Benchmarking Crisis in Social Media Analytics: A Solution for the Data Sharing Problem

Original Manuscript

Social Science Computer Review
1-27


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393211012268
journals.sagepub.com/home/ssc


Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem

Dennis Assenmacher[1,2], Derek Weber[3,4], Mike Preuss[5], André Calero Valdez[6], Alison Bradshaw[7], Björn Ross[8], Stefano Cresci[9], Heike Trautmann[1], Frank Neumann[3], and Christian Grimme[1]

Abstract

Computational social science uses computational and statistical methods in order to evaluate social interaction. The public availability of data sets is thus a necessary precondition for reliable and replicable research. These data allow researchers to benchmark the computational methods they develop, test the generalizability of their findings, and build confidence in their results. When social media data are concerned, data sharing is often restricted for legal or privacy reasons, which makes the comparison of methods and the replicability of research results infeasible. Social media analytics research, consequently, faces an integrity crisis. How is it possible to create trust in computational or statistical analyses, when they cannot be validated by third parties? In this work, we explore this well-known, yet little discussed, problem for social media analytics. We investigate how this problem can be solved by looking at related computational research areas. Moreover, we propose and implement a prototype to address the problem in the form of a new evaluation framework that enables the comparison of algorithms without the need to exchange data directly, while maintaining flexibility for the algorithm design.

Keywords

social media analytics, benchmarking, social computing, reproducibility

[1] University of Münster, Germany
[2] Queensland University of Technology (QUT), Brisbane, Australia
[3] University of Adelaide, South Australia, Australia
[4] Defence Science and Technology Group, Department of Defence, Canberra, Australian Capital Territory, Australia
[5] Universiteit Leiden, the Netherlands
[6] RWTH Aachen University, Germany
[7] Alison Bradshaw Legal, Glenside, South Australia, Australia
[8] University of Edinburgh, United Kingdom
[9] Institute of Informatics and Telematics (IIT-CNR), Pisa, Italy

Corresponding Author:
Dennis Assenmacher, University of Münster, Münster 48149, Germany.
Email: dennis.assenmacher@wi.uni-muenster.de

Social media platforms such as Twitter, Facebook, and Instagram have become an integral part of modern life. Nowadays, people use these platforms not only for private social interaction but also for civic, educational, and political purposes, such as following global or local news and for participating in discussions of controversial political and social issues in an open manner (Zhang et al., 2010). Consequently, journalists have long employed these platforms as channels for identifying trending stories and topics and as an outlet for their own professional content; politicians consider these channels valuable for listening "to the people," distributing their views and policies and running their campaigns; and "the people" share content and their opinions, participating in the global exchange of information. However, the social media ecosystem is also a data vault, which contains valuable information about all participants and their behavior. Commercial players use social media as marketing platforms, building customer profiles from the data they hold in order to generate profits through targeted advertising; the financial sector tries to infer sentiment on stocks or information on (future) trading behavior in order to optimize portfolios and profit; and researchers consider these platforms as environments in which to observe human behaviors from a variety of perspectives and scales, from the opinions of individuals to the evolution of groups, up to societal processes.

Particularly since the first reports of political astroturfing in 2010 (Metaxas & Mustafaraj, 2012), the focus of research has shifted to the "dark side" of social media. Instead of strengthening participation, social media is increasingly considered the medium of choice for spreading disinformation, fake news and propaganda, for running automated accounts (e.g., social bots), and for manipulating society in general (Cresci, 2020; Da San Martino et al., 2020; Persily & Tucker, 2020). An enormous amount of research based on the analysis of data gathered from social media platforms has been published. This research has often addressed fundamental questions of human interaction and behavior, as well as implications for politics, government policies and regulations, the economy, and society more broadly. This suggests the rise of a new branch of data-driven research, which is increasingly important in decision making and for understanding online human behavior, including how it relates to off-line behavior.

It is therefore imperative that members of this emerging research community constantly reflect on their methodological foundations as a way to ensure good scientific practice. This work provides a critical view of this topic, specifically focusing on the comparison and benchmarking of computational methods on social media data and the replicability of research results. The authors of this work are active representatives of the social media analytics community and perform research on fake news, cyberhate, the analysis of social media content and sentiment, artificial intelligence, and automation in social media, as well as on manipulation and disinformation campaigns. Many of the authors met at the 2020 Human Computer Interaction International Conference held (virtually) in Copenhagen, Denmark, and discussed the topic of replicability and comparability of research using social media data. This discussion highlighted a serious problem with two aspects in the field, which both relate to the availability of data, namely the ability to access representative data and the ability to easily share it to facilitate benchmarking. One may ask: *How could there be a problem with data? There is so much data available on social media platforms! And once you have a copy, is it not available for all?* The answer to these questions is sadly very simple: *Almost all research in social media completely depends on the goodwill of the (commercial) platforms to share or provide data and permit its use*.

The targets of research in social media analytics are the commercial platforms and their users, and the relevant data reside within the physical and legal domain of the platforms themselves. As a consequence, access is provided and controlled by the platforms alone. It is made available (and constrained) via mechanisms such as Application Programming Interfaces (APIs) and data streams, often with a monetary cost required for greater access. Sharing of entire data sets is commonly forbidden by the platforms' terms and conditions or at least very restricted.[1] Furthermore, data

provided by an API may not be the same when queried by another researcher (e.g., Twitter's 1% sample API; Joseph et al., 2014; Paik & Lin, 2015; Weber et al., 2020). In addition to the data held by social media platforms constantly growing, the existing data changes over time as users and the platforms modify, remove, and otherwise interact with posts, accounts, and associated data. Additionally, platforms rarely disclose the sampling strategies they use for the data they provide via their APIs (Morstatter et al., 2013).

A further consideration is the danger of bias in social media studies resulting from differences in the availability of data between platforms (Persily & Tucker, 2020). For example, Twitter data have been traditionally easier to access for researchers, and although Twitter activity contributes a significant portion of online discussion, it is by no means the only major platform. If research focuses on Twitter simply because of the availability of data, the research community will struggle to provide a nonbiased view of the online world. Results will suffer from the streetlight effect: looking for lost keys where a streetlight is, instead of where they have been lost.

Consequently, researchers usually need to gather data sets individually or through small collaborations and are then forced to ground their theories and evaluations solely on these unique collections. The difficulty in exchanging data due to licensing restrictions hinders the comparison of methods and the replication of experiments. The reported results lack a common frame of reference. In the long term, the lack of transparency and comparability in our community may lead to a loss of confidence in published results. For these reasons, this article makes two major contributions:

1. We direct the attention of the social media analytics community to this rarely addressed problem and highlight the difficulties of exchanging data and the resulting lack of benchmarks. From our perspective, the problem is not sufficiently covered by the ongoing discussion of a "reproducibility crisis" as mentioned by Baker (2016), Amrhein et al. (2017), and Cockburn et al. (2020). The problem is far more fundamental and has to be addressed to ensure trust in results from this field of research.
2. We provide a conceptual framework to solve the problem of data exchange for benchmarking purposes, which is currently massively restricted by platform licensing. Rather than requiring the exchange of data, we propose establishing a distributed benchmarking architecture that leaves data in the possession of the collector while enabling comparison of the data by exchanging algorithms instead. In addition to this, we provide a specification and a prototype implementation. These additions enable more informed discussion of the advantages and potential challenges in establishing a benchmarking standard acceptable to the community.

This work is structured as follows: We initially provide examples of different data-centric social media analytics scenarios the authors have experienced. This helps the reader to understand the problem in context. The problem is then formally discussed and defined. Subsequently, existing benchmarking practices are considered. From there, we derive a set of requirements and develop a conceptual model, which is realized as a benchmarking prototype, including a discussion of its potential and limitations. Finally, we highlight the benefits of the approach as well as future paths toward a common benchmarking initiative in the social media analytics community.

Anecdotes From the Community and How a Framework Would Improve the Situation

Before we provide a concise description of the problem considered in this work, we provide some examples of cases related to techniques and methods used (and the associated challenges faced) by researchers in the field of social media analytics. Based on the experiences described and literature

discussing the reproducibility crisis, we then formalize the challenge of data exchange in the next section.

Stream Clustering

Stream clustering of textual data is an emerging topic in the field of social media analytics. The overwhelming amount of data produced on social media platforms justifies stream clustering having its own discipline in the machine learning research community, a discipline focused on the development of new ideas and algorithms in an unsupervised setting. In recent years, several new algorithms have been proposed by the community, each with its own advantages and disadvantages (Carnein et al., 2017; Kumar et al., 2020; Yin et al., 2018). A fundamental problem that not only hinders the acceptance but also the progress of such new ideas is the limited capacity for replicated evaluation, which is due to a lack of both accessible data sets and algorithm implementations. A few standard data sets of textual data exist, such as Reuters or 20 Newsgroup (Lang, 1995; Lewis et al., 2004), but these were never expected to be analyzed within a stream setting. Therefore, as mentioned, researchers gather and evaluate their own unique data sets (Assenmacher, Adam, et al., 2020; Assenmacher, Clever, et al., 2020). While this is a valid approach in theory, it comes with problematic side effects. Most importantly, the raw data cannot be freely shared, and thus to benchmark a new approach against existing methods, the responsibility falls to the new method's developer to find and/or implement and then execute the existing algorithms themselves. This is often hindered by (1) a lack of public implementations and/or (2) the risk of misunderstanding the implementation and the resulting misconfiguration. While in theory an algorithm's implementation should be publicly available, in practice, there are good reasons for code not to be shared, such as when the algorithm is used commercially. Results published from such comparative evaluations (if any) should, therefore, be considered with caution.

