A Two-Phase Framework for Detecting Manipulation Campaigns in Social Media

Dennis Assenmacher(✉), Lena Clever, Janina Susanne Pohl, Heike Trautmann, and Christian Grimme

Information Systems and Statistics, University of Münster, Münster, Germany
{dennis.assenmacher,lena.clever, janina.pohl, heike.trautmann,christian.grimme}@uni-muenster.de

Abstract. The identification of coordinated campaigns within Social Media is a complex task that is often hindered by missing labels and large amounts of data that have to be processed. We propose a new two-phase framework that uses unsupervised stream clustering for detecting suspicious trends over time in a first step. Afterwards, traditional offline analyses are applied to distinguish between normal trend evolution and malicious manipulation attempts. We demonstrate the applicability of our framework in the context of the final days of the Brexit in 2019/2020.

Keywords: Social campaign detection · Stream clustering · Unsupervised learning

1 Introduction

Social media has become an important infrastructure for modern information sharing and networking. In most developed countries, the majority of people are already connected via one or multiple platforms [6]. Even more important, decision makers like politicians or multiplicators like journalists are also an integral part of social media networks. These groups function as bridge between the social media ecosystem and the offline world outside social media. While politicians try to get in touch with the sentiment of public debates about their programs or decisions, journalists try to pick up stories and use the public sphere as additional outlet.

Quite logically, social media has become a central platform for campaigns. Politicians try to reach the public with their ideas, but in contrast to former media types, users can also reach politicians directly. Both can also try to initiate societal debates by placing topics. And when journalists pick up these topics because they seem of critical importance in social media, their reach goes even beyond the boundaries of the social media ecosystem.

As such it is of utmost importance not only for journalists but for the whole society to provide some transparency on campaigns in social media. This shall provide insights into the origins of and motivations behind an observed topic: is a

G. Meiselwitz (Ed.): HCII 2020, LNCS 12194, pp. 201–214, 2020.
https://doi.org/10.1007/978-3-030-49570-1_14

campaign organic or orchestrated (automatic as well as human-driven), i.e., who is participating in these campaigns? What means are employed when placing a topic?

These questions go beyond the challenge of classifying single accounts as social bots or humans. We have to consider interaction of actors and thus the complete (or a representative sample of the) data stream, which is produced on a social media platform. These analyses do no longer focus on singular accounts or a group of users but on the content produced over time. Clearly, the corpus of data that needs to be analyzed is far too large for human manual inspection. But also classical methods of data analysis are not capable to store all data and process it in real time. Real-time detection of possible campaigns, however, is necessary to not lag behind with analysis, when topics reach critical popularity. At the same time, we still need to verify whether campaigns are organic or artificial. This decision can usually not be made ad-hoc and often needs a deeper, sometimes even forensic analysis of campaign data.

In order to address both challenges at the same time, we propose a two-phase framework which supports both campaign and trend detection and a-posteriori in-depth analysis of respective data. Our idea integrates a stream-based unsupervised detection of critical topics and an independent, offline, and extendable analytics environment. This allows to instantly identify upcoming and important topics and subsequently analyze and verify their artificial character. Note that this approach should be considered as a human-in-the-loop support tool, where no automatic decision on a campaign's quality is made. In principle, it is designed to enable detection and transparent analysis of current topics in many contexts, either the discovery of new and interesting topics or the fight against manipulation via artificial campaigns.

The rest of this work is structured as follows: the next section will summarize related research in the context of this work and then Sect. 3 will detail the two-step framework's concept proposed in this paper. Section 4 shows the application of our framework in the context of the Brexit discussion two months before and at the final Brexit date at the end of January 2020. Finally, Sect. 5 summarizes and discusses the results of our work and provides some future perspectives.

