Information Systems Frontiers (2020) 22:395–409
https://doi.org/10.1007/s10796-018-9867-2

# Going Back in Time to Predict the Future - The Complex Role of the Data Collection Period in Social Media Analytics

Stefan Stieglitz[1] · Christian Meske[2] · Björn Ross[1] · Milad Mirbabaie[1]

Published online: 8 June 2018

**Abstract**
In the context of events that involve public voting, such as televised competitions or elections, it has increasingly been recognized that communication data from social media is related to the outcome. Existing studies mainly analyse the number of messages and their sentiment, yet the role of different data collection periods has not been examined sufficiently. We collected Twitter data in 2015 and 2016 to examine the relationship between the audience voting of the Eurovision Song Contest and predictors based on quantity and emotions, and compared the results of using data from before and during the event. We found that the choice of time period greatly affected the results obtained. Data collected prior to the event exhibited a much stronger association with the final ranking than data collected during the event. In addition, the model based on pre-event data in 2015 showed considerable accuracy in predicting the 2016 results, illustrating the usefulness of social media data for predicting the outcomes of events outside social media.

**Keywords** Social media analytics · Time period · Predictive analytics · Eurovision · Twitter

## 1 Introduction

Social media enables users to discuss their opinions, spread information, and let others be part of their thoughts and experiences (Larosiliere et al. 2017; Stieglitz et al. 2018a, b). Twitter, as a microblogging platform with a focus on information sharing and mobile usage, is of increasing importance for the live channelling of media events (Highfield et al. 2013; Kim et al. 2015; Vaccari et al. 2015). In comparison to posts on social networking sites such as Facebook, tweets not only reach friends and direct followers, but also people searching for a specific hashtag mentioned in the tweet. Therefore, every Twitter user is able to start a discussion, or to follow and join debates that are already in progress (Bruns and Burgess 2011; Marwick and boyd 2011). As of the third quarter of 2017, Twitter has approximately 30 million monthly active users (Statista 2017). Tweets are well suited to keep followers updated even on rapidly changing situations such as crises (Gruber et al. 2015; Eriksson and Olsson 2016; Pond 2016). The platform has been extensively researched in the crisis communication domain (Oh et al. 2013; Stieglitz et al. 2017; Stieglitz et al. 2018a), in political communication (Greene and Cunningham 2013; Stieglitz and Dang-Xuan 2013; Valenzuela and Bachmann 2015; Nulty et al. 2016), and in the fields of business and marketing (Dijkmans et al. 2015; Spence et al. 2016).

Twitter data can be collected through its application programming interface (API). The gathered data can be used e.g. for analysing the intentions of voters during elections and deriving conclusions from the findings (Maldonado and Sierra 2015), predicting the stock market (Nann et al. 2013; Nofer and Hinz 2015) or making sales forecasts (Benthaus and Skodda 2015). One particular phenomenon is the use of Twitter to connect and support conversations between television viewers. When examining the Eurovision Song Contest (ESC), a large European media event and music competition, Highfield et al. (2013) found that Twitter is an unofficial extension of the event and is used by viewers all over Europe to communicate in real time. In 2015, Twitter and Eurovision cooperated officially (Storvik-Green

✉ Stefan Stieglitz
stefan.stieglitz@uni-due.de

Christian Meske
christian.meske@fu-berlin.de

Björn Ross
bjoern.ross@uni-due.de

Milad Mirbabaie
milad.mirbabaie@uni-due.de

[1] University of Duisburg-Essen, Duisburg, Germany

[2] Freie Universität Berlin, Berlin, Germany

2015a), suggesting that tweets may reflect the opinions of the viewers. Since the event involves a televoting system where viewers vote for their favourite entries, tweets may be useful to determine who is likely to win. A survey of the literature with respect to viable predictors revealed that the volume and the expressed sentiment of social media data were commonly found to increase the predictive power of applied models across most of the areas of research (Schoen et al. 2013; Ceron et al. 2014; Nguyen et al. 2015). However, some of the results were contradictory, and the previous studies typically only used a single data set collected over a relatively short time period. It is unclear whether these relationships persist across several years.

Previous research provides an indication that the period of data collection plays an important role in the quality of the predictions (Mishne and Glance 2006). The influence of this choice remains to be discussed.

In summary, in the context of media events, little is known about the general nature of predicting the outcomes of this type of event, using a combination of different predictors such as volume, sentiments, or the role of the time period considered.

In this paper, we seek to address the current lack of methodological knowledge, by investigating Eurovision Song Contests 2015 and 2016 as media events and the power of a model that combines tweet volume and sentiment to predict the events' audience voting. Overall, this study aims to answer two guiding research questions:

1. What is the relationship between tweet volume, tweet sentiment and the outcome of a media event, using the example of the Eurovision Song Contest audience ranking?
2. What influence does the choice of time period have on the results obtained in response to the above research question?

To answer the above research questions, we use ordinal logistic regression and statistical hypothesis tests. The data for fitting the model was collected before and during the 2015 contest, but before the results were announced. In addition to that, one year later, we collected data on the 2016 contest. The benefit of this second data set is twofold. First, it allows us to measure the predictive accuracy of the regression model by evaluating how well the 2015 model forecast the 2016 outcome. Secondly, it allows us to test the same hypotheses as in 2015 a second time, using the 2016 data instead. This replication of our own study allows us to see if the previous findings could be reproduced one year later. This procedure results in a unique combination of an explanatory goal with a predictive evaluation: hypotheses are derived theoretically and tested statistically, while at the same time we produce and test a true forecast based on data collected a year in advance.

The remainder of this paper proceeds as follows: In section 2, we present literature on the prediction of public events based on social media data analysis. In this section, we derive four hypotheses from the literature that address the above research questions. Section 3 describes the research design, including information regarding data collection and methodology. We then present the statistical results and model evaluation in section 4, followed by a discussion of the tested hypotheses, their implications for practice and the limitations of our approach in section 5. The paper ends with a conclusion in section 6, including an outlook to further research.

## 2 Related Work and Hypotheses

### 2.1 Social Media Analytics and Event Forecasts

In recent years, social media platforms have evolved into powerful communication tools and information sources in daily life. The analysis of the data requires a systematic approach to gather, prepare and analyse the data for a specific purpose (Stieglitz et al. 2014; Wahyudi et al. 2018). Social Media Analytics is a research field that addresses this issue (Stieglitz et al. 2018b). By using mixed approaches, such as social network analyses and sentiment analyses, one can, for example, identify influential actors in the communication and the type of content they create (Golbeck et al. 2017). This information can be used for detecting topics (Chinnov et al. 2015), measuring the reputation of a company (Dijkmans et al. 2015; Spence et al. 2016), analysing trends (Kaschesky et al. 2013) or managing events (Calderon et al. 2014).

The analysis of social media data reveals information about an event, which can be used for improving decision-making in various domains for specific purposes (Cheong and Lee 2011; Rudra et al. 2018). In crisis situations, data can be collected and the content can be analysed (Oh et al. 2013; Stieglitz et al. 2017; Avvenuti et al. 2018; Stieglitz et al. 2018a), e.g. to detect and categorize earthquakes in a specific area (Sakaki et al. 2010; Imran et al. 2015) or to identify an epidemic in its early stages (Li and Cardie 2013). In political communication, it is possible to identify and categorize the actors according to their political affiliation (Greene and Cunningham 2013) and analyse the emotionality of their content (Stieglitz and Dang-Xuan 2013; Valenzuela and Bachmann 2015; Nulty et al. 2016).

The presence of all this data has naturally given rise to the desire to study past behaviour, and ultimately, to forecast the future (Huberty 2015). Predicting with social media data in general works well when online behaviour accurately reflects offline behaviour. During big media events, Twitter often serves as a "backchannel" that the audience uses to discuss the event (Highfield et al. 2013). As a consequence, tweets can mirror the behaviours and opinions of the crowd.