Hate Speech Detection

In recent years, an increased demand for sophisticated hate speech (abusive language) detection mechanisms has been observed, especially from news media sites confronted with the challenge of moderating toxic and hateful content in their websites' comment sections (Fortuna & Nunes, 2018; Niemann et al., 2020; Riehle et al., 2020). Typically, moderators have to manually inspect and delete comments before they are made public. This obligation is frequently further enforced by legislation. To train and validate sophisticated deep-learning models, for example, that have emerged over the last decade, a large amount of annotated data are needed to deal with content at scale. Moreover, only the data owned by the platforms can be used and thus, because they are hard to exchange, the problem of comparability arises. A neural architecture may work well for one data set, but generalizability can rarely be demonstrated, let alone guaranteed. Therefore, it is of great value to determine how an existing model performs on other, well-known data sets. This is often not feasible because of the platform's constraints on the sharing of data, and because of privacy regulations, which vary across the world. The few existing data sets that are commonly used for benchmarking purposes are small and suffer from biases despite the great amount of effort that goes into their creation and curation (Wiegand et al., 2019).

Social Bot and Disinformation Campaign Research

The endeavor of detecting social bots (automated agents that are designed to support disinformation, manipulation, or propaganda campaigns and otherwise simulate human behavior (Al-Rawi, 2019; Cresci, 2020; Grimme, Preuss, et al., 2017) and disinformation campaigns addresses both

computational detection methods and the social/political relevance of these methods simultaneously and has led to an enormous body of research literature (Bastos & Mercea, 2019; Cresci et al., 2019; Ferrara, 2017; Grimme, Assenmacher, et al., 2017; Hagen et al., 2020; Kollanyi et al., 2016; Neudert, 2017; Nizzoli et al., 2021; Shao et al., 2018; Weber & Neumann, 2020).

The proposed methods range from very simple decision rules (e.g., defining highly active accounts as social bots; Howard & Kollanyi, 2016; Kollanyi et al., 2016; Neudert, 2017) to exploratory (e.g., Hegelich & Janetzko, 2016; Grimme, Assenmacher, et al., 2017; Ross et al., 2018) or machine learning approaches (Assenmacher, Adam, et al., 2020; Cresci et al., 2018; Mazza et al., 2019; Varol et al., 2017) for detecting bots and campaigns up to interactive human-in-the-loop techniques (Assenmacher, Clever, et al., 2020; Grimme et al., 2018). Naturally, all works ground their findings in social media data, streamed or accessed otherwise via the appropriate APIs. According to good scientific practice, authors state the sources of data, access methods, parameters for filtering and other configuration information, and the amount of data gathered within a specific time window (often millions of posts). What they usually cannot provide is the data itself. According to the licensing conditions of most social media platforms, access to public content via APIs is permitted but is mostly restricted to private use. Additionally, the data provided are required to be updated regularly to account for changes caused by the deletion of accounts or posts, including retrospectively.

A further constraint of Twitter's conditions, in particular, is that although it permits the public release of the identifiers (IDs) of tweets and user profiles on which research is based, there is no provision for sharing snapshots of the same entities as they change over time.

The natural consequence of these data-sharing limitations is that most newly proposed detection techniques are evaluated on new ad hoc data sets, rather than on existing, well-known, reference ones. This inevitably hinders replicability and comparability of results.

Discussion

The issue of licensing data has led to two common practices in the community: (1) Authors provide no data sets at all but only the filtering and time window parameters they used for retrieving the data via the platforms API. Theoretically, this information is sufficient to access the same data, assuming the sampling strategy of the platform is known (or can be seeded) and the retroactive access to the (unchanged) data is possible. (2) Authors provide a data set, however, it does not contain the original posts themselves but only the unique IDs of the considered posts. Using them, others could—again in theory—access the same data as long as it has not been changed (e.g., posts being removed or interacted with, accounts being deleted). The latter approach (known as rehydrating) seems to be acceptable for high ranked journals (e.g., *Nature Communications*), which otherwise enforce data transparency and replicability of results, perhaps because it is understood that no generally acceptable alternative currently exists.

Very few researchers openly provide complete data sets. Patrick Warren and Darrel Linvill published Twitter data extracted from the Russian IRA troll factory,[2] which was collected from the Twitter Firehose by the Clemson University Social Media Listening Center. It is, however, important to mention that these data contain only tweets of account IDs that were published by the U.S. Congress during the investigation of potential social media–driven manipulation in the 2016 Presidential election. Twitter now offers a number of election-related data sets for research, which, though rich in detail, do not consist of the raw data.[3] Another positive example for data sharing is a small collection of data available from the University of Indiana on the Botometer website,[4] which contains small but complete data sets. Some of those data sets, however, are also restricted to manually identified bots and human accounts. Beyond that there are few exceptions of publicly available data sets for machine learning tasks. These include Wikipedia,[5] Reddit (provided in

databases maintained by Jason Baumgartner[6]), and 4 years of Gab data (Fair & Wesslen, 2019). Importantly, there is a distinct lack of available data sets from other prominent platforms, such as Facebook, Instagram, and WhatsApp, despite ongoing concerns regarding the dissemination of misinformation and disinformation on those platforms.

The problem that arises with the lack of shared, complete, and unchanged data sets is that most research in this area remains anecdotal and preliminary. First, replicability is almost impossible for new algorithms on new data sets, and second, comparison of new technical approaches (e.g., detection mechanisms) is again hindered due to a lack of common benchmark data. This significantly weakens such research in the context of scientific discussion and trust in the results more broadly.

Problem Definition

The previous examples highlight a problem in the development of computational methods and algorithms for social media data that brings to mind the familiar discussion of reproducibility and replicability of experimental research. In this section, we briefly review this well-known problem and its terminology before showing that the problem in social media analytics has different roots. While the replication crisis is caused by, among other things, publication bias and a lack of research skills (Amrhein et al., 2017), the benchmarking crisis in social media analytics results from a lack of sufficiently large, high-quality data sets on which computational methods can be compared, which is a consequence of restrictions on data sharing (Ruths & Pfeffer, 2014). Plesser's historical summary (Plesser, 2017) of the discussion of reproducibility and replicability in the context of computer science states that a broader revision of experimental methodology started with the work of Claerbout and Karrenbach in the early 1990s (Claerbout & Karrenbach, 1992). These authors defined "reproducibility" as the validation of results by the same researchers on a data set, while "replication" was considered to mean that other researchers obtain (relatively) similar results by redoing a documented experiment. King used replicability as the common term to summarize all variants of redoing experiments (King, 1995). Similar to Claerabout and Karrenbach, replicability and reproducibility are used by the Association of Computing Machinery (ACM, 2020) badging system, which is applied when certifying research artifacts submitted by authors as part of their publications. These different definitions suggest that three categories exist, namely

- *repeatability*: the same team repeats the same experiment using the same data and experimental setup;
- *reproducibility*: a different team reproduces the same experimental results using the same data and experimental setup; and
- *replicability*: a different team replicates consistent results using new data and possibly a different experimental setup.

Specifically, the definitions of reproducibility and replicability are in line with the definition provided in a multisciences perspective report and best practice guidelines on both aspects by the National Academies of Science, Engineering and Medicine (Fineberg et al., 2019).

Terminology very recently used by Cockburn et al. (2020); published in *Communications of the ACM*—the same publisher as the badging system) proposes reversed definitions for reproducibility and replicability. Baker (2016), in contrast, uses the term reproduction to refer to all repetitions of experiments (regardless of by whom they are conducted and whether on original or new data), when she writes of a reproducibility crisis.

Although the terminology in previous works may vary, the basic idea of all relevant discussions is similar: Researchers should reflect on their choices of methods when working with data and

experiments. While the natural and social sciences have always been confronted with the challenges of experimentation and unreliable data, the focus of computer science has only gradually shifted from deterministic computation (with little experimentation) toward simulation, random experiments, approximation, and (big) data analytics (Bartz-Beielstein, 2006; Bartz-Beielstein et al., 2010; Fineberg et al., 2019). Once computational methods are applied to social data, computer scientists are confronted with challenges similar to the natural and social sciences. Specifically, the field of social media analytics is working at the boundaries of political science, social science, and computer science, and must address the same issues faced in empirical social science research, as succinctly stated by King (1995): "At its most fundamental, if the empirical basis for an article or book cannot be reproduced, of what use to the discipline are the conclusions? What purpose does an article like this serve?" (p. 445).

Apart from possible general shortcomings in the evaluation and analytics methodology (e.g., as highlighted and discussed in Hutton & Henderson, 2015, 2018; Ioannidis, 2005; Lewandowsky & Oberauer, 2020; for general advice on improving reproductibility and replicability, see also Fineberg et al., 2019), the methodological problem of social media analytics often lies in its empirical foundation—the available data. As shown in a very recent commentary by researchers from the social media analytics domain (Pasquetto et al., 2020), the availability of empirical data strongly depends on the access provided by the social media platforms. The sharing of data among researchers is constrained by license terms defined and imposed by these platforms. Although researchers may be allowed to access data and analyze it locally, the sharing of gathered data is limited to reduced or aggregated artifacts or even completely prohibited (Bruns, 2019). These limitations are mostly not related to privacy issues or data protection frameworks (like the General Data Protection Regulation in the domain of the European Union) but originate from the legitimate commercial interests of the platforms. (Open) Sharing of data among researchers may also give competitors access to the data and result in them profiting from it. Offering some degree of openness, the platforms allow limited access to their data for academics, however, similar to the case of Twitter, they often restrict sharing of collected data to object IDs only (which refer to the objects on the platform, meaning the full data needs to be retrieved again) or only to data sets of potentially irrelevant sizes (e.g., <50,000 objects per day[7]).

Another aspect that affects the collection of data and their usability across institutional boundaries, and thus has a direct impact on the comparability of methodological approaches, is certainly the different handling of data according to protocols and regulations implemented by Institutional Review Boards (IRBs). The regulations (e.g., regarding privacy or data exchange), which are set by different IRBs, can directly influence the usability of the collected data across institutions. On the one hand, censoring important features before data is shared can reduce their value for others dramatically, while on the other hand, the process of data preparation before publication can be tedious and is infeasible for many research groups. Several studies—not only in the context of social media research—address this issue and demonstrate the heterogeneous landscape of IRB regulations and protocols (Moreno et al., 2013; Taylor & Pagliari, 2018; Timmers et al., 2020).

