016 and https://*Botometer*.iuni.iu.edu/#!/faq for more info on evaluated features and scores

[5] The relevance of this measure and its potential influence on the estimated bot scores will be discussed in section 3 of this paper (see: *3. Limits of the approch*)

[6] The API endpoint can be found, along with its documentation, on https://market.mashape.com/OSoMe/*Botometer*

[7] The code written may soon be released as a proper *R* package for public use

2. Data analysis

As proven effective in prior studies utilising the *Botometer* bot scores, the threshold to label an account as bot was set at 0.5 (Bessi & Ferrara, 2016; Davis et al., 2016). Figure 2 shows the distribution of bot score returned by *Botometer*. Interestingly, most of the probability mass is centered at around 0.2, with a pronounced bump at around 0.3. In contrast to the U.S. Presidential election, a smaller amount of Twitter bots seems to have been active in the German federal election, with only a small bump at 0.6 that suggesting users exhibiting clear bot characteristics (for comparison, see figure 1 in Bessi & Ferrara, 2016). The relatively large amount of users scoring between 0.4 and 0.6 (with most of the probability mass still below 0.4) also suggests that a high number of users were classified as uncertain by the *Botometer* API.[8] In total numbers, 1,632 of the 11,897 users (~14%) had a botscore between 0.4 and 0.6. Generally speaking, the overall bot scores were lower in the Germen federal election than in the U.S. Presidential election (see Bessi & Ferrara, 2016 for comparison).

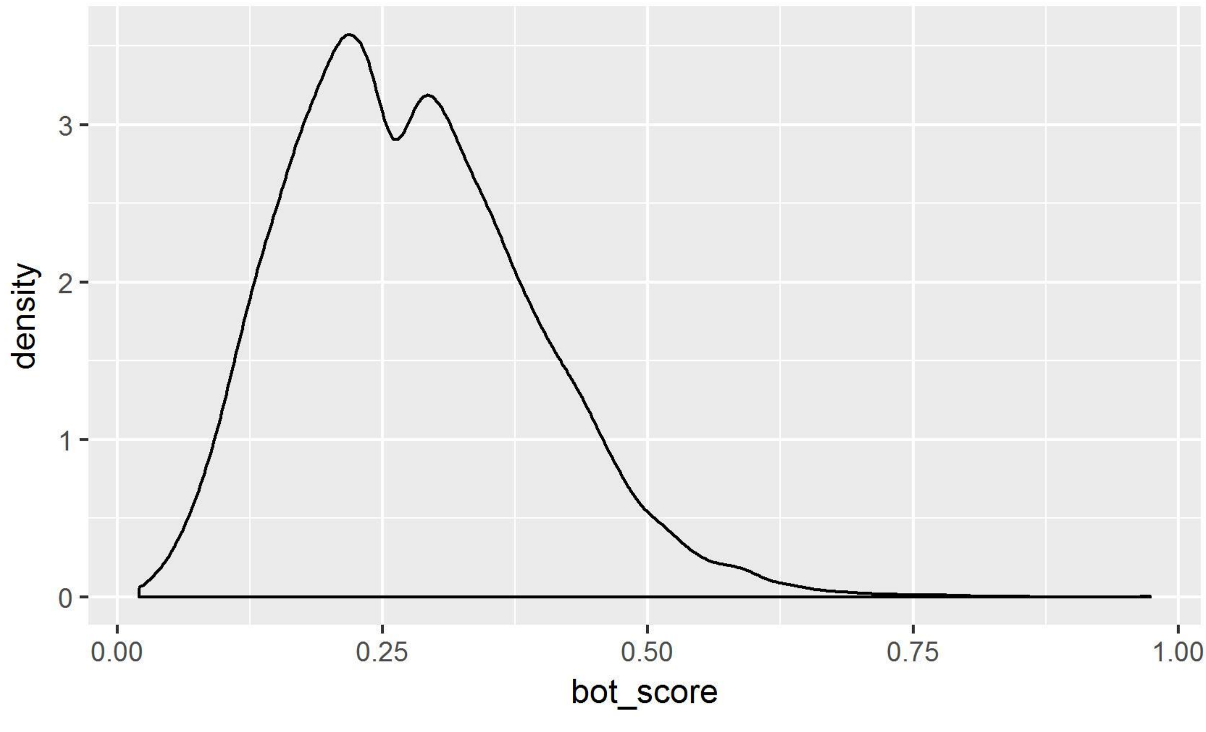

***Figure 2**: Distribution of the probability density of bot scores assigned to the 11,897 active users in the sample*

[8] As the *Botometer* FAQ states: "A middle-of-the-road score like this [40-60%] is a signal that our classifier is uncertain about the classification."
https://*Botometer*.iuni.iu.edu/#!/faq

