A Trust-Motivated Framework for Assessing Governments Engagement with Citizens on Social Media

Ilaria Bonavita†, Emily Maitland†
†Trilateral Research Ltd.
One Knightsbridge Green (5th Floor)
London, SW1X 7QA, UK
ilaria.bonavita@trilateralresearch.com, emily.maitland@trilateralresearch.com

Abstract
Social media can act as a sentinel to inform governmental agencies of citizens’ attitudes and to detect signs of distrust. A number of studies have used qualitative research methods to investigate the role of social media in building and maintaining trust relationships between governments and citizens. Identifying quantitative indicators of government trustworthiness on social media is a difficult and little-explored question. Although engagement metrics have been used to gather insights on governmental use of social media, they are usually not grounded on theoretical trust frameworks. In this study, we propose to characterise the governmental presence on social media through a set of indicators, computed directly and automatically from social media data, that can be mapped to dimensions of trust drawn from trust literature. An application of our framework on two public health authorities’ Twitter handles is also presented.

Introduction
Trust in official institutions by citizens is essential for the efficient and effective implementation of policies by a government. This is particularly relevant in times of crises and uncertainties as a decline in trust can lead to lower rates of compliance with the rules and regulations designed to face these difficulties (Bargain and Aminjonov 2020). The COVID-19 Pandemic has fostered governments to evaluate more closely their ability to establish a trustworthy relationship with their audience and to identify strategies to strengthen this relation (Abrams and Greenhawt 2020). The online channel has emerged as the main medium for governments to communicate with citizens with social media acting as a tool for information delivery, public sensing and reassuring citizens while securing their support (Malecki, Keating, and Safdar 2021, Tangcharoensathien et al. 2020). Identifying the mechanisms of trust formation and ways to measure trust on social media have become relevant questions that leading health institutions need to answer in order to provide guidance on risk communication to local and national authorities. Existing studies attempting to measure trust in institutions largely rely on indirect measurements such as surveys (Lazarus et al. 2020, Fetzer et al. 2020) and, apart from isolated cases (Wong et al. 2021), do not focus explicitly on social media. On the other hand, studies investigating public trust by analysing social media are mostly qualitative (Limaye et al. 2020, Wang et al. 2021). Despite the importance of these works in highlighting different aspects of the citizen-government relation on social media, due to their qualitative nature, they do not offer metrics that can provide visual and numeric summarisation of some dimensions of trust and that can be quickly computed from accessible social media data. Metrics to measure engagement with stakeholders on social media have been applied in the e-governance field (e.g., (Haro-de Rosario, Sáez-Martín, and del Carmen Caba-Pérez 2018)) and they can be leveraged for addressing the problem of trust measurement but they are usually not based on theories of trust.

Our study aims to bridge these gaps and investigate an area that has received little attention concerned with the identification of quantitative indicators from social media data that are grounded in theoretic trust frameworks. Drawing on traditional literature on production of trust and on qualitative studies investigating trust in social media, we propose a methodology to gather data from an official social media account and automatically extract indicators from the account’s shared content and interactions. These indicators are then mapped to dimensions of trust that have been identified as relevant in the web domain. Indicators and related visuals can be compiled in a report providing a way to evaluate the social media presence of governmental agencies through the lens of trust theories.

The remainder of the paper is organised as follows: we first introduce some notions of trust categorisation and social media engagement in e-governments; secondly, we present a framework that, drawing on literature findings, summarises and translates them into quantitative social media indicators; we then show an application of this framework to the Twitter accounts of two UK official health stakeholders; finally, we discuss our findings together with the implications, limitations and future perspectives of our study. In the following, we use the term citizens to refer to the social media audience of an official governmental account. We are, however, aware that social media users are not representative of all the demographics of citizens (Greenwood, Perrin, and Duggan 2016).
Background and related work

Measuring trust in e-governments There exists a vast literature advancing models and conceptions of trust. From a practical perspective, a relevant question is how trust is produced and how it can be maintained. One of the classifications most used in studies that tried to address this question is the one proposed by Zucker (1986) who identifies three modes of trust production: 1) characteristic-based trust which is tied to personal characteristics, such as family background and ethnicity; 2) process-based trust, which is produced through repeated exchanges; and 3) institutional-based trust, which is produced through formal institutional processes, such as professional certification and government regulation. Thomas (1998) identified the last two modes as the most significant for the creation of trust in government.

