Social Sciences & Humanities Open 2 (2020) 100060

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.elsevier.com/locate/ssaho

Detection of self-reported experiences with corruption on twitter using unsupervised machine learning

Jiawei Li [a,b,c], Wen-Hao Chen [d], Qing Xu [a,b,c], Neal Shah [a,b,c], Jillian C. Kohler [e,f], Tim K. Mackey [a,b,c,e,g,*]

[a] Department of Healthcare Research and Policy, University of California, San Diego - Extension, USA
[b] Department of Anesthesiology, University of California, San Diego School of Medicine, San Diego, CA, USA
[c] Global Health Policy Institute, San Diego, CA, USA
[d] Department of Computer Science and Engineering, University of California, San Diego, Jacobs School of Engineering, USA
[e] WHO Collaborating Centre for Governance, Transparency and Accountability in the Pharmaceutical Sector, University of Toronto, Ontario, Canada
[f] Leslie Dan School of Pharmacy, Dalla Lana School of Public Health and Munk School of Global Affairs, University of Toronto, Ontario, Canada
[g] Division of Infectious Disease and Global Public Health, University of California, San Diego School of Medicine, Department of Medicine, San Diego, CA, USA

ARTICLE INFO

Keywords:
Corruption
Big data
Natural language processing
Machine learning
Healthcare
Bribery
Sustainable development goals

ABSTRACT

Background: Corruption is a significant challenge to the future of human development, economic progress, and population health in the post millennium. Corruption, in its different forms of bribery, fraud, waste, collusion, and illicit financial flows, not only leads to waste but can also erode trust in government and public systems. Corruption is also complex and globalized with different forms of corruption occurring across different countries and multiple industries. One critical tool to leverage in the fight against corruption is the use of innovative technologies such as machine learning.

Methods: In this study, we deployed an unsupervised machine learning methodology using natural language processing to collect and analyze data from the popular social media platform Twitter with the aims of detecting self-reported experiences with corruption, including in the health sector. We collected data from the Twitter public API for keywords associated with corruption and used the biterm topic model to extract themes from the entire corpus of Tweets in order to detect user-generated messages reporting or discussing experiences with corruption.

Results: We analyzed 22, 180, 425 tweets filtered for corruption-related keywords from January–May 2019. Using a combination of NLP and manual annotation, we detected 2383 tweets from 1556 users that included self-reporting of corruption for two dominant themes: police bribery and healthcare corruption. Overall, we found a small number of users actively reporting experiences with corruption, identified users located in countries that are perceived as having higher levels of corruption by their citizens, and found that the majority of messages included reports of users' own experiences and/or documentation of corruption.

Conclusion: Though technology is not a "silver bullet" that can entirely address the multifaceted nature of global corruption, this study demonstrates its potential utility as a force for good to enable better detection, characterize forms of corruption in different sectors, and hopefully inform future anti-corruption efforts. Additionally, the UN Sustainable Development Goals, with shared goals of fighting corruption, improving population health, encouraging technology adoption, and fostering multistakeholder partnerships, may serve as a critical governance space to catalyze technology-driven anti-corruption approaches.

1. Introduction

Corruption, defined as "the misuse of entrusted power for private gain", is a global wicked problem that cuts across all sectors, including energy, construction, transportation and storage, public procurement, politics, and healthcare (Mackey et al., 2016; TI, 2006). According to the World Economic Forum, corruption, bribery, illicit financial flows, and theft and tax evasion, have an estimated USD 1.26 trillion cost per annum

* Corresponding author. Global Health Policy Institute, 8950 Villa La Jolla Drive, Suite A124, San Diego, CA, 92037, USA.
E-mail address: tmackey@ucsd.edu (T.K. Mackey).

https://doi.org/10.1016/j.ssaho.2020.100060
Received 20 March 2020; Received in revised form 17 June 2020; Accepted 31 August 2020
Available online 22 September 2020

for developing countries ("Corruption costs developing countries $1.26 trillion every year - yet half of EMEA think it's acceptable," n.d.). Additionally, approximately three-quarters of the 180 countries measured by the Transparency International Corruption Perceptions index (CPI), an index that scores countries on how corrupt their public sectors are perceived among their public with ratings ranging from 10 [very clean] to 0 [highly corrupt], score lower than five on the index.