2 Related Work

Social media has been discussed as environment for disinformation, manipulation, or deception for more than a decade [8] and since the Brexit decision in 2016 as well as the election of Donald Trump for president of the United States, social media is considered an important infrastructure for manipulating societies [3,22]. Much effort has been put into the (computer-aided) detection of automation in social media. *Social Bots* are considered very potent actors in the distribution of disinformation [10,11,14,18], and consequently, detection techniques for social bots have been (and still are) an important topic of research [9,10,13,20]. While research started with a focus on the classification of single accounts as bots- or human-driven, some recent publications emphasize the importance of detecting collaboration of multiple actors [9,12]. An exceptionally early proposal was

made by Lee [16] already in 2014 to discriminate campaigns into *organic* and *non-organic* ones. While the first arises from classic human interaction in social media the latter type of campaigns is promoted by artificial or automated mechanisms or purchased and supported by the social platform [16].

Campaign detection started with offline analysis of network data and topologies, the clustering of posted or shared content, and the investigation of topics' temporal development. All applied techniques and extracted features mainly aimed for supporting or enabling machine learning approaches. More recent detection approaches afterwards focused on the application of machine learning in campaign detection in order to identify characteristic patterns of organic and non-organic campaigns [10,20].

However, there are some major disadvantages of (supervised) machine learning approaches in this context:

1. Models have to be trained using labelled data. Especially for campaigns in social media, this kind of data is usually not sufficiently available. An insufficient data base, however, makes the approaches imprecise.
2. The learned patterns can only capture the characteristics found in available input and learning data. That is, the machine learning approaches may become outdated and inflexible regarding new kinds of orchestrated campaigns.

There is some recent work [7,9,23] which addresses the application of unsupervised detection methods like clustering and network analysis as solutions to some of the issues. These approaches do not need initial training and can detect unknown characteristics. However, as correctly pointed out in [23], these methods are computationally too complex to handle the observed amount of social media content in real-time.

In this work, we pick up a proposal we recently made, i.e. using stream-clustering approaches for topic detection [2] and apply it as a first step in a two-phase analysis process. We propose the augmentation of the detection of campaign candidates with a subsequent analysis phase. In this second phase, previous mentioned established group- or single account analysis can be applied to verify or reject whether a campaign is malicious or not and to possibly detect responsible actors in this campaign. As such, we consider this work as a step towards an integration of modern classification approaches into campaign detection for fast *and* precise transparency in social media communication.

3 The Two-Phase Framework for Detection and Analysis

In the following, we introduce our two-phase framework for automated campaign detection. The framework is depicted in Fig. 1. Within the first phase, the incoming text data stream (e.g. Twitter stream) is processed into $tfidf$ vectors and aggregated via the `textClust` algorithm [5]. The algorithm handles text stream data and clusters similar documents together into so-called micro-clusters, which

represent the recently discussed topics of the stream. Additionally, a micro-cluster filtering is applied. By this, topics, which behave suspiciously in terms of their development over time, are extracted. In the second phase, these topics can be further analyzed, via numerous metrics and visual representations of text (meta-) data.

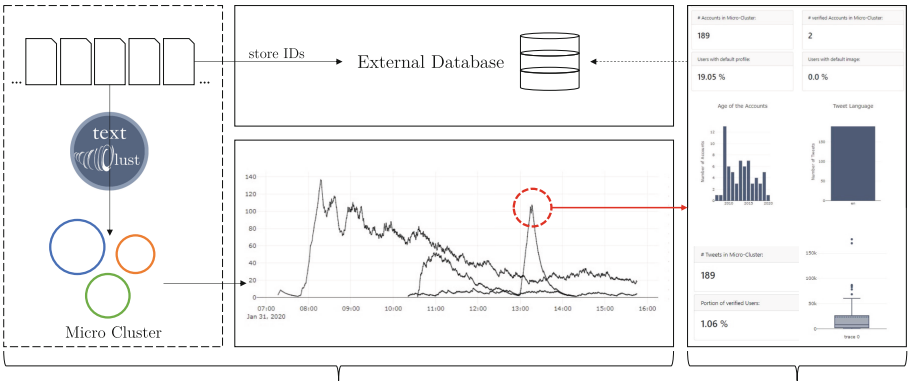

Fig. 1. 2-phase framework for analyzing suspicious cluster evolution

3.1 Phase 1: Text Stream Clustering

Stream clustering algorithms apply clustering on potentially unbounded data streams in an online fashion. The fact that the stream is potentially unbounded makes it impossible to store the complete data for calculations [4,19]. Due to this, observations can be processed only once. As the complete range of the data is not known in advance, the stream clustering algorithm needs to be able to adjust clusters online and in real-time.