Therefore, search engine queries or posts on microblogging systems have been used to forecast the spread of epidemics (Chen et al. 2016; Li et al. 2016), the behaviour of the stock market (Bollen et al. 2011; He et al. 2016) and the outcomes of elections (Tumasjan et al. 2010; Burnap et al. 2016; Charles and Reid 2016) with varying degrees of success. Successful forecasts have been made in the context of predicting movies' box office revenues (Asur and Huberman 2010). In the context of media events, only little research exists in the context of predicting an event's outcome. One of the few exceptions is the work of Ciulla et al. (2012), which predicted the winner of the TV show "American Idol". Pavlyshenko (2013) focused on a set-theoretic model of key tags for Twitter messages. The author considers the possibility of applying the theories of frequent sets, association rules, and formal concept analysis to event forecasting using tweet mining and also focused on Eurovision.

Focusing on Twitter specifically, recent research has pointed out two main variables that are relevant for predictions: the number of tweets in a certain time on a specific topic (Ciulla et al. 2012), and the sentiment expressed in those tweets (Kaya and Conley 2016). These findings will be discussed in the following section and are the basis of our hypotheses.

## 2.2 The Predictive Power of Volume and Sentiment in Different Time Periods

Especially in the context of predicting political elections, the number of tweets as the single predictor has proved to be a popular choice among researchers. The "one tweet, one vote" heuristic assumes that a greater attention to a candidate on Twitter is correlated with a higher likelihood of electoral success. DiGrazia et al. (2013) suggests that metrics based on message volume alone could contribute value to election predictions in US House races. In the 2008 US presidential primaries, the number of Facebook supporters was enough to predict the result successfully (Williams and Gulati 2008). Sang and Bos (2012) analysed the possibility of simply counting Twitter messages, which include mentions of political parties and predict the election outcome of a Dutch senate election in 2011. On the other hand, there is an ongoing debate on whether social media is an unbiased representation of public opinion that can be used to predict election results. For example, in British Columbia's 2001 provincial election, the number of mentions on Internet message boards did not indicate the relative strength of the parties (Jansen and Koop 2006). Yet, in the context of media events, Ciulla et al. (2012) have shown that simple measures quantifying the popularity of the American Idol participants on Twitter strongly correlated with their performances in terms of votes. In summary, previous research has shown contradictory results. Most studies were based on data from only one event, a serious limitation. A hypothesis that was true a year ago is not necessarily true today. We consider the ESC a media event similar to "American Idol" and therefore propose the following hypothesis:

> H1: There is a consistent, replicable positive relationship between the number of artist-related tweets and a better artist ranking in the audience voting.

Kaya and Conley (2016) analysed the accuracy of sentiment analysis for predicting the winner of a contest in a TV show. The authors discuss the possibility of conducting a sentiment analysis to predict events and compared several lexicons using a frequency-based statistical classification and k-means. Focusing on Twitter as a platform for generating social media data, Stieglitz and Dang-Xuan (2013) have shown that sentiment positively correlates with the number of retweets, as well as with the speed of retweeting (time between original tweet and first retweet). Tweets that express emotions disseminate more quickly through the Twitter network. Besides a high level of cognitive involvement, certain emotions such as anger, anxiety, awe, or amusement might trigger a high level of physiological arousal (Berger 2011), which has been shown to be a driver of information sharing (Berger 2011; Berger and Milkman 2012). Content that evokes high arousal is more viral, while the reverse is true as well (Stieglitz and Dang-Xuan 2013). It follows that the tweets' sentiment is of importance in the emergence of events because it factors into the awareness, recall, and judgement of information (Fox 2008; Kinsinger and Schacter 2008) as well as the motivation associated with information behaviour. As for predictions, Tumasjan et al. (2010) found that the tweets' sentiment is correlated with the electors' preferences in the political context. Moreover, O'Connor et al. (2010) reported similar results. Furthermore, Thelwall et al. (2011) state that the expressed sentiment on Twitter around an event does not change very much. However, most of these results have only been tested in a single time period. It is questionable whether they can be generalized. These findings could support predictions only if the relationship between mood on social media and event outcomes is shown to be relatively stable. Therefore, we propose the following second hypothesis:

> H2: There is a consistent, replicable positive relationship between the sentiment of artist-related tweets and a better artist ranking in the audience voting.

There have been a few attempts at combining both variables, volume and sentiment, for improving prediction results. Zhang et al. (2011) examined the role of volume and sentiment in the field of macroeconomics by analysing Twitter posts to predict stock market indicators such as the Dow Jones. They found that for posts including specific emotive words such as *hope*, *worry* and *fear*, the total number is more predictive of stock indices than the number and proportion of their forwarding times and original authors' followers. Importantly, they state that checking Twitter for emotional outbursts of any kind predicts how the stock market will be going the next day. O'Connor et al. (2010) proposed a simple model, wherein the share of mentions of John McCain or Barack Obama and the sentiment attached to those mentions could provide a leading indicator of performance in presidential polling. Among the popular research topics, the prediction of a movie's box office revenue, in our opinion, comes closest to the prediction of a traditional media event, because it carries similar characteristics. First, both topics have an entertainment character. Second, in both topics, people are influenced by their environment. Various criteria, such as written reviews by other people, could have an effect on buying decisions or votes.

Asur and Huberman (2010) found that analysing sentiment improved the prediction of the success of movies – in comparison to only measuring the volume of tweets. From a marketing perspective, previous research in the context of product sales studied the relationship between the valence and volume of the electronic word of mouth and sales (Liu 2006; Li and Hitt 2008; Archak et al. 2011). These results provide further motivation for the combination of both variables into a predictive model. Yet, many of these studies are a decade old or more. It is unclear whether changes in usage habits have led to changes in these relationships. An important contribution of our hypotheses H1 and H2 lies with the consideration of the two data sets (2015 and 2016) in one analysis. Although both hypotheses have been tested in other contexts, we argue that analysing both over two years and thus comparing both events enriches our findings. It allows us to say whether their statistical significance is replicable and their effect size consistent.

Finally, there are few empirical results on the question of which time period should be chosen for data collection. According to Yu and Kak (2012), there are no guidelines yet for choosing a reasonable and accurate time window. In many studies, researchers do not point out why a specific time period was chosen during which to collect social media content, a tendency that becomes problematic when the results depend on the time period considered (Jungherr et al. 2012). For the case of movie box office, the number of positive references after the film was released correlates more with the event result (prediction of success of the movie) than the total count in the pre-event period. Hence, in this case the total count from the post-event period seems to be the better "predictor" (Mishne and Glance 2006). However, a model that uses data from after an event can obviously not be used to predict its outcome in advance. Moreover, in the case of a public event such as Eurovision, the theory of avoidance of cognitive dissonance as well as self-presentation (Festinger and Carlsmith 1959; Schlenker and Goldman 1982) should be considered. Based on this theory, we assume that people who already follow the semi-finals, and form as well as disclose their opinions at an early stage are less likely to change their previously established opinions during the main event. This behaviour can be explained by the tendency to avoid cognitive processes that lead to an internal dissonance in the case that the initially favoured contestant performs worse during the main event. Such effects can be increased through a perceived external pressure due to social influence, for instance if the earlier opinion was expressed visibly to other human beings (Baumeister and Tice 1984), e.g. via social media. At the same time, we expect that people who already participated in pre-event discussions are more likely to also participate in the final audience voting process, due to an increased engagement and interest in the topic. Therefore, we assume that data from the pre-event phase will represent the audience better than the tweets during the event. Hence, we posit:

> H3: The explanatory power of artist-related tweets from prior to the event is higher than for those from during the event. This relationship is also valid across more than one year.

## 3 Research Design

### 3.1 Event Description and Data Collection

The ESC has been held annually since 1956, and is one of the largest music competitions and longest-running television shows in the world (Georgiou 2008). Held annually in May, the live contest attracts millions of viewers both from competing countries and from around the world (Storvik-Green 2015b). It follows a consistent format: competing countries each select a contestant and song to represent them. Participating countries have established various procedures to select their candidates (e.g. national contests such as the Swedish "Melodienfestivalen"). Members of the European Broadcasting Association are generally allowed to send a contestant to the semi-finals. During each of the two semi-finals, ten participating countries are selected by the public audience via telephone and Internet for the finals. Additionally, the "big 5" (Germany, France, UK, Italy, Spain), as well as the host country automatically qualify, leading to 26 contestants in the finals. Since 2015, Australia has also participated due to the popularity of the contest there. In 2015, the country was automatically qualified, increasing the number of participants to 27; since 2016, they have participated in the semi-finals like European participants. During the finals, people from each country can vote for other countries (not for their home country). Besides this public telephone voting, national juries also vote in secret. The way public votes and jury votes are combined to select the final ranking is occasionally changed, but our analysis only considers the public votes. The jury votes are not published until after the public has finished voting, and are therefore very unlikely to be reflected in tweets.

