Information Propagation Speed and Patterns in Social Networks: A Case Study Analysis of German Tweets

Raad Bin Tareaf*, Philipp Berger, Patrick Hennig, Sebastian Koall, Jan Kohstall, Christoph Meinel

Hasso Plattner Institute, University of Potsdam, Potsdam, Germany.

* Corresponding author. Tel.: (+49) 331 5509 549; email: raad.bintareaf@hpi.uni-potsdam.de
Manuscript submitted September 10, 2017; accepted November 12, 2017.
doi: 10.17706/jcp.13.7.761-770

Abstract: In this paper, we present our experiences in analyzing Twitter data. The analysis has shown that information diffuses over time through the Twitter network in certain patterns. Furthermore, it has shown those friend relationships significantly influence the information propagation speed on Twitter. Since it was launched in 2006, the microblogging service grew tremendously. Tweets are sent by users all around the world. Results show that there are two major patterns. While these patterns accommodate us to understand the diffusion of information through Twitter in an even better plan, the analysis of friend networks provides information on who influences the network, concerning the number of re-tweets and the time between a tweet and its re-tweets. The approaches have been evaluated both technically, based on how certain a topic matches one of the patterns and how prominent friends are compared to other users, and conceptually, based on existing, well-known approaches in measuring the speed and scale of information diffusion on Twitter.

Key words: Friends network, pattern detection, propagation speed, twitter analysis.

1. Introduction

Twitter is a microblogging and network website where users read and write millions of short messages on a variety of topics every day. The network was launched in March 2006. Only half a year later a public API was added. The REST API allows crawling a huge amount of structured data and the corresponding metadata in a short amount of time. For every tweet, there is information available about its creation time and the connections to other tweets. Therefore Twitter is a good starting point for analyses about information propagation. Information propagation is the process of information getting diffused in a network and reaching individuals through interactions. Information speed is the amount of time needed in order to make such an interaction happen. The number of interactions happening at the same time is called scale. [1] Understanding which individuals cause a scaling topic in a short time enables trend prediction and influencing.

Online social networks are graph based, which means that the basic structure of such a network can be displayed as a graph. Nodes represent user and edges their interactions. A node is called infected if it already received a piece of information. By interacting with further nodes it diffuses an information, they become infected too and a cascade is built. Those cascades are called the information range. In this paper, information propagation patterns will be presented. This patterns can be used to determine how a topic's importance will develop during a day. Besides, it will describe how to detect a network of friends which

propagates information faster and who are the most valuable members of that network.

2. Related Work

Influencers can be identified by calculating the page rank for a set of Twitter users. It is shown by Kwak *et al.* [2] that a user's influence can be calculated by the number of his retweets too. Furthermore, the paper compares trending topics on Twitter to other media and shows that major news is most likely to trend on Twitter too.

Cha *et al.* [3] present indegree, retweets and mentions as measures of influence for users on Twitter. They interpret the dynamics of individuals influence over time in two scopes. Initially, they tracked the popularity of top influentials over a protracted period and inspect how well they preserve their ranks. After that, they focused on users who raised their influence on a specific topic over a short time period, in order to understand what behaviors make ordinary individuals influential. It is shown that influencers hold their status among different topics. Hennig *et al.* [4] use information retrieval approaches to identify trends inside unstructured blog data. They state that tags as part of blog meta information and links influence whether some topic causes a trend. Information diffusion has been analyzed by Yang *et al.* [1] based on Twitter interactions using mentions (@<username > feature) of Twitter. Diffusion is divided into three major properties: speed, scale and range. While the overall retweeting speed on Twitter already has been analyzed there is the lack of information concerning differences between local and worldwide diffused tweets. By applying common information retrieval approaches like the HITS algorithm or the Clique Percolation Method (CPM) onto Twitter data, Java *et al.* [5] analyze the structure of user groups on Twitter and how information diffuse through the network. Hubs and authorities are identified in order to reconstruct groups out of the given data. The paper of Naaman *et al.* [6] interpreted geographical areas of trends and computed features for different categories of trends. There is a difference between so-called exogenous trends which are coming from the outside of the network like an earthquake or other breaking news and endogenous trends which are activities only in twitter and not correspond to external events. For example, a post of a famous celebrity.