The stream clustering algorithm can be divided into two phases: In the online phase, micro-clusters are derived directly from the incoming observations. A micro-cluster is an aggregation of observations, which are locally dense. While the concrete observations are discarded after the distance calculations, the clusters are stored as representation of the actual data distribution. In the offline-phase, the respective micro-clusters can be clustered on-demand via traditional clustering techniques. This phase is independent from the online phase and can be scheduled on demand at any point in time. As here only the limited number of micro-clusters, as a representation of the original data is used, the calculations can be done by using the data multiple times.

In contrast to incremental clustering algorithms, stream clustering algorithms must be able to deal with the explicit notion of time. The complete range of data is not known at the beginning and the distribution of the stream data may

change over time (which is known as concept drift). Therefore, micro-clusters need mechanisms to adapt to changes in the data stream. To simulate a temporal drift, micro-clusters are usually weighted. The weight ensures that clusters, which are not updated by new observations for a while, will be decayed slowly. If the weight falls below a threshold, the cluster is removed completely.

textClust: The idea of micro-clusters as representation of stream data was originally designed for numeric data. Nevertheless, the idea can be transformed to textual data as well [1].

For our experiments, we use the textClust algorithm [5]. Within the textClust algorithm, the produced micro-clusters mc are represented as 4-tuples:

$$mc = (w, t, TF, ID)$$

The relative importance of a micro-cluster is reflected by its tokens t (namely most describing words) and its weight w. The weight is increased by 1 each time a new observation is allocated to the cluster. To be able to detect concept-drifts and account for temporal changes, the weight is exponentially decayed at each time step by

$$f(w) = w * 2^{-\lambda(t_{now}-t)},$$

where λ denotes the fading factor, t_{now} the current time and t the time the specific micro-cluster was last updated. A cleanup procedure is applied every t_{gap} time steps where all micro-clusters below a predefined threshold are removed from the clustering result. The same applies for all tokens within a respective micro-cluster.

The term frequency of representative cluster words as n-grams is denoted in the tf vector. Distance calculations between two micro-clusters using the cosine similarity are based on the $tfidf$ vectors. Note, that the $tfidf$ representation extends the traditional term frequency by weighting down words that appear in many documents, as they are considered to be less important. For every new observation, first a new micro-cluster is created and second, the distance to all other micro-clusters is calculated. If the new micro-cluster is in small distance (below a certain threshold r) to one of the existing micro-clusters, it is merged with the respective cluster. Otherwise, the new micro-cluster remains and is added to the set of all micro-clusters.

The similarity of two $tfidf$ vectors is calculated via the adjusted cosine-similarity. Within this metric, the average weight of the micro-cluster is taken into account. Therefore, each token (within a certain cluster) is weighted relative to the average weight. Let A and B represent two $tfidf$ vectors from two different micro-clusters. The adjusted cosine similarity between them with their respective means μ_A and μ_B is then defined as follows:

$$cos(\alpha) = \frac{\sum_i (A_i - \mu_A)(B_i - \mu_B)}{\sqrt{\sum_i (A_i - \mu_A)^2} \cdot \sqrt{\sum_i (B_i - \mu_B)^2}}$$

The fourth element within the micro-cluster definition ID captures the post IDs, which relate to the corresponding texts within a cluster. The post ID vector

is irrelevant within the clustering phase, but gets important in the second phase of the framework, when suspicious stream data is analyzed in more detail.

Micro-cluster Monitoring to Detect Campaigns: A micro-cluster represents a topic discussed in the text stream. Each cluster consists of tokens, which describe the content, as well as a weight, which represents the importance (number of associated text instances) of the cluster.

Next to the overall topic monitoring of the incoming stream data, we are especially interested in suspicious stream behavior. The identification of rapidly arising and growing clusters might be of interest in the field of trend or campaign detection. Especially, since we are interested in non-organic campaigns, driven by bots or trolls, the temporal evolution of the campaign can be used as an indication for unusual behavior [21]. Since it is not feasible to manually inspect the complete number of micro-clusters over time, an automated filtering step has to be applied. In an earlier work, we already proposed a method that reduces the number of micro-clusters by focusing on micro-clusters that do exhibit a significant change of weights within the last cleanup procedure [2].