Corruption is global in its reach with no respect for geopolitical borders, manifests in several different forms (including petty and grand corruption), and can also infiltrate various types of organizations, communities and institutions, including in developed and developing countries alike (Gupta, Davoodi, & Tiongson, 2000; Lalountas, Manolas, & Vavouras, 2011; Mackey et al., 2016). Corruption also impacts human rights. The United Nations Office of the High Commissioner on Human Rights stated that corruption is an "enormous obstacle to the realization of all human rights". Corruption can also foster inequity as it skews how resources are distributed and creates barriers to public services and goods, such as essential medicines (Kohler, Chang Pico, Vian, & Mackey, 2018; Mackey & Liang, 2012; Vian, 2008). Lastly, corruption is a potential barrier to achieving global development goals, such as the 2030 United Nations' Sustainable Development Goals (SDGs).

Importantly, the clandestine nature, scope, and diversity of corruption introduces challenges for prevention, detection, and enforcement. Anti-corruption efforts are influenced by risk characteristics that are specific to particular industrial sectors, which include factors such as the amount of spending in the public sector, market uncertainty, information asymmetry, and the inherent complexity and globalization of international markets (Mackey et al., 2016; Mackey, Kohler, Lewis, & Vian, 2017; Vian, 2008). This is further complicated by social-cultural factors that differ across countries when it comes to the definition of a corrupt act and what are acceptable or tolerable forms of corruption and graft (Mackey, 2019; Nishtar, 2010; Vian, 2008). In certain sectors, such as healthcare, the negative impact of corruption goes well beyond economic loss, impacting security, patient safety, and population health outcomes, especially for the world's most poor and vulnerable (Mackey et al., 2016; 2017b).

However, detecting and measuring the specific characteristics, scope, and major trends involving global corruption is difficult, with accurate estimates on its impact difficult to establish (Mackey, 2019). Specifically, difficulties in accurately measuring corruption are compounded by issues such as imprecise definitions, lack of sufficient protection for reporters of corruption (e.g. whistleblowers), and the added dimension of the opacity of criminal corrupt acts (Mackey & Liang, 2012; Reynolds & McKee, 2010). When corruption is multijurisdictional and transnational in nature, crossing more than one country, region, or population, addressing it demands cooperation and coordination across multiple stakeholders who need to agree upon priorities, goals, and shared anti-corruption activities in order to be effective.

Fortunately, shared objectives of combating global corruption are central pillars of the UN SDGs, a global governance mechanism that may hold the key to combating corruption while also providing a platform to integrate forms of emerging technology to help modernize the fight against corruption. This is explained below.

1.1. Corruption and the United Nations' Sustainable Development Goals

As the successor of the Millennium Development Goals (MDGs), which focused on poverty alleviation by 2015, the United Nations' SDGs represent a global agenda for sustainable peace and prosperity that are much more comprehensive than the MDGs. The MDGs originally included eight goals focused on eradicating poverty and hunger, achieving universal primary education, promoting gender equality, ensuring environmental sustainability, and addressing key global health issues (e.g. reducing child mortality, improving maternal health, and combating key infectious diseases). In contrast, the SDGs comprise of broader set of 17 core SDGs, along with their associated 169 targets and 232 indicators (compared to only 8 goals and 21 targets for the MDGs).

Chief thematic areas of the more expansive SDGs are addressing poverty, hunger, education, economic growth and development, urbanization, inequality, sustainability and the environment, marine conservation, affordable and clean energy, climate change, science and technology, and importantly justice (SDG 16) and health and well-being (SDG 3). Specific to addressing corruption, Goal 16 ("promote peaceful and inclusive societies for sustainable development") includes Target 16.5, which calls for substantially reducing corruption and bribery in all their forms. Measuring progress towards reducing corruption are indicators 16.5.1 and 16.5.2 that examine the proportion of persons and businesses that have paid or were asked to pay a bribe to a public official.

In addition to Goal 16 that directly addresses corruption, other SDG goals can provide incentives for adoption of anti-corruption initiatives coordinated across multiple actors. This includes SDG Goal 17 that focuses on multistakeholder partnerships and can act as a catalyst for creating a shared governance framework that links anti-corruption goal SDG16 to other SDG goals and targets that are complimentary. Specifically, targets 17.14 (enhancing policy coherence) and 17.16 (enhancing global partnership around SDGs by complementing with multistakeholder partnerships) can unify programmatic activities and UN agency workstreams.

