Extracting Account Attributes for Analyzing Influence on Twitter

Johan Fernquist, Ola Svenonius, Lisa Kaati, Fredrik Johansson
Swedish Defence Research Agency (FOI)
Kista, Sweden
firstname.lastname@foi.se

***Abstract*—The last years has witnessed a surge of auto-generated content on social media. While many uses are legitimate, bots have also been deployed in influence operations to manipulate election results, affect public opinion in a desired direction, or to divert attention from a specific event or phenomenon. Today, many approaches exist to automatically identify bot-like behaviour in order to curb illegitimate influence operations. While progress has been made, existing models are exceedingly complex and nontransparent, rendering validation and model testing difficult.**

We present a transparent and parsimonious method to study influence operations on Twitter. We define nine different attributes that can be used to describe and reason about different characteristics of a Twitter account. The attributes can be used to group accounts that have similar characteristics and the result can be used to identify accounts that are likely to be used to influence public opinion. The method has been tested on a Twitter data set consisting of 66,000 accounts. Clustering the accounts based on the proposed features show promising results for separating between different groups of reference accounts.

***Index Terms*—bots, Twitter, bot detection, social media, impact, influence**

I. Introduction

In the past years, bot and influence detection on social media has become a big area of research. Initially bots were designed to automate online processes that were too time consuming for a person to do, but has since then been used for a variety of different purposes. For example, bots have been widely used for advertising campaigns by sharing links in different social media trying to generate traffic to a website. Bots have also been a popular way to help with customer contacts on websites and is often the first step when a customer needs help. In that case, bots are used to relieve the staff in customer support.

In later years, bots have been used in influence operations where they have been used to convey a message. With the use of bots, a message or topic can appear to be more discussed, accepted, and mainstream than it actually is and can therefore in the long term have an impact on public opinion. When a post on social media has a large number of likes and shares, it can for many people appear as genuine and truthful information even though the content of the post is false.

Bots have also been used during political election campaigns, with the purpose to either support or malign certain parties or candidates. In the 2017 German federal election, researchers found evidence that that communities on Twitter contained social bots that had an obvious tendency for a certain party [20]. Similar results was found in the Swedish general elections in 2018 where several researchers found bot-like activity on Twitter in discussions about the election [11], [15].

The most notable case were bots have been used with the aim to influence public opinion before an election is when the Russian government-linked organization *Internet Research Agency* (IRA) used bots to influence the 2016 United States election. Twitter identified 3,814 accounts linked to IRA that posted 175,993 tweets on Twitter [3]. In February 2018, 1.4 million people were informed by Twitter that they may have engaged with the IRA-linked accounts during the election period. These accounts were by many referred to as *trolls* [6], [10], [23], a term that is often used to describe fake identities that are motivated politically and used to interact with ordinary users on social networks [13].

The utilization of social media for foreign policy goals is increasingly seen as a problem both for global tech corporations and defence analysts across the globe [16]. Given political liberties such as free speech in Western democracies, control over the Internet is neither possible nor desirable. Instead, a pressing need to identify influence operation in an early stage is getting stronger as new such events continue to unfold. Today, several services and tools for detecting bots are available online. Examples of such services are: Twitter Audit [2], Bot Sentinel [1], and Botometer (earlier called BotOrNot) [9]. All the mentioned services works similarly: the user can select a Twitter account to see whether the account is likely to be a bot or a genuine account.

A problem with these services is the lack of information about why an account is identified as a bot (or not). The services are built on machine learning technologies which in many cases does not provide information about the process for classifying an account as a bot or not. This is problematic because replication of the models is nearly impossible, making it difficult for other researchers to estimate the accuracy of the models. Another problem with these approaches is that there is no commonly accepted definition of what a *bot* is. In [13], Gorwa and Guilbeault provide a typology of bots and describe different types of bots such as web robots, chat bots, spambots, social bots, and sock puppets/trolls. Spambots are for example bots that post on online comment sections and spread advertisements or malware, while sock puppets are fake identities used to interact with ordinary users on social networks [13].

EISIC 2019, November 26-27, 2019, Oulu, Finland
978-1-7281-6735-0/19/$31.00 © 2019 IEEE

In the bot detecting services mentioned above, the result of a bot classification does not provide the user with information about what features or attributes that is making an account being classified as bot or not. In Botometer [9], the user can get an understanding of what type of features (content, sentiment, friend, network etc.) that were triggered to classify the account, but not the actual features which triggered the model which makes it unmotivated to claim that a certain account is an actual bot.

The lack of transparency is problematic for several reasons. One problem is that it does not allow us to estimate the number of false positives and/or negatives other than on artificial training data. A higher transparency would allow the research community to fine-tune the bot recognition and validate the results on new data sets. The limitations of existing models is highlighted in a blogpost by Michael Kreil, who criticises several studies for being flawed while writing about social bots in the US and Brexit elections [17]. Kreil states that the bot definitions are misguided and that we have to advance our understanding of bots to the point where we can distinguish between bots (as in software bots) and social bots. The aim of this paper is to take a first step in an attempt to accomplish this.