In addition to storing only the actual weight w of the cluster, the weight before the last update w_{last} is included for calculating the difference $\Delta_w = w - w_{last}$ within $tgap$ cluster updates. Based on this, the average weight change $\mu_w = \frac{\sum_i \Delta_{w_i}}{k}$ of all micro-clusters k, as well as the respective standard deviation $\sigma_w = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\Delta_{w_i} - \mu)^2}$, can be computed. The Chebyshev's inequality is used to determine clusters with unusual weight patterns [17]. The inequality states that:

$$P(|X - \mu| \geq t \cdot \sigma) \leq \frac{1}{t^2},$$

where X is a random variable with expected value μ, standard deviation σ and t any positive number. To ensure a feasible amount of clusters to (manually) analyze in a second step, we chose 6σ ($t = 6$) as threshold. The parameter setting can be adjusted depending on the context, as well as the underlying data. With this parameter setting about 3% of the micro-clusters are selected for further analysis, which is (in this case) a suitable amount for further investigations. The set of clusters of further interest I is thereby defined as:

$$I = \{mc | \, |\Delta_w - \mu_w| \geq 6 \cdot \sigma_w\}$$

3.2 Phase 2: Offline Analysis of Suspicious Clusters

Within the first phase of the framework, textual stream content is clustered and suspicious cluster evolution is filtered online and in real-time. In a second offline phase, suspicious clusters can be further examined. Here, all kinds of (computationally) expensive analyses can be applied. On the one hand, the micro-cluster content can be examined by the help of the stored cluster tokens. On the other hand, the user is able to gather meta-data via the ID vector of the suspicious

micro-cluster. As the ID vector captures all post IDs of the respective cluster, the Twitter REST API can be used to extract post meta data, e.g. the author ID or name. Further, meta data about the author can be gathered simultaneously. With the meta data the user is able to enrich the underlying data enormously. Especially for the detection of non-organic campaigns, further information about the human user is indispensable.

Authors of a micro-cluster can be analyzed regarding the age of their accounts, their post behavior, as well as their number of followers and followees. In the second phase of the micro-cluster analysis, visual representations can help to identify non-normal behavior. A dashboard can extremely help to visualize underlying structures in data and meta data of the post and accounts. Exploring e.g. the number of distinct accounts responsible for a micro-cluster, or checking the average age of the accounts, could help to identify social bots.

Furthermore, established bot detection methods can be applied. A well-known example for a bot detection method, which could be easily applied when the author ID is known, is the Botometer approach [20]. This tool gives an indication, whether an account is presumable steered by a human or a bot, by taking several meta data into account. Applying algorithms like the Botometer in the second phase of the framework can help to give an impression of the origin of the campaign and may help to detect non-organic campaigns.

In this work a first prototype of our dashboard is used for evaluation purposes[1] (see Fig. 2). We only rely on simple offline metrics which can be directly extracted from the tweets gathered during our experiments. Within the dashboard a variety of data and meta data can be visualized. For a first setting, we implemented figures and metrics representing the number of distinct accounts, the age of the accounts, such as the number of followers, and the percentage of verified accounts contributing within the specific topic. Further, we show how many an which posts are contained in this cluster at which point in time. This list is not exhaustive and can be complemented and customized. Up to now, we do not utilize additional supervised methods such as Botometer and leave this open for future research.

4 Case Study and Evaluation

In this work we exemplary demonstrate our framework in the context of the Brexit movement. Fur this purpose, we collected Twitter data by utilizing the platform's Streaming API. Twitter proclaims that the API provides 1% of the global traffic produced by the platform. Preliminary experiments showed that by filtering specific hashtags (in this case we only filter out tweets containing the term `Brexit`), we are able to obtain almost a complete conversation history [7]. More precisely, we collected data in late 2019, before the Brexit (between 20th and the 27th of November) and on the actual Brexit day on the first of February. We explicitly removed retweets from our analysis since we want to identify trends

[1] A python implementation of `textClust` and the corresponding dashboard can be downloaded here: https://textclust.com/.

