frontiers in Big Data

MINI REVIEW
published: 11 September 2018
doi: 10.3389/fdata.2018.00002

Data Analytics Applications for Streaming Data From Social Media: What to Predict?

Frank Emmert-Streib[1,2]*, *Olli P. Yli-Harja*[2,3] *and Matthias Dehmer*[4,5,6]

[1] *Predictive Medicine and Data Analytics Lab, Department of Signal Processing, Tampere University of Technology, Tampere, Finland,* [2] *Institute of Biosciences and Medical Technology, Tampere, Finland,* [3] *Institute for Systems Biology, Seattle, WA, United States,* [4] *Department for Biomedical Computer Science and Mechatronics, UMIT - The Health and Lifesciences University, Hall in Tyrol, Austria,* [5] *Faculty for Management, Institute for Intelligent Production, University of Applied Sciences Upper Austria, Steyr, Austria,* [6] *College of Computer and Control Engineering, Nankai University, Tianjin, China*

Social media in general provide great opportunities for mining massive amounts of text, image, and video-based data. However, what questions can be addressed from analyzing such data? In this review, we are focusing on microblogging services and discuss applications of streaming data from the scientific literature. We will focus on text-based approaches because they represent by far the largest cohort of studies and we present a taxonomy of studied problems.

Keywords: social media, data analytics, prediction model, forecasting, big data, computational social science, scientometrics, data science

OPEN ACCESS

Edited by:
Dongwon Lee,
Pennsylvania State University,
United States

Reviewed by:
Seungwon Yang,
Louisiana State University,
United States
Lei Li,
Hefei University of Technology, China
Jingrui He,
Arizona State University, United States

***Correspondence:**
Frank Emmert-Streib
v@bio-complexity.com

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *01 June 2018*
Accepted: *02 August 2018*
Published: *11 September 2018*

Citation:
Emmert-Streib F, Yli-Harja OP and Dehmer M (2018) Data Analytics Applications for Streaming Data From Social Media: What to Predict? Front. Big Data 1:2. doi: 10.3389/fdata.2018.00002

1. INTRODUCTION

The establishment of the World Wide Web (WWW) in the 1990s revolutionized the communication between people in many different and profound ways affecting our professional and social life alike. One particular consequence of the WWW has been the creation of social media that provide a forum for the direct exchange of digital information in the form of texts, photos, or videos, e.g., via blogs, microblogs, photo sharing, video sharing, social bookmarking, virtual worlds, social gaming, or social networking web pages. The top sites such as Twitter, Facebook, LinkedIn, and Google+ are used by hundreds of millions of active users worldwide. In the following, we will focus on text-based social networking services for microblogging that are publicly accessible. This excludes Instagram (image-based) and Youtube (video-based) but also Whatsapp (not publicly accessible chats) from our considerations.

Due to the relatively brief history of the WWW and the social networking services there is still a severe lack of understanding what, e.g., the information provided by microblogs can be used for. For this reason, we provide a review of the literature with a focus on application areas of prediction models that have been developed so far for analyzing data from microblogging services.

By prediction models we mean methods that aim at forecasting new events rather than merely summarizing or describing information contained in data. For instance, among the first studied questions of social media were investigations related to the topological structure of social networks. Specifically, the degree distribution, the community structure and motifs of acquaintance networks representing the "friendships" among members of social networking services, corresponding to nodes in such graphs, have been investigated (Java et al., 2007; Aparicio et al., 2015). Such studies are more descriptive in nature. Instead, in this review we present an overview of the literature that use social media data for classification, regression, or time series prediction problems.

2. GENERAL APPLICATION FIELDS AND NUMBER OF PUBLICATIONS

We are starting our review my demonstrating that the field of social media analytics is of great interdisciplinary interest occupying already today a large share in the literature.

In order to show this, we are using the Web of Science (WoS) (Clarivate Analytics, 2009) database, which is an online subscription-based citation indexing service operated by Clarivate Analytics. WoS contains comprehensive information about published scientific articles in all areas. We used WoS searching for articles containing the name of a microblog either in the title, abstract, or as a keyword we found: Twitter: 16614, Facebook: 15483, Tumblr: 175, GNU social (previously known as StatusNet and Laconica): 72, Plurk: 56. From this we conclude that the by far most frequently investigated microblogs in the literature are Twitter and Facebook. For this reason, we will focus on these in the following.