In this paper, a new approach to analyze influence operations on Twitter is presented. Instead of classifying an account as a bot or not we study a number of different attributes that Twitter accounts have, and show that we can group accounts with similar attributes. This makes it possible to reason around different types of accounts and group accounts with similar values on the attributes.

The attributes are related to influence, degree of automation and diversity in produced messages. Specific combinations of the attribute values will make accounts more or less interesting to study when trying to detect influence operations. The attributes are explainable and not complex feature vectors customized for machine learning algorithms. Using this approach it seems to be possible to develop a method where an account (instead of just being classified as bot or not) now can be classified as likely to be a spambot, a sockpuppet, an automatic feed or a genuine user. The results can be used to study and detect potential influence operations.

II. Influential Behaviour on Social Media

As discussed above, there are no commonly accepted classification framework for automatically generated user content. Before any behavioural classification can be done, it is therefore important to define online influence, what type of behaviour could be expected from influential actors, and what basic characteristics can be used to describe this behaviour.

'Influence' is a term that broadly refers to a change in attitude or behaviour that would otherwise not have taken place, i.e. synonymous to the *third face of power*, as discussed by Lukes [18]. Influence is thus a variant of power, but excludes the idea of force. It therefore must rely on suggestion, manipulation of information, persuasion, and compliance of the subject [7], [14]. 'Influence operations', in turn, are "activities conducted ... to influence the perceptions, behaviour and decisions of target groups" to the benefit of some group or actor [21, p. 14]. Most relevant to our interests here are influence operation orchestrated by foreign powers to influence e.g. elections.

Attempting to construct a model of online influence, including but not limited to bots, is a challenging task because the actors behind such activities do not wish to be recognized. Stealth is part of the manipulation taking place and therefore creating yet another typology of bots makes little sense. Instead we reverse the concept and look closer on the target groups of influence operations. The aim with an influence operation is, as stated above, to make a group of people (in this case Twitter users) behave differently than they otherwise would have. In order to understand how such influence works we briefly turn to the psychology of online influence. Previous research by Ahn [4], Cialdini [7], Cialdini and Goldstein [8], Moreno [19], and Winter [27] focused on how users on social media typically receive information, how they can be deceived or persuaded, and what kind of manipulation typically would not work. The results from these works is that the classical principles of influence, as formulated in [7],[1] are only partly at work in online environments. Thereby the anonymity aspect of social media is crucial: an anonymous communicating agent will be received differently than a known one [4]. Secondly, a significant social dimension of social media relies on peer-to-peer validation, and such validation may also be a strong source of influence [14], [19]. Third, users are typically most influenced either from views that adhere to their own or expert views [27]. Especially in Winter's study, reasonable argument play an important role, as long as they validate in-group views (for politically uninformed users) and as long as they do not contradict one's previous knowledge (for well-informed users) [27]. Taking these points into consideration, a model that seeks to map influence operations would have to consider *anonymity*, *social validation*, *how messages spread* across existing networks and how users *interact*. Below we describe a set of attributes that seeks to accommodate these various aspects of online influence.

III. Attributes

To describe and reason about different characteristics of a Twitter account we have chosen to study a number of different attributes. The attributes that we study are:

- Anonymity
- Popularity
- Confirmation
- Spread
- Interaction with others
- Posting intensity
- Posting automation
- Network focus
- Topic variation

[1] Reciprocation, consistency, social proof, liking, authority, and scarcity

Next to anonymity and interaction with others, which are self-evident, popularity, confirmation, and spread refer to social validation activities. Network focus and topic variation represents the way that messages spread. In order to detect automated behaviour, we included two attributes with that aim: posting intensity and posting automation. Together they allow us – in theory – to isolate the behaviour that we would expect from influential accounts. Each attribute is described in detail below.

A. Anonymity

The attribute 'anonymity' measures how anonymous the owner of an account is. The measurement of this attribute is based on the degree in which a user or an organization behind an account can be positively identified. If an account is verified by Twitter (meaning that Twitter lets people know that an account of public interest is authentic [25]), the owner of the account is not considered to be anonymous at all. In this case the identity of the user behind the account has been verified by Twitter. If the account has a URL *and* a location in the profile we consider the account more anonymous than a verified account but still less anonymous than if the account only includes a URL *or* a location. If an account does not contain a URL *or* a location, or is unverified, the level of anonymity is as high as it can get.