As a consequence of platform licensing and instead of dealing with often tedious regulations for data sharing, researchers gather data via these platforms but refrain from sharing it openly. Only very few data sets are openly available as discussed above. In other words, the lack of commonly shared data is a central problem not only for reproducibility and replicability but also for comparability of competing or complementary approaches, leading to these specific issues:

1. *Sampling of data is not transparent*: As the amount of social media data is enormous, researchers can rarely make use of more than a sample of the data available. Depending on the sampling strategy employed by the platform when data are requested (which may not be transparent or unbiased; Morstatter et al., 2013), the returned data set for a given time

window may differ. Although existing streaming data competitions[8] require competitors to work on streams that adhere to similar filter and access conditions, they clearly disclaim that the results may not be repeatable and that the stream data are not reusable by others.

2. *Data changes over time*: Even if researchers could have access to all data (including retrospectively), comparability of the data is not guaranteed. Platforms modify their data over time as part of curation. When propaganda campaigns or malicious activity is detected, user accounts or posts are often deleted. For researchers who develop automation or campaign detection methods, for those attempting emotion detection and for those who develop network analysis approaches, this can dramatically affect their results (Holzmann et al., 2018; Weber et al., 2020). Removing the data that their work is specifically targeting may change the performance of their approaches, potentially unfairly underrating them. It is important to distinguish this problem from another problem in social media analytics, namely, the ephemeral nature of some data collection approaches and research findings. The ever-changing nature of the social media landscape means that data sets from different time periods may give very different answers to the same research questions (Stieglitz et al., 2020), and scholars may find that it has become impossible to ask the same questions of new data because the necessary API no longer exists (Bruns, 2019), the researcher's code no longer works, or results are outdated by the time the paper is published (Munger, 2019).
3. *Comparison with others becomes impractical or outright impossible*: Considering the above problems regarding the access and sharing of data, it is not surprising that very little comparison of analytic methodology occurs in the literature. Even limited comparisons require a significant time to be carried out and researchers often prefer to invest time in developing new techniques rather than on comparing existing ones. Measures are defined and are usually not compared directly but instead are validated through the use of increasingly larger data sets, except in those few cases where research groups are large enough that they can collate and publish on such data sets and share them internally (e.g., Pacheco et al., 2021). As a consequence, approaches devised by researchers tend to remain isolated and are largely self-validated. Using these methods to support research on societal models or to justify political decisions is at least daring if not dangerous.

In the long run, the lack of data sharing, reproducibility, comparability, and competition based on common problems with common data sets (also referred to here as benchmarking) will lead to a loss of confidence in this research field and its results, and we could call this our own "benchmarking crisis" or "methodology crisis." Without any methodological foundation with known metrics to compare on common benchmark data, generalization is impossible and the whole research area may become unreliable.

Although the related term "reproducibility crisis" has been mentioned many times before (Baker, 2016; Bruns, 2013; Hutton & Henderson, 2015; Pasquetto et al., 2020), besides regulatory proposals, no practical solution for the specific problem facing social media analytics has been provided. Some authors have previously offered high-level frameworks, as well as workflows (Hutton & Henderson, 2018), for data documentation or scientific guidelines (ACM, 2020) for reproducibility in general, but a solution for the specific data exchange problem and a framework for benchmarking social media analytics tools is still lacking. As such, apart from clearly identifying this problem, our work provides a conceptual model that provides the means for *establishing benchmarks as the data-related foundation of social media analytics*. Clearly, other research communities that rely on empirical data use benchmarking approaches to have a common foundation. Thus, before we present our proposed approach, its requirements, the conceptual model, and its specification, we will review several approaches to data sharing and enabling comparability in related scientific areas.

Benchmarking Approaches in the Wild

From the computer science perspective, benchmarking describes the continuous comparison of processes and methods to gold standard approaches in order to systematically close the performance gap of methods.[9] The basic principle is to identify differences to the state-of-the-art and to provide opportunity for improvement. The classical process of benchmarking comprises (a) the selection of a method (product, process, etc.) to be compared, (b) the selection of competitors, (c) the acquisition of comparison data, (d) the analysis of performance gaps and the identification of problem sources as well as (e) the identification of improvement steps (which could then be implemented).

The ability to benchmark computational methods requires that all of these five steps can be performed and that comparison is enabled by data and method exchange, while the shortcomings or superiority of methods are measurable with performance indicators. We have already discussed that the sharing of data is one of the major obstacles in social media analytics benchmarking, as is the sharing of algorithms and processes in other domains. As such, the topic of benchmarking is not at all new in computer science, however, there are varying approaches to different scenarios. In the following, we address some of these aspects from different areas in computer science—from optimization research to AI algorithm development. Note that this is not a comprehensive survey on all available benchmarking approaches from all areas in computer science but rather a collection of spotlights that highlight different issues in benchmarking. The selection of topics is based on fields the authors have worked in. As such, these areas also inspired the framework presented in this work.

Open Science

In the aftermath of the reproducibility project[10] and the resulting failure of many studies to be replicated, the growing need to make scientific endeavors more transparent and reproducible was identified. One approach was to identify central repositories for storing project information, data, and analysis code. Most famously, the Open Science Framework (OSF[11]) provides a service targeted at researchers that allows the sharing of different project parts with different collaborators, reviewers, or the public, each with different levels of access. A core benefit lies in the ease of use and the research-centric processes. For example, the OSF provides specific means for sharing projects for review, that is, projects are shared anonymously and persistently but are only available through a private link. Another feature is the ability to fork (or duplicate) projects to facilitate replication efforts. Many digital analysis tools interact well with the data API of the OSF (Calero Valdez, 2020). Another important criterion for choosing a repository is the financial operating plan. Here, the OSF has secured funding to operate for more than 50 years as of today.[12] Several other repositories exist for other research fields (e.g., nature maintains a collection of repositories[13]).

However, many of these repositories are data-centric and focus predominantly on sharing project data over code and other information. The repositories thus require sufficient rights to share the data.

An Early Benchmarking Effort

Around 1980, a series of tournaments was held in the interests of studying the evolution of persistent cooperative behavior through computational simulation (Axelrod & Hamilton, 1981). In these tournaments, strategies (algorithms) were pitched against one another to win an iterated game of the Prisoner's Dilemma, at a time when traditional game theory did not have a way to explain the altruism exhibited by many species, where individuals would act against their own interests to help others. One iteration of the game pits two individuals against each other, who can either cooperate or defect. If they both cooperate, they both benefit moderately (e.g., they each receive five points); if one defects and the other cooperates, the defector benefits greatly (e.g., they receive 10 points) but

the cooperator receives nothing; and if they both defect, neither receives anything. In the absence of iteration, the obvious strategy is to defect in the hope that the other individual cooperates. In the iterative scenario, much more representative of creatures surviving in communities in nature, other strategies can be more effective over time. By inviting other researchers to submit strategies, it was determined that tit for tat (cooperate in the first round, then do whatever the other did in the previous round) was the most effective strategy. A number of valuable contributions were made to the field of evolutionary biology based on these studies, including the description of a natural mechanism by which altruism could be explained, which further contributed to other theories (Dawkins, 1989).

In this scenario, the benchmark element is the common platform on which the strategies were executed and compared, and the performance metrics were the results of the individual competitions. The Prisoner's dilemma has a simple specification, so it would have been commensurately easy for participants to ensure their strategies complied with the rules of the game. Conceptually, however, this scenario shares many elements with the current requirements for algorithm benchmarks despite the extra requirements imposed by modern computational tasks, such as classification and optimization.

Benchmarking of Algorithms in Optimization

The area of benchmarking in optimization is crucial to compare algorithmic approaches, and a wide range of benchmarking competitions have been established in the optimization literature. Benchmarks set standards in terms of areas of interest for given problems where algorithms are desired to perform well (Ansótegui et al., 2019) and should contain important properties of considered optimization problems (Zamuda et al., 2018). They allow the capture of different characteristics of underlying problems and distinguish algorithms in terms of properties where they perform well or poorly. For example, in the area of satisfiability, SAT competitions (Audemard et al., 2020) have significantly advanced the knowledge on the performance of SAT solvers and led to significant algorithmic improvements over the years. Similarly, in the area of evolutionary computation, benchmark competitions at conferences, such as IEEE CEC[14] and ACM GECCO,[15] have set new standards in terms of comparability of algorithms and their evaluation on a large variety of problem classes. While the aforementioned competitions and benchmarks mainly focus on the comparability of algorithms on defined problem classes (i.e., the problem structure can usually be formulated as an equation and the benchmark data are no secret), other competitions specifically address the robustness of methods and approaches with respect to unknown and changing environments. In optimization, these benchmarks are termed as "black-box" problems and it is considered as important, that algorithm designers do not have insights into the problem characteristics and data. The BBComp[16] benchmark competition "aims to close this gap by providing an algorithm testbed that is truly a black box to participants." Thus, this testbed encapsulates a range of problems unknown to the algorithm designer. They are only informed about problem properties like the dimension and bounds on variables. For solving the problem, only a (usually small) budget of black-box queries is provided simulating real-world restrictions like costly evaluation of solutions. As such, we find a slightly related setting to our social media analytics scenario, as in social media analytics data sharing is restricted. However, this restriction is artificially created: The problem data is hidden behind a web service interface to regulate access from the outside.[17] However, all queried data (in this case, simple function values) are openly provided; they are in principle allowed to be shared, which would not be allowed for social media raw data provided by many platforms. The BBComp approach nevertheless provides a partial solution to our problem. We can encapsulate and standardize data access via an API for arbitrary methods. Additionally, we need to ensure that data are not leaving the local domain of the data holder.

Selected AI and Data Science Domains

In the following, we aggregate and address several scientific areas that strongly rely on learning by using large amounts of data and data streams. This comprises recommender systems, machine learning approaches, and AI methods in games.