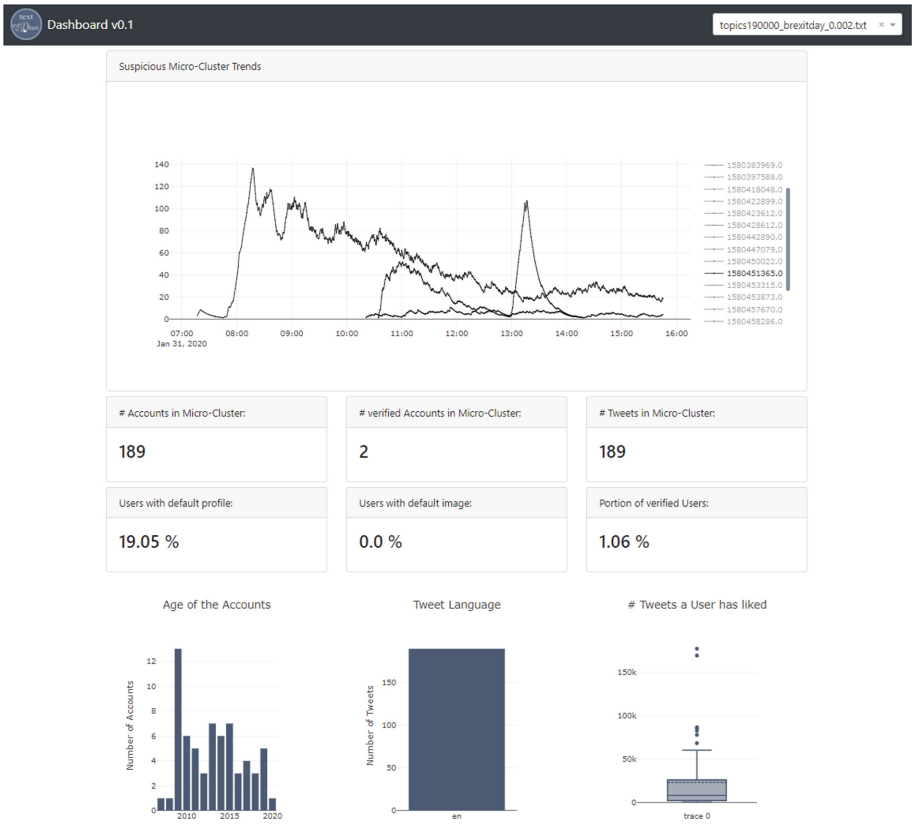

Fig. 2. Dashboard prototype to evaluate micro-cluster trends in the second phase

only based on original content excluding simply exaggerated trends based on retweet cascades [15]. In total we gathered roughly 1.3 million tweets, which were clustered by our `textClust` algorithm.

As specified in Sect. 3.1, the `textClust` algorithm requires some parameters that have to be set in advance. Especially λ, r and t_{gap} do highly influence the final clustering result. The λ parameter affects how fast micro-clusters fade out over time and is thus responsible for the overall lifetime of a topic. While a small value ensures that micro-clusters, which are not frequently updated, are not immediately discarded from the set of all micro-clusters, a larger value dismisses them rigorously. Also, t_{gap} influences which clusters are discarded since a larger value leaves more time for potential micro-cluster updates (and cleaning). The distance threshold r affects the granularity of micro-clusters. While a large value merges $tfidf$ vectors which are not necessarily very similar to each other (and therefore may represent different topics), a small value only merges sentences which are almost identical. The choice of suitable parameters does highly depend on the underlying data set. Therefore, we cannot rely on best-practice parameter