The ESC generates a considerable amount of attention and traffic on social media. In 2015, Eurovision cooperated with Twitter, who supported the conversation around the event by implementing "hashflags". If a hashtag of a participating country (e.g. #GER for

**Table 1** Hashtags and names of participating countries

| 2015<br>27 countries | 2016<br>26 countries |
|---|---|
| #ALB OR Albania, #ARM OR Armenia, #AUS OR Australia, #AUT OR Austria, #AZE OR Azerbaijan, #BEL OR Belgium, #CYP OR Cyprus, #ESP OR Spain, #EST OR Estonia, #FRA OR France, #GBR OR United Kingdom, #GEO OR Georgia, #GER OR Germany, #GRE OR Greece, #HUN OR Hungary, #ISR OR Israel, #ITA OR Italy, #LAT OR Latvia, #LTU OR Lithuania, #MNE OR Montenegro, #NOR OR Norway, #POL OR Poland, #ROM OR Romania, #RUS OR Russia, #SLO OR Slovenia, #SRB OR Serbia, #SWE OR Sweden | #ARM OR Armenia, #AUS OR Australia, #AUT OR Austria, #AZE OR Azerbaijan, #BEL OR Belgium, #BUL OR Bulgaria, #CRO OR Croatia, #CYP OR Cyprus, #CZE OR Czech Republic, #ESP OR Spain, #FRA OR France, #GBR OR United Kingdom, #GEO OR Georgia, #GER OR Germany, #HUN OR Hungary, #ISR OR Israel, #ITA OR Italy, #LAT OR Latvia, #LTU OR Lithuania, #MLT OR Malta, #NED OR Netherlands, #POL OR Poland, #RUS or Russia, #SRB OR Serbia, #SWE OR Sweden, #UKR OR Ukraine |

Germany) was included in a tweet, the corresponding country flag appeared in the shape of a heart.

For our empirical analysis, we used tweets from Twitter in connection to the ESC. For each year, we examined the four-day period from the morning of the first semi-final up to the main event, just before the release of the first results (see Table 2). We searched for tweets containing at least one of the following four keywords: *#esc, esc2015, eurovision, esc15.*[1] The hashtag *#esc* was presented by the organizers as an official hashtag on Twitter. A preliminary examination of the Twitter conversation resulted in the addition of the keywords *esc2015*, *Eurovision* and *esc15*, which were frequently used in the context of the event in both years. We decided not to consider tweets containing the names of the countries or artists without mentioning Eurovision to avoid tweets that do not refer to the competition.

We considered a tweet to be about a country if it contained either the official country hashtag, the "hashflag", or the country's name. In 2016, unlike 2015, Twitter did not officially support the hashflags. Some Twitter users nevertheless continued to use the same hashtags as the year before, while other switched to using the name of the country. Therefore, we filtered the 2016 data set by the same hashtags as 2015 where applicable. We did not search for countries that only participated in the semi-finals, because they do not appear in the audience voting in the final. Table 1 shows the hashtags and keywords searched.

The 2015 data set consisted of a total of 689,287 tweets by 198,372 users that included both a Eurovision-related key word and a country. The tweets contained 862,882 instances of a country being mentioned either in name or as a hashflag. The 2016 data consists of 960,870 tweets by 242,063 users and 1,089,718 country mentions. The number of tweets increased in comparison to 2015: While the number of tweets using a hashflag dropped slightly from 482,050 to 462,208, the number of tweets mentioning a country by name increased by more than twofold from 258,478 to 580,777. This change in usage habits may have taken place because the official support for hashflags was discontinued.

We split each dataset into two different periods to examine H3. As shown in Table 2, the first time period was chosen so as to cover the two semi-finals and the run-up to the final but not the actual event itself. The second time period was chosen to cover the artists' performances, but excludes the release of the results.

We did not consider country mentions by users known to be from the same country as the entry. Eurovision watchers cannot vote for their own country, but they might tweet about it. For example, the Swedish contestant might enjoy strong support from Sweden even if the artist's performance is mediocre. To remove the effects of such patriotic tweets, we considered the locations entered by users as part of their profile descriptions. If this location contained the name of the country being mentioned, this mention was excluded from the analysis (2015: 4.0% of cases, 2016: 3.8% of cases). As a result, 828,542 country mentions in 2015 were used in the analysis, and 1,048,137 in 2016.

We counted the number of remaining tweets $|T_c|$for each country $c$, and used the tool SentiStrength to determine the tweets' emotions. It has already proven useful in classifying emotions on platforms such as MySpace and Twitter (Thelwall et al. 2011; Mousavizadeh et al. 2015; Debortoli et al. 2016; Wu et al. 2016). The sentiment analysis algorithm labels every tweet with a positive and a negative score. The positive score *pos* and the negative score *neg* each vary in the integer range [1; 5] (from not positive to strongly positive, or from not negative to strongly negative). To analyse the sentiment, we determined the *sentiment polarity* (Pfitzner et al. 2012) of each tweet and calculated the mean sentiment polarity for each country to be able to compare them (cf. Table 3).

Additionally, we calculated the ratio of positive to negative tweets for each country (Asur and Huberman 2010). Each

[1] A search for a word such as "eurovision" also returns all tweets that use the word as a hashtag, i.e. "#eurovision".

**Table 2** Time periods (in Central European Summer Time) and descriptive statistics

| | Time period | No. of tweets | No. of users |
|---|---|---|---|
| 2015 | | | |
| Entire period | 19 May 9:00 AM – 23 May 11:40 PM | 689,287 | 198,372 |
| Pre-event period | 19 May 9:00 AM – 23 May 9:00 PM | 219,488 | 70,807 |
| Event period | 23 May 9:00 PM – 23 May 11:40 PM | 469,803 | 149,643 |
| 2016 | | | |
| Entire period | 10 May 9:00 AM – 14 May 11:40 PM | 960,870 | 242,063 |
| Pre-event period | 10 May 9:00 AM – 14 May 9:00 PM | 217,035 | 74,503 |
| Event period | 14 May 9:00 PM – 14 May 11:40 PM | 743,838 | 193,018 |

tweet was considered positive if its positive score is higher than its negative score and vice versa. A *ratio* of greater than 1 means that there were more positive tweets than negative ones, and a ratio between zero and one means that there were more negative tweets. This variable reflects a different facet than polarity. For example, the mean sentiment polarity of tweets surrounding a country could be positive but their ratio less than 1 if there is a small number of highly enthusiastic comments but the overall reception is somewhat unfavourable. The differences in measures of sentiment might also help explain the apparent inconsistency in results between many research studies which focused on the volume and valence of reviews and their effect on product sales: Liu (2006) found that volume increases sales more than valence, while Chintagunta et al. (2010) found that volume increases sales less than valence (Rosario et al. 2016). To examine this difference, we used both.

### 3.2 Use of the 2016 data

In addition to addressing our research questions, we use the 2016 data to assess the predictive accuracy of our model. This course of action allows us to infer whether our conclusions about the impact of the data collection period have implications for predictive modelling. The literature review showed several such previous attempts at predicting Eurovision. It is important to note that models with high explanatory power do not necessarily display high predictive accuracy on unseen data. In addition, predictive accuracy cannot be inferred from explanatory measures such as $R^2$. We therefore have to employ a separate set of evaluation measures and a separate set of data to measure predictive accuracy (Shmueli and Koppius 2011). Fortunately, it was possible for us to collect such a data set in 2016.