The level of anonymity (A) is defined as:

$$A = \begin{cases} 0 & \text{if user is verified} \\ 1 & \text{if user has URL and location} \\ 2 & \text{if user has URL or location} \\ 3 & \text{else} \end{cases}$$

Other measurements for anonymity would have been possible, such as custom adaptions of the account page or the existence of an account photo. However, the existence of neither a profile picture, nor custom backgrounds are considered to affect the level of anonymity. This is because a large amount of non default profile pictures does not show an actual human being but rather cartoons, memes and such. For those accounts that want to appear as genuine, it is not hard to use a random (or synthetically generated) picture of a person and claim to be that person.

B. Popularity

The attribute 'popularity' (P) measures how popular an account is. Popularity is often measured by the number of followers [5], [24], and we have decided to use the same measure in our popularity attribute.

$$P = number\ of\ followers$$

C. Confirmation

'Confirmation' is a measurement of how other Twitter users accept and agree with an account. It highlights one of the core functions of social media, i.e. sentiment attribution to user-generated content by other users [26]. One way to display agreement on Twitter is to like a post; another is to post a positive comment. Twitter does not provide a list of comments related to specific posts, which makes it difficult to measure the amount of positive comments. Therefore, we use the average number of likes to measure confirmation. Some accounts are very popular and always receive a lot of likes from loyal followers. Other accounts are not so popular but in some cases manage to publish a tweet that goes viral and receives many likes. To take both cases into consideration the confirmation attribute C is computed as follows:

$$C = \text{\# likes for most liked tweet} + \frac{\text{\# total received likes}}{\text{\# total published tweets}}$$

D. Spread

The attribute 'spread' is a measurement of how often the tweets from an account are retweeted (republished) and how widely they are spread on Twitter. A tweet that has been originally published by account A and then retweeted by account B appears as a published post in the feed of account B, with information that the tweet is a retweet of user A. All further retweets (and other interactions such as likes and comments) for the retweet in account B's feed are still counted for on the original post of account A. As in the case with confirmation, some very popular accounts might always receive a lot of retweets, and others might have some tweets that are retweeted a lot. The spread (S) attribute is therefore analogous to confirmation, as described above, and measured as follows:

$$S = \text{\# retweets for most retweeted tweet} + \frac{\text{\# total received retweets}}{\text{\# own tweets authored}}$$

In the formula above, the user's retweets are excluded from the count of own tweets that have been authored by the user.

E. Interaction with others

'Interaction with others' is a measure of how much an account interacts with other accounts. This is potentially an important measure when analyzing influence operations since accounts that want to spread a certain message would most likely try to interact with as many other accounts as possible. The attribute is a sum of four ratios: between own tweets and *retweets* of other users, own tweets and *replies* to others, own tweets and *mentions* of other users, and the share of own tweets containing *hashtags*. Retweets are included in the attribute since retweeting is a way of republishing another account's content and therefore interacting with other users. Replying is a straight forward way to interact with others since it is done by a comment to someone else's tweet. Mentioning someone else's account name is, similarly to a reply, a direct form of interaction. In this case a notification will be sent that they have been mentioned in a tweet. Finally, using hashtags is a way of marking your tweet as a certain topic that makes it searchable for others. Interaction with others (I) is thus computed as follows:

$$I = \frac{\textit{\# retweets made}}{\textit{\# tweets analyzed}} + \frac{\textit{\# own tweets with account mentions}}{\textit{\# own tweets}} + \frac{\textit{\# own tweets which are replies}}{\textit{\# own tweets}} + \frac{\textit{\# own tweets with hashtags}}{\textit{\# own tweets}}$$

F. Posting intensity

'Posting intensity' (PI) is a measurement of how many tweets on an average that an account publishes daily:

$$PI = \frac{\textit{\# total published tweets}}{\textit{\# days since created}}$$

G. Posting automation

The attribute 'posting automation' (PA) reflects how automated the posting behaviour of an accounts is. Certain account types such as news channels are more active during certain minutes of an hour or even in certain seconds of a minute. That kind of activity pattern could indicate that an account is operated by a software since it is not reasonable to believe that a real person would tweet on the same minute every hour. Some accounts have a strict posting scheme and posts on the same minutes while other accounts posts on the same second. To get a measurement of posting automation, we use two vectors with 60 elements each - one for every minute and one for every second. Each vector contains the number of posts that are posted on a specific minute and second. The value of posting automation is the maximum variance of these two vectors.

$$PA = \max(\text{Var}\,(\textit{posting minutes}), \text{Var}\,(\textit{posting seconds}))$$

H. Network focus

The 'network focus' attribute is a measurement of an account's retweeting behaviour. For influence operations, it is possible that several accounts can be used to favor a single account by only retweeting messages from the same account over and over again. In that case, the retweeting accounts end up with a very low network focus score, using our measurement. Retweet bots (that can be payed to retweet a tweet) are usually retweeting several different accounts and would instead end up with a higher score. We calculate the network focus attribute (NF) as:

$$NF = \begin{cases} 0 & \text{if } \textit{\#retweets made} \leq x \\ \dfrac{\textit{\# unique retweeted users}}{\textit{\# retweets made}} & \text{else} \end{cases}$$

x is a constant set so that a small number of retweets made does not make the account appear to have a high network focus due to small number of retweets made.