In **Figure 1A**. an overview of scientific fields is shown as tagged to published articles containing the keyword Twitter or Facebook, either in the title, the abstract, or as a keyword. It is not surprising that most publications are computer science or social science related. However, also quite a large fraction of papers comes from medicine, management & business, and even arts & humanities. Interestingly, the fraction of psychology related publications is rather low despite the fact that intuitively one would name this field first due to the personal nature of tweets and Facebook postings. One reason for this underrepresentation may be related to computational obstacles psychologists need to overcome when they want to analyze social media data because available tools may not allow to tackle targeted research questions as conceived by psychologists.

In **Figure 1B**, we show the number of published articles containing the keywords Twitter, Facebook, "machine learning" or "artificial intelligence." For papers containing the words Twitter or Facebook these numbers are total numbers, for "machine learning" and "artificial intelligence" these numbers are subtracted by the minimal number of published papers in these fields between 2006 and 2016. For "machine learning" this number is 3266 and for "artificial intelligence" it is 12560. By subtracting these numbers we shifted both curves downward (baseline shift) to make all four curves comparable with each other due to the fact that articles investigating Twitter or Facebook commenced only around 2008 whereas the work in machine learning and artificial intelligence goes much further back. In this sense, the curves shown for machine learning and artificial intelligence provide only information about *new research directions* as started around 2008. From this comparison we learn that the proportion of social media related publications compared to all articles involving machine learning or artificial intelligence is amazingly high, making it about 1/4 in 2016. Another tendency we can observe is that the number of Twitter related publications is overtaking Facebook since 2013. We did not include the years 2017 and 2018 in **Figure 1B**. because the counts in WoS are still incomplete but also for these years we find this trend to continue (data not shown).

3. APPLICATIONS

3.1. Specific Scientific Application Fields

The idea of utilizing data from social media for making predictions has generated great interest (Kalampokis et al., 2013; Schoen et al., 2013). The question is what can one predict based on such data? Prominent examples for such studies are prediction models that investigated the emotional constitution of people (Fernandez et al., 2012; Kross et al., 2013; Ortigosa et al., 2014), personal traits and characters (Kosinski et al., 2013), stock market behavior (Bollen et al., 2011; Siganos et al., 2014), election results (Alonso and Vilares, 2016; Tumasjan et al., 2011).

Further examples are consumer behavior (Ringelhan et al., 2015) , public health (Sinnenberg et al., 2017), opinion flow (Wu et al., 2014), sharing cascades (Kupavskii et al., 2012; Cheng et al., 2014), account classification (Chu et al., 2010, 2012; Dickerson et al., 2014), conflicts among friends (Liu and Weber, 2014), demographics of users (Culotta et al., 2015), mental health (Guntuku et al., 2017), heart disease (Eichstaedt et al., 2015), tourism (information search and decision-making behaviors) (Zeng and Gerritsen, 2014), word-of-mouth (WOM) or consumer reviews (Zhang et al., 2012), box-office revenue of movies (Asur and Huberman, 2010), levels of rainfall (Lampos and Cristianini, 2012), earthquakes (Sakaki et al., 2010), theoretical implications introduced by social media (Kane et al., 2014). In **Table 1** we provide a comprehensive overview of many important questions that have been studied using social media data. We would like to note that here we emphasized the "What to predict" aspect of these studies by highlighting the questions that have been addressed.

As one can see from **Table 1** there are many different questions studied so far. In order to organize these publications, we introduce a taxonomy to categorize these publications according to a few major variables. In **Figure 1** we give a graphical summary of our taxonomy. Overall, these questions fall into seven different fields (E, Economy; G, Geophysics; H, Health; M, Management; S, Sociology; Ps, Psychology; Po, Politology) covering almost all science areas. In this figure, we provide furthermore information about four additional layers, namely (I) the time horizon of the prediction (horizon) for making predictions about the future (F) or the present (P), (II) the level of prediction (level) for macro (Ma) and micro (Mi) level predictions, (III) the time of prediction (time) for batch (Ba) and real-time (Rt) predictions, and for (IV) making spatial (Sp) or non-spatial (Ns) predictions. Each of these layers will be discussed in the following sections.