3. Concept

The basis of friend networks and trend pattern detection are the definitions of information speed and scale. In this paper the definition of information scale from [1] will be used, therefore the scale is defined by the number of retweets for an original tweet. Furthermore, information speed will be measured in tweets per hour in the pattern detection.

3.1. Trends

As a first step to analyze the dataset, we built histogram for hashtags to visualize the spreading of information. To do so, we settle all tweets which were posted during the same hour inside one chunk. We decided to use one-hour segment since we found it as a reasonable measurement, it is not to small like 10 minutes which would lead to huge deviation and would make it difficult to discover a trend and it is no larger than one-hour that would hide information and reduce the results of the analysis Analyzing histograms lead us to realize that there are different trends. Some only affect users in one region and stay local. Those have a propagation pattern where the number of tweets decreases during the local night time. While other trend which have global influence do not decrease because another time zone is awake at this time and tweeting about it. On target of this paper is to outline how to detect if a trend is locally or globally propagated.

3.2. Friends Network

In contrast to [1] where mentioning users inside a twee considered important and users mentioning each other are named friends, this paper will identify the importance of user by building a network of retweets relation. Someone will be called a friend if he retweets another user and the other on retweets the first. The higher number of interactions indicate a stronger tie between those users and could result in lower response time to tweets of friends in the entire network.

3.3. HITS Algorithm and Friends

Based on the concept of friends, HITS Algorithm is able to calculate how much influence one friend has. Due to the fact that every relation between friends is bidirectional, calculate hub and authority score are equal. Alternatively, this score can be calculated on tweet-retweet relations between users The HITS algorithm calculates authority and hub score in iterations. Each step the scores are calculated as shown I equation 1 and 3. Afterwards each score is normalized as I equation 2 and 4 in order to create converging HITS scores That way the overall sum of the scores is always one, which makes it easier to estimate the relative influence of one node.

$$newAuthorityScore = \sum_{i=1}^{n-1} incomeEdges[i] \tag{1}$$

$$authorityScore = hubScore / \sum_{i=1}^{n} authorityScore[i] \tag{2}$$

$$newHubScore = \sum_{i=1}^{n-1} outgoingEdges[i] \tag{3}$$

$$hubScore = hubScore / \sum_{i=1}^{n} hubScores[i] \tag{4}$$

4. Implementation

As a first step, we got the only information we were interested in by developing an outlier removal for the analysis. Some of the outliers are much older than this current peak we were looking at. The implemented outlier filtering cuts off all tweets in a certain time interval before the actual peak. By iterative examination, we found that taking three days before and after led to convenient results. To determine if a trend is local or global, we developed indicators which would support the certainty of whether a particular hashtag diffuses locally or globally.

4.1. Indicators

To determine the propagation speed of the tweets, the number of tweets per hour are examined. The speed of the trend is determined by the number of tweets/re-tweets per hour, typically, the higher number of tweets increases the trend propagation speed. The time points from i - 1 to i are investigated, and the tweet is deemed local or global only if the speed increases past the 51% time mark. Generally, there is now a significant change in the hashtag status, enough to trigger the process of categorizing the said hashtag by using the following four indicators:

- *Day-Night Cycle Indicator:* Trends which are only valid for a certain region or country follows a characteristic day-night cycle. So the topic becomes popular in the morning and then towards the night fewer people tweet about it since most of them are going to sleep. Therefore, we analyze if the trend is following the local day-night pattern. If there are increases in tweeting amount for a hashtag at the night, this will indicate that there are people from a different time zone which also tweeting about this topic during their daytime. So it is broadly popular.