I. Topic variation

The attribute 'topic variation' is a measurement of how much an account focuses on a specific topic, or whether it is moving between different subject areas. This measurement is likely higher for accounts such as retweet bots that are payed to retweet different users every time, since they will not retweet messages that focus on a single topic. To measure topic variation we use the number of different hashtags used by a specific account, while controlling for how often the account uses hashtags in general. Using several different hashtags indicates a higher variation in topics. Topic variation (TV) is calculated as:

$$TV = \frac{\textit{\# unique hashtags used}}{\textit{\# hashtags ever used}} \cdot \frac{\textit{\# tweets with hashtags}}{\textit{\# tweets analyzed}}$$

The attribute is designed to produce a lower value for accounts with a low ratio of tweets containing hashtags. This prevents accounts with only a few tweets containing hashtags to get a high topic variation value.

In sum, these nine account attributes were constructed to allow us to transparently distinguish between different types of accounts. Given the aim to identify influence operations in real time, our interest naturally focus on rather anonymous accounts with a high posting frequency. Given our knowledge of earlier influence operations on social media, it is likely that accounts with high scores on posting automation in combination with a strong network focus will be of interest to us. However, the advantage of our approach is that it does not necessarily discriminate between human (trolls, sockpuppets) and automated accounts (spambots, social bots). The interesting factor is the accounts' behavior, and when we apply the attributes to a Twitter data set, the hypothesis is that certain types of accounts will stand out in the analysis.

IV. Implementing the Attributes on Twitter Data

To get an understanding on how the different attributes can be used to study groups of accounts on Twitter we have tested the attributes using data from Twitter. A random sample of all tweets for a few days in February 2019 were downloaded using the free Twitter Streaming API. From the sample, a set of active accounts were selected and downloaded. A sample of the the most popular Twitter users from [12] was also included. In total, a set of 90,000 Twitter accounts were downloaded and used to test the attributes described above.

When the attributes are calculated we only consider the last 250 tweets of each account. If an account has published less than 250 tweets, the calculations are based on all of the account's tweets.

A. Anonymity on Twitter

Anonymity can be expressed with values from 0 to 3. In Figure 1, we can see how the values of anonymity is distributed in our sample accounts. It is clear that a majority of the accounts are not verified, hence most of the accounts have

an attribute value ranging from 1 to 3.[2] The majority of the accounts are scoring 2 in anonymity, meaning that in our sample it is most common to have a URL or a location in the profile.

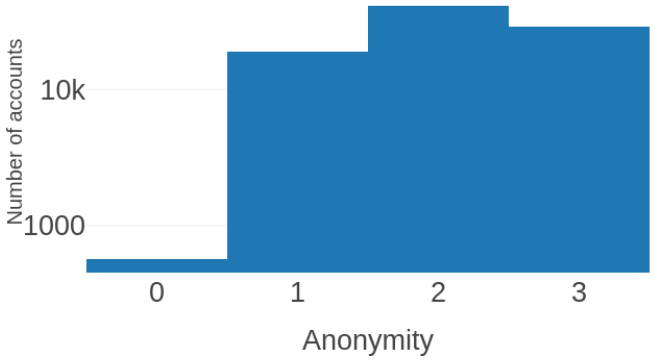

Fig. 1. Histogram of the anonymity distribution for our sample accounts

B. Popularity on Twitter

The attribute 'popularity' can range from 0 to the maximal amount of followers that an account have on Twitter. The popularity attribute for our sample is shown in Figure 2. Around 40 accounts that had more than 3 million followers were left out to get a better view of the popularity distribution. It is clear that the majority of the accounts has a limited amount of followers and a low degree of popularity. Relatively few accounts in our sample have more than one million followers.

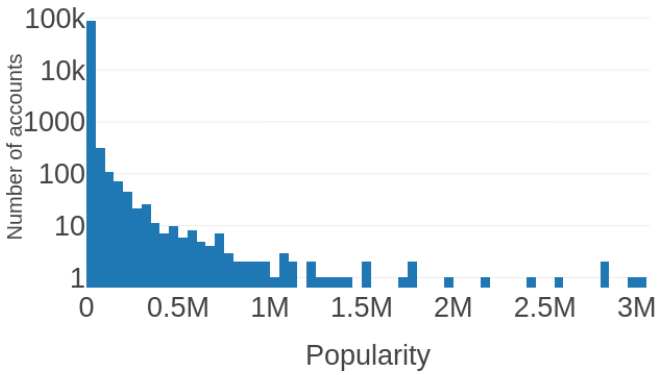

Fig. 2. Histogram of the popularity attribute distribution for our sample accounts

C. Confirmation on Twitter

'Confirmation' measures how much other Twitter users agree with an account. In our calculations we use the number of likes for most liked tweet and mean received likes for the 250 latest tweets. The result is shown in Figure 3. As can be seen in the figure, the shape of the distribution is the same as for popularity. Almost 50 accounts with a confirmation value higher than 50,000 is left out from the figure.