One area missing from the above (see **Figure 2**) were studies in humanities. By performing a WoS search looking for articles containing the words Twitter/Facebook, humanities, and prediction/forecast we found no results. However, we found articles (54) searching for Twitter/Facebook and humanities. Interestingly, these articles are descriptive rather than predictive in nature. Examples for such studies are (Vainio and Holmberg, 2017). In Lee et al. (2017) and Vainio and Holmberg (2017) the authors studied who tweeted scientific articles with at least one Finnish author/co-author and that had high altmetric counts on Twitter and in Lee et al. (2017) the use of Twitter by scholars in the digital humanities was studied for informal

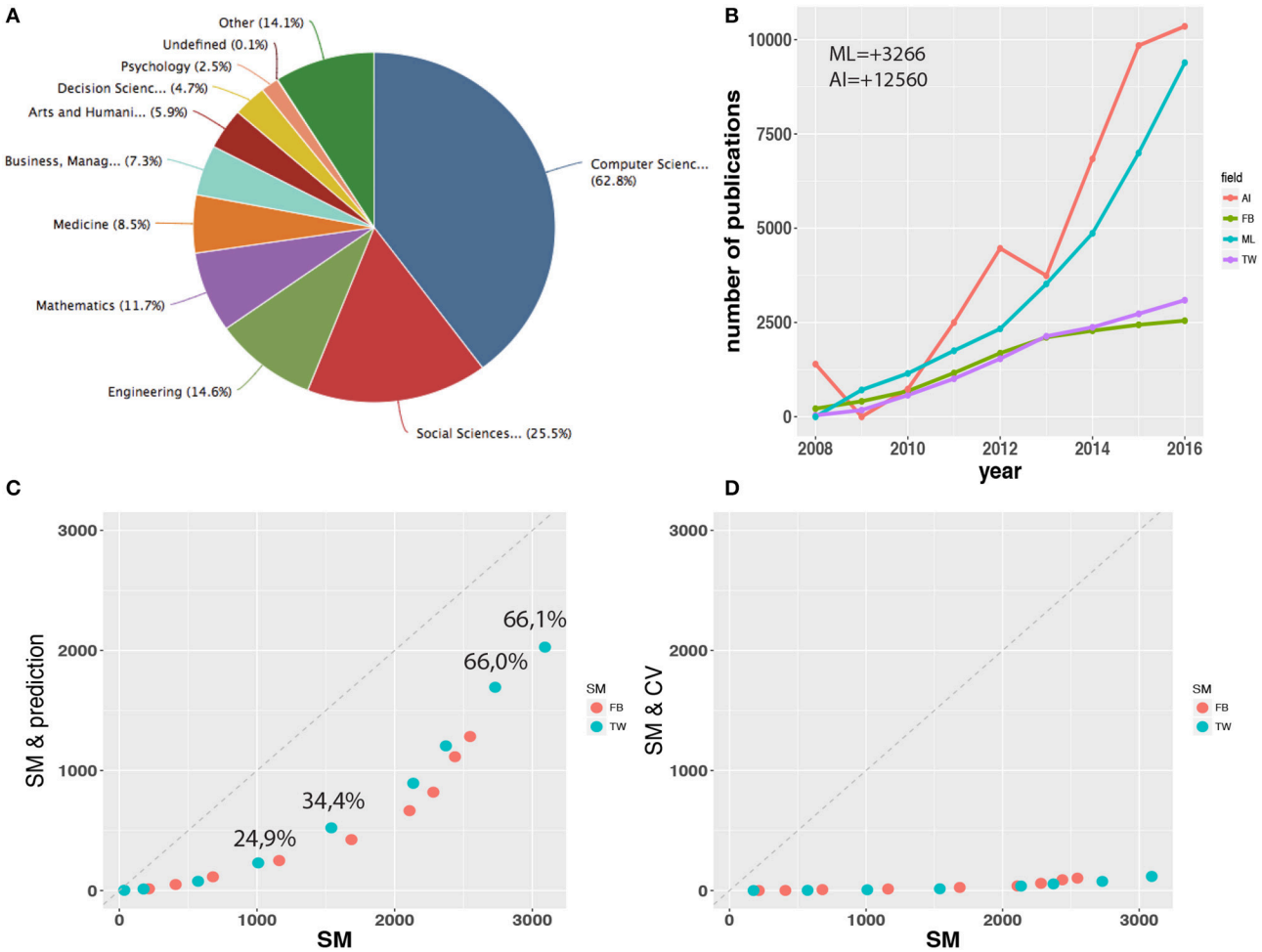

FIGURE 1 | (A) Scientific fields of published articles investigating Twitter (TW) or Facebook (FB). **(B)** The number of published articles containing the keywords Twitter, Facebook, "machine learning" (ML) or "artificial intelligence" (AI). The numbers ML = +3266 and AI = +12560 indicate the baseline shift for ML and AI. **(C)** Scatter plot comparing articles containing the social media (SM) keyword Twitter or Facebook with articles containing additionally "prediction" or "forecast." The shown "forecast." The shown percentages are for Twitter giving the fraction of prediction related publications referred to all publications. **(D)** Similar to **(C)**, but now containing additionally the keywords "cross validation" (CV) or "resampling".

scholarly communication. Those and similar papers performed a descriptive statistical analysis but no predictions were made.

3.2. Time Horizon of the Forecasting

There are two different types of prediction models used in the literature with respect to the prediction itself. The first type predicts the future and the second prediction type predicts the present. The former type is naturally understood because this is what is usually implied by a prediction or a forecast, namely that it should tell us something about the near or far future. For this reason, almost all of the above studies are from this type. However, the second type is unconventional because neither in classical statistics nor machine learning such predictions are made. An example in our context is the prediction of rainfall levels Lampos and Cristianini (2012). Here the idea is to use Twitter users as sort of *social sensors* that report real-world events instantaneously. Another example is the prediction of earthquakes (Sakaki et al., 2010). In the literature such predictions are called *nowcasting* or *predicting the present* (Schoen et al., 2013).

3.3. Macro- vs. Micro-Level Predictions