**Table 3** Overview of variables

| Name | Definition |
|---|---|
| Number of tweets | $n_c = \|T_c\|$ |
| Sentiment polarity | $polarity_c = \frac{1}{\|T_c\|} \sum_{t \in T_c} pos_t - neg_t$ |
| Sentiment ratio | $ratio_c = \frac{\{t \in T_c : pos_t > neg_t\}}{\{t \in T_c : neg_t > pos_t\}}$ |
| Audience rank | $rank_c \in \{1 \dots 27\}$ |

Another advantage of the 2016 data set is that it allows us to challenge the conclusions from the 2015 data by replicating the results. There have been prominent calls for showing the replicability of results in several disciplines, from medicine (Ioannidis 2014) to the social sciences (Schmidt 2009), after several previously accepted findings were discovered to be spurious (perhaps most famously, the suggested link between MMR vaccine and autism; see Taylor et al. 1999). The fraction of spurious findings in the research literature was estimated theoretically (Ioannidis 2005) and empirically to be more than half. In a particularly prominent attempt to replicate 100 psychological studies, 97 of the original studies had statistically significant results but only 36 of the replications did (Nosek et al. 2015). Replication has also been called for in the IS literature as a potential aid in ensuring that findings can be generalized (Cheng et al. 2016). In the context of social media, where platforms change considerably over time, it is especially important to show that results can be replicated across time periods (Ruths and Pfeffer 2014). We therefore also used the 2016 data set to conduct a replication study. The replication was conducted by following the same explanatory modelling and hypothesis testing procedure as 2015 on the independent sample and comparing the results.

### 3.3 Choice of Model and Evaluation Criteria

An ordinal logistic regression was conducted to explain the variation in the outcome variable *rank* from the predictor variables *number of tweets* and *mean tweet sentiment strength*. The model can be stated as (Harrell 2015):

$$\Pr[Y \geq j | X] = \frac{1}{1 + \exp\left[-\left(\alpha_j + X\beta\right)\right]}$$

where $Y$ is the outcome (*rank*), $X$ is the vector of predictors, and the intercepts $\alpha_j$ and coefficient vector $\beta$ are estimated by a maximum likelihood fitting procedure. Ordinal logistic regression (OLR) does not assume the dependent variable to be interval-scaled, and therefore does not claim a linear relationship between the number of tweets or sentiment and the rank like an ordinary least squares (OLS) approach would. OLR also does not make assumptions about the distribution of the dependent variable conditional on the values of the independent variables. Our setup for OLR is different from typical OLR setups in that the dependent variable takes a different value for each observation, instead of there being several observations per "class". However, another advantage of OLR is that it can handle this case (Harrell 2015). We used the statistical software package R (R Core Team 2016) and the R package rms (Harrell 2017) to perform the required calculations.

Prior to carrying out the regression analysis, the number of tweets was log-transformed and both predictors were standardized by subtracting the sample mean and dividing by the sample standard deviation. We log-transformed the number of tweets because we hypothesize a percent increase in the odds of obtaining a better rank to be associated with a percent increase in the number of tweets, instead of an absolute increase in the number of tweets. The standardization allows us to ignore changes in the overall number of tweets or mean sentiment over the years. For the purpose of explaining variation in the ranking, only the relative differences between the numbers and sentiments of tweets about participants in any given year matter.

Evaluation measures for ordinal data differ from the ones typically used for nominal data (such as accuracy, precision and recall) or interval-scaled data (such as MAPE and PRESS). As an explanatory measure of model fit on the original data, we use Nagelkerke's pseudo $R^2$ and Wald's Z test for the individual coefficients as well as the likelihood ratio model fit test for the entire model. As a measure of predictive accuracy on unseen data, compare the predicted means from the ordinal regression model to the true ranks using Spearman's ρ and Kendall's τ, and the mean absolute difference (MAD). The former two are widely used measures of association for ordinal data (Harrell 2015). The latter has the advantage of being easy to interpret.

In summary, we first carried out the described explanatory modelling procedure using ordinal logistic regression based on the 2015 data set. We then used 2016 data for two distinct purposes, the evaluation of the predictive accuracy of the 2015 model and the replication of the results of the hypothesis tests, and we used different evaluation criteria for each goal. Table 4 summarizes this research design.

## 4 Findings

### 4.1 Voting Results

Table 5 shows the results of the voting in the 2015 ESC. In 2015, each country awarded 12, 10 and 8 to 1 points to other countries. Entries ranked outside a country's top ten choices did not receive points. As always, countries were not allowed to vote for their own entry, and the points awarded by a country were calculated from two separate votes, the jury vote and the televoting. The latter was carried out via the official app, text messages and phone calls. Both were weighted equally to calculate the points awarded, but the results of the two votes were also published separately.[2] As mentioned above, we use placements derived from the audience voting alone as the dependent variable because we analyse the communication of that very audience on Twitter. As shown in Table 5, the televoting results differ slightly from the total ranking. For example, Sweden won according to the aggregated result, but without the jury votes, Italy would have won.

### 4.2 Descriptive Statistics

To provide an overview of the measured variables, Table 6 shows summary statistics. The number of tweets is higher for countries with a good ranking (Spearman's ρ = −0.64; the correlation is negative because a better rank is represented by a lower number). The countries with the highest number of English-language tweets are Australia, Sweden and Russia. Romania, Cyprus and Montenegro account for the lowest volume.

For sentiment polarity, the range of possible values is −5 to +5, where 0 indicates balanced sentiment. Mean sentiment was positive for all countries. The rank correlation shows that countries with a better ranking clearly tended to have more positive tweets (Spearman's ρ = −0.43 for polarity). Latvia, Lithuania and Belgium have the most positive mean sentiment; Australia, France and Hungary have the lowest mean.

### 4.3 The Choice of Time Period, Part I

We fit the ordinal logistic regression model to the entire data set to examine H1 and H2 regarding the influence of the number of tweets and sentiment on the audience rank. Since two different variables for measuring sentiment have been discussed in the literature, we fit two separate models, one using sentiment polarity, and one using the positive-to-negative ratio. The model with sentiment polarity resulted in a much better fit (cf. Table 7). To investigate the impact of the choice of time period on the results (H3), we fit two more models, one to the data collected before the event, up to

[2] https://www.eurovision.tv/page/results Accessed 29 November 2017.

**Table 4** Summary of research design

| Research goal | Evaluation measures | Data set | Purpose | Relevant sections |
|---|---|---|---|---|
| Explanatory (Test hypotheses) | Pseudo $R^2$, *p* values from Z test and model fit likelihood ratio test | 2015 | Hypothesis testing | 4.3 |
| | | 2016 | Replication | 4.5 |
| Predictive (Assess predictive accuracy) | Spearman's ρ, Kendall's τ, Mean Absolute Difference (MAD) | 2015 | Training | 4.4 |
| | | 2016 | Testing | 4.4 |

9 pm the night of the event, and another to from the data collected during the event. The results are reported in Table 7.

The difference in pseudo $R^2$ between the models is substantial. For all models, the likelihood ratio test rejects the null hypothesis that the model fit is no better than chance ($p < 0.0001$). However, the model that uses the data from the pre-event period is much better than the model from during the event. An inspection of the variables reveals that the association of the variable *number of tweets* and the final ranking is much weaker in the period of the event (Spearman's $\rho = -.40$) than in the whole data set. In contrast, the association is slightly higher for the pre-event period ($\rho = -.86$).

Although the models are not reported here in full for the sake of brevity, this finding is the same when sentiment ratio is used in the model instead of polarity (pseudo $R^2 = .715$ for pre-event tweets, and pseudo $R^2 = .193$ for tweets from during the event). However, since sentiment polarity leads to a better fit, we use this variable throughout the rest of the analysis.

## 4.4 Evaluation of Predictive Accuracy: Predicting the 2016 Ranking

The above measures of model fit can only be calculated in retrospection, once the results are known, because the parameters were chosen to achieve the best fit on the available data. We next use the model that was fit to the 2015 data to predict the results of Eurovision 2016, in order to evaluate its predictive accuracy.

The 2016 event saw a slight change in the voting system. Both the juries and the audience from each country now each awarded 12, 10 and 8 to 1 points to other countries. Again, voters could not vote for their own country. The ranking derived from the audience voting alone was different from the final ranking, and we attempted to predict the audience ranking. Table 8 shows both rankings. Table 9 shows summary statistics for 2016.