Benchmarking in recommender systems. Recommender systems are software systems that help identify items that are relevant to the current user utilizing the previous choices by all users (Adomavicius & Tuzhilin, 2005; Resnick & Varian, 1997). They recommend suitable additions to products in an online shopping basket (e.g., at Amazon) or recommend movies to watch after finishing another (e.g., at Netflix). Since the performance of these algorithms directly influences business income, the performance of such algorithms is crucial to the owner of a recommender system. Nothing demonstrates this importance more than the 2006 Netflix prize. Netflix issued a US$1 million prize for the best recommender system outperforming their own algorithm. For this purpose, Netflix shared 2 gigabytes of user data in the form of user-generated movie ratings.[18] Participants would submit their algorithm, which they could test on a subset of the data. At the end of the competition, all algorithms were tested on the full data set. Interestingly, the main finding from this experiment was that the core metric used in benchmarking (i.e., accuracy) was not the most useful metric for a real-life recommender system. The best algorithm was able to very accurately predict the ratings of bad and hated movies, which contributes to a high accuracy but provides no real business value (McNee et al., 2006). New metrics were designed.

While such data are seemingly harmless to share with the public, risks of identity disclosure from harmless data are known for recommender systems (Ramakrishnan et al., 2001). Other data sets have since been added to the standard benchmarking suite of recommender systems, such as the MovieLens data, Jester joke data, last.fm music data, or Wikipedia data. Many others can be found on Kaggle. This platform even offers a feature where data holders can create their own competitions, by uploading the data and creating a task, which can be then solved by external participants (this however is currently limited to supervised tasks with predefined evaluation metrics.[19,20]

Data from recommender systems share many similarities with social media data. They are user-generated, they are time-dependent, and large data sets are owned by corporations. However, given the benefit to the field as a whole, companies have been willing to publicly release benchmarking data sets after anonymization. Releasing benchmark data sets for social media might be more complicated, though, since the core business of social media companies is the data itself, while that of companies such as Netflix and Amazon is based on product selling (hence the focus on effective product recommendation).

Another key difference in data from a recommender system lies in the ease of anonymization. Removing data columns from a recommender system that carry identifying information leaves plenty of other information (e.g., structural) for algorithms to analyze and thus to contribute to good performance. In social media, on the other hand, textual data can leave the author exposed, making it hard to algorithmically anonymize such data.

Game AI research. Since around 2017, it became clear that the heavily deep learning–oriented thriving branches of AI have to deal with reproducibility problems. For a few years, many deep-learning methods have been experimentally investigated and published with single runs (no repetition, no statistics). The main argument of the authors is usually that repetitions take too long because single runs already mean hours or days. Similar reasons had been used to do an experimental comparison of algorithms without proper methodology also in other areas of computer science, for example, in meta-heuristics/evolutionary optimization. After some years of discussion, researchers in that field have largely converged to a common toolset that basically consists of structured experimentation

and statistics (Bartz-Beielstein et al., 2010). The AI field is also moving gradually in a similar direction, starting with some authors complaining about the current situation and suggesting some fixes. As the Atari Learning Environment is one of the most popular game-related benchmarks of recent years, the work of Machado et al. (2018) is especially interesting in this respect.

Setting up benchmarks in game AI often means to provide a whole environment with problem instances and baseline algorithms connected by a defined API. All this is usually open sourced, enabling all potential users to check for mistakes, and test and use single components for themselves, thereby accelerating the speed of these developments enormously. Well-known examples in this direction are the OpenAI gym[21] and Deep Mind's OpenSpiel.[22] More specialized interfaces/benchmarks as PySc2 for StarCraft[23] target a single game only but set up a huge number of defined problems of different complexity levels and come with a large amount of data that can be used to learn from human approaches.

Two recent approaches of benchmarks that have actually been designed as competitions are especially interesting here:

- The Generative Design in Minecraft competition[24] provides data (in the form of levels) to the user who has to come up with an algorithmic approach in order to produce a visually appealing generated settlement, which is in turn rated by a jury.
- The Obstacle Tower Challenge (OTC[25]; Juliani et al., 2019) aimed to develop a flexible agent, which can handle a lot of different situations, such as frequently occur in jump-and-run games, from finding keys for doors to playing complex minigames inside the game (e.g., Sokoban). Participants were provided with test levels to set up their methods, and then were required to submit agents that were then benchmarked on slightly different levels on machines belonging to the competition host.

In contrast to the situation in social media analysis, data availability is not the major difficulty here, although, as in the example of PySc2, a huge bulk of human play data is exchanged. Anonymization may be much easier for game data, as text information is irrelevant for the analysis of single- or two-player games. The main issue to tackle is rather establishing a fair comparison. Runtime comparisons on different hardware and also on slightly different software platforms are inherently unfair. In the case of the OTC, this is resolved by collecting the agents and running all of them in one physical environment.

Computer vision. In computer science domains and especially in domains that benefit significantly from machine learning, in general, such as computer vision, the performance of new machine learning approaches and algorithms is always evaluated with publicly available benchmarking data sets. In computer vision, a plethora of standardized data sets exists that tackle different problems. From character recognition (MNIST), over object detection, and classification (ImageNet) to age prediction (IMDB-WIKI), data sets are freely available and can be accessed by researchers from all over the world (LeCun et al., 2010; Rothe et al., 2018; Russakovsky et al., 2015). There exist several competitions in this domain as well. However, in general, the data are usually available for download and only the target variable is kept secret. The Kaggle platform is used in this context as well.

(Automated) machine learning. The goal of Automated Machine Learning (AutoML; Hutter et al., 2018) is to create supervised algorithms that work well on unknown data sets and can tune themselves automatically without human intervention. Here, the situation is comparable to what was described in the previous section as restricted data access. Especially in this domain, a centralized competition platform exists, to which source code of external algorithms can be uploaded: CodaLab.[26] A researcher can use the CodaLab platform to create competitions, to which competitors

Table 1. Overview of the Functionality of Preexisting Benchmarking Frameworks in Other Computation Domains.

Benchmarking Framework	Data Disclosure	Hosting	Complexity	Extensible	Social Media Focus	Lightweight	Programming Language
Kaggle Comp	Yes	External	Low	No	No	Yes	Any
BBComp	Partly	Internal	Medium	Yes	No	No	Any
Codalab	Yes/no	Both	High	Yes	No	No	Any
Ours	No	Internal	Low	Yes	Yes	Yes	Any

submit their code for evaluation, thus ensuring the code is not manually fine-tuned against given data sets.

Natural language processing (NLP). NLP has a long history of multiple teams of researchers competing on challenges, often the prediction of human ratings (annotations) of text passages. These competitions are also known as shared tasks. An example of this is the SemEval series of tasks, which has run since 1998.[27] The organizers, who are often leading researchers in the field, acquire the annotated data and distribute it to the participants. If the data are sensitive, participants are usually asked not to share it with others, but the degree to which this can be enforced is limited. First, the participants receive data with annotations, later, they receive the test set without annotations. They then submit their predictions for the test data to the organizers who evaluate it and present the results. Today, CodaLab is often used for this evaluation step: Participants upload their output file to the centralized platform but no code.[28] Hosting a task requires considerable effort, but it is prestigious. Participants submit papers in which they present their approach, and the organizers summarize the results in a paper.

Benchmarking overview. To summarize the properties of the benchmarking approaches discussed here, Table 1 presents the capabilities of current benchmarking frameworks along the dimensions important for the social media domain and for solving the challenges, which we identified in the Problem Definition section. Each framework must first and foremost solve the third problem of enabling comparability. Thus, it is important to consider the aspect of data disclosure as well as the aspect of complexity. While the former is the prerequisite for formal compliance with existing licensing conditions, the latter aspect is necessary for wide acceptance and interdisciplinary use of the framework. The greater the effort required to install an instance of the framework, the more it will be confined to technology-savvy communities. This should clearly be avoided. Along with these requirements come hosting and lightweight framework necessities. Extensibility and maximum flexibility in programming language support will enable wide use in many disciplines of social media analytics.

With the establishment of such a framework—especially with the inclusion of nondisclosure and local hosting requirements—the previously defined problems of deviating sampling of data used for comparisons and their mutation over time (by the platforms and other influences) are at least partially solved. Archiving data once collected and making it available eliminates uncertain sampling influences when data are obtained by others. It simultaneously detaches the data from the temporal context of the collection. The data are therefore available unchanged in the state they were at the time of collection. Nevertheless, the decentralized hosting of the data makes it possible to offer different data sets, which may well vary in terms of sampling, as a basis for benchmarking. The resulting diversity of data samples can be used to evaluate and increase the robustness of analytical approaches.

With this discussion in mind, Codalab can be considered as the only existing suitable candidate from Table 1. This is because social media data currently cannot be disclosed to external parties and thus cannot be uploaded to an external web service; Kaggle Comp and BBComp do not offer a solution to this problem. Codalab offers the option for the researcher to host the complete platform on their own infrastructure, thus entirely within their own domain. However, as mentioned before and indicated in the table, the framework is highly complex to manage and is equipped with a plethora of features that are simply not relevant for the current task: the indirect supply of social media data that cannot be shared.

Thus, in the following, we propose a specialized but flexible and lightweight framework, which addresses the discussed problems and enables social media researchers from the computer and social sciences to share data for comparison and benchmarking.

Framework Requirements

Before we specify the requirements for a framework that aims to overcome the issues discussed above, we specify the important roles that participate in such an environment. First, the *data holder*, as the name implies, has access to data that cannot be shared. In our proposed framework, this role is responsible for hosting the evaluation infrastructure. Second, there exists a role representing those who want to run an algorithm on the data without having access to it. We refer to this role as the *researcher*.

Why are the existing approaches not sufficient to overcome the issue presented? As discussed in previous sections, data accessibility in social media analytics is the limiting factor, similar to the situation in the AutoML domain. In many of the existing competitions, for example, in NLP, all participants are given access to the training data. In cases where this should be avoided, such as in the AutoML example, high setup costs are involved. Setting up the CodaLab platform, for example, on one's own server is far from simple and requires complex configuration. It has many features that might be useful when the goal is to host a competition for dozens of simultaneously competing participants, but the use case we address is fundamentally different.