Another distinction in the predictions is with respect to the level of the prediction. The majority of articles makes predictions on a macro-level for which individual Twitter or Facebook users are irrelevant. Instead, what is important is the aggregation of users into categories. Examples for this is, e.g., predicting outcome of elections or box-office success of movies (Asur and Huberman, 2010; Alonso and Vilares, 2016; Tumasjan et al., 2011). In contrast, predictions on the micro-level make predictions for Twitter or Facebook users themselves. Examples are predicting the personality (Golbeck et al., 2011; Quercia et al., 2011; Hughes et al., 2012; Youyou et al., 2015) or human mobility (Jurdak et al., 2015).

TABLE 1 | An overview of questions addressing "What do predict" with social media data.

"What to predict"	References
Bot detection (account classification)	Chu et al. (2010, 2012); Dickerson et al. (2014)
Box-office revenue of movies	Asur and Huberman (2010)
Company value	Luo and Zhang (2013)
Conflicts among friends	Liu and Weber (2014)
Consumer behavior	Ringelhan et al. (2015)
Crime incidents	Gerber (2014); Aghababaei and Makrehchi (2016)
Demographics of users	Culotta et al. (2015)
Earthquakes	Sakaki et al. (2010)
Election results	Alonso and Vilares (2016); Tumasjan et al. (2011)
Emotional constitution of people	Fernandez et al. (2012); Kross et al. (2013); Ortigosa et al. (2014)
Epidemic of infection disease	Santillana et al. (2015)
Fake news	Gupta et al. (2013); Conroy et al. (2015)
Heart disease	Eichstaedt et al. (2015)
Mental health	De Choudhury et al. (2013); Guntuku et al. (2017)
Popularity of news	Bandari et al. (2012)
Movie ratings	Oghina et al. (2012)
Opinion flow	Wu et al. (2014)
Personal traits and characters	Kosinski et al. (2013)
Public health	Robillard et al. (2013); Sinnenberg et al. (2017)
Sharing cascades	Kupavskii et al. (2012); Cheng et al. (2014)
Stock market behavior	Bollen et al. (2011); Siganos et al. (2014)
Rainfall levels	Lampos and Cristianini (2012)
Suicide rates	Won et al. (2013)
Tourism	Zeng and Gerritsen (2014)
Word-of-mouth (WOM) or consumer reviews	Zhang et al. (2012)

3.4. Batch vs. Real-Time Predictions

The difference between batch and real-time models is that in the former case data are gathered off-line and then one prediction is made. In the latter case this process is iterated multiple times and data are generated on-line. Examples for batch predictions are election forecasts whereas real-time predictions forecast the political opinion continuously (Alonso and Vilares, 2016; Tumasjan et al., 2011). In general, the need for developing a real-time model depends on the application one is aiming at. For instance, if one intends to predict the outbreak of an epidemic of an infection disease this needs to be done in a real-time manner because there is not one scheduled event to occur one wants to predict but there is all the time a possibility for the outbreak to happen (Robillard et al., 2013; Santillana et al., 2015). Another example is the prediction of stock market values (Bollen et al., 2011; Siganos et al., 2014).