We calculated the predictions for 2016 from tweets before the event and using the 2015 model parameters. The mean absolute deviation between the raw predictions and the actual ranks is 4.88. We compared our model against the random guessing baseline using a number of common statistics for the comparison of ordinal data as well as MAD. For each of the statistics, we calculated the median and upper and lower confidence interval bounds using a Monte Carlo approximation (1.000.000 iterations). Table 10 shows that the model compares favourably to the baseline according to all statistics. The probability of obtaining a MAD as good as the one from our model through random guessing is less than 0.3%. We conclude that our predictions are clearly better than random.

## 4.5 The Choice of Time Period, Part II: Replicating the 2015 results

We constructed the same linear regression model as previously, this time using the 2016 data set. Table 11 summarizes the results.

**Table 5** Final Ranking Eurovision 2015

| Country | Total | Audience | Country | Total | Audience | Country | Total | Audience |
|---|---|---|---|---|---|---|---|---|
| Sweden | 1 | 3 | Serbia | 10 | 10 | Romania | 15 | 12 |
| Russia | 2 | 2 | Georgia | 11 | 13 | Armenia | 16 | 11 |
| Italy | 3 | 1 | Azerbaijan | 12 | 14 | Albania | 17 | 9 |
| Belgium | 4 | 4 | Montenegro | 13 | 18 | Lithuania | 18 | 16 |
| Australia | 5 | 6 | Slovenia | 14 | 19 | Greece | 19 | 21 |
| Latvia | 6 | 8 | Romania | 15 | 12 | Hungary | 20 | 22 |
| Estonia | 7 | 5 | Armenia | 16 | 11 | Spain | 21 | 20 |
| Norway | 8 | 17 | Albania | 17 | 9 | Cyprus | 22 | 23 |
| Israel | 9 | 7 | Lithuania | 18 | 16 | Poland | 23 | 15 |

**Table 6** Descriptive statistics and correlations (Spearman's ρ) for the 2015 data

| | Before event | | During event | | Entire time frame | |
|---|---|---|---|---|---|---|
| Descriptive statistics | | | | | | |
| Variable | Mean | Std. dev. | Mean | Std. dev. | Mean | Std. dev. |
| No. of tweets | 11790.7 | 4769.0 | 18896.2 | 10288.0 | 30686.7 | 13703.5 |
| Sentiment polarity | 0.780 | 0.143 | 0.488 | 0.200 | 0.610 | 0.150 |
| Sentiment ratio | 8.913 | 3.100 | 3.896 | 1.750 | 5.005 | 1.769 |
| Correlations | | | | | | |
| | Sentiment polarity | Sentiment ratio | Sentiment polarity | Sentiment ratio | Sentiment polarity | Sentiment ratio |
| No. of tweets | .278 | .118 | .199 | .404 | .211 | .312 |
| Sentiment polarity | | .659 | | .896 | | .907 |

The overall fit, as measured by the pseudo $R^2$, is not as good as 2015. A closer look at the individual time periods (cf. Table 11) reveals that again, the fit is much better when only tweets from before the event are included in the analysis. The standardized regression coefficients are especially well-suited for comparing the two years. For the variable *number of tweets*, they are fairly similar to the ones observed in 2015. For the entire period and the pre-event period, they are within one standard deviation. However, for the variable sentiment polarity, they are close to zero. In stark contrast to 2015, the sentiment of the tweets about a country and its final ranking were seemingly unrelated (Spearman's ρ = −0.10).

# 5 Discussion

## 5.1 Results

After removing patriotic mentions of one's own country, social media data reveals the relative strengths of contestants. The ordinal logistic regression, using the number of tweets and sentiment from the entire period in 2015, is a reasonably good fit. The model likelihood ratio test is significant ($p < .0001$). In addition, the standardized regression coefficient indicates a strong association between the number of tweets predictor and the outcome. These results support hypothesis H1, a finding consistent with previous research (Williams and Gulati 2008; Tumasjan et al. 2010; Ciulla et al. 2012; DiGrazia et al. 2013).

As for Hypothesis H2, the variable *sentiment polarity* is significant in the model ($p = .0234$). The coefficient indicates a strong association. The hypothesis can be considered tentatively supported. Again, this result is in line with similar findings reported in the past (Asur and Huberman 2010; Gayo-Avello et al. 2011; Mehndiratta et al. 2014).

With respect to the time period, the data gathered in the pre-event period proved to be of much greater use (pseudo $R^2 = .740$) than the tweets that were posted during the event (pseudo $R^2 = .209$). Hence, the time period for data collection had a considerable influence on the results. The tweets posted before Eurovision are valuable indicators of artist popularity among those loyal fans who will later, during the event, spend money to vote, and these fans do not change their opinion overnight. For entertainers and executives, this means that

**Table 7** Eurovision 2015: Ordinal logistic regression model summaries for the different time periods (dependent variable: audience rank)

| | β | SE β | Wald Z | p |
|---|---|---|---|---|
| Entire period ($R^2$ (Nagelkerke) = .431; LR $\chi^2 = 15.19$, $p = 0.0005$) | | | | |
| Log(no. of tweets) | −1.3733 | 0.4380 | −3.14 | .0017 |
| Sentiment ratio | −0.4086 | 0.3645 | −1.12 | .2624 |
| Entire period ($R^2$ (Nagelkerke) = .511; LR $\chi^2 = 19.29$, $p < 0.0001$) | | | | |
| Log(no. of tweets) | −1.5130 | 0.4292 | −3.53 | .0004 |
| Sentiment polarity | −0.8122 | 0.3576 | −2.27 | .0231 |
| Pre-event period ($R^2$ (Nagelkerke) = .740; LR $\chi^2 = 36.23$, p < 0.0001) | | | | |
| Log(no. of tweets) | −3.2220 | 0.7420 | −4.34 | < .0001 |
| Sentiment polarity | −0.6254 | 0.4069 | −1.54 | .1243 |
| Event period ($R^2$ (Nagelkerke) = .209; LR $\chi^2 = 6.31$, $p = 0.0425$) | | | | |
| Log(no. of tweets) | −0.6945 | 0.3645 | −1.19 | .0567 |
| Sentiment polarity | −0.3765 | 0.3351 | −1.12 | .2612 |

**Table 8** Final ranking of Eurovision 2016

| Country | Total | Audience | Country | Total | Audience | Country | Total | Audience |
|---|---|---|---|---|---|---|---|---|
| Ukraine | 1 | 2 | Belgium | 10 | 16 | Israel | 14 | 22 |
| Australia | 2 | 4 | Netherlands | 11 | 17 | Latvia | 15 | 13 |
| Russia | 3 | 1 | Malta | 12 | 21 | Italy | 16 | 18 |
| Bulgaria | 4 | 5 | Austria | 13 | 8 | Azerbaijan | 17 | 12 |
| Sweden | 5 | 6 | Israel | 14 | 22 | Serbia | 18 | 11 |
| France | 6 | 9 | Latvia | 15 | 13 | Hungary | 19 | 14 |
| Armenia | 7 | 7 | Italy | 16 | 18 | Georgia | 20 | 20 |
| Poland | 8 | 3 | Azerbaijan | 17 | 12 | Cyprus | 21 | 15 |
| Lithuania | 9 | 10 | Serbia | 18 | 11 | Spain | 22 | 23 |

the performances at the main event do not seem to matter much when it comes to convincing the audience to cast a vote, and therefore the performances at smaller events leading up to it should be prioritized, as should building a positive image early on. For researchers, it means that the choice of time period is crucial and should be justified well instead of being made arbitrarily. Alternatively, studies should be conducted repeatedly using different data sets to demonstrate that the results do not depend on the time of data collection. Especially the higher usefulness of pre-event data should be tested in other contexts. For example, in the context of an election, our results would suggest that the opinions on social media expressed by those who already know which candidates they like days before the event might be more indicative of the final outcome than tweets from during the election.

We considered both sentiment polarity (Pfitzner et al. 2012) and sentiment ratio (Asur and Huberman 2010) to quantify the emotionality in tweets. Polarity resulted in a better fit. This result confirms the suspicion that differences in sentiment measures might help explain inconsistent previous results (Rosario et al. 2016). It also implies that researchers should pay close attention to the choice of sentiment measure, as it may considerably affect model fit. However, in our model, the main finding regarding the time periods was observed with both sentiment measures. These results strengthen our conclusions from the time periods.