- *Night-Inactivity Indicator:* Less exclusive than the Day-Night Circle Indicator is the Night-Inactivity Indicator. Since some trends go up and down several times during the day, they fail the Day-Night Cycle Indicator which assumes that it only raises during the day and a number of tweets drop during the night. So we observed that there is no raise during a night, which would indicate that it is also popular in another time zone. All trends which are covered by the Day-Night Cycle Indicator are as well considered local by the Night-Inactivity Indicator.
- *Language Indicator:* Another indicator is to detect the used language. For example, a language like German which is not so far distributed globally allow to see that the trend is most popular there. Indeed, the language indicator explores the language distribution of the tweets and when more than 80% of the tweets are having one language, then it is a genuine indicator that it is only trending where the language is spoken. Since English is spoken all over the world this does not apply for English.
- *Short-Day Trend Indicator:* Some trends are extremely short that they do not even satisfy the whole day night cycle. Only a few hours after their growth, the amount of tweets starts to shrink. To identify those tweets, we analyze how long the difference between the first big jump in popularity and the decline at the end is. If it is less than twelve hours, we assume that most likely the trend is local since a global trend stays much longer active because of the active users in different time zones.

4.2. Detecting Local to Global Transformation

A trend which persists from local to global starts in one country or region of origin and then it becomes broadly popular and people all around the world start tweeting about it. To track when this change happens, our algorithm goes iteratively through the tweets and checks for every hour if there is a significant change in the tweets which indicates that there is a change from local to global. The factors which are taken into consideration are the following:

- *Increasing Quantity:* since the trend is becoming global, more users are involved in and more users are tweeting about it as well as the number of tweets strongly increased. Therefore, we examine the amount of tweets at time stamp i and $i + 1$ and compare the values for a significant increase. Also, we take the previous values into consideration and compare it to the average value of the tweets in the time passed before $i + 1$.
- *Time irregularities:* in order to distinguish local/- global trend, the time of the hashtag *expansion* is important. So if it is at the local night time we can assume that it is from users in another time zone which are still awake and this means that the trend is popular in more than one region. For example, if a German hashtag being retweeted heavily at 03:00 am, this implies that there is a possibility of being expanded globally and other time zone regions such as EDT in the United States are tweeting about it in their normal day time. This indicates together with the other factors that the trend is becoming global.
- *Language Distribution:* another aspect we are taking into consideration is the language *distribution*. A local trend has a different language distribution than a global trend. So we analyze how the trend evolves over time and assumes that the more the language distribution is changing towards English the more global the trend is becoming. As before, this only applies if English is not the primary language in the beginning.

5. Dataset Description

Overall 1.218.250 Tweets from 785.671 different users were collected between the 2nd and 31st of January 2017. During that time 167 hashtags were tracked over a period of at least 24 hours per hashtag. Furthermore the dataset contains 124 non-hashtag search terms. If a crawled tweet was a retweet, then original tweet is added to the database in order to have more relations between tweets. Fig. 1 shows 20

hashtags of the dataset with the most tweets in descending order. There are two with around 20,000 tweets, at the lower edge of our dataset there are hashtags with around 100 tweets (e.g.: #BOCKOE, #BBBT).