Leveraging on Thomas theoretic framework, several more recent studies have investigated the effect that government use of digital technologies has on citizens’ trust. The majority of these studies attempt to measure trust in government with surveys and they focus on citizens’ use of government websites. Notable is the work of Tolbert and Mossberger (2006) that analysing the Pew survey data found a statistically significant relationship between trust and use of a local government web site. The main contribution of the study is leveraging on Thomas framework to suggest six aspects of e-governments that can lead to increased trust: responsiveness, accessibility (related to process-based trust); transparency, responsibility (related to institution-based trust); efficacy, efficiency and participation (related to both trust modes).

Measuring trust in governments on social media Few studies explicitly addressed the measurement of trust in government on social media. Park et al. (2015) developed a structural equation model to analyse antecedents and formation of citizen trust and they empirically tested it through questionnaires from Korean population who used the government’s social media service. The results suggested that the utilization of Twitter by a leading officer is important for developing citizens’ trust in Twitter as a government-to-citizens (G2C) communication medium.

Measurement of trust directly from social media data is a largely unexplored field and, despite the availability of theoretical frameworks like the ones above mentioned, translating them to quantitative indicators is an open question. Azab and ElSherif (2018) is the only study we found addressing this issue. In their work, they suggested different trust indicators obtainable from Facebook data and they tested the proposed framework on accounts of three Egyptian ministries.

Metrics of citizens’ engagement with governments Although not directly based on theories of trust, relevant for our work is the research line investigating the ways with which citizens engage with governments on social media. Most of these studies highlight the potential of social media to increase government perceived transparency, participation and collaboration with citizens (e.g., Lee and Kwak 2012) and attempt to measure these dimensions with content analysis or empirical methods (e.g., Mergel 2013). The studies providing quantitative measurements are based on Bonsón, Royo, and Ratkai (2017) who investigated push and pull activities, communication and networking strategies of e-governments with an engagement metric accounting for the popularity, commitment and virality of posts shared by governments on Facebook.

Motivation for this work Our framework extends the metrics suggested by Bonsoón, by building upon the trust framework proposed by Tolbert and Mossberger. Azab and ElSherif is the closest to our work due to its attempt to model social media data with a framework of trust. However, differently than in our study, their items were obtained through a commercial tool and the computation of some of them cannot be automated. Another limitation of their framework lies in the unbalanced number of items for each trust dimension which makes dimension scores not easily comparable. In our study, particular effort was put in designing a balanced framework and assuring non-redundancy of the items, to avoid over counting some trust aspects. Another difference is that we show an application of the framework on Twitter rather than on Facebook. With this work, we intend to provide a lightweight, open-source and interpretable set of metrics to be used by public agencies or citizens and to be improved thanks to end-users’ and developers’ feedback. Due to the collaborative aims of our work, Twitter seemed the most suitable platform since their API allows to freely access a wider range of data than Facebook and there exists a wide and active community of social media mining practitioners working with Twitter data from whose inputs our framework can benefit.

Proposed framework

Using Tolbert and Mossberger (2006) as a reference framework, we consider the seven dimensions of trust presented below and, for each dimension, we identify quantitative indicators that relate to it that can provide useful insight on social media usage and interaction of an official account.

Process-Based trust
• Responsiveness. A government can use social media to easily address citizens’ requests. This process results in an increased perception that government cares about its citizens and hence in increased citizens’ trust. This component can be assessed by checking if an account allows direct messages and if it replies to them.

• Accessibility. Government social media accounts have the potential to make information and services from different agencies accessible to citizens, ideally, in all the languages that are relevant to their audience. We propose to translate this item into indicators aimed at 1) assessing the presence of other public stakeholders in the network of followers of the account of interest and 2) identify if the account shares content in relevant languages and/or content translation mechanisms are enabled.

Institution-Based trust
• Transparency. Government accounts can use social media to inform citizens about their activities. By textual analysis of the posts and URLs, we can infer if reports, agendas, meeting records are referenced. We can also determine
whether a government account engages with other government agencies or accounts, communicating in a fashion that not only signposts social media users to other official or legitimate sources but also demonstrates accountability and openness in its communications.