## 5.2 Evaluation and Replication

Our model is useful for predicting the results of future installments of the ESC. Importantly, the calculation of the prediction only required the 2015 data (tweets and results), and tweets collected prior to the event in 2016. It is therefore possible, with this method, to calculate and publish the prediction before the event begins. This true prediction of the future is in contrast to "predictions" whose calculation requires knowledge of the final result or data collected after the event, e.g. when researchers only report $R^2$ or other measures of fit (Shmueli and Koppius 2011) to evaluate a model, or calculate the correlation between a film's opening weekend revenue and the sentiment of blog posts written after the opening weekend (Mishne and Glance 2006).

One of the roles of predictive analytics in research is to assess the relevance of scientific models and theories in

**Table 9** Descriptive statistics and correlations (Spearman's ρ) for the 2016 data

| | Before event | | During event | | Entire time frame | |
|---|---|---|---|---|---|---|
| Descriptive statistics | | | | | | |
| Variable | Mean | Std. dev. | Mean | Std. dev. | Mean | Std. dev. |
| No. of tweets | 9618.9 | 5836.11 | 30,694.3 | 14,984.5 | 40,313.0 | 19,067.1 |
| Sentiment polarity | 0.579 | 0.176 | 0.356 | 0.194 | 0.415 | 0.165 |
| Sentiment ratio | 4.802 | 1.964 | 2.951 | 1.398 | 3.216 | 1.273 |
| Correlations | | | | | | |
| | Sentiment polarity | Sentiment ratio | Sentiment polarity | Sentiment ratio | Sentiment polarity | Sentiment ratio |
| No. of tweets | .102 | −.001 | −.099 | −.051 | −.101 | .028 |
| Sentiment polarity | | .888 | | .950 | | .936 |

**Table 10** Evaluation of predictive accuracy (using 2015 data for training, and 2016 data for testing)

| | MAD | Spearman's ρ | Kendall's τ |
|---|---|---|---|
| Random (95% upper CI) | 10.77 | 0.39 | 0.27 |
| Random (median) | 8.69 | 0.00 | 0.00 |
| Random (95% lower CI) | 6.46 | −0.39 | −0.27 |
| Tweets + Sentiment | 4.88 | 0.60 | 0.43 |

practice (Shmueli and Koppius 2011). As demonstrated in the related work section, prior research has shown for a large variety of events that the number and sentiment of tweets are correlated with the outcomes. Our research confirms that this relationship is useful in practice and can be exploited to build a workable predictive model. We focused on Eurovision, which is convenient because the outcomes are published so others can reproduce our research more easily. Yet, given the large number of publications that have shown this relationship for a wide variety of events, there is good reason to believe that our method can be applied to other data with more immediate implications for business, such as product sales.

Our results also demonstrate the importance of replication studies. The 2016 replication confirmed some, but not all of the hypotheses that were confirmed by the 2015 data (see Table 12). More precisely, the 2016 data provide additional evidence for hypotheses H1 ($p = .0005$) and H3 but do not support H2 ($p = .8539$). These results further demonstrate the relationship between the number of tweets and voting results, as well as the influence of the time period considered in the course of data collection. However, they call the usefulness of the predictor sentiment polarity into question. In that sense, they are inconsistent with previous research that found sentiment a valuable predictor.

**Table 11** Eurovision 2016: Linear regression model summaries for the different time periods (dependent variable: audience rank)

| | β | SE β | Wald Z | p |
|---|---|---|---|---|
| Entire period ($R^2$ (Nagelkerke) = .425; LR $\chi^2 = 14.36$, $p = 0.0008$) | | | | |
| Log(no. of tweets) | −1.6861 | 0.4854 | −3.47 | .0005 |
| Sentiment polarity | −0.0734 | 0.3986 | −0.18 | .8539 |
| Pre-event period ($R^2$ (Nagelkerke) = .483; LR $\chi^2 = 17.12$, $p = 0.0002$) | | | | |
| Log(no. of tweets) | −1.7821 | 0.4673 | −3.81 | .0001 |
| Sentiment polarity | 0.4488 | 0.4419 | 1.02 | .3098 |
| Event period ($R^2$ (Nagelkerke) = .267; LR $\chi^2 = 8.07$, $p = 0.0177$) | | | | |
| Log(no. of tweets) | −1.1708 | 0.4147 | −2.82 | .0048 |
| Sentiment polarity | 0.0735 | 0.4001 | 0.18 | .8542 |

If the association between social media sentiment and votes cannot be established with certainty, this has important implications for practice. It may mean that current methods of measuring tweet sentiment do not capture user opinion accurately enough. Even the most up-to-date machine learning methods for three-way (positive/neutral/negative) sentiment classification, most of which make use of deep neural networks and word embeddings, only achieve an accuracy of 60–70% (Nakov et al. 2016).

## 6 Limitations

An important limitation of our data collection approach is due to the selection of hashtags and keywords. We decided to track only the official hashtags for each country and mentions of the country's name but might have missed relevant content as a result. Of course, this limitation is common to all analyses of subsets of Twitter data.

Secondly, such a simple model with only two predictors is, of course, not an adequate causal model for the outcome of a complex real-world event such as Eurovision, especially one with such an intricate voting procedure. In the absence of information on the actual number of votes, we had to rely on the rank as a crude approximation. However, we were only interested in examining the relationship between outcome and variables based on social media data, and what matters in the context of this study is that the association with the number of tweets (H1) was observed consistently, in both years, and that it was so strong, while another (H2) could not be replicated.

In addition, opinions expressed by Twitter users do not necessarily reflect the thoughts of the actual voters since the two groups are usually not identical. However, as long as their opinions correlate, one can be used to infer the other to some extent. The empirical results of our analyses indicate that there is indeed an association strong enough to be useful for making predictions in this context.

The goal of our research was not to maximize predictive power. Instead we assessed the accuracy of an explanatory model. We have shown that incorporating social media data is likely to improve the performance of an existing model that uses information from outside social media, and the time of data collection plays a crucial role.

## 7 Conclusion

Media events like the ESC generate a great deal of buzz on social media and Twitter in particular. The competition takes place each year in a similar manner, making it comparable from one year to another. Furthermore, since 2015, the

**Table 12** Overview of Hypotheses and results

| Hypothesis | Results 2015 | Results 2016 | Conclusion |
|---|---|---|---|
| H1: There is a consistent, replicable positive relationship between the number of artist-related tweets and a better artist ranking in the audience voting. | Supported | Supported | Supported |
| H2: There is a consistent, replicable positive relationship between the sentiment of artist-related tweets and a better artist ranking in the audience voting. | Supported | Not supported | Not supported |
| H3: The explanatory power of artist-related tweets from prior to the event is higher than for those from during the event. This relationship is also valid across more than one year. | Supported | Supported | Supported |

connection between Twitter and Eurovision has been stronger than ever due to the introduction of "hashflags" as well as an increasing Twitter usage in general (e.g. via mobile devices). Because of these aspects, we consider the ESC a unique opportunity to gain a better understanding of the explanatory and predictive power of social media data in the context of media events. In particular, we analysed tweets regarding their volume and expressed sentiment to examine their relationship with the results of the televoting, and to forecast the 2016 ranking using a predictive model trained on 2015 data. The volume of tweets related to an artist alone is a significant predictor of the artist's ranking, a hypothesis supported by both the 2015 and the 2016 data. Our second hypothesis, however, regarding the sentiment expressed in those tweets, was not consistently supported. The third hypothesis combines the two variables, positing that a model using both of them is more accurate than if either of them is considered independently. In addition to this, we examined whether the timeframe in which those tweets were posted is of any significant influence to the predictive power of such a model (as proposed in H3). Our hypothesis that the data gathered before an event provides better results than the data gathered during an event in such an analysis was repeatedly confirmed.