Keeping both sides—the researcher (who does not have access to the data) and the data holder—incentivized is another major challenge when it comes to the specification of such an evaluation platform. In order to motivate data holders to host an evaluation architecture, it should be easy to deploy and configure. This requirement holds for the researcher side, too. By keeping the overhead of additional configuration and implementation as low as possible, the incentive for researchers to push their algorithms to the platform can be maximized. Therefore, the framework should be as lightweight as possible so that it can be quickly deployed by many data holders in the domain.

Another fundamental requirement is the free choice of the programming language used by the researcher and, as a result, the independence of any development environment and module dependencies from the benchmarking system. While there are definitely some preferred programming languages and packages for scientific computing, there is still a plethora of algorithms (mostly baseline approaches) implemented in what is now considered legacy programming languages. Therefore, the framework should be constructed in a way that allows any software product to be utilized, without further restrictions, to the greatest extent possible.

While one of the major advantages of any such standardized framework is the evaluation of data without direct access, another positive effect is comparability between different methods or approaches for a given task. The creation of trust in the correctness and validity of the calculated metrics is of utmost importance to ensure widespread acceptance in the research community.

To summarize, an evaluation framework for social media analytics should fulfill at least the following requirements:

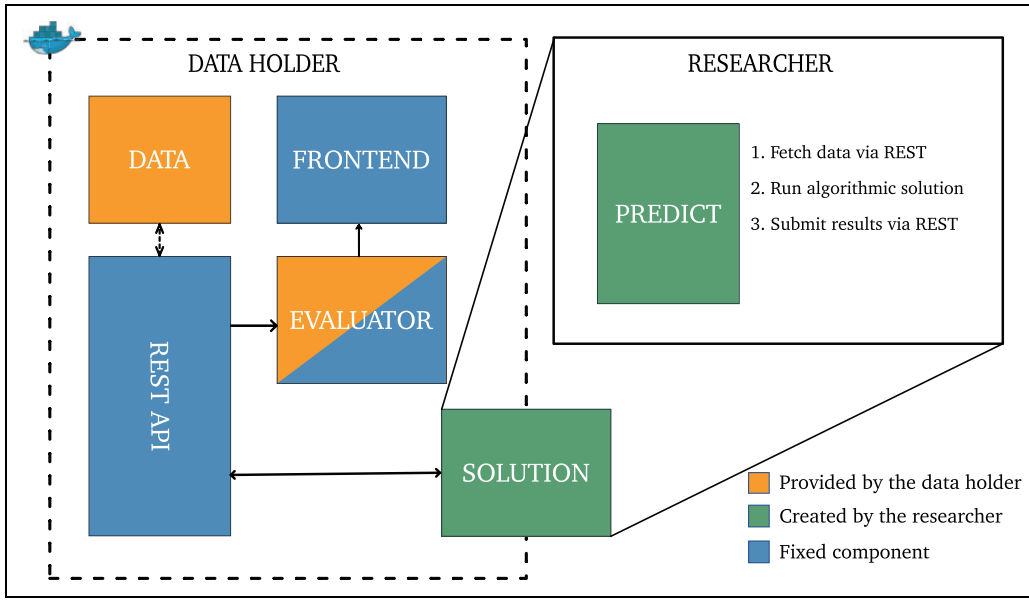

Figure 1. Proposed evaluation framework.

1. It should be easy and intuitive to set up, configure, and deploy on the data holders side.
2. There should be minimal overhead for researcher engagement (including implementation and submission to evaluation platforms).
3. There should be few restrictions on the researcher's choice of programming language and libraries.
4. The evaluation platform environment should be safe and should ensure data integrity.
5. The trustworthiness of evaluation results should be ensured.

Proposed Framework

To fulfill the requirements above while minimizing additional implementation effort, we utilize operating system–level virtualization techniques. Figure 1 displays the core components of the proposed architecture from both the data holder and researcher perspectives. We now briefly describe the overall architecture and each of its components. Then, we provide an overview of how the framework is intended to be used in practice by examining a workflow example.

Framework Architecture

The data holder side consists of four components, which can be categorized into three classes. The first class includes the fixed components, highlighted in blue, which are preimplemented modules that will rarely need to be customized. These are the front end and the REST (Representational State Transfer) API service. The REST API service implements a unified interface through which the researcher's solution component accesses data and arranges for the evaluation of its results—that is, it provides the single communication point through which information can be exchanged between the researcher's solution and the data holder. The front end is a web-based dashboard that serves multiple purposes. First, it facilitates the submission of new solution components by the researcher.

We use "solution" to refer to a Docker image which includes all the code, libraries, and dependencies that are needed to run the client's algorithm. Second, it provides a central location for visualizing the results of evaluating researchers' submissions. This could be a leaderboard with different evaluation metrics (e.g., accuracy, recall, precision). Again, both of these components are regarded as clearly defined noncustomizable components that rarely need changing.

The second class includes two customizable components hosted by the data holder: the evaluator and the data. The evaluator is responsible for conducting the evaluation task on the specified data set (in case more than one data set is available). Obviously, the data holder is not only responsible for defining the evaluation task but also for specifying which metrics are appropriate. While the specification of data formats for submission and the evaluation strategy are the responsibility of the data holder, we want to highlight that there are several standard tasks in social media analytics that can be implemented in advance, leaving only the specification of the underlying data set to be customized. A common task is supervised classification (e.g., classifying bots and human-driven accounts and the detection of hate speech). We argue, therefore, that the evaluator should be regarded as a semi-customizable component. For some problems, the data holder just has to "switch" the underlying data set, while for more unique tasks (e.g., measuring the extent of inauthentic coordination), additional programming effort may be required to define the evaluation procedure.

The third class holds the solution component, which has to be specified by the researcher. Using the APIs available on the data holder side, it accesses the data holder's data sets, and executes its algorithm on the retrieved data, and then arranges for the evaluation of its algorithm's results on separate validation data (also made available by the data holder). In principle, the solution component is a Docker container that holds an implementation of an algorithmic task-solver, the associated dependencies, which is provided with a means to communicate with the RESTful services described above when the container runs. No direct communication occurs between the evaluator and the solution. Data retrieval and submission of predictions for evaluation will only be allowed through the RESTful service. There are several reasons for this method of communication; most importantly, using REST allows the researcher solution to be written in any programming language as long as it can make HTTP requests. Additionally, it defines the methods to access and submit solutions, making the implementation and associated configuration as simple as possible. Other than the API, the only knowledge the solution component requires is the address and port of the server component within the provided virtual network (and this information can be predefined or introduced via runtime parameters).

Example Workflow

Any time there are research questions that require access to a privileged data set to be answered, our proposed architecture could be employed. Social media data sets fit that criteria, but the applicability of this approach is much broader. We now provide an example workflow based on a simple mock use case representative of the kinds mentioned as anecdotes. In our example, we assume that a research group has used the Twitter API to crawl public tweets and has classified them as hate or nonhate speech via a crowdsourcing annotation study, thereby producing a valuable labeled data set. Moreover, they have also developed a new deep-learning approach that automatically (without human intervention) tunes itself and classifies the data correctly, outperforming existing state-of-the-art algorithms on this new ground-truth data.

The research group plans to publish the new approach in an internationally recognized journal. Therefore, they make the source code available to the research community on a code repository platform such as Github. To prove the validity of their claims, the research group now has to show that their results are (a) reproducible, that is, other research groups can recreate the computed

performance results and that (b) the proposed solution outperforms existing approaches on the data. In this scenario, both requirements cannot be fulfilled simply due to the fact that the research data cannot be shared with the community. Consequently, potential reviewers of the article have to put unconditional trust in the validity of the claims.

Our proposed framework can be used by the research group (now: data holder) to facilitate this desired comparability and increase the trust of the conducted research. The data holder therefore will need to install the data set and customize an appropriate evaluator. Since the task itself is a textbook example of a supervised binary classification problem, a predefined evaluator template can be used and the only configuration needed specifies the structure of the given data set. Next, the framework has to be started at the data holder's site. This launches the front end, the evaluator, and the RESTful server providing the API. All components are hosted in individual virtual Docker containers, and the Docker environment facilitates communication within a virtual and isolated network. Only the front end needs to be externally accessible to permit the submission of new solutions and allow inspection of the current leaderboard.

After the data holder environment is deployed, (external) researchers can submit their solution components via a web interface. Subsequently, the platform arranges the solution components (i.e., the submitted Docker containers) to be run, so that it can train and test its algorithm on the data as necessary. When deployed on the data holder side, in the last step, the solution component locally fetches the validation data from the API and evaluates the trained model/solution on it. After evaluation, the results and performance metrics are passed back to the solution submitter and posted to the leaderboard with the consent of the submitter. Besides the fact that the new data set is now made implicitly available to the research community, a profound advantage of this procedure is that external researchers can now quickly reproduce the proposed method's results by downloading the code from Github and submitting it to the platform. Under the assumption of a broad acceptance of our approach, each platform instance can now be considered as a representation of the status quo (as well as the complete history) of task performances on specific data sets. This makes the overall academic process much easier, as the scoreboard (with the access date) can be cited as proof of the proposed approach's superiority. In the long term, this also may lead to a situation where it is considered good academic practice (and even required by high-tier outlets) to provide such proof within submitted publications.

The complete process of solution submission and deployment is shown in Figure 2. A more in-depth specification of the proposed prototype, including a blueprint of the RESTful API as well as the evaluator component, can be found in the Online Appendix.