- **Authenticity.** This item is not included in any of the frameworks from previous studies but we believe that, in the context of social media, is essential for an official account to provide evidence of its authenticity (which is linked with the perception of a trustworthy institution). We suggest determining if the account is verified, if in its description the URLs to the official website is provided and if the official website points to the social media account of the agency. Although this is usually the case for accounts with a consolidated presence on social media, some gaps may be identified for governments that are new to the adoption of web technologies.

**Process and Institution-Based trust**

- **Effectiveness.** Government use of social media is effective if citizens find relevant the content shared and find answers to their questions (increasing process-based trust). An effective use of technology from the government may also give a favourable impression to citizens hence increasing the institution-based trust. As indicators for this item, we propose to quantify the content shared by the government account which links to official resources. To measure the government’s effectiveness in addressing citizens’ questions, we identify the presence of URLs in the government replies to comments. This can be interpreted as a commitment from the government to provide detailed answers while attempting to overcome the length limitations of social media messages. An additional indicator of effectiveness can be the length of discussion threads in which the government addresses user comments (if the government can address the comments effectively, there would be no need to go into long discussions). The automatic extraction of this information is potentially time-consuming and we did not consider it in our study.

- **Efficiency.** Social media offers governments the chance to promptly address citizens’ concerns. Timing and frequency of provision of information can be further optimised by the use of automated mechanisms. We propose to quantify this item by retrieving the average reply time to citizens’ comments and by assessing whether the account shares posts with sufficient regularity. Analysis of the account posting patterns can provide further insights on efficient social media management and on use of automated posting. However, discerning a human from an automated posting behaviour with unambiguous and reliable metrics is a challenging task and an active area of study. We leave this assessment for follow up work and in the current study, we provide a visualisation of the posts shared by the account per day and hour.

- **Participation.** Social media provide a way for citizens to participate in the political discussion and for governments to directly address citizens’ comments. To assess the participation level of a government account, we can identify whether the government replies to comments and estimate the citizens’ interest in interacting with the government through engagement metrics. We suggest to use the metrics of popularity (P), commitment (C), virality (V) and engagement (E) of a post proposed by Bonson and Ratkai (2013), averaged for all posts and normalised by 1000 followers (equation 1). These metrics are easy to interpret and compute and have been useful to measure engagement with e-governments in other studies (Bonsón, Royo, and Ratkai 2017, Silva et al. 2019).

\[
\begin{align*}
\text{Popularity} (P) &= \frac{\text{total likes}}{\text{total posts}} \\
\text{Commitment} (C) &= \frac{\text{total comments}}{\text{total posts}} \\
\text{Virality} (V) &= \frac{\text{total re-shared posts}}{\text{total posts}} \\
\text{Engagement} (E) &= \frac{P + C + V}{\text{total followers}} \times 1000
\end{align*}
\]

**Experiment**

To test the framework proposed above in a real setting, we monitored for one week the Twitter accounts of two official British health agencies, Public Health England and National Health Service England (accounts name PHE_uk and NHSEngland respectively) and computed quantitative indicators for each of the framework dimensions. These indicators can be automatically compiled into an overall report on a specific account, covering a particular time-frame. Computation of these indicators relies only on data accessible with a free Twitter developer account.

**Data collection**

Data from the account timelines were collected through a standard Twitter rest API, accessed via tweepy Python library (Roesslein 2020). To access the API we make use of the free developer account which limits the queries to the previous seven days. All data accessed via the Twitter APIs are subject to Twitter’s data policy and developer agreements (Twitter, Inc. retrieved online: April 2021). For the current work, we only count and aggregate data over a given time period. None of the content of any account is stored and no identifying information can be determined from the final set of computed indicators.

**Method**

We outline here the details for computing the indicators; although specific to Twitter, this section can provide a useful guidance for researchers interested in applying the framework on other social media. The indicators are summarised in Table 1.