One fundamental contribution of our research that goes beyond existing studies is that it has made apparent how dependent the results of social media-based analyses are on the chosen time frame. This is true on a small scale, when one decides when to start and when to finish data collection for a particular event. It is also true on the larger scale, since the results obtained in 2015 and 2016 differed considerably. If results fail to generalize between two instalments of the same competition, there is obviously cause for concern.

Our approach nevertheless yielded a prediction model whose accuracy was apparently unaffected by this result. For practical applications, the issue may not be as serious as one might think. Social media has been established as a useful source of information for forecasting. Still, more research is needed on the circumstances that determine the usefulness of individual predictors, and to examine which results generalize. Only if we compare different data sets from various time periods and possibly different social media, we will be able to identify the patterns that are spurious or short-lived, and the ones that hold up.

## References

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science, 57*(8), 1485–1509.

Asur, S., & Huberman, B.A. (2010). Predicting the future with social media. In *Proceedings of the Web Intelligence and Intelligent Agent Technology 2010 IEEE/WIC/ACM International Conference*, 492–499.

Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., & Tesconi, M. (2018). CrisMap: a Big Data Crisis Mapping System Based on Damage Detection and Geoparsing. *Information Systems Frontiers*, 1–19.

Baumeister, R. F., & Tice, D. M. (1984). Role of self-presentation and choice in cognitive dissonance under forced compliance: Necessary or sufficient causes? *Journal of Personality and Social Psychology, 46*, 5–13.

Benthaus, J., & Skodda, C. (2015). Investigating Consumer Information Search Behavior and Consumer Emotions to Improve Sales Forecasting. In *Proceedings of the 21st Americas Conference on Information Systems,* Puerto Rico.

Berger, J. (2011). Arousal Increases Social Transmission of Information. *Psychological Science, 22*, 891–893.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*, 192–205.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*, 1–8.

Bruns, A., & Burgess, J.E. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference* 2011.

Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams. (2016). M.140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies, 41*, 230–233.

Calderon, N.A., Arias-Hernandez, R., & Fisher, B. (2014). Studying Animation for Real-Time Visual Analytics: A Design Study of Social Media Analytics in Emergency Management. In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS) 2014* (pp. 1364-73). IEEE.

Ceron, A., Curini, L., Iacus, S., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*, 340–358.

Charles, C. A. D., & Reid, G. (2016). Forecasting the 2016 General Election in Jamaica. *Commonwealth & Comparative Politics, 54*(4), 449–477.

Chen, J., Chen, H., Wu, Z., Hu, D., & Pan, J. Z. (2016). Forecasting smog-related health hazard based on social media and physical sensor. *Information Systems, 64*, 281–291.

Cheng, Z., Dimoka, A., & Pavlou, P. (2016). Context may be King, but generalizability is the Emperor! *Journal of Information Technology, 31*, 257–264.

Cheong, M., & Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers, 13*(1), 45–59.

Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., & Trautmann, H. (2015). An Overview of Topic Discovery in Twitter Communication through Social Media Analytics. In *Proceedings of the 21st Americas Conference on Information Systems,* Puerto Rico.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. *Marketing Science, 29*, 944–957.

Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perraand, N., & Vespignani, A. (2012). Beating the news using social media: the case study of American Idol. *EPJ Data Science, 1*(8), 1–11.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, 39.

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS One, 8*.

Dijkmans, C., Kerkhof, P., & Beukeboom, C. J. (2015). A stage to engage: Social media use and corporate reputation. *Tourism Management, 47*, 58–67.

Eriksson, M., & Olsson, E. K. (2016). Facebook and Twitter in Crisis Communication: A Comparative Study of Crisis Communication Professionals and Citizens. *Journal of Contingencies and Crisis Management, 24*(4), 198–208.

Festinger, L., & Carlsmith, M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*, 203–210.

Fox, E. (2008). *Emotion Science: Cognitive and Neuroscientific Approaches to Understanding Human Emotions*. Basingstoke: Palgrave Macmillan.

Gayo-Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM-2011)* (pp. 490–493). Menlo Park: The AAAI Press.

Georgiou, M. (2008). In the end, Germany will always resort to hot pants: watching Europe singing, constructing the stereotype. *Popular Communication, 6*, 141–154.

Golbeck, J., Gerhard, J., O'Colman, F., & O'Colman, R. (2017). Scaling Up Integrated Structural and Content-Based Network Analysis. *Information Systems Frontiers*, 1–12.

Greene, D., & Cunningham, P. (2013). Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th Annual ACM Web Science Conference,* 118–121.

Gruber, D. A., Smerek, R. E., Thomas-Hunt, M. C., & James, E. H. (2015). The real-time power of Twitter: Crisis management and leadership in an age of social media. *Business Horizons, 58*, 163–172.

Harrell, F. E. (2015). *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer.

Harrell, F. E. (2017). rms: Regression Modeling Strategies. https://CRAN.R-project.org/package=rms Accessed 29 November 2017.

He, W., Guo, L., Shen, J., & Akula, V. (2016). Social Media-Based Forecasting: A Case Study of Tweets and Stock Prices in the Financial Services Industry. *Journal of Organizational and End User Computing, 28*, 74–91.

Highfield, T., Harrington, S., & Bruns, A. (2013). Twitter as a technology for audiencing and fandom: The #Eurovision phenomenon. *Information. Communications Society, 16*, 315–339.

Huberty, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. *International Journal of Forecasting, 31*, 992–1007.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys, 47*.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine, 2*(8).

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS Medicine, 11*(10).

Jansen, H. J., & Koop, R. (2006). Pundits, ideologues, and the ranters: The British Columbia election online. *Canadian Journal of Communication, 30*(4), 613–632.

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review*, 30, 229–234.

Kaschesky, M., Sobkowicz, P., Hernandez Lobato, J. M., Bouchard, G., Archambeau, C., Scharioth, N., et al. (2013). Bringing Representativeness into Social Media Monitoring and Analysis. In *Proceedings of the 2013 46th Hawaii International Conference on System Sciences (HICSS).* IEEE.

Kaya, M., & Conley, S. (2016). Comparison of sentiment lexicon development techniques for event prediction. *Social Network Analysis and Mining, 6*, 1–13.

Kim, J.W., Kim, D., Keegan, B., Kim, J.H., Kim, S., & Oh, A. (2015). *Social Media Dynamics of Global Co-presence During the 2014 FIFA World Cup*. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Kinsinger, E. A., & Schacter, D. L. (2008). Memory and emotion. In M. Lewis, J. A. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of Emotions* (pp. 601–617). New York: The Guildford Press.

Larosiliere, G., Carter, L., & Meske, C. (2017). How does the world connect? Exploring the global diffusion of social network sites. *Journal of the Association for Information Science and Technology, 68*(8), 1875–1885.

Li, J., & Cardie, C. (2013). Early Stage Influenza Detection from Twitter. arXiv:1309.7340 [cs.SI]. https://arxiv.org/abs/1309.7340.

Li, X., & Hitt, L. M. (2008). Self-Selection and Information Role of Online Product Reviews. *Information Systems Research, 19*(4), 456–474.

Li, E. Y., Tung, C., & Chang, S. (2016). The wisdom of crowds in action: Forecasting epidemic diseases with a web-based prediction market system. *International Journal of Medical Informatics, 92*, 35–43.

Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact of Box Office Revenue. *Journal of Marketing, 70*(July), 74–89.

Maldonado, M., & Sierra, V. (2015). Can social media predict voter intention in elections? The case of the 2012 Dominican Republic Presidential Election. In *Proceedings of the 21st Americas Conference on Information Systems,* Puerto Rico.

Marwick, A. E., & boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*, 114–133.

Mehndiratta, P., Sachdeva, S., Sachdeva, P., & Sehgal, Y. (2014). Elections Again, Twitter May Help!!! A Large Scale Study for

Predicting Election Results Using Twitter. In Srinivasa S., Mehta S. (eds) *Big Data Analytics. BDA 2014. Lecture Notes in Computer Science, vol 8883* (pp. 133-144). Cham: Springer.

Mishne, G., & Glance, N.S. (2006). Predicting Movie Sales from Blogger Sentiment. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs,* 155-158.