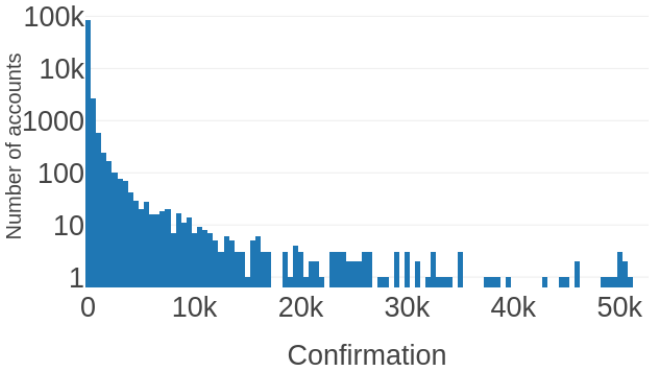

Fig. 3. Histogram of the confirmation attribute distribution for the sample accounts

D. Spread on Twitter

The attribute 'spread' measures the amount of retweets. We use the 250 latest tweets to calculate the number of retweets for most retweeted tweet and the mean amount of retweets. The distribution is shown in Figure 4. The 50 accounts with highest spread are not included in the figure.

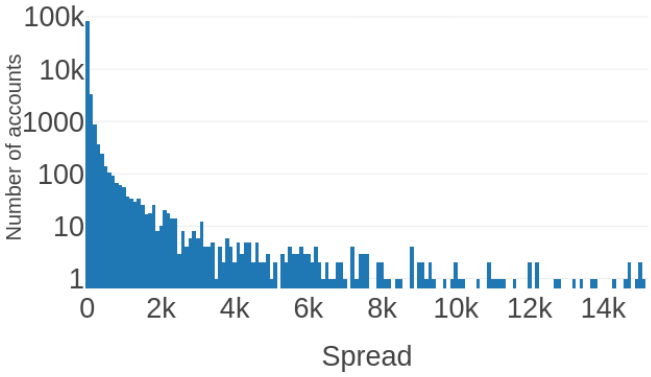

Fig. 4. Histogram of the spread attribute distribution for our sample accounts

E. Interaction with others on Twitter

The interaction attribute measures of how much an accounts is interacting with other accounts. In Figure 5, the distribution for the interaction attribute in our sample is shown. The most common values for interaction are 0 or 1. An interaction value of 0 means that the account never retweets nor replies, uses hashtags or mentions. An interaction value of 1 means that that the account is either always retweeting or always using hashtags, mentions or replying, or a combination of the latter three. An interaction value larger than 1 means that the account

[2]Twitter has currently halted its verification process. See: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts (accessed April 15, 2019)

is doing a combination of at least two of retweeting, using hashtags, mentioning or replying.

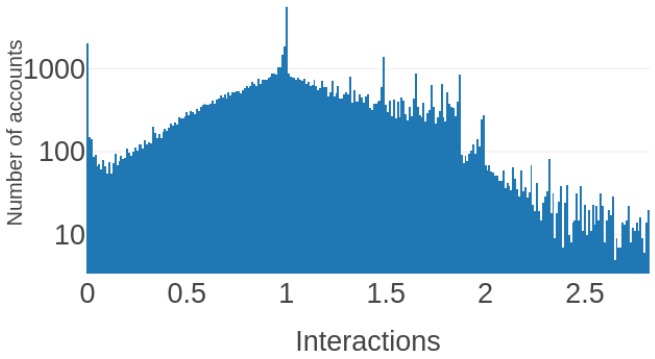

Fig. 5. Histogram of the interaction attribute distribution for our sample accounts

F. Posting intensity on Twitter

The amount of tweets that are published by an account per day provides the posting intensity. Figure 6 shows how many tweets per day the accounts in our sample have posted. The 30 accounts that posted most tweets per day are left out from the figure. Note that here we are using the total number of published tweets, not only the 250 latest tweets.