With the threshold set at 0.5, 389 of the 11,897 users were classified as bots. It is to be noted, that 878 users could not be classified by *Botometer* due to various erros in the process. Interestingly, the dominant causes for this classification as "unknown" were "User not found", suggesting a deleted account, and "User suspended", indicating that the user has been suspended for violating Twitter's user guidelines. Full statistics on the user classification can be found in Table 1. The high amount of suspended and potentially deleted users (with 762 combined almost twice as high as the number of bots detected in the sample) may indicate two things: 1) that a number of these accounts were bots suspended by Twitter for exhibiting obvious bot behaviour; 2) that bots were deleted by their creators after their task of partaking in the online discussion on the German federal election was done. Since *Botometer* can only check live, that is, the momentarily available data on any given user, these speculations cannot be verified as the bot detection was run only after the sample was collected.

Table 1: *User classification of active users in the sample (total 11,897 active users). Values rounded*

Classification	**Users**	**% of active users**
Human	10,630	89.4%
Bot	389	3.3%
Unknown, of which:	878	7.4%
"User not found"	495	4.2%
Suspended	267	2.2%
Other errors	116	1%

With bots identified, the analysis' next point of interest is the amount of tweets generated by bots respectively humans. Table 2 shows the total and percental amount of tweets as well as the average amount of tweets during the observation by user classification in the subsample of active users (containing a total of 1,533,311 tweets). As expected, bots have a considerable higher average output of tweets than human users, with the suspended users leading by far. This suggests that bots were indeed used to disrupt the online discussion by spreading a high amount of (mis-)information and potentially trying to emulate grass-roots movements by sheer volume. It also indicates that "suspended" users may have been suspended due to their spam behaviour, which may also point to less sophisticated bots being suspended.

Table 2: *Tweets of active users and average tweets by user classification (total 1,533,311 tweets). Values rounded*

Classification	**Tweets**	**% of tweets**	**Avg. Tweets/User**
Human	1,330,793	86.8%	125.2
Bot	62,459	4.1%	160.6
Unknown, of which:	140,059	9.1%	159.5
"User not found"	65,392	4.3%	132.1
Suspended	55,160	3.6%	206.6
Other errors	19,507	1.3%	168.2

To investigate the spatio-temporal dynamics of tweet production in bots and humans, the timeline of tweet volume was visualised for the observation period. As figure 3 shows, human and bot behaviour differs significantly, not only in total volume. While human tweeting dynamics follow pronounced circadian rhythms, the bots' output stays relatively constant.