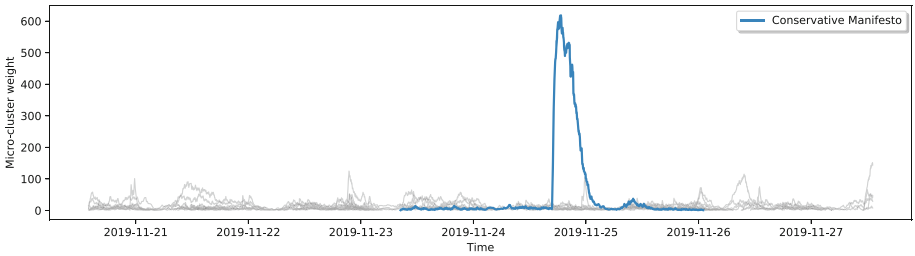

Fig. 3. Large micro-cluster that emerged from promoted Twitter campaign

settings. In context of our data set we systematically tested different parameter combinations. We found that λ also influences the number of identified trends. Since the Brexit day itself was very popular on Twitter with more than one million Tweets only on that day, we set a higher λ in this scenario. Therefore we decided to set λ to 0.001 (November) and 0.002 (Brexit day) respectively. We set t_{gap} to a fixed value of 100 and specified the distance threshold rather generously as 0.6. For all our experiments we used term-fading (fading according to elapsed time and not number of observations) to compensate variances in the stream throughput due to day/night cycles.

4.1 Identification of Promoted Tweets

A quantitative evaluation of our approach is almost infeasible due to missing ground-truth data. In this proof-of-concept analysis we show that our framework is actually able to detect trending content within the Twitter stream. When we inspected the filtered micro-clusters from the data gathered between the 20th and the 27th of November, we identified one micro-cluster which exhibits a significantly higher cluster weight than all other ones (see Fig. 3). Consequently, we inspected this micro-cluster more in-depth, utilizing our Dashboard prototype. In total 1900 Tweets are assigned to that specific micro-cluster, with 1850 unique users. This implies that this unusual peak cannot be explained by single spamming accounts. However, we found that the message which was tweeted by all these different accounts is always exactly the same, motivating people to vote for the Conservative party to get the Brexit done (see Fig. 4). It has to be again emphasized that we explicitly excluded retweets from our clustering. Therefore, the observed phenomenon is an unusual distribution pattern. Since we have access to the original Tweet IDs, we inspected the Tweet more in detail. Interestingly, each of the Tweets in question consists of an additional button by which people are able to easily share the same content on their profile (via a new original Tweet) with one click. Further investigation revealed that this so-called *call-to-action* button is one feature of Twitter intended for businesses to reach their customers. Surprisingly, this feature also seems to be used in political context and has significant impact on the global conversation stream of that topic.

Fig. 4. Call-to-action button for promoted tweets

Despite the high cluster weight, the trend lasted only a few hours and completely faded out afterwards.

4.2 Organic vs. Non-organic Trends

While our filtering approach during the first phase drastically reduces the number of interesting micro-clusters, it is not guaranteed that all of them do exhibit non-organic trends that should be classified as malicious. In context of the actual Brexit day (first of February 2020), we exemplary show how normal evolving trends can be distinguished from non-organic ones and how the second phase of our framework supports this differentiation. Within Fig. 5, we display three micro-clusters which all represent different topics that were discussed on Twitter that day. The blue trace represents a micro-cluster, containing tweets where users simply wished a happy Brexit day (similar to birthday wishes). As it can be inspected in the Figure, the trend (increase of the micro-cluster weight) started approximately at 8:00 AM with its peak 15 min later. This is not surprising, since it simply reflects that people started posting about the Brexit after they woke up (at nighttime the tweet throughput is significantly smaller than during the day). After the peak of the micro-cluster it slowly fades out until the end of the day, implying that the throughput of newly arriving tweets decreases over time.

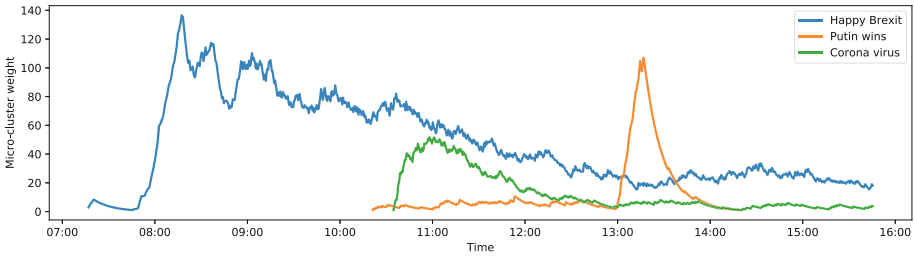

Fig. 5. Organic and non-organic micro-cluster trends at the Brexit day (Color figure online)

The green trace corresponds to a micro-cluster that summarized tweets about the first two cases of the corona virus in Britain which coincidentally happened at the same day. Again, the micro-cluster was created and immediately increases in it's weight. Afterwards, similar to the Happy Brexit micro-cluster, the weight is slowly faded out during the day. The last micro-cluster established at 13PM and captures tweets about Putin which subliminally imply his involvement in the Brexit and that he finally wins. In contrast to the other two clusters, we observe a sharp weight edge with rapid fading after peaking.