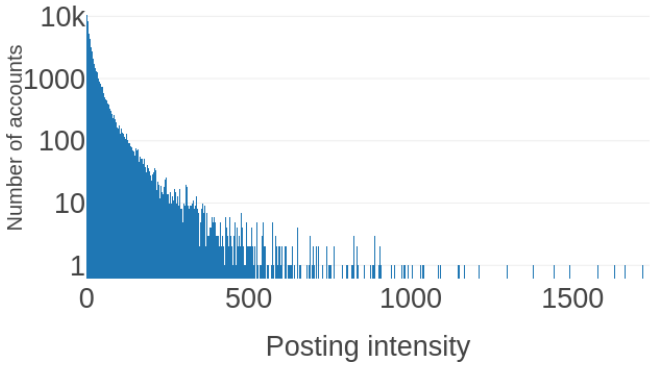

Fig. 6. Histogram of the posting intensity attribute distribution for our sample accounts

G. Posting automation

'Posting automation' reflects how automated the posting behaviour of an accounts is. In Figure 7, the value of posting automation is shown. The most common value for posting automation is 0 which means that there are no signals of automatic posting. In the figure, two peaks at 512 and 1024 can be noticed. These peaks show accounts that post at exactly the same minute or second every time, this indicates that the accounts might be subject to software automation.

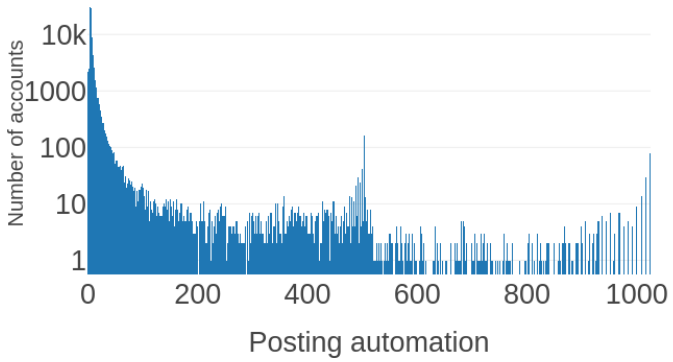

Fig. 7. Histogram of the posting automation attribute distribution for our sample accounts

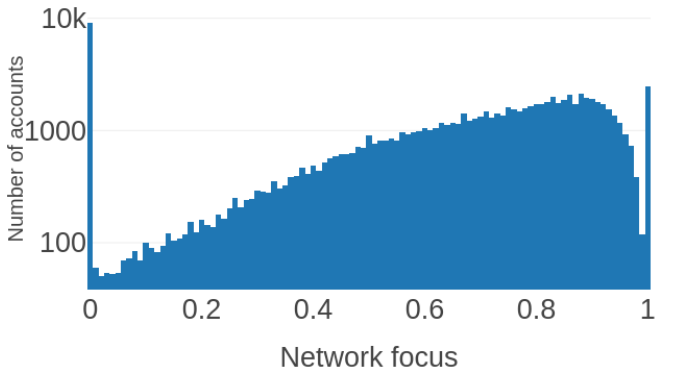

Fig. 8. Histogram of the network focus attribute distribution for our sample accounts

H. Network focus on Twitter

The 'network focus' attribute measures an accounts interaction with other accounts. Figure 8 shows the distribution of the network focus attribute in our sample. The most common value for our population is 0 which means that the account has not been retweeting more than 5 (in our case, threshold x is set to 5) times for the latest 250 tweets. A value of 1 means that the account has retweeted at least 5 times and the retweets are from different accounts every time.

Since some accounts that have done just a few retweets and retweeted several different accounts will end up with a high network focus score even though they might not have been used as retweet bots, we decided to set a minimum of 5 retweets made to calculate the network focus or otherwise the value is just set to 0. Accounts with a value near 0 are those accounts which have retweeted the same account with a high frequency.

I. Topic variation on Twitter

'Topic variation' measures how much an account is discussing a specific topic. Figure 9 shows the distribution for the topic variation in our sample. The most common value is

0 meaning that the accounts have not been using hashtags at all.

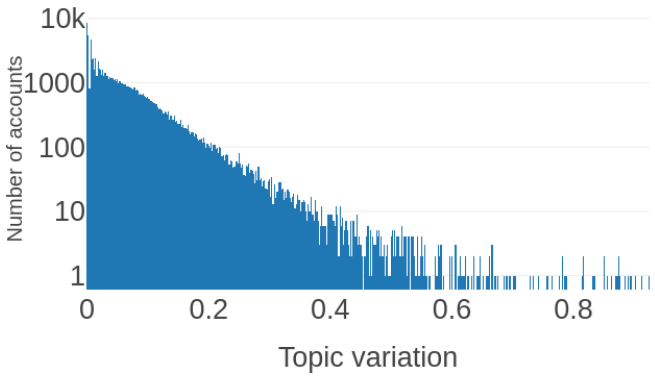

Fig. 9. Histogram of the topic variation attribute distribution for our sample accounts

V. Visualizing accounts with similar attributes

Once the introduced attributes have been calculated for a dataset of interest, the idea is that the analyst can use these attributes to explore the dataset and to identify accounts with certain characteristics of interest. To exemplify one way in which such an analysis can be undertaken, we have visualized the downloaded Twitter accounts as well as a number of reference accounts.