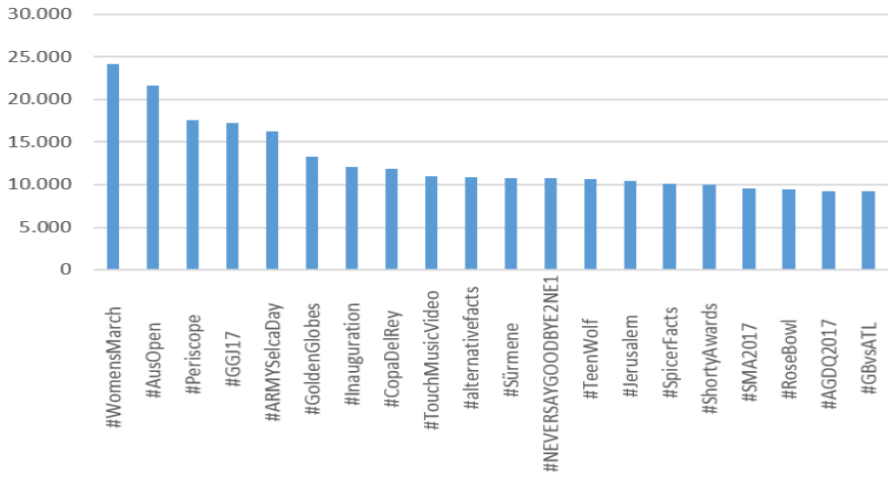

Fig. 1. 20 hashtags with most tweets.

Table 1. Chosen Location and Their WOEIDs

Location	WOEID
Worldwide	1
Australia	23424748
Germany	23424829
USA	23424977

5.1. Data

The Tweets are crawled from four different locations all over the world. Table 1 shows the chosen locations and their corresponding WOEIDs. Those ids were created by Yahoo and aim to provide n unique identifier for any place of the world. Places like countries as well as cities and villages have unique ids.

5.2. Influence of Social Media Bots

The users with the most tweets we found in our dataset are bots. They usually have a simple task like posting weather forecasts or the results of football games. Those bots are common and they have one task and fulfill it. Another kind of bots is Social Media Bots where they try to influence users by either posting fake information or other conspiracy theory. Since our aim is to analyse and measure the speed of trendy hashtags from human users over the time, we intentionally ignored those bots and only considered with 167 hashtags from valid users.

5.3. Data Crawler

The data crawler is a script which was executed 15 minutes, to match the refresh timespan of the Twitter API. At the beginning of each day ten trends were chosen randomly out of the list available through of the API. Resulting in 291 different trends crawled in 30 days. The remaining nine trends were chosen multiple times on different days.

5.4. Global and Local Patterns

As shown in Fig. 2 the trend “#Diekmann” is varying up and down with the local day and night times in

Germany. Since Kai Diekmann is mostly known in Germany, we found that "#Diekmann" did not spread in the same way in USA. Analyses of the dataset confirmed that it was only a German local trend. In Fig. 3 the same trend is analysed in case of United- States country. Considering the same time slot that we analysed "#Diekmann" within Germany, EDT time considered in USA. Specifically, when the time was around 13:00 in Germany, US time was almost 07:00 am. The propagation and the usage of the "#Diekmann" in Twitter shows a noticeable different respectively to time, language and location. The discovering of this pattern led us to look for hashtags which do not fit with that. For example the hashtag "#DavidBowie" visualized in Fig. 4 is not matching with German day and night cycles, it stays up the whole time and since he was globally famous we manually considered as it as a global trend.

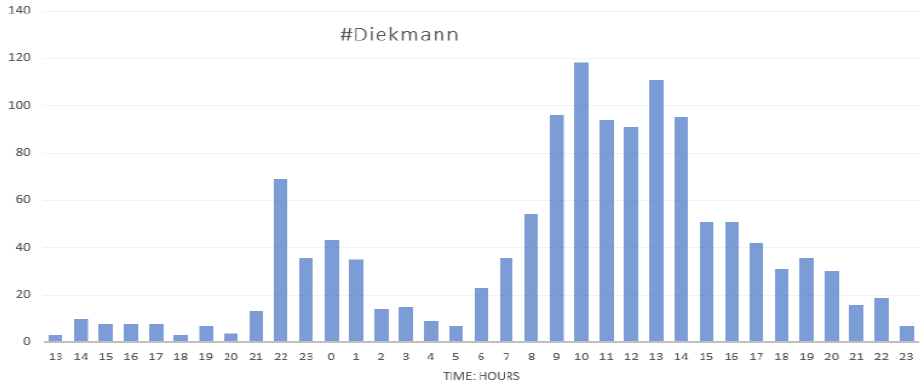

Fig. 2. Tweets per hour for # Diekmann in Germany.