3.5. Non-spatial vs. Spatial Predictions

A final distinction of prediction models relates to non-spatial vs spatial predictions. A non-spatial prediction makes a forecast for the population as a whole, e.g., the outcome of an election (Alonso and Vilares, 2016; Tumasjan et al., 2011). In contrast, a spatial prediction makes a forecast for, e.g., all municipalities of a country. In this sense predictions in the former case can be considered as *scalar* whereas in the latter case they are *multivariate*. In order to accomplish a spatial prediction, usually information about the geolocation of the users is utilized. This information may be either directly available, or needs to be inferred from the content of the microblogs.

4. DISCUSSION

As we have shown in **Figure 1B**, the interest in studying data from social media increases every year. However, also the proportion of prediction related publications increases every year. In order to see this we show **Figure 1C**. In this scatter plot we show results we obtained from a WoS search for articles containing the social media (SM) keyword Twitter or Facebook (x-axis) and for articles containing additionally the keywords "prediction" or "forecast" (y-axis). The fraction of the values on the y-axis to the values on the x-axis, i.e., y_i/x_i, gives the percentage of prediction related publications compared to all publications. In **Figure 1C**. the shows values are for Twitter (values for Facebook are similar). Due to the fact that the number of publications increases every year, as can be seen from **Figure 1B**,the x-axis in this figure is proportional to the publication year and, hence, one can see that the fraction of prediction related publications increases over the years reaching currently well over 60%.

4.1. Gaps in the Literature

When collecting the articles for this review we noticed that despite the fact that all considered publications utilize prediction models, only a small fraction of these make an attempt to ensure the statistical soundness of the models. As a simple indicator for this omission we searched the WoS for articles containing the keywords Twitter or Facebook and for articles that contain the keywords Twitter and cross validation or Twitter and resampling (similarly for Facebook). The result of these searches is shown as a scatter plot in **Figure 1D**. The shown pairs correspond to the same publication year and y-axis label SM & CV is an abrieviation for our second search query. This figure confirms our perception indicating that only a small fraction of all articles applies resampling methods in order to quantify the uncertainty in the data and to guard against overfitting. Given the fact that the analyzed social media data are "big," resampling methods can always be applied. Overall, this indicates a possible problem that would require further analysis.

4.2. Potential Future Developments

4.2.1. Data Integration

The vast majority of studies analyzed only data from social media. However, a combination of such data with external data would allow to address further questions. For instance, health related studies could benefit from *integrating data* from disease databases, e.g., Online Mendelian Inheritance in Man (OMIM) (OMI, 2007), Gene Ontology (Ashburner et al., 2000),

T: Twitter
F: Facebook

E: Economy
G: Geophysics
H: Health
M: Management
S: Sociology
Ps: Psychology
Po: Politology

P: Present
F: Future

Ma: Macro
Mi: Micro

Ba: Batch
Rt: Real-time

Sp: Spatial
Ns: Non-spatial

Topics of predictions	Social media	Applications	Horizon	Level	Time	Spatial
emotional constitution of people	T, F	Ps	P	Mi	Ba	Ns, Sp
user geolocations	T	Ps	P	Mi	Ba	Sp
popularity of tweets	T	S, Ps	P	Mi	Ba	Sp
outcome of public events	T	S, Ps	F	Ma	Ba, Rt	Ns
crime incidents	T	S, Ps	P	Ma	Rt	Sp
breaking news detection	T	S	P	Ma	Rt	Ns
sharing cascades	T, F	S	F	Mi	Ba	Sp
conflicts	T	S	F	Mi	Ba	Sp
account classification	T	S	P	Mi	Ba	Sp
demographics of users	T	S	P	Mi	Ba	Sp
stock market shares	T, F	E, Ps	F	Ma	Rt	Ns
political opinions	T, F	Po	F, P	Mi	Rt	Sp
election results	T	Po	F	Ma	Ba, Rt	Ns
infectious diseases	T	H	F, P	Ma	Rt	Sp
heart disease	T	H, Ps	F, P	Mi	Ba, Rt	Sp
mental health	T	H	F, P	Mi	Ba, Rt	Sp
substance abuse	T, F	H, Ps	F, P	Ma, Mi	Ba, Rt	Sp
box-office revenues for movies	T, F	M	F	Ma	Ba, Rt	Ns
rainfall levels	T	G	P	Ma	Rt	Sp
earthqukes	T	G	P	Ma	Rt	Sp