Mousavizadeh, M., Koohikamali, M., & Salehan, M. (2015). The Effect of Central and Peripheral Cues on Online Review Helpfulness: A Comparison between Functional and Expressive Products. In *Proceedings of the 36th International Conference on Information Systems,* Fort Worth.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). Semeval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation,* San Diego, US.

Nann, S., Krauss, J., & Schroder, D. (2013). Predictive Analytics On Public Data-The Case Of Stock Markets. In *Proceedings of the 21st European Conference on Information Systems, Utrecht, Netherlands* (Vol. 102, pp. 1–12).

Nguyen, T., Shiraia, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*, 9603–9611.

Nofer, M., & Hinz, O. (2015). Using Twitter to predict the Stock Market-Where is the mood effect? *Business & Information Systems Engineering, 57*, 229–242.

Nosek, B. A., et al. (2015). Estimating the reproducibility of psychological science. *Science, 349*. https://doi.org/10.1126/science.aac4716.

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the European Parliament. *Electoral Studies, 44*, 429–444.

O'Connor, B., Balasubramanyan, R., & Routledge, B. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-2010)* (pp. 122–129). Menlo Park: The AAAI Press.

Oh, O., Agrawal, M., & Rao, R. (2013). Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. *MIS Quarterly, 37*(2), 407–426.

Pavlyshenko, B. (2013). Forecasting of events by Tweet data Mining. arXiv:1310.3499 [cs.SI]. https://arxiv.org/abs/1310.3499.

Pfitzner, R., Garas, A., & Schweitzer, F. (2012). Emotional Divergence Influences Information Spreading in Twitter. *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM-2012)* (pp. 543-546). Menlo Park, CA: The AAAI Press.

Pond, P. (2016). The space between us: Twitter and crisis communication. *International Journal of Disaster Resilience in the Built Environment, 7*(1), 40–48.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ Accessed 29 November 2017.

Rosario, A. B., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research, 53*, 297–318.

Rudra, K., Sharma, A., Ganguly, N., & Imran, M. (2018). Classifying and Summarizing Information from Microblogs During Epidemics. *Information Systems Frontiers*, 1–16.

Ruths, D., & Pfeffer, J. (2014). Social media for Large Studies of Behavior. *Science, 346*, 1063–1064.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, 851–860.

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, 53–60.

Schlenker, B. R., & Goldman, H. J. (1982). Attitude change as a self-presentation tactic following attitude consistent behavior: Effects of role and choice. *Social Psychology Quarterly, 45*(2), 92–99.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research, 23*, 528–543.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly, 35*, 553–572.

Spence, P. R., Sellnow-Richmond, D. D., Sellnow, T. L., & Lachlan, K. A. (2016). Social media and corporate reputation during crises: the viability of video-sharing websites for providing counter-messages to traditional broadcast news. *Journal of Applied Communication Research, 44*(3), 199–215.

Statista (2017). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017 (in millions). https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users Accessed 29 November 2017.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems, 29*, 217–248.

Stieglitz, S., Bruns, A., & Krüger, N. (2014). Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems. *Business & Information Systems Engineering, 6*(2), 89–96.

Stieglitz, S., Bunker, D., Mirbabaie, M., & Ehnis, C. (2017). Sense-Making in Social Media During Extreme Events. *Journal of Contingencies and Crisis Management, 26*(1), 4–15.

Stieglitz, S., Mirbabaie, M., & Milde, M. (2018a). Social Positions and Collective Sense-making in Crisis Communication. *International Journal of Human-Computer Interaction, 34*(4), 328–355.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018b). Social Media Analytics – Challenges in Topic Discovery, Data Collection, and Data Preparation. *International Journal of Information Management, 39*, 156–168.

Storvik-Green, S. (2015a). #Eurovision Twitter hashflags go live! eurovision.tv. https://eurovision.tv/story/eurovision-twitter-hashflags-golive. Accessed 04 Jun 2018.

Storvik-Green, S. (2015b). Nearly 200 million people watch Eurovision 2015. eurovision.tv. https://eurovision.tv/story/nearly-200-million-people-watcheurovision-2015. Accessed 04 Jun 2018.

Taylor, B., Miller, E., Farrington, C. P., Petropoulos, M., Favot-Mayaud, I., Li, J., et al. (1999). Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *The Lancet, 353*, 2026–2029.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter Events. *Journal of the American Society for Information Science and Technology, 62*, 406–418.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-2010)* (pp. 178–185). Menlo Park: The AAAI Press.

Vaccari, C., Chadwick, A., & O'Loughlin, B. (2015). Dual Screening the Political: Media Events, Social Media, and Citizen Engagement. *Journal of Communication, 65*, 1041–1061.

Valenzuela, S., & Bachmann, I. (2015). Pride, Anger, and Cross-cutting Talk: A Three-Country Study of Emotions and Disagreement in Informal Political Discussions. *International Journal of Public Opinion Research, 27*(4), 544–564.

Wahyudi, A., Kuk, G., & Janssen, M. (2018). A Process Pattern Model for Tackling and Improving Big Data Quality. *Information Systems Frontiers*, 1–13.

Williams, C., & Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. In *Annual Meeting of the American Political Science Association,* Boston, MA.

Wu, J., Srite, M., & Deng, S. (2016). Tweet, Favorite, and Envy. In *Proceedings of the 22nd Americas Conference on Information Systems,* San Diego.

Yu, S., & Kak, S. (2012). A survey of prediction using social media. arXiv:1203.1647 [cs.SI]. https://arxiv.org/abs/1203.1647.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia–Social and Behavioral Sciences, 26*, 55–62.

**Stefan Stieglitz** is Professor of Professional Communication in Electronic Media / Social Media at University of Duisburg-Essen, Germany. Previously, he held a position as an assistant professor at the University of Münster. In 2015 and 2016 Stefan was a Visiting Professor in the Discipline of Business Information Systems at the University of Sydney Business School. Stefan is a principal investigator of the research training group "User-centred Social Media" and director and founder of the Competence Center Connected Organization. He received his PhD from University of Potsdam. In his research he investigates digital communication and collaboration with an emphasis on social media analytics. Moreover, he analyses user behavior and technology adaption of collaborative information systems in organizational contexts. His work has been published in reputable journals including Journal of Management Information System, Business & Information Systems Engineering, International Journal of Information Management, and Management Information Systems Quarterly Executive. His work has been recognized with one of the AIS Best Information Systems Publications Award in 2016.

**Christian Meske** is an Assistant Professor of Digital Transformation and Strategic Information Management at the Department of Information Systems, Freie Universität Berlin. He is also a funded member of the Einstein Center Digital Future (Berlin). Christian received his PhD from the Department of Information Systems at the University of Münster and studied Business Economics at the University of Potsdam. He was a Visiting Scholar at the University of Sydney Business School as well as Florida State University. Christian has published in reputable journals and conferences such as Business & Information Systems Engineering (BISE), Business Process Management Journal (BPMJ), Journal of the Association for Information Science and Technology and International Conference on Information Systems (ICIS). His work on the adoption and usage of communication and collaboration technologies has been recognized with one of the AIS Best Information Systems Publications Award in 2016.

**Björn Ross** is a research associate in the Research Group Professional Communication in Electronic Media/Social Media at the Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen (Germany). He is pursuing a PhD there with a focus on the diffusion of information on public social networking sites, and he also explores how information from social media can help predict events outside social media. He has published in the International Journal of Information Management (IJIM) and presented his work at conferences including the European Conference on Information Systems (ECIS) and the Hawaii International Conference on System Sciences (HICSS), where it was recognized with a Best Paper Award. His research is funded by the German Research Foundation (DFG) as part of the Research Training Group User-Centred Social Media.

**Milad Mirbabaie** is a research associate at the Department of Computer Science and Applied Cognitive Science and member of the Research Training Group "User-Centred Social Media" at the Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen (Germany). He studied Information Systems at the University of Hamburg and is currently a PhD student at the University of Münster (Germany), Department of Information Systems. He was a Visiting Scholar at the University of Sydney Business School in the years 2014, 2016, and 2017. His work on the use of social media for crisis management, on social media analytics and on sense-making in social media communication has been awarded with the Claudio Ciborra Award for the most innovative research article at the European Conference on Information Systems in 2017.