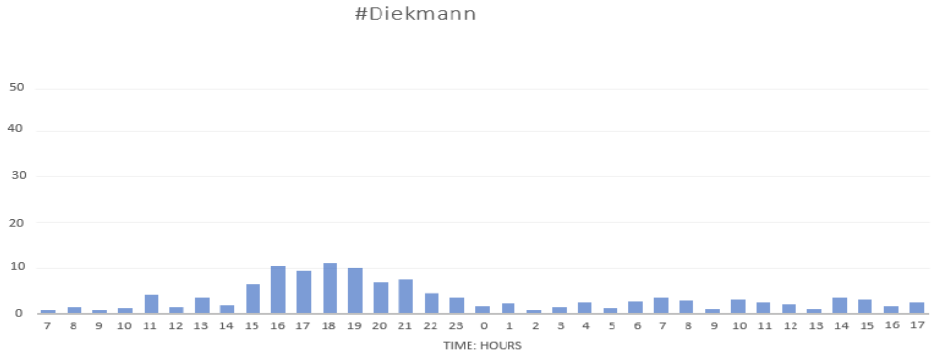

Fig. 3. Tweets per hour for #Diekmann in USA.

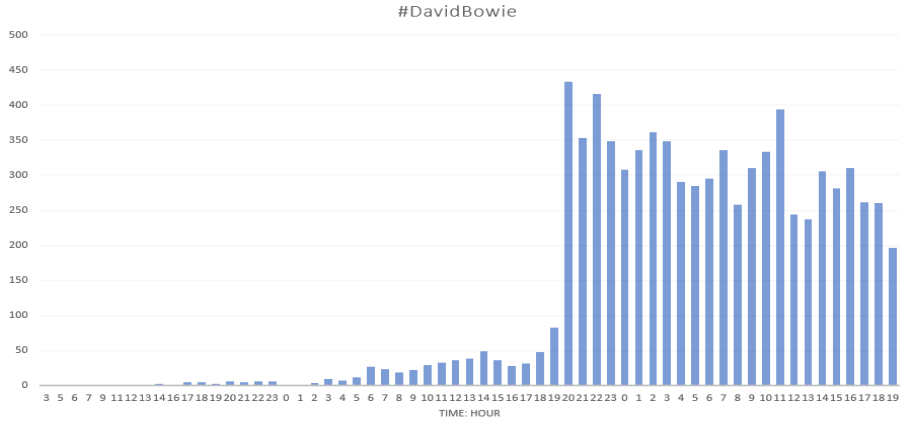

Fig. 4. Tweets per hour for #DavidBowie.

Table 2. Average Values of the Dataset Compared to Friend Networks

	Overall Dataset	Friends network
User Count	785,671	1117
Tweet Count	1,218,250	15,856
Retweet Count	769,551	30,739
Average retweet time (in hour)	10,08	7,21

5.5. Friends Network

Based on the definition described in Section 3, friend networks inside the data were identified. Due to the fact that retweets were included into the dataset, some extremely old tweets are in the dataset. To not influence average values with this outliers for the following calculations tweets posted before.

01.01.2017 are excluded from any calculation. Including those tweets friend network are in average 583.7 % faster in retweeting. This value does not make much sense and only leads to the conclusion, that friend networks tend to not retweet older tweets. As shown in Table 2 a relatively small percentage (0.14 %) of friends posted 1.3 % of all tweets. The percentage distribution shifts even further towards the friend networks for the number of retweets. They cause 4 % of all retweets. In average this is done 39 % faster than by a common user. This statistics lead to the conclusion that friend networks are the most active user group on Twitter.