While the first cluster is an appropriate example for an organic trend that naturally arose due to the topic relevance, the last two both are not easy to interpret, since they contain controversial content that may originate from targeted opinion manipulation. Again, we utilized our Dashboard prototype to inspect those micro-clusters more in-depth. The corona virus cluster in total consisted of about 300 tweets. All tweets were posted by different authors who mainly originate from the UK. Also, the actual content of the tweets differed from each other. Although the term corona virus was always included in the tweet, the wording was always different. However, most tweets embedded an external URL, which linked to a BBC article which was published one day before[2]. Using these insights, we conclude that the corona virus trend evolved also in an organic manner and was triggered by the newspaper article. Lastly, we inspect the cluster about Putin. Here, we observe completely different meta-data: First, all of the 320 tweets that were assigned to that cluster only originated from 60 accounts. Further inspection of the different users revealed that 124 tweets (almost 40% of the cluster tweets) were produced by one single account. The message which was posted by that account was always the same. The only difference was that each tweet mentioned different political individuals. Hence, we deduce that this micro-cluster resulted from a dedicated spamming attack by one single account. For crossvalidation, we used the Botometer service to check whether this specific account can be classified as a bot (automated program). Although the content score is slightly higher than average, Botometer classifies the account as human. However, as we already stated in preliminary work, the Botometer system can be

[2] The article can be accessed here: https://www.bbc.com/news/health-51325192.

fooled and it is furthermore not of the uttermost importance to identify whether an account is automated or not. The overall goal should be the identification of malicious coordinated campaigns, executed by humans or non-humans [12].

5 Discussion and Future Work

In this work we proposed a new two-phase framework that is capable of identifying artificially created and organic trends on social media stream data. By utilizing unsupervised stream clustering combined with an additional filtering approach, we can circumvent the problem of missing ground-truth data during the first online phase and simultaneously reduce the amount of unimportant data that has to be inspected manually. Within a second offline phase, we use meta-information that was persisted to secondary memory during clustering to get additional insights into the cluster contents. Within a Dashboard prototype the information is aggregated to valuable KPIs. Our experiments show that our framework is capable of identifying different types of trends. Ranging from simple spammers to coordination via multiple accounts, we revealed organic and non-organic trends that highly affected the overall discussion about the Brexit. We realize that the second offline step is necessary to get reliable insights regarding the type of trend and to verify or reject whether a campaign is malicious or not.

While we currently only employ simple aggregation metrics within the second phase of our framework, there is a lot of room for applying additional, more sophisticated analyses such as the identification of user networks. Upcoming research should also focus on optimal parameter configuration. Ideally, parameters should be automatically adjusted during the online phase. The insights from different cluster evolution can also be used to produce ground-truth data within a semi-supervised setting. Via the cluster filtering method, information of suspicious post development and account meta data is gathered. After validation, this data might serve as ground-truth in supervised campaign detection approaches.

Acknowledgements. The research leading to these results received funding by the Federal Ministry of Education and Research, Germany (Project: PropStop, FKZ 16KIS0495K), the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014–2020, Project: MODERAT!, No. CM-2-2-036a), and the Ministry of Culture and Science of the federal state of North Rhine-Westphalia (Project: DemoResil, FKZ 005-1709-0001, EFRE-0801431). All authors appreciate the support of the European Research Center for Information Systems (ERCIS).