Discussion and Perspectives

In this work, we shed light on a problem that is well known by almost every researcher working in the field of social media analytics: the limited ability of researchers to meaningfully compare computational methods due to restrictions when it comes to data sharing. We illustrated this issue by describing the resulting limitations in various subdomains of the field. A further investigation of how benchmarking and data comparability are handled in other computational research areas revealed that the problem of such restricted access to data is almost unique to social media analytics. As an exception, we found that in the field of automated machine learning, competitions are organized in such a way that competitors do not have direct access to data. In those competitions, code is externally executed by the data holder—usually, a centralized infrastructure that hosts the competition. We adopted the core idea and conceptualized a distributed framework that enables benchmarking without data exchange. Instead, data holders encapsulate data in a lightweight environment that allows the evaluation of the respective data by external methods without publishing the data. However, we are still at the beginning of dealing with this inherent domain problem. Although

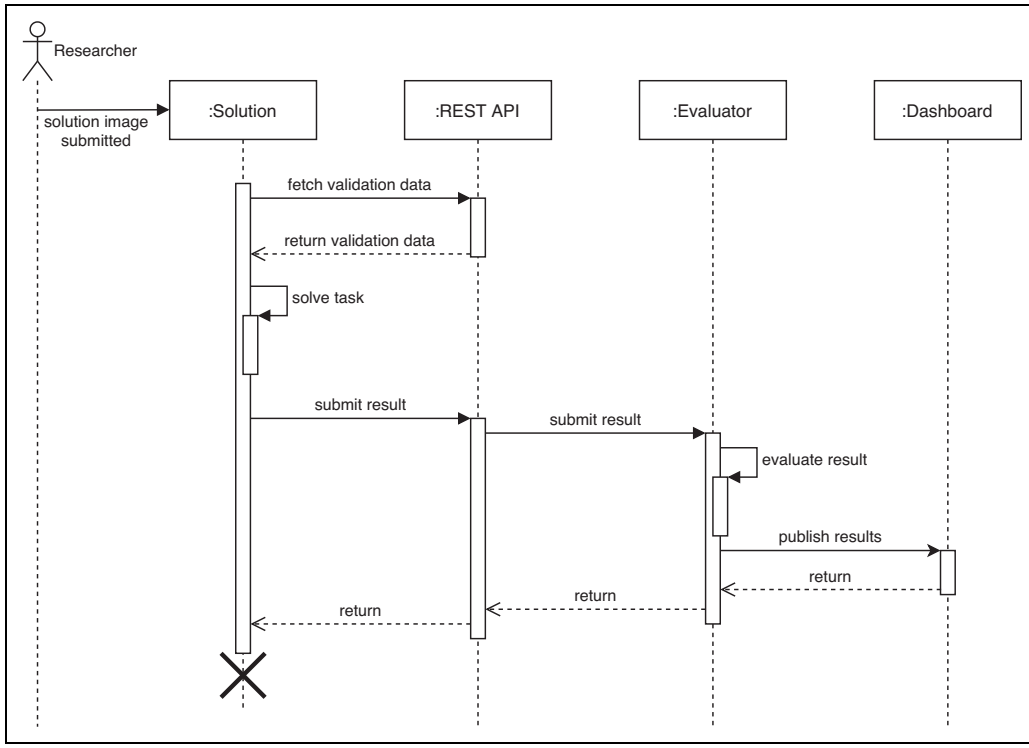

Figure 2. Sequence diagram to submit a new algorithm.

our proposed solution tackles a plethora of the identified issues, it is still a prototype that does not solve the whole problem. Therefore, it is necessary that the community adapts and develops the concept. Here, we reflect on the remaining issues and discuss implications and consequences.

Creating Trust

One of the main concerns that comes with our proposed distributed evaluation architecture is the creation of trust into the validity of the results generated at each data holder's site. Since data science is a competitive field and the success of new algorithms is highly dependent on how well they perform, there may be an incentive for researchers hosting their own data sets to manipulate evaluation results, such that their algorithm performs best when compared with other approaches. Certainly, it is almost impossible to avoid such individual manipulation attempts or to prevent them. However, our framework's terms of use can require component customization and other modifications to be made public. We argue that, in any case, with increased acceptance of the framework and a plethora of instances, hosted by different research groups, a network of trust will emerge organically. If an algorithm only works well on data hosted by the developer of the method, this will raise questions of overfitting, whereas an algorithm that performs well on multiple instances of our framework will be more likely to be trusted.

Conceptual Issues

The proposed concept may raise various detailed questions that require attention in future work, but also in the concrete implementation (beyond the prototype).

One key issue with our framework is the limited capacity to distribute the computational load. Since the data cannot leave the data holder's side, all computations have to be done on hardware that is owned by the data holder. Especially for computationally intensive tasks, this remains a problem that is beyond the scope of this work. The issue of computation time is also strongly connected with the issue of creating incentives for all participants. Why should the data holder execute external algorithmic approaches on their machines, when these are created by different research groups (which would be equivalent to competitors)? Science can be incentivized by measures that fit the Open Science approach discussed before. It could be requested that used benchmarking data be cited in the work on new methods. On the one hand, this would lead to the corresponding scientific recognition of infrastructure operation and data collection. On the other hand, this is also a mechanism for identifying particularly suitable benchmarks within the community and rewarding their continued existence. At the same time, the distributed operation of benchmarking instances according to the proposed concept relieves each individual scientist of the burden of archiving vast amounts of benchmarking data and still keeping and maintaining them for later verification of results (even if not publicly accessible).

Additionally, there is an incentive for social media platforms themselves to host such infrastructure: Besides benefits for marketing, it may also provide opportunities for not losing control of the narrative depending on what researchers discover and how they report it. A model similar to that used by military and national security think tanks forming the link between the military and national security agencies as well as industry has been suggested (Persily & Tucker, 2020) but with the specific requirement that researchers need to maintain their independence from any social media platforms to whose data they get privileged access (Persily & Tucker, 2020, pp. 325–326).

Of course, we cannot deny that this benchmarking concept was designed with specific use cases in mind. It is therefore an important task to examine the concept for further use cases and at the same time to suggest suitable measures for evaluating results. For example, it is certainly easier to propose benchmarks for binary classification (e.g., for the detection of social bots) than to include campaign detection in a suitable benchmark, since there is usually no ground truth available. Nevertheless, suitable metrics can be defined to compare achieved results. It is then necessary to agree on a multitude of metrics rather than a single metric, which can be used in the community to evaluate such use cases.

In theory, it is also possible in our framework to train models on external data without the data being shared. If there is a sound basis for this (with regard to terms and conditions as well as ethical and privacy-related considerations), the trained models may be returned and further enhanced with more specific training. Naturally, it must be ensured that the trained models cannot be reverse engineered and do not consist of the identity function, which would simply reveal the raw data.

Legal Issues

The approach described in this article will overcome licensing limitations from the owners of data sources (including but not limited to social media platforms), which prevent data sets collected from them from being shared with third parties. Once collected, the data are required to remain with the data holder and not be shared or transferred, as this will contravene the terms of collection.

Any researcher use of the architecture will need to be subject to some basic terms and conditions of access, at the very least to ensure the data source's conditions are maintained. If the data set is subject to ethical restraints on its use (which will depend on the manner of its collection), the restraints will need to be made public and the researcher self-evaluate their compliance with the restraints or sign up to the relevant ethics protocol, as appropriate, potentially as part of a registration process. Manual data holder-side evaluation of applications to join the ethics protocol (i.e., case-by-case assessment by a person) may be required, and this will increase the overheads imposed

on the data holder (or at least their institution). Similarly, privacy restraints on access to or use of the data need to be identified and, again, the researchers self-assess compliance. For particularly sensitive data, this may have to take place manually or by negotiation between researchers. Compliance with privacy rules is complicated by cross-jurisdictional access—what may be acceptable in a researcher's location may not be permitted at the data holder's location.

If the data are configured in a way that hides sensitive portions, then any analysis must respect that configuration and the researcher must agree not to act in a way to expose that sensitive data. A prohibition on deanonymizing deliberately anonymized data would be an example of this.

Any other restraints on the data set as collected will need to be included in the terms of access and notified to the researcher and the researcher agree to honor the limitations. This can be done contractually through a registration process and can be supported by automatic code analysis to check that submissions adhere to technical limitations (e.g., not attempting to open sockets to hosts other than the data holder). The data holder's computational infrastructure may also have conditions to be imposed on access, such as rate of access, times of access, storage and processing limits, and constraints from the hosting institutions.

Going Beyond the Proposed Concept

The proposed concept offers a much-needed step in the direction of more transparent, comparable, that is, replicable science on social media analytics. Of course, many of our suggestions require further discussions in the community. However, we are convinced that there is broad agreement in the community on the necessity of a common benchmarking approach. While some properties of the concept may need further development, we encourage the community to participate in an agile fashion pushing this concept toward an accepted benchmarking environment that can constantly adapt to changes in regulations or access rules of social media platforms. We expect that the further development of this concept will have to address four major challenges (besides the issues described above):

1. *Trust:* We have discussed several ideas to prevent abuse in such a system. As there are multiple risks for abuse in an environment that may run (potentially) arbitrary code, security and data protection issues will be of central importance to this framework. However, not every person on the globe would require access to such a system. A trust-based approach using a central registry—possibly using ORCID as IDs—seems feasible, also providing a citeable reference for achieved performances on various data sets. Alternatively, assuming some homomorphic encryption algorithm exists for a given problem, distributed processing in a group of institutional supercomputers is also feasible.
2. *Resources and costs:* We proposed a distributed benchmarking framework that depends on the willingness of data holders to contribute their data to the community. Therefore, the lightweight character of the software setup and incentives for investing resources into the evaluation of external approaches are necessary. In addition to the technical setup of a benchmarking server, the execution of submitted algorithms may be costly and thus has to be rewarded with access to other benchmarking environments. This will automatically lead to a future discussion of centralized versus decentralized arrangements. Both have to be evaluated, however, we believe that the incentives and business models for a centralized architecture should be identified first, while decentralized collaboration can be framed as a longer term grassroots movement to advance scientific methodology.
3. *Institutionalization:* Considering aspects of centralization for the proposed benchmarking infrastructure, scientific institutions may collaborate to establish a common social media data grid that enables the distributed use of individually collected data in combination with

available institutional (and possibly idle) computing resources. However, this approach potentially also provides a model for the social media platforms themselves to implement server infrastructure for making discreet and fixed data sets available to researchers while retaining overall control of the data, potentially saving the original researcher engaging in collection in the first place. Researcher then merely would need to specify the query criteria and the platform could automatically construct the appropriate results from its data holdings. This proposition is effective only if the social media platform creates snapshots of the data set; otherwise, the original issue of nonreplicable data remains, as the platform curates its broader data accounts (e.g., through deleting posts and users as appropriate), and could be hard to distinguish from the monetized tiers of data access. In addition, the commercial approach carries the risk of a renewed lack of transparency. Moreover, the control over the benchmarking data lies exclusively with the respective platform, which again leads to a dependency of the scientific data basis on commercially operating companies.