A. Reference accounts

To get an understanding of the values of the attributes in different groups of accounts we have included a set of reference accounts. The reference accounts are accounts that have certain known characteristics, a description of the different accounts types are provided below.

- **Automatic feed**: Services which are publishing their posts automatically. Examples of this kind of services might be news channels automatically publishing news stories when they are published on a website. Automatic feeds can also for example be authorities that either continuously or once in an hour publish posts about the latest events related to the authority.
- **Comedians**: A group of known Swedish comedians.
- **Politicians**: A group of known Swedish politicians.
- **Pornbots**: Accounts which have a highly sexual appearance in their profile trying to get visitors to a specific website often mentioned in the profile.
- **Opinion-formers**: Swedish people with a high participation in different kind of political discussions on Twitter.
- **Journalists**: A group of known Swedish journalists.
- **Retweet bots**: Accounts connected to paying services which have been used to retweet posts of different accounts.

B. Visualizing

By projecting Twitter accounts based on the nine different attributes onto a 2D-space, it is possible to find groups of accounts that have similar characteristics. We have used *T-distributed Stochastic Neighbor Embedding* (t-SNE) from the Python package Scikit learn [22] to visualize how the different reference accounts are positioned in relation to each other if we project the accounts onto a 2D-space.

In Figure 10 we can see how the different account types are positioned in the 2D-projection. Some of the accounts are more tightly positioned than others. The retweet bots and especially the automatic feed accounts have positioned themselves relatively isolated compared to the other reference accounts. The accounts with the automatic feeds are quite few but clearly separated from the other. For the comedians, the politicians, the pornbots and the opinion-formers, the situation is different. These accounts seem to behave in a similar way, which is somewhat expected. Politicians and comedians seek to maximize visibility and impact of their content and to raise awareness for specific issues. It is not uncommon for comedians to use political material in their work. The grey accounts in the figure are unclassified accounts, belonging to the 90,000 Twitter accounts that were used to test the different attributes. We believe that unidentified accounts that end up together with reference accounts of interest may be good candidates for further analysis.

VI. Conclusions and Future Work

This paper presents a set of attributes that can be used to classify different types of accounts on Twitter. The aim with the attributes is to be able to differentiate not only between different types of bot-like behaviour, but between other types of behaviour as well. Ultimately we are interested in identifying ongoing, illegitimate influence operations (as opposed to e.g. commercials and advertising). A total of nine different attributes describes different characteristics of account behaviour. These attributes are Anonymity, Popularity, Confirmation, Spread, Interaction with others, Posting intensity, Posting automation, Network focus, and Topic variation. Hopefully, these nine attributes are sufficient to describe the relevant behavioural aspects needed to distinguish between the types of accounts that are most relevant for our aims.

To get an understanding of the presence of the different attributes we have applied the attributes on a Twitter data set. By using t-SNE to visualize the accounts we obtained some clues about how well the attributes work when confronted with real-world data. For example, we noticed that accounts classified as automatic feeds and retweet bots can be distinguished from other accounts. However, further development and refinement is necessary. The attributes presented in the paper are only the initial steps of our method for identifying different kinds of accounts. For future work, we will start investigating how the account types can be separated or clustered using cluster- and classification algorithms. We will also try different approaches to identify interesting accounts based on the proposed attributes. There is a need of more

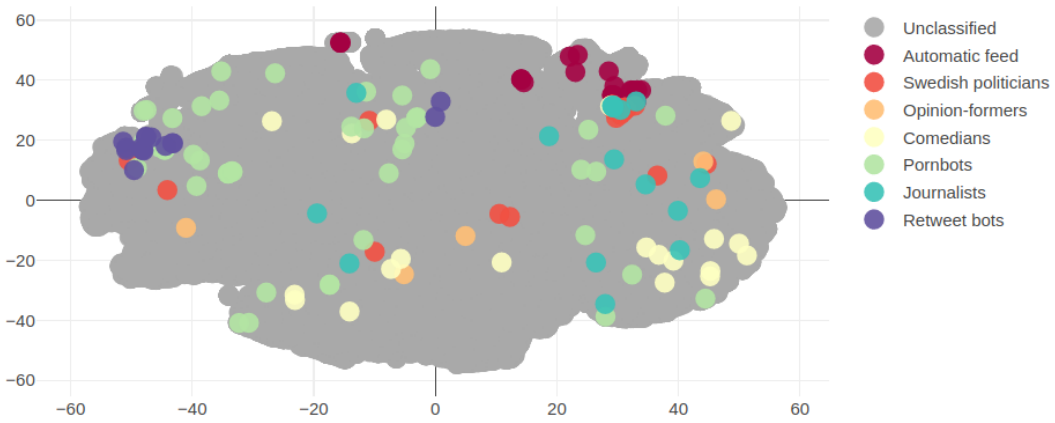

Fig. 10. 2D-projection of our sampled accounts

reference accounts from different kinds of bots, for example retweet bots and like bots from different providers. It might also be the case that additional attributes have to be considered in order to assure that different groups of accounts can be separated.