**Authenticity indicators.** Twitter already allows for accounts to be ‘verified’ if they can demonstrate to Twitter that they are authentic. This can be assessed directly from the ‘User’ object. The User object also has fields for a description (provided by the account operator), URLs and a location, all of which can be directly accessed through the API. If a URL pointing to an official domain (in our case,
we considered official URLs ending with `.gov,.gc.ca,.mil` or `.nhs.uk` is provided, we crawl the page and retrieve whether the account Twitter page is referenced there.

**Responsiveness indicators.** To assess if the account allows direct messages (DMs) we attempt to send a via `tweepy` a DM, asking for general information on the account owner. In the case of positive feedback (the API returns an error code if this action is not allowed), we request the DMs from our inbox until a reply from the monitored account is found.

**Accessibility indicators.** Tweet objects include a field `lang` proving the language of the tweet as identified by Twitter. We use this information to infer whether an account posts in more than one language (this is particularly relevant for countries in which multiple languages are commonly spoken). To identify whether the account has access to information shared by other official agencies, we collect the list of the followed accounts and extract those listing an official URL in their account details or description.

**Transparency indicators.** To determine whether the account shared transparent content we check the URLs and the meta-data of the web pages linked in the account tweets for the presence of keywords such as ‘report’, ‘programme’, ‘agenda’. Furthermore, we retrieve the mentions of other official agencies in the recent tweets to ensure that the account is engaging and communicating with the wider community of official accounts, demonstrating openness in its communications.

**Effectiveness indicators.** We use the presence of links in the account replies to users’ comments to determine whether the account answers citizens’ questions with additional information. To infer whether the account shares content that is relevant for its citizens, we check through recent tweets for URLs pointing to governmental or ‘official’ domains.

**Efficiency indicators.** To assess whether the account maintains a good level of activity, we check how often the account posts to its own social media timeline between the hours of 06:00 and 21:00 in the local timezone. We compute whether the account posts an average of 1 post per 3 hours to ensure that citizens have access to timely and up-to-date news or information from the agency. We additionally compute the average reply time with which the account addresses comments. We considered 2 hours a reasonable threshold of efficient account management. These thresholds can be considered parameters and made more or less conservative based on needs.

**Participation indicators.** As indicators of participation in the political discussion (from both the official agencies and its audience perspective), we consider the citizens’ engagement (1) rate with the content generated by the account and whether the account replies to citizens’ comment. Owing to the limitations of the Twitter API, there is no robust method to check whether any given tweet has been replied to, only the number of replies. To check whether an account replies to comments, we can check whether any of the most recent Tweets share a property indicating that they are a reply to another Tweet. We then check that these ‘reply’ Tweets are replies to accounts other than the current account. As long as the recent Tweets are checked frequently enough, no reply Tweets should be missed. Since reply Tweets contain the id number of the Tweet they are in reply to, it is also possible to look up the Tweet that is being replied to. Deciding which percentage defines a good engagement rate is not a trivial subject and there exists only vague general references mostly focused on the industry domain. Based on these references, we opted to consider 0.2 a threshold of good engagement. As other of the proposed indicators, this value can be adjusted to the cases of interest.

In addition to computing the above indicators, we can also produce a descriptive report that tracks the evolution of engagement, since this is a time-varying quantity, as well as the activity of the account over the previous seven days. Examples of activity and engagement are shown in Fig. 1 and Fig. 3 respectively.

<table>
<thead>
<tr>
<th>Trust dimension</th>
<th>Indicators</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authenticity</td>
<td>Is the account verified?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account give an official page which links back to the account?</td>
<td></td>
</tr>
<tr>
<td>Responsiveness</td>
<td>Does the account allow direct messages?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account reply to direct messages?</td>
<td></td>
</tr>
<tr>
<td>Accessibility</td>
<td>Does the account post in multiple languages?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account follow other official accounts?</td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td>Does the account share transparent content (meetings/agendas)?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account mention other official accounts?</td>
<td></td>
</tr>
<tr>
<td>Effectiveness</td>
<td>Do more than X% of tweets link to an official resource (X=10)?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account answer citizen’s questions with additional information?</td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td>Does the account post regularly, at least K times per hour (K=0.33)?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account reply within N hours to comments (N=2)?</td>
<td></td>
</tr>
<tr>
<td>Participation</td>
<td>Is the engagement rate above the threshold T (T=0.2)?</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Does the account reply to comments?</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The proposed framework to map social media indicators to trust dimensions.
Findings and Discussion