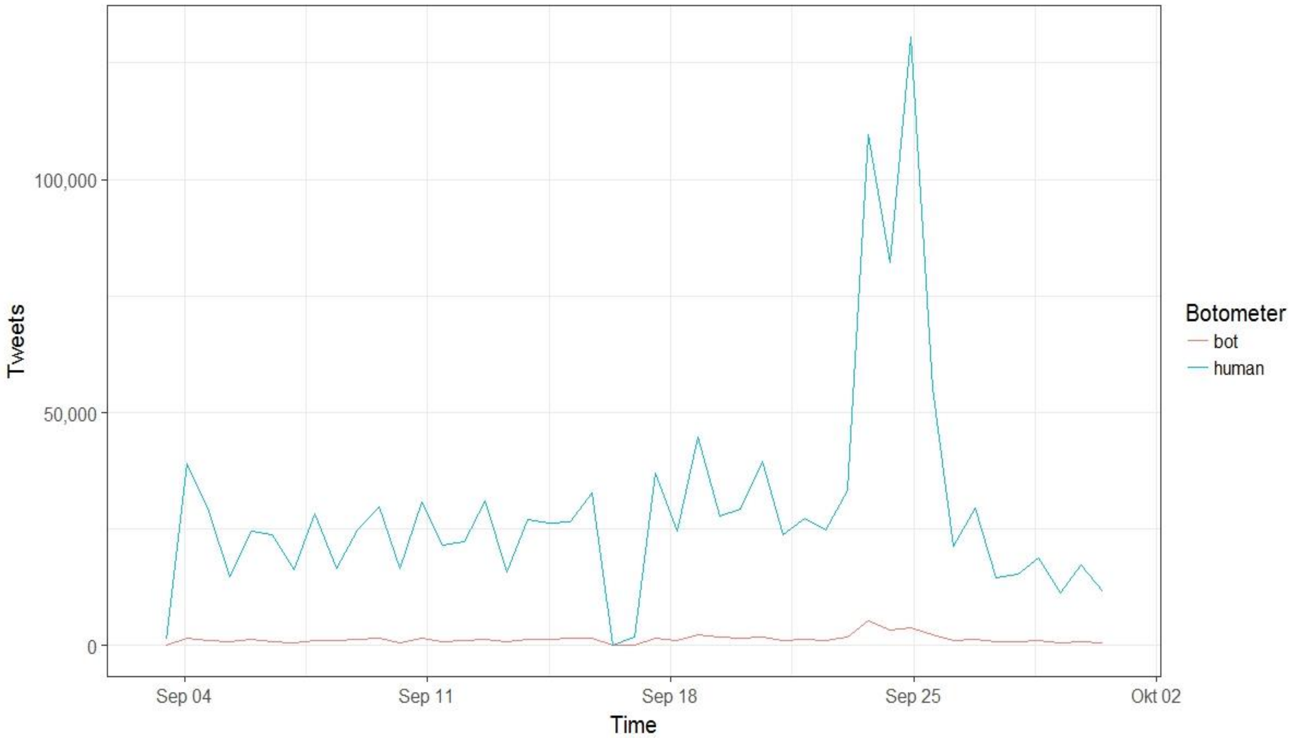

***Figure 3**: Timeline of the volume tweets generated by humans and bots during the observation period. Election day: 24th of September.*
Please note: the pronounced drop in tweet volume on September 16th is caused by a loss of data for that day

While both bots and humans activity react to the "Kanzlerduell" (a TV debate between top chancellor candidates Merkel and Schulz) at the night of the 3rd September, bot activity drops sharply after election day on the 24th of September. The pronounced spike of human activity on the 25th of September can easily be explained by the election results being confirmed – a topic the bots seem way less interested in. On the other hand side, the bots exhibit the highest activity right before and on election day, likely in an effort to mobilize voters. These findings confirm observations of bot behaviour during the 2016 U.S. Presidential election, that bots react less to external influences than humans and instead focus strongly on their own agenda (Bessi & Ferrara, 2016). Please note that the pronounced drop in all activity on September 16th is caused by the aforementioned streaming error, causing almost all tweets for that day being lost.

Finally, tweet / retweet ratios for both humans and bots were examined, as table 3 shows. Interestingly, human users have a higher retweet ratio than bots. This means that, counter-intuitively, bots create more original tweets than humans. It needs to be noted, however, that tweeting the exact same tweet multiple times does *not* count as a retweet, since it is not classified as such by Twitter. Exploratory investigation of the sample suggests that bots tend to tweet the exact same tweet multiple times (even in short intervals), each time generating a "new" tweet with a unique tweet-id. Further content analysis of the sample would be required to investigate and take this issue into account when considering tweet / retweet ratios for bots and humans.

Table 3: *Tweets, retweets and re-tweet/tweet ratio of active users by user classification (total 1,533,311 tweets). Values rounded*

Classification	Tweets	Retweets	Ratio
Human	1,330,793	949,087	0.71
Bot	62,459	36,764	0.59

3. Limitations of the approach

Even though the amount of actually identified bots during the observation period is relatively low, it is a first step to monitor the bot population in online discussions. In this respect, it is not unlikely that the bot population is indeed lower in the German federal election than in the U.S. Presidential election and a reasonable amount of twitter bots was identified in this paper. Considering Germany's technological lag in the use, distribution and relevance of social media relative to the United States however, it is likely that the bot population in German social media discussions will grow in the future. At the same time, it must be noted that a number of limiting factors in the tools employed may have negative effect on the distribution of bot scores in the sample, methodically underestimating the prevalence of bots in the German federal election.

Possibly the biggest cause of inaccuracy in estimating bot scores with *Botometer* lies in the language specific features of the application. With both *Content* and *Sentiment* features not being usable for the German sample, two of six categories were disregarded, naturally reducing accuracy in the prediction. Furthermore, it seems that the universal score almost always predicts a *lower* bot score than it would if language specific features were included. That means that there is profound possibility that *Botometer* would systematically underestimate bot scores in non-English contexts. This circumstance might also be an explanation for the relatively low amount of bots identified in the sample. It might also explain why a considerable amount of users (around 14%) is classified in the range between 0.4 and 0.6 and can thus be considered uncertain in their classification. If we were to assume that the universal score underestimates a given user's bot score by just 0.1, the rise on the total amount of bots in this sample would be considerable. In order to verify this thesis, however, *Botometer*'s language specific features would need to be readjusted for other languages. This task would not only require a large amount of ressources, but also remains in the hands of its creators. A different approach would be to collect large samples of universal and language-specific bot scores for english-speaking users, estimate the approximate difference between the scores and adjust the treshold for bot classification in other languages accordingly. Both approaches however are out of scope for this paper.

The second constraint that must be noted concerns the relatively large number of suspended and "not found" (that is, possibly deleted) users. Since it is unlikely that the accounts were checked for their bot scores before, it is not possible to classify them with *Botometer* anymore. That means that the assumption, that a certain number of these accounts may have been deleted or suspended *exactly because* they were bots cannot be verified. If that was the cause, however, it would also add to the underestimation of the prevalence of bots in the sample. In order to test the assumption, a specialised set of tools would be required to retrieve the likelihood of these accounts being bots from *within* the sample. These methods, however, would likely be less accurate since they necessarily lack access to most metadata and network variables utilised by *Botometer*, since only tweets concerning the election are found in the sample.