FIGURE 2 | Taxonomy of questions that have been investigated so far by prediction models. Overall, these questions fall into seven different applications. E, Economy; G, Geophysics; H, Health; M, Management; S, Sociology; Ps, Psychology; Po, Politology. In addition, distinctions are made regarding the horizon, level, time, and spatial nature of the predictions (see main text for details).

or DrugBank (Wishart et al., 2007). This approach enables also in a natural way the extension of text mining approaches because the external information may be utilized in form of dictionaries, e.g., lists of words from a specific category, that can be used to perform a guided sentiment analysis.

Support for our argument for using external information is provided by Ciulla et al. (2012). The authors found that information provided by tweets alone is not sufficient in order to predict the outcome of a social event (the winner of American Idol) but tweets need to be complemented with information about the geographic location of the tweets.

Another purpose for data integration could be for increasing prediction accuracy and reducing prediction errors. This could be accomplished by utilizing different, independent sources of social media data. In this way one could also naturally obtain quantitative estimates for the variability in the data.

4.2.2. Social Networks

A further direction to explore could be the utilization of social networks (Wasserman and Faust, 1994). An example area where this could be of relevance is studies about infectious outbreaks. The reason for this is that an infection can only spread by human contacts. However, usually, this human contact network is not known. As an approximation for such a human contact network one could utilize data from social media to infer such a network. The simplest way to do this could be by utilizing the information "who is a follower of whom" which can be directly extracted from Twitter. However, one can go beyond these follower networks by also constructing semantic networks. The semantic networks could be constructed from estimating the similarity, e.g., among Twitter users based on the content of their tweets and conditioned on metadata. As a result, the information from these different networks could be integrated leading to characteristic spatial scores of the twitter activity and content in specific area.

4.2.3. Deep Learning

Finally, it will be interesting to see if new machine learning and artificial intelligence methods, above all deep learning methods (Hinton et al., 2006; Bengio et al., 2009; LeCun et al., 2015), e.g., deep neural networks, deep decision trees or deep belief networks, will change the *type of questions* addressed with social media data. So far, deep learning methods have found ample applications in image recognition, audio classification, genomics and text mining, e.g., (Lee et al., 2009; Alipanahi et al., 2015; Jiang et al., 2015; He et al., 2016), however, for social media mining we cannot observe from the current literature that new "What to predict" questions have emerged. Instead, familiar questions

are studied with these new methodologies focusing on "How to predict." Maybe, more experience is needed until scientists find new questions that can be raised with such computer- and data-intense approaches.

5. CONCLUSIONS

In this paper we surveyed the literature of prediction models for social media with a focus on the questions that have been addressed so far. Since we are observing a transition from descriptive to predictive studies in the last years (see **Figure 1C**) a taxonomy of such questions is a natural first step in understanding the capabilities of social media. We anticipate this trend to continue and the diversity of question to increase. However, a necessity for the latter is a better comprehension of the data social media provide by exploring their limitations and possibilities with respect to statistical models.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

MD thanks the Austrian Science Funds for supporting this work (project P30031).

REFERENCES

Aghababaei, S., and Makrehchi, M. (2016). "Mining social media content for crime prediction," in *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on* (Omaha, NE: IEEE), 526–531.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Alonso, M., and Vilares, D. (2016). A review on political analysis and social media. *Procesamiento Leng. Nat.* 56, 13–23.

Aparicio, S., Villazón-Terrazas, J., and Álvarez, G. (2015). A model for scale-free networks: application to twitter. *Entropy* 17, 5848–5867. doi: 10.3390/e17085848

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Asur, S., and Huberman, B. A. (2010). "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01; WI-IAT '10* (Washington, DC: IEEE Computer Society), 492–499. doi: 10.1109/WI-IAT.2010.63

Bandari, R., Asur, S., and Huberman, B. A. (2012). "The pulse of news in social media: Forecasting popularity," in *ICWSM, Vol. 12*, 26–33.