The ‘spider chart’ shown in Fig. 2 plots the indicators of the current framework for the ‘PHE.uk’ account. The chart for the ‘NHSEngland’ is the same, suggesting that at this level, both accounts exhibit similar overall performance and behaviour on Twitter. These charts allow for quick assessment of which dimensions the account is performing well in, as well as identification of which dimensions could be improved. For ‘PHE.uk’, we see that authenticity, accessibility and transparency all have maximum score. This is owing to the fact that the agency is verified on Twitter, and has good links between the Twitter account and the official URL. Contrastingly, since the account does not allow DMs, the score is zero for responsiveness. Neither of the accounts analysed here appears to respond to citizens’ comments which decreases the score for both participation (responding to citizens’ comments) and efficiency (responding within a given time-frame). This also decreases the score for effectiveness, since the account does not reply to citizen questions with links or resources containing more information.

More fine-grained differences in performance can be seen in Fig. 3, which plots the engagement Eq. (1) across several days. While ‘NHSEngland’ appears to have consistent engagement above the threshold for this time period, ‘PHE.uk’ is below the threshold until later in the week. While a longer time-series of such data is needed to make robust conclusions, we note that changes in the engagement metric occur as new social media posts enter and exit the rolling time window over which the engagement is calculated. We note that in the time-period we observed, the major contribution to the engagement $E$ comes from the popularity $P$, with this component being at least an order of magnitude larger than the others, commitment and virality. In general, we expect that the relative importance of any of these metrics may be particular to a given social media platform.

Our findings show that the set of suggested indicators can provide intuitive measurements and visualisations of an account’s social media behaviour within a literature-based trust framework. The indicators are obtained with free and accessible tools and their implementation is time and computationally inexpensive. However, our study presents several limitations.

We comment that it is difficult to establish ‘baseline’ or expected values for the indicators we have computed. That is, when measuring a value such as engagement, without historical data to compare with, or reference values from peer institutions, it is challenging to assess whether the absolute values or the variations we observe in Fig. 3 are within an expected margin or not. Further, the current framework provides no means to control for confounding effects such as the citizen trust in government, or the digital literacy of the government or the citizen population. Further, it is known that Twitter users are not representative of all citizens (Sloan 2017), and this is a crucial factor in assessing the social media interactions between official agencies and citizens, especially in the context of ‘trust’, since there are multiple demographics that are critically under-represented on social media. Further, as noted above, different social media platforms may have different ‘temperatures’ or styles of communica-
tion that engender distinct approaches to the interaction between official accounts and citizens’ accounts.

Further Work
In a follow-up paper, we intend to carry out cross-validation of this framework with previous studies discussing public trust in government or official institutions. Since it is resource intensive and time-consuming to make robust measurements of public trust, it would be more feasible to compare the output of this framework to other studies conducted in the domain of public trust. Another possibility is to consider a comparison between the metrics included here and public opinion polls. Monthly trackers such as those carried out by YouGov (YouGov retrieved online: May 2021) provide a measure of public opinion of bodies such as the NHS, and may form an initial point of reference for public trust. That is, we would expect that highly trusted institutions would score well on our proposed trust indicators.

A limitation of the current framework is that scores in each dimension are binary. Using continuous scores would allow for more fine-grained insights and allow for greater distinction between accounts. Another way to achieve a similar result would be to include more indicators for each dimension and normalise the sum of the scores so that all dimensions are given an equal or appropriate weight.

Some additional indicators could include the use NLP techniques to analyse the content of the citizen communications to an agency, such as detecting the presence of questions, or counting terms that could indicate praise/criticism as a way to gauge public opinion of the agency, as expressed on the social media platform. Additionally, the analysis of extra information on the types of posts the agency shares could be of interest. This might include listing the types of content shared, such as videos and images, or details of the type of content that is posted, such as blogs or reports about scientific findings.

Acknowledgments
The paper here presented is based on research undertaken as part of the European Commission funded project STAMINA (Grant Agreement 883441).

References


YouGov. retrieved online: May 2021. NHS national service opinion poll. URL: https://yougov.co.uk/topics/politics/trackers/how-good-or-bad-are-national-nhs-services.