5.6. HITS Algorithm Results

Further analyses of the dataset have shown that top scoring users get retweeted faster, with more distinct retweeters than any other user inside the dataset. While some users get retweeted faster, they have less retweets or their tweets are always posted by the same users. Others have more retweets but in average they are retweeted significantly slower. Considering every retweet as a relation between two users and calculating HITS scores based on the resulting graph, produces high authority scores for users who get retweeted often as shown in Table 3. As already mentioned, even though the top scoring users for this graph might be of interest, their tweets get retweeted slower than the ones of friends.

Table 3. Top Ten Authority and Hub Scores of Friends

User ID	Authority/Hub Score
755616083621457920	0.1117
752062144938516482	0.0999
722700148590178304	0.0878
731127235856371713	0.0875
741978219008647168	0.0803
743403845145944064	0.0793
722703213724569601	0.0784
782441976683790338	0.0776
764704294524194816	0.0725
755618679769870336	0.0575

6. Observation

6.1. Dataset

We have chosen Twitter trends in order to have a topic which actually causes tweets. A pitfall of this approach is that some trends already are next to becoming irrelevant and none is tweeting about it. This problem can be partially solved by considering the original tweets of retweets as part of the database too. While we had a variety of different hashtags, many of them did not have an equal amount of tweets like explained in Section 5. One major problem of this fact is, that results of the HITS scores cannot be compared between different hashtags, due to the widely varying amounts of tweets per hashtag. Overall, the dataset was useful to identify the patterns and enabled assumptions about friend networks, which nevertheless need to be testified on further data. For the analysis of local trends, the dataset is fitting since it was collected each time with one special localization so that the local trends are covered quite well. Since global trends are at many more places, the dataset does not contain enough global trends to generalize them. Another limitation is the Twitter API so that we could not crawl the whole global trends.

6.2. External Influencers

The trend "#DavidBowie" is identified as a global trend and diffused in an almost normally distributed pattern. The main reason is probably that the trend was caused by an influence coming from outside of Twitter. David Bowie died on the tenth of January in 2016. The trending hashtag on the ninth and tenth of January in 2017 is just a result of his obit. The presented approach cannot take external influences into account, therefore certain hashtags may show up, which can be identified, but there will be a lack of explanations about how they developed over time. Other external influences are events like catastrophes which then become so overwhelming in their dimension that other trends do not get much attention and in those moment everyone is only tweeting about the event.

6.3. Trend Detection

The trend detection delivers for German trends good results which show that most of the trends where are started in Germany also stay in Germany. This can be of course due to the language barrier which makes it impossible for foreigners to understand what German trends are about but also because of the strong localization aspects. Many trends refer directly to some event happening in Germany. For example, a TV series displayed or news about German personalities. All those trends make it little interesting for people outside of Germany to tweet about it. Our algorithm detects those trends precisely. Analyzing the dataset for trends we marked as local, the Night Inactivity Indicator is legitimate. There are fewer trends which match with the Day Night Cycle Indicator since many trends are not active long enough to fulfill a whole cycle. So usually they decline again during the day. Most of those trends that do not last long are covered by the short day trend indicator. For a valid prognosis, if a trend is local it is necessary that either the Day Night Cycle Indicator or the Night Inactivity Indicator and either the Short Day Trend Indicator or the Language Indicator are true. This is fitting for most of the local hashtags in the dataset.

7. Future Work

In the paper we presented indicators to detect local trends and when local trends become global. This work can be enhanced by also analyzing, whether a trend is global. Non local trends cannot automatically be categorized as global. There are certain propagation patterns, which are neither of both categories and cannot be generalized since they are unique for a hashtag. It has to be determined, if a trend is truly global and therefore it effects the whole world and everybody is posting about it. The HITS Algorithm could only detect fast tweets with a comparatively large amount of tweets. Therefore, other ranking algorithms for network nodes like the page rank algorithm have to be evaluated. Furthermore, there is a need for some

kind of scoring functions, that can take retweet time, retweet amount and number of distinct retweeters into account. This function would enable the user to choose between different approaches, if he values one of the factors more than the others.