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127 . doi: 10.1561/2200000006

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *J. Comput. Sci.* 2, 1–8. doi: 10.1016/j.jocs.2010.12.007

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). "Can cascades be predicted?," in *Proceedings of the 23rd International Conference on World Wide Web* (Seoul: ACM), 925–936.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2010). "Who is tweeting on twitter: human, bot, or cyborg?," in *Proceedings of the 26th Annual Computer Security Applications Conference* (Austin, TX: ACM), 21–30.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: are you a human, bot, or cyborg? *IEEE Trans. Depend. Secure Comput.* 9, 811–824. doi: 10.1109/TDSC.2012.75

Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., and Vespignani, A. (2012). Beating the news using social media: the case study of american idol. *EPJ Data Sci.* 1:8. doi: 10.1140/epjds8

Clarivate Analytics (2009). *Web of Science*. Available online at: https://en.wikipedia.org/wiki/Clarivate_Analytics

Conroy, N.J., Rubin. V. L., and Chen, Y. (2015). "Automatic deception detection: methods for finding fake news," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (St. Louis, MO: American Society for Information Science), 82.

Culotta, A., Kumar, N. R., Cutler, J. (2015). "Predicting the demographics of twitter users from website traffic data," in *AAAI* (Austin, TX), 72–78.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. in *ICWSM, Vol.13*, 1–10.

Dickerson, J. P., Kagan, V., and Subrahmanian, V. (2014). "Using sentiment to detect bots on twitter: Are humans more opinionated than bots?," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (IEEE), 620–627.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., et al. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychol. Sci.* 26, 159–169. doi: 10.1177/0956797614557867

Fernandez, K.C., Levinson, C. A., and Rodebaugh, T. L. (2012). Profiling: predicting social anxiety from facebook profiles. *Soc. Psychol. Pers. Sci.* 3, 706–713. doi: 10.1177/1948550611434967

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decis. Support Syst.* 61, 115–125. doi: 10.1016/j.dss.2014.02.003

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). "Predicting personality from twitter," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (Boston, MA), 149–156. doi: 10.1109/PASSAT/SocialCom.2011.33

Guntuku, S., Yaden, D., Kern, M., Ungar, L., Eichstaedt, J. (2017). Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* 18, 43–49. doi: 10.1016/j.cobeha.2017.07.005

Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. (2013). "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd International Conference on World Wide Web.* (Rio de Janeiro: ACM), 729–736.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hughes, D. J., Rowe, M., Batey, M., and Lee, A. (2012). A tale of two sites: twitter vs. facebook and the personality predictors of social media usage. *Comput. Hum. Behav.* 28, 561–569. doi: 10.1016/j.chb.2011.11.001

Java, A., Song, X., Finin, T., and Tseng, B. (2007). "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (San Jose, CA: ACM), 56–65.

Jiang, Z., Li, L., Huang, D., and Jin, L. (2015). "Training word embeddings for deep learning in biomedical text mining tasks," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on* (Washington, DC: IEEE), 625–628.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PLOS ONE* 10:e37027. doi: 10.1371/journal.pone.0131469

Kalampokis, E., Tambouris, E., and Tarabanis, K. (2013). Understanding the predictive power of social media. *Inter. Res.* 23, 544–559. doi: 10.1108/IntR-06-2012-0114

Kane, G., Alavi, M., Labianca, G., and Borgatti, S. (2014). What's different about social media networks? a framework and research agenda. *MIS Q.* 38, 275–304. Available online at: https://misq.org/what-s-different-about-social-media-networks-a-framework-and-research-agenda.html

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5802–5805. doi: 10.1073/pnas.1218772110

Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., et al. (2013). . Facebook use predicts declines in subjective well-being in young adults. *PLOS ONE* 8:e69841 . doi: 10.1371/journal.pone.0069841

Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., et al. (2012) . "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, HI: ACM), 2335–2338.

Lampos, V., and Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.* 72, 1–22. doi: 10.1145/2337542.2337557

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436.

Lee, H., Pham, P., Largman, Y., and Ng, A. Y. (2009). "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 1096–1104.