References

1. Aggarwal, C.C.: Mining text and social streams: a review. SIGKDD Explor. Newsl. **15**(2), 9–19 (2014). https://doi.org/10.1145/2641190.2641194
2. Assenmacher, D., Adam, L., Trautmann, H., Grimme, C.: Semi-automatic campaign detection by means of text stream clustering. In: Proceedings of the Thirty-Three International Florida Artificial Intelligence Research Society Conference (FLAIRS 2020), Florida, USA. AAAI Press (2020). accepted
3. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion. First Monday 21(11) (2016). https://doi.org/10.5210/fm.v21i11.7090
4. Carnein, M., Assenmacher, D., Trautmann, H.: An empirical comparison of stream clustering algorithms. In: Proceedings of the ACM International Conference on Computing Frontiers (CF 2017), pp. 361–365. ACM (2017). https://doi.org/10.1145/3075564.3078887
5. Carnein, M., Assenmacher, D., Trautmann, H.: Stream clustering of chat messages with applications to twitch streams. In: de Cesare, S., Frank, U. (eds.) ER 2017. LNCS, vol. 10651, pp. 79–88. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70625-2_8
6. Chaffey, D.: Global social media research summary (2019). https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. Accessed 21 Feb 2020
7. Chen, Z., Subramanian, D.: An unsupervised approach to detect spam campaigns that use botnets on twitter. CoRR abs/1804.05232 (2018). http://arxiv.org/abs/1804.05232
8. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? pp. 21–30 (2010). https://doi.org/10.1145/1920261.1920265
9. Cresci, S., Petrocchi, M., Spognardi, A., Tognazzi, S.: On the capability of evolved spambots to evade detection via genetic engineering. Online Soc. Netw. Media **9**, 1–16 (2019). https://doi.org/10.1016/j.osnem.2018.10.005. http://www.sciencedirect.com/science/article/pii/S246869641830065X
10. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of promoted social media campaigns (2016). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13034
11. Fredheim, R.: Putin's bot army - part one: a bit about bots (2013). http://quantifyingmemory.blogspot.co.uk/2013/06/putins-bots-part-one-bit-about-bots.html
12. Grimme, C., Assenmacher, D., Adam, L.: Changing perspectives: is it sufficient to detect social bots? In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10913, pp. 445–461. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91521-0_32
13. Grimme, C., Preuss, M., Adam, L., Trautmann, H.: Social bots: human-like by means of human control? Big Data **5**(4), 279–293 (2017)
14. Hegelich, S., Janetzko, D.: Are social bots on twitter political actors? empirical evidence from a Ukrainian social botnet. In: International AAAI Conference on Web and Social Media, pp. 579–582 (2016). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015
15. Kessling, P., Grimme, C.: Analysis of account engagement in onsetting twitter message cascades. In: Grimme, C., Preuss, M., Takes, F.W., Waldherr, A. (eds.) MISDOOM 2019. LNCS, vol. 12021, pp. 115–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39627-5_10

16. Lee, K., Caverlee, J., Cheng, Z., Sui, D.Z.: Campaign extraction from social media. ACM Trans. Intell. Syst. Technol. 5(1) (2014). https://doi.org/10.1145/2542182.2542191
17. Mood, A.M., Graybill, F.A., Boes, D.C.: Introduction to the Theory of Statistics, 3rd edn. McGraw-Hill, New York (1974)
18. Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., Stieglitz, S.: Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. Eur. J. Inf. Syst. 1–19 (2019). https://doi.org/10.1080/0960085X.2018.1560920
19. Silva, J.A., Faria, E.R., Barros, R.C., Hruschka, E.R., Carvalho, A.C.D., Gama, J.: Data stream clustering: a survey. ACM Comput. Surv. **46**(1), 13:1–13:31 (2013). https://doi.org/10.1145/2522968.2522981
20. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: International AAAI Conference on Web and Social Media, pp. 280–289. AAAI (2017). https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817
21. Varol, O., Ferrara, E., Menczer, F., Flammini, A.: Early detection of promoted campaigns on social media. EPJ Data Sci. **6**(1), 13 (2017). https://doi.org/10.1140/epjds/s13688-017-0111-y
22. Woolley, S.: Automating power: social bot interference in global politics. First Monday **21**(4)(2016)
23. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with artificial intelligence to counter social bots. Hum. Behav. Emerg. Technol. **1**(1), 48–61 (2019). https://doi.org/10.1002/hbe2.115. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.115