8. Conclusion

In this paper, we showed several ways to analyze trends on Twitter. The characterization of trends in local and global allows to discover how big the influence and how fast the propagation speed of a trend is and in which way a similar trend will evolve. This allows as well to differentiate the influence of a tweet depending on its trend pattern. Friend networks are the most active part of the Twitter network and cause relative to the numbers large amount of traffic. Friends can, once they are identified, be used as an initial point for any marketing campaign. Those networks will tend to diffuse information faster and on a bigger scale. Results of the HITS algorithm have shown that it only provides a limited set of predictions. Even though it can derive which Twitter user will cause a fast scaling number of tweets, no ranking predictions between them can be made yet.

References

[1] Yang, J., & Counts, S. (2010). Predicting the speed, scale, and range of information diffusion in Twitter. *ICWSM*, *10,* 355-358.

[2] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web* (pp. 591-600). ACM.

[3] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, *10(10- 17),* 30.

[4] Hennig, P., Berger, P., Lehmann, C., Mascher, A., & Meinel, C. (2014). Accelerate the detection of trends by using sentiment analysis within the blogosphere. *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 503- 508). IEEE.

[5] Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (pp. 56-65). ACM.

[6] Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the Association for Information Science and Technology, 62(5)*, 902-918.

Raad Bin Tareaf is a Ph.D student at Hasso-Plattner-Institute, Potsdam, Germany, born in 1991, got a bachelor degree in computer science form German-Jordanian University in Jordan. He wrote his bachelor thesis in cooperation with Continental AG in Regensburg, Germany about automotive platforms integration using OSLC integration standards. He did an internship on big scale software testing at Robert Bosch GmbH in Stuttgart, Germany. After he received his bachelor's degree, he studied masters of enterprise system engineering at princess Sumaya University for technology in Jordan. Since his master studies, he is focusing on the topic of social media analytics especially by investigating the previous used algorithms in that domain and by trying to come up with new algorithms that can specify and predict the future form social media data in order to facilitate the process of decision making.

Philipp Berger is a Ph.D student at the Hasso-Plattner-Institute, Germany, investigating new algorithms for analyzing social media channels. Currently, his focus is the identification of important authors and topics among weblogs. His research results are part of a joint research project called BlogIntelligence that offers a web prototype portal

(blog-intelligence.com). Philipp, born in Berlin, earned his bachelor and master degree in IT systems engineering at the Hasso-Plattner-Institute. During his studies he went abroad for an internship on product feature extraction of product reviews at SAP Labs, India. Philipp successfully launched his start-up, check https://nexenio.com/.

Patrick Hennig was born in 1988 in the city of Walldurn, Germany, got a bachelor degree (B.Sc) in IT-systems engineering from Hasso-Plattner-Institute in Potsdam, Germany. He wrote his bachelor thesis in cooperation with SES Astra, Belgium and SAP Research Pretoria, South Africa about project management with SCRUM in small scale projects. After he received his bachelor's degree he studied abroad at University of Coimbra, Portugal in computer engineering. Since his masters studies (M.Sc.) in IT-systems engineering and the master thesis about identifying trends based on the blogosphere at Hasso-Plattner-Institute in Potsdam, Germany, he is focusing on the topic of social media monitoring especially by working on an analysis framework for blogs, integrated into a web portal (blog-intelligence.com), with the objective to leverage content- and context-related structures and dynamics in the blogosphere. Particularly he is working on the challenge to do real-time analyses on this big set of data. Currently, Patrick is a Ph.D student at Hasso-Plattner-Institut in Potsdam, Germany. Patrick successfully launched his start-up, check https://nexenio.com/.