Lee, M., Yoon, H., Smith, M., Park, H., and Park, H. (2017). Mapping a twitter scholarly communication network: a case of the association of internet researchers? conference. *Scientometrics* 112, 767–797. doi: 10.1007/s11192-017-2413-z

Liu, Z., and Weber, I. (2014). "Predicting ideological friends and foes in twitter conflicts," in *Proceedings of the 23rd International Conference on World Wide Web* (Seoul: ACM), 575–576.

Luo, X., and Zhang, J.(2013). How do consumer buzz and traffic in social media marketing predict the value of the firm? *J. Manage. Inform. Syst.* 30, 213–238. doi: 10.2753/MIS0742-1222300208

Oghina, A., Breuss, M., Tsagkias, M., and de Rijke, M. (2012). "Predicting imdb movie ratings using social media," in *European Conference on Information Retrieval* (Barcelona: Springer), 503–507.

OMI (2007). *Online Mendelian Inheritance in Man, OMIM (TM)*.

Ortigosa, A., Carro, R. M., and Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *J. Comput. Syst. Sci.* 80, 57–71. doi: 10.1016/j.jcss.2013.03.008

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). "Our twitter profiles, our selves: predicting personality with twitter," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (IEEE), 180–185.

Ringelhan, S., Wollersheim, J., and Welpe, I. M. (2015). I Like, I Cite? Do facebook likes predict the impact of scientific work? *PLOS ONE* 10:e0134389. doi: 10.1371/journal.pone.0134389

Robillard, J. M., Johnson, T. W., Hennessey, C., Beattie, B. L., Illes, J. (2013). Aging 2.0: health information about dementia on twitter. *PLoS ONE* 8:e69861. doi: 10.1371/journal.pone.0069861

Sakaki, T., Okazaki, M., and 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, WWW '10* (Raleigh, NC: ACM), 851–860. doi: 10.1145/1772690.1772777

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., and Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput. Biol.* 11:e1004513. doi: 10.1371/journal.pcbi.1004513

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., and Gloor, P. (2013). The power of prediction with social media. *Inter. Res.* 23, 528–543. doi: 10.1108/IntR-06-2013-0115

Siganos, A., Vagenas-Nanos, E., and Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *J. Econ. Behav. Organ.* 107, 730–743. doi: 10.1016/j.jebo.2014.06.004

Sinnenberg, L., Buttenheim, A., Padrez, K., Mancheno, C., Ungar, L., and Merchant, R. (2017). Twitter as a tool for health research: A systematic review. *Am. J. Public Health* 107, e1–e8. doi: 10.2105/AJPH.2016.303512

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2011). Election forecasts with twitter: how 140 characters reflect the political landscape. *Soc. Sci. Comput. Rev.* 29, 402–418. doi: 10.1177/0894439310386557

Vainio, J., and Holmberg, K. (2017). Highly tweeted science articles: who tweets them? an analysis of twitter user profile descriptions. *Scientometrics* 112, 345–366. doi: 10.1007/s11192-017-2368-0

Wasserman, S., and Faust, K. (1994) . *Social Network Analysis*. Cambridge; New York, NY: Cambridge University Press.

Wishart, D.S., Knox. C., Guo. A.C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2007). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906. doi: 10.1093/nar/gkm958

Won, H.-H., Myung, W., Song, G.-Y., Lee, W.-H., Kim, J.-W., Carroll, B. J., et al. (2013). Predicting national suicide numbers with social media data. *PLoS ONE* 8:e61809. doi: 10.1371/journal.pone.0061809

Wu, Y., Liu, S., Yan, K., Liu, M., and Wu, F. (2014). Opinionflow: visual analysis of opinion diffusion on social media. *IEEE Trans. Vis. Comput. Graph.* 20, 1763–1772. doi: 10.1109/TVCG.2014.2346920

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1036–1040. doi: 10.1073/pnas.1418680112

Zeng, B., and Gerritsen, R. (2014). What do we know about social media in tourism? a review. *Tour. Manage. Perspect.* 10, 27–36. doi: 10.1016/j.tmp.2014.01.001

Zhang, Z., Li, X., and Chen, Y. (2012) . Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews. *ACM Trans. Manage. Inform. Syst.* 3:23. doi: 10.1145/2151163.2151168

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Emmert-Streib, Yli-Harja and Dehmer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.