4. *Unification:* By creating a unified access path to individually held data, the developed framework directly contributes to the comparability of methods and thus to a normalization of the research environment. At the same time, however, the proposal—assuming its global use—can also lead to a harmonization of ethical and regulatory protocols, which up to now IRBs have had to consider on a case-by-case basis. In the short term, it can eliminate the need to release data for data exchange, leading to less time-consuming preparation for (usually very restrictive) data exchange. Since the data provided never leave the local domain, it is sufficient to comply with the individual regulations and protocols for data collection. In the medium term, we could envision and speculate on an evolutionary process leading to the homogenization of IRB policies. Assuming widespread adoption of the framework, the policies of the most accepted providers (based on the most data and best information) will prevail and may determine IRB policies with respect to the collection of social media data. If nothing else, the incentive underlying this evolutionary process could be a citation requirement of data used for a comparative study.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dennis Assenmacher, Christian Grimme, Mike Preuss, and Heike Trautmann acknowledge support by the German Federal Ministry of Education and Research (FKZ 16KIS0495K), the Ministry of Culture and Science of the German State of North Rhine-Westphalia (FKZ 005-1709-0001, EFRE-0801431, and FKZ 005-1709-0006), and the European Research Center for Information Systems (ERCIS). Dennis Assenmacher and Christian Grimme are additionally supported by the DAAD PPP Germany–Australia 2020 project ID 57511656. Stefano Cresci acknowledges funding by the EU H2020 Program under the scheme INFRAIA-01-2018-2019: Research and Innovation action grant agreement #871042 SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics.

Supplemental Material

The supplemental material is available in the online version of the article.

Notes

1. For example, Twitter's terms stipulate that a researcher may only share 50,000 complete tweets per day per individual recipient, or 1.5 m tweet IDs in a 30-day period (https://developer.twitter.com/en/developer-terms/policy).
2. https://github.com/fivethirtyeight/russian-troll-tweets/
3. https://about.twitter.com/en_us/advocacy/elections-integrity.html#data
4. https://botometer.iuni.iu.edu/bot-repository/datasets.html
5. https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research#Twitter_and_tweets
6. https://pushshift.io/
7. https://developer.twitter.com/en/developer-terms/agreement-and-policy
8. For example, https://github.com/lintool/twitter-tools/wiki/TREC-2015-Track-Guidelines
9. Clearly, the understanding of benchmarking is similar for the comparison of products, services, and workflows in economy.
10. https://osf.io/ezcuj/
11. http://www.osf.io
12. As noted on the Open Science Framework FAQ, as of February 15, 2021, https://help.osf.io/hc/en-us/articles/360019737894-FAQs
13. https://www.nature.com/sdata/policies/repositories
14. https://wcci2020.org/competitions/
15. https://gecco-2020.sigevo.org/index.html/Competitions
16. https://www.ini.rub.de/PEOPLE/glasmtbl/projects/bbcomp/index.html
17. https://www.ini.rub.de/PEOPLE/glasmtbl/projects/bbcomp/documentation.html
18. The data are now available at https://www.kaggle.com/netflix-inc/netflix-prize-data
19. https://www.kaggle.com/search?q=recommender
20. https://www.kaggle.com/c/about/inclass/
21. https://gym.openai.com/
22. https://deepmind.com/research/open-source/openspiel
23. https://github.com/deepmind/pysc2
24. http://gendesignmc.engineering.nyu.edu/
25. https://www.aicrowd.com/challenges/unity-obstacle-tower-challenge/
26. https://github.com/codalab/codalab-competitions
27. https://semeval.github.io/
28. http://alt.qcri.org/semeval2020/index.php?id=codalab

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749.

Al-Rawi, A. (2019). Gatekeeping fake news discourses on mainstream media versus social media. *Social Science Computer Review*, *37*(6), 687–704.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat ($p > 0.05$): Significance thresholds and the crisis of unreplicable research. *Peer Journal*, *5*, e3544.

Ansótegui, C., Bonet, M. L., Giráldez-Cru, J., Levy, J, & Simon, L. (2019). Community structure in industrial SAT instances. *Journal of Artificial Intelligence Research*, *66*, 443–472.

Assenmacher, D., Adam, L., Trautmann, H., & Grimme, C. (2020). Towards real-time and unsupervised campaign detection in social media. In R. Bartak & E. Bell (Eds.), *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference* (pp. 303–307). AAAI Press.

Assenmacher, D., Clever, L., Pohl, J. S., Trautmann, H., & Grimme, C. (2020, July 19–24). A two-phase framework for detecting manipulation campaigns in social media. In G. Meiselwitz (Ed.), *Social computing and social media: Design, ethics, user behavior, and social network analysis—12th International Conference, SCSM 2020* (Held as part of the 22nd HCI International Conference, HCII 2020, Proceedings, Part I, Volume 12194 of Lecture Notes in Computer Science; pp. 201–214). Springer.

Association of Computing Machinery (2020). Artifact review and badging. https://www.acm.org/publications/policies/artifact-review-and-badging-current

Audemard, G., Paulevé, L., & Simon, L. (2020). SAT heritage: A community-driven effort for archiving, building and running more than thousand SAT solvers. In L. Pulina & M. Seidl (Eds.), *SAT (Volume 12178 of Lecture Notes in Computer Science*; pp. 107–113). Springer.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*(26), 353–366.

Bartz-Beielstein, T. (2006). *Experimental research in evolutionary computation—The new experimentalism. Natural Computing Series*. Springer.

Bartz-Beielstein, T., Chiarandini, M., Paquete, L., & Preuss, M. (Eds.). (2010). *Experimental methods for the analysis of optimization algorithms*. Springer.

Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38–54.

Bruns, A. (2013). Faster than the speed of print: Reconciling "big data" social media analysis and academic scholarship. *First Monday*, *18*(10). https://doi.org/10.5210/fm.v18i10.4879

Bruns, A. (2019). After the "Apicalypse": Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566.

Calero Valdez, A. (2020). Making reproducible research simple using RMarkdown and the OSF. In *International conference on human-computer interaction* (pp. 27–44). https://doi.org/10.1007/978-3-030-49570-1_3

Carnein, M., Assenmacher, D., & Trautmann, H. (2017). Stream clustering of chat messages with applications to twitch streams. In S. de Cesare & F. Ulrich (Eds.), *Proceedings of the 36th international conference on conceptual modeling (ER'17)* (pp. 79–88). Springer International Publishing.

Claerbout, J. F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. In *SEG technical program expanded abstracts 1992* (pp. 601–604). https://doi.org/10.1190/1.1822162

Cockburn, A., Dragicevic, P., Besancon, L, & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79.

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 61–72.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter. *ACM Transactions on the Web (TWEB)*, *13*(2), 1–27.

Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2018). From reaction to proaction: Unexplored ways to the detection of evolving spambots. In *The web conference 2018 (WWW'18)* (pp. 1469–1470). https://doi.org/10.1145/3184558.3191595

Da San Martino, G., Cresci, S., Barrón-Cedenõ, A., Yu, S., Di Pietro, R., & Nakov, P. (2020). A survey on computational propaganda detection. In *The 29th international joint conference on artificial intelligence (IJCAI'20)* (pp. 4826–4832). https://doi.org/10.24963/ijcai.2020/672

Dawkins, R. (1989). *The selfish gene*. Oxford University Press.

Fair, G., & Wesslen, R. (2019). Shouting into the void: A database of the alternative social media platform gab. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 608–610.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, *22*(8). https://doi.org/10.5210/fm.v22i8.8005

Fineberg, H. V., Allison, D. B., Barba, L. A., Chong, D., Freire, J., Gabrielse, G., Gatsonis, C., Hall, E., Jordan, T. H., Scheufele, D. A., Stodden, V., Wilson, T., & Wood, W. (2019). *Reproducibility and replicability in science*. National Academies Press.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 0–30.

Grimme, C., Assenmacher, D., & Adam, L. (2018). Changing perspectives: Is it sufficient to detect social bots? In G. Meiselwitz (Ed.), *Social computing and social media: User experience and behavior* (pp. 445–461). Springer International Publishing.

Grimme, C., Assenmacher, D., Adam, L., Preuss, M., & Stockdiek, J. F. H. L. (2017). *Bundestagswahl 2017: Social-media-Angriff auf das #kanzlerduell? (Technical report 2017.1). Project PropStop*. www.propstop.de.

Grimme, C., Preuss, M., Adam, L., & Trautmann, H. (2017). Social bots: Human-like by means of human control? *Big Data*, *5*(4), 279–293.