One aspect that needs to be addressed is the issue of attribute manipulation. Detection tools are susceptible to manipulation in order to disguise automated or semi-automated activity. The approach presented here is flexible in the sense that parameters and thresholds can be adjusted to accommodate changes in account behaviour. Additionally, cloaking large-scale operations would require additional resources, which would raise the cost of such operations substantially.

VII. Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation action program under grant agreement no. 832921.

References

[1] Bot sentinel. https://botsentinel.com/. Accessed: 2019-04-08.
[2] Twitteraudit. https://www.twitteraudit.com/. Accessed: 2019-04-08.
[3] Update on Twitters review of the 2016 US election. https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html. Accessed: 2019-04-04.
[4] T. K. Ahn, R. Huckfeldt, and J. B. Ryan. Communication, influence, and informational asymmetries among voters. 31(5):763–787.
[5] N. Angelovska. Top 10 Facebook fan pages and Instagram accounts in 2018–Cristiano Ronaldo takes the lead, January 2019. Accessed: 2019-04-08.
[6] M. Burgess. We finally know the full extent of Russia's Twitter trolling campaign, October 2018. Accessed: 2019-04-04.
[7] R. B. Cialdini. *Influence: science and practice*. Pearson Education [u.a.], 5. ed., internat. ed edition. OCLC: 605532873.
[8] R. B. Cialdini and N. J. Goldstein. Social influence: Compliance and conformity. 55(1):591–621.
[9] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. *CoRR*, abs/1602.00975, 2016.
[10] B. Elgin. Twitter revises data on Russian trolls and their 2017 activity., February 2019. Accessed: 2019-04-04.
[11] J. Fernquist, L. Kaati, and R. Schroeder. Political bots and the Swedish general election. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 124–129. IEEE, 2018.
[12] Friend or follow. Twitter: Most followers. Accessed: 2019-04-08.
[13] R. Gorwa and D. Guilbeault. Understanding bots for policy and research: Challenges, methods, and solutions. *CoRR*, abs/1801.06863, 2018.
[14] R. E. Guadagno, N. L. Muscanell, L. M. Rice, and N. Roberts. Social influence online: The impact of social validation and likability on compliance. *Psychology of Popular Media Culture*, 2(1):51–60, Jan. 2013.
[15] F. Hedman, F. Sivnert, B. Kollanyi, V. Narayanan, L.-M. Neudert, and P. N. Howard. News and political information consumption in sweden: Mapping the 2018 Swedish general election on Twitter. Data Memo 2018.3. Oxford, UK: Project on Computational Propaganda, November 2018.
[16] S. Hegelich and D. Janetzko. Are Social Bots on Twitter Political Actors? Empirical Evicende from a Ukranian Social Botnet. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, pages 579–582, Cologne, Germany, 2016. AAAI.
[17] M. Kreil. The social bot research of Oxford and Co. is flawed., December 2018. Accessed: 2019-04-08.
[18] S. Lukes. *Power: A Radical View*, volume 2. Palgrave Macmillan, New York, 2005.
[19] M. A. Moreno, R. Kota, S. Schoohs, and J. M. Whitehill. The facebook influence model: A concept mapping approach. 16(7):504–511.
[20] F. Morstatter, Y. Shao, A. Galstyan, and S. Karunasekera. From alt-right to alt-rechts: Twitter analysis of the 2017 German federal election. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 621–628, 2018.
[21] J. Pamment, H. Nothhaft, H. Agardh-Twetman, and A. Fjllhed. Countering Information Influence Activities: The State of the Art. MSB1261, Department of Strategic Communication, Lund University, Lund, July 2018.
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[23] A. Romano. Twitter released 9 million tweets from one Russian troll farm. heres what we learned., October 2018. Accessed: 2019-04-04.
[24] Statista. Instagram accounts with the most followers worldwide as of april 2019 (in millions), April 2019. Accessed: 2019-04-08.
[25] Twitter. Twitter: About verified accounts. Accessed: 2019-04-08.
[26] J. van Dijck and T. Poell. Understanding Social Media Logic. *Media and Communication*, 1(1):2–14, 2013.
[27] S. Winter, C. Bruckner, and N. C. Kramer. They came, they liked, they commented: Social influence on Facebook news channels. *Cyberpsychology, Behavior, and Social Networking*, 18(8):431–436, Aug. 2015.