Finally, it needs to be noted that the scope of the study was limited in its resources, limiting its observation period to a relatively small amount of time. Further, more sustained studies need to be conducted in order to accurately gauge the prevalence of social bots in the German twittersphere. This paper can only be a step in this direction.

Conclusion

With Twitter becoming increasingly approved as a medium for both information diffusion and discussion by users and institutions alike, its potential for both democratic discourse and manipulation is rising. The capability of bots, automated programs posing as humans in social media, to manipulate online discourse, public opinion and political legitimacy bears great risk for the democratic potential of social media. In light of the recent rise of populists both in the U.S. and Europe, these risks are higher than ever and call for the application of reliable tools to identify bots. This paper was a first approach to apply one of these tools, *Botometer*, to estimate the prevalence of bots in the German federal election.

In identifying bots in a sample of more than 2 million tweets related to the election, it is, for now, likely that Twitter bots in the German twittersphere are, for the time being, a fringe phenomenon. Considering the natural lag in social implementation for new media between the U.S. and Germany, however, it is highly unlikely this will not change dramatically in the near future. Furthermore, it is possible that the tools developed for the English twittersphere underestimate the prevalence of bots in non-English contexts. In light of both these assertions, it is important to sharpen the tools of bot detection for not only the German, but also the European context. With populism rising anew, it is more important than ever to identify attempts of deliberate manipulation, spreading of misinformation and delegitimization in the democratic discourse.

By mapping their behaviour, it could also be shown that bots differ in their activity patterns. Further research will need to be conducted to fully confirm the findings of previous research on Twitter bots for the German twittersphere. Suspicions, that populist parties such as the AfD may have a higher following of bots spreading their agenda and deliberate misinformation will have to be scrutinised. Luckily, the sample of tweets acquired for this research holds plenty of potential for further research. This research on bot prevalence can only be a first step in fully understanding the inner workings of the twittersphere, German or not.

Bibliography

Bessi, A., & Ferrara, E. (2016, November 7). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday, Vol. 21, No. 11*.

Chadwick, A. (2013). Symphonic Consonance in Campaign Communication: Reinterpreting Obama for America. In A. Chadwick, *The Hybrid Media System: Politics and Power.* Oxford Scholarship Online.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. *Developers Day Workshop at World Wide Web Conference.* Montreal.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM vol.59 no.7*, pp. 96-104.

Howard, P., & Hussain, M. (2011, July). The Upheavals in Egypt and Tunisia: The Role of Digital Media. *Journal of Democracy Vol. 22*, pp. 35-48.

Lilleker, D. G., Trenscher, J., & Štětka, V. (2015). Towards hypermedia campaigning? Perceptions of new media's importance for campaigning by party strategists in comparative perspective. *Information, Communication & Society, 18:7*, pp. 747-765.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pp. 400-408.

Murthy, D. (2012, December). Towards a Sociological Understanding of Social Media: Theorizing Twitter. *Sociology vol. 46 no. 6*, pp. 1059-1073.

OSoME. (n.d.). *Botometer FAQ*. Retrieved January 5, 2018, from https://*Botometer*.iuni.iu.edu/#!/faq

Owen, D. (2014). New Media and Political Campaigns. In K. a. Kenski, *The Oxford Handbook of Political Communication.* Oxford University Press.

Scholz & Friends, Ubermetrics, & Datenlabor Berlin. (2017). *Wer hat den Social Media Wahlkampf gewonnen?* Retrieved Dezember 10, 2017, from https://s-f.com/wer-hat-den-social-media-wahlkampf-gewonnen/

Appendix: hashtags used for generating the sample

Related to:	Hashtags:
German federal election (Bundestagswahl)	#BTW #BTW17 #Bundestagswahl #Bundestagswahl2017
SPD (Social Democratic Party)	#SPD #ZeitFürMartin #Schulz
CDU (Christian Democratic Union)	#CDU #Merkel
CSU (Christion Social Union)	#CSU #KlarfürunserLand
B90/Grüne (Green Party)	#Gruene #Gruenen #DieGruenen #DarumGruen
AfD (Alternative for Germany)	#AfD #TrauDichDeutschland #Weidel
Die Linke (Left PA	#LINKE #dieLinke
Piraten (Pirate Party)	#Piraten #FreuDichAufsNeuland
FDP (Free Democratic Party)	#FDP #DenkenWirNeu #Lindner
NPD (National Democratic Party)	#NPD
Die PARTEI (The PARTY)	#DiePARTEI