Hagen, L., Neely, S., Keller, T. E., Scharf, R., & Vasquez, F. E. (2020). Rise of the machines? Examining the influence of social bots on a political discussion network. *Social Science Computer Review*. https://doi.org/10.1177/0894439320908190

Hegelich, S., & Janetzko, D. (2016). Are social bots on twitter political actors? Empirical evidence from a Ukrainian social botnet. In *International AAAI conference on web and social media* (pp. 579–582). AAAI Press. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015

Holzmann, H., Anand, A., & Khosla, M. (2018). Delusive PageRank in incomplete graphs. In *COMPLEX NETWORKS (1) (Volume 812 of studies in computational intelligence;* pp. 104–117). Springer. https://doi.org/10.1007/978-3-030-05411-3_9

Howard, P. N., & Kollanyi, B. (2016). Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. CoRR, abs/1606.06356. https://dx.doi.org/10.2139/ssrn.2798311

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2018). *Automated machine learning: Methods, systems, challenges*. Springer. http://automl.org/book

Hutton, L., & Henderson, T. (2015). Making social media research reproducible. In *Proceedings of the ICWSM workshop on standards and practices in large-scale social media research*. https://ojs.aaai.org/index.php/ICWSM/article/view/14685

Hutton, L., & Henderson, T. (2018). Toward reproducibility in online social network research. *IEEE Transactions on Emerging Topics in Computing*, *6*(1), 156–167.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Joseph, K., Landwehr, P. M., & Carley, K. M. (2014). Two 1%s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. In W. G. Kennedy, N. Agarwal, & S. J. Yang (Eds.), *Social computing, behavioral-cultural modeling and prediction* (pp. 75– 83). Springer International Publishing.

Juliani, A., Khalifa, A., Berges, V., Harper, J., Teng, E., Henry, H., Crespi, A., Togelius, J., & Lange, D. (2019, August 10–16). Obstacle tower: A generalization challenge in vision, control, and planning. In S. Kraus (Ed.), *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019* (pp. 2684–2691). http://ijcai.org.

King, G. (1995). Replication, replication. *PS: Political Science; Politics*, *28*(3), 444–452.

Kollanyi, B., Howard, P. N., & Woolley, S. C. (2016). *Bots and automation over twitter during the US election* (Technical report data memo 2016.4). Project on Computational Propaganda. www.politicalbots.org

Kumar, J., Shao, J., Uddin, S., & Ali, W. (2020). An online semantic-enhanced Dirichlet model for short text stream clustering. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 766–776). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.70

Lang, K. (1995). NewsWeeder: Learning to filter netnews. In A. Prieditis & S. Russell (Eds.), *Machine learning proceedings 1995* (pp. 331–339). Morgan Kaufmann.

LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST handwritten digit database. *ATT Labs*. http://yann.lecun.com/exdb/mnist

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*(1), 358.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, *5*(April), 361–397.

Machado, M. C., Bellemare, M. G., Talvitie, E., Veness, J., Hausknecht, M., & Bowling, M. (2018). Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, *61*, 523–562.

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting temporal patterns for botnet detection on Twitter. In *The 11th international ACM Web Science conference (WebSci'19)* (pp. 183–192). https://doi.org/10.1145/3292522.3326015

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on human factors in computing systems* (pp. 1097–1101). https://doi.org/10.1145/1125451.1125659

Metaxas, P. T., & Mustafaraj, E. (2012). Social media and the elections. *Science*, *338*(6106), 472–473.

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, Behavior, and Social Networking*, *16*(9), 708–713.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, & I. Soboroff (Eds.), *Proceedings of the seventh international conference on weblogs and social media, ICWSM 2013*. The AAAI Press. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6071/6379

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119859294

Neudert, L.-M. N. (2017). *Computational propaganda in Germany: A cautionary tale* [Technical report]. Project on Computational Propaganda. www.politicalbots.org

Niemann, M., Riehle, D. M., Brunk, J., & Becker, J. (2020). What is abusive language? Integrating different views on abusive language for machine learning. In *Proceedings of the 1st multidisciplinary international symposium on disinformation in open online media, MISDOOM 2019* (pp. 59–73). Springer. https://doi.org/10.1007/978-3-030-39627-5_6

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., & Tesconi, M. (2021). Coordinated behavior on social media in 2019 UK general election. In *The 15th international AAAI conference on web and social media (ICWSM'21)*. https://arxiv.org/abs/2008.08370v2

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media. In *The 15th international AAAI conference on web and social media (ICWSM'21)*. https://arxiv.org/abs/2001.05658v2

Paik, J. H., & Lin, J. (2015). Do multiple listeners to the public Twitter sample stream receive the same tweets? In *TAIA* (p. 4). https://cs.uwaterloo.ca/∼jimmylin/publications/Paik_Lin_TAIA2015.pdf

Pasquetto, I., Swire-Thompson, B., Amazeen, M., Benevenuto, F., Brashier, N., Bond, R., Bozarth, L., Budak, C., Ecker, U., Fazio, L. K., Ferrara, E., Flanagin, A., Flammini, A., Freelon, D., Grinberg, N., Hertwig, R., Jamieson, K., Joseph, K., Jones, J., & Yang, K. (2020). Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-49

Persily, N., & Tucker, J. A. (Eds.). (2020). *Social media and democracy*. Cambridge University Press.

Plesser, H. E. (2017). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, *11*, 76.

Ramakrishnan, N., Keller, B. J., Mirza, B. J., Grama, A. Y., & Karypis, G. (2001). Privacy risks in recommender systems. *IEEE Internet Computing*, *5*(6), 54.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58.

Riehle, D. M., Niemann, M., Brunk, J., Assenmacher, D., Trautmann, H., & Becker, J. (2020). Building an integrated comment moderation system—Towards a semi-automatic moderation tool. In *Proceedings of the HCI International 2020*. https://doi.org/10.1007/978-3-030-49576-3_6

Ross, B., Brachten, F., Stieglitz, S., Wikstrom, P., Moon, B., Munch, F. V, & Bruns, A. (2018). Social bots in a commercial context—A case study on SoundCloud. In *Proceedings of the 26th European conference on information systems (ECIS2018)*. https://aisel.aisnet.org/ecis2018_rip/52

Rothe, R., Timofte, R., & Gool, L. V. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2–4), 144–157.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787.

Stieglitz, S., Meske, C., Ross, B., & Mirbabaie, M. (2020). Going back in time to predict the future—The complex role of the data collection period in social media analytics. *Information Systems Frontiers*, *22*(2), 395–409.

Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, *14*(2), 1–39.

Timmers, M., van Dijck, J. T. J. M., van Wijk, R. P. J., Legrand, V., van Veen, E., Maas, A., Menon, D., Citerio, G., Stocchetti, N., & Kompanje, E. (2020). How do 66 European institutional review boards approve one protocol for an international prospective observational study on traumatic brain injury? Experiences from the CENTER-TBI study. *BMC Medical Ethics*, *21*, 36.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *International AAAI conference on web and social media* (pp. 280–289). AAAI. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817

Weber, D., Nasim, M., Mitchell, L., & Falzon, L. (2020). A method to evaluate the reliability of social media data for social network analysis. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 317–321). https://doi.org/10.1109/ASONAM49781.2020.9381461

Weber, D., & Neumann, F. (2020). Who's in the gang? Revealing coordinating communities in social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 89–93). https://doi.org/10.1109/ASONAM49781.2020.9381418

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: The problem of biased datasets. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (Vol. *1*, pp. 602–608). Association for Computational Linguistics.

Yin, J., Chao, D., Liu, Z., Zhang, W., Yu, X., & Wang, J. (2018). Model-based clustering of short text streams. In Y. Guo & F. Farooq (Eds.), *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, KDD '18* (pp. 2634–2642). Association for Computing Machinery.

Zamuda, A., Nicolau, M., & Zarges, C. (2018). A black-box discrete optimization benchmarking (BB-DOB) pipeline survey: Taxonomy, evaluation, and ranking. In H. Aguirre (Ed.), *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 1777–1782). ACM.

Zhang, W., Johnson, T. J., Seltzer, T., & Bichard, S. L. (2010). The revolution will be networked: The influence of social networking sites on political attitudes and behavior. *Social Science Computer Review*, *28*(1), 75–92.

Author Biographies

Dennis Assenmacher is a PhD student at the group of Data Science: Statistics and Optimization at the Department of Information Systems, University of Münster, Germany. His research focuses on social media analytics as well as supervised and unsupervised data stream learning.

Derek Weber is a senior Defense Scientist undertaking a PhD at the University of Adelaide, Australia. His research regards the amplification of influence through online coordinated behavior, particularly in discussions of political and social issues.

Mike Preuss is an assistant professor at LIACS, the computer science institute of Universiteit Leiden in the Netherlands. He is involved in different ways of using (game) AI algorithms for solving real-world problems, in disinformation, chemistry, applied optimization, and for improving games themselves.

André Calero Valdez is a senior researcher and research group leader at the Human-Computer Interaction Center at RWTH Aachen University. His research focuses on the interface between humans and AI in complex networked settings addressing emergent phenomena such as opinion formation, disinformation, and hate speech.

Alison Bradshaw is a commercial lawyer specializing in intellectual property and information technology. She has over 20 years' experience in the field of technology commercialization.

Björn Ross is a lecturer in Computational Social Science at the University of Edinburgh School of Informatics. In his research, he uses computational methods to study social media and related technologies. He was recently awarded the Stafford Beer Medal for the best paper published in the European Journal of Information Systems in 2019.

Stefano Cresci is a researcher at IIT-CNR in Pisa, Italy. His work mainly involves social media analysis, with a particular focus on information disorder, social bots, disinformation. He was recently awarded the IEEE Computer Society Italy Section Chapter 2018 PhD Thesis Award, the 2019 IEEE Next-Generation Data Scientist Award, and the 2020 ERCIM Cor Baayen Young Researcher Award.

Heike Trautmann is professor of Data Science: Statistics and Optimization at the Department of Information Systems, University of Münster, Germany. She is also director of the European Research Center for Information Systems (ERCIS) and head of the ERCIS competence center Social Media Analytics. Her group mainly focuses on automated algorithm selection and configuration, data stream mining, social media analytics, and (multi-objective) evolutionary optimization.

Frank Neumann is a professor and leader of the Optimisation and Logistics Group at the School of Computer Science, The University of Adelaide, Australia. In his work, he considers algorithmic approaches in particular for combinatorial and multi-objective optimization problems and focuses on theoretical aspects of evolutionary computation as well as high impact applications in the areas of renewable energy, logistics, and mining.

Christian Grimme is an associate professor for Information Systems at the University of Münster, Department of Information Systems. His research interests are computational propaganda detection in social media, large-scale data analysis, as well as decision support and optimization.