Finally, SDG9 ("build resilient infrastructure, promote inclusive and sustainable industrialization) encourages technology adoption, R&D, and access to information and communication technologies, all tools that can be leveraged to combat corruption (Joudaki et al., 2015a; Mackey & Nayyar, 2017; Shim & Eom, 2008).

1.2. Addressing corruption with technology

The utility of different forms of technology to combat corruption is gaining increased attention including from organizations such as the Organization for Economic Co-operation and Development (OECD), the Asia-Pacific Economic Cooperation (APEC), the World Economic Forum, the International Monetary Fund, and the United Nations Development Programme (UNDP). Though there are tools available to address corruption and fraud in the context of improving transparency in public administration, enabling enhanced financial management, establishing procedures to address conflicts of interest, and other solutions used in forensic auditing and law enforcement, emerging digital technologies that can more systematically address corruption in specific sectors are also now being encouraged (Mackey & Liang, 2012; Transparency International, Global Corruption Report 2006, 2005, pp. 1–378).

Key to the potential utility of new forms of anti-corruption technology are increasing access to and transparency of information in a way that can be used to detect and ultimately reduce the presence of corruption (Banning-Lover, 2016; Holeman, Cookson, & Pagliari, 2016; Silveria, 2016; "Technology against corruption," 2013). Innovative technologies that are now being positioned to address corruption include big data analytics, mobile and website applications, artificial intelligence and machine learning, distributed ledger technology also known as blockchain technology, technology to combat misinformation online, and civic technologies used to encourage greater participation in government (including e-government and open government initiatives) (Banning--Lover, 2016; Mackey, Kohler, et al., 2017; Silveria, 2016; "Technology against corruption," 2013). The primary use cases for these technologies include creating new channels to report corruption, monitoring for corruption and fraud, collecting and aggregating corruption-related data (e.g. media reports of corruption), and social and citizen anti-corruption education and mobilization (Wickberg, 2013).

Specifically, big data analytics can be leveraged to detect trends and patterns from different forms of data (e.g. financial flows), while also introducing the potential to enable real-time monitoring for better fraud and corruption prevention (Silveria, 2016). For example, multilateral development banks and national governments can use data mining to audit and track payments to detect patterns of collusion and falsified

information in areas such as public procurement (e.g. abnormal bidding patterns or bid awards), particularly if these solutions are integrated with open data sources (such as Openspending.org that tracks government and corporate financial transactions) and other e-governance and e-procurement systems (Silveria, 2016; Wickberg, 2013). This can also be applied to the healthcare sector, where data mining can lead to identification of specific healthcare claims or providers that fit a fraud or abuse "profile," enabled through analysis of large-scale reimbursement claims datasets or in the procurement of medicines, which is an area that is highly susceptible to corruption (Herland, Bauder, & Khoshgoftaar, 2018; Herland, Khoshgoftaar, & Bauder, 2018; Joudaki et al., 2015c; 2015b). Similarly, blockchain technology could be used to cryptographically validate data written to a distributed ledger and ensure a final and immutable record of transactions, including potential use of cryptocurrencies to avoid financial fraud and abuse (Aldaz-Carroll & Aldaz-Caroll, 2018; "Promise and peril: blockchain, Bitcoin and the fight against corruption," 2018).

Mobile technologies can also be used to empower citizens to report and "crowdsource" concerns about corruption and fraud, including sending relevant evidence (e.g. financial documents or even photos of incomplete construction projects) or digitally recording proof of a bribe (Silveria, 2016). Machine learning and predictive modeling can be used for anomaly detection, identifying and flagging improper payments or high-risk transactions/contracts, and enabling process automation (Silveria, 2016). However, it should be noted that generally in order to ensure artificial intelligence (AI) solutions are effective, there is a need for sufficient data to train models for accurate classification, a requirement directly tied to the push for transparency and open data sources through e-government and open government initiatives (Silveria, 2016).

Finally, social networking sites are now a global phenomenon, with several popular social media platforms, such as Facebook and Twitter, acting as a digital public square where hundreds of millions of users' express opinions about important social issues. Increased digital engagement in the social media public square has coincided with the rise of an area of study known as "infoveillance", which is defined as "the science of distribution and determinants of information in an electronic medium" (Eysenbach, 2009). Reflecting this utility, the use of social media in disciplines of social sciences, political science, and public health is growing, where large volumes of social media messages can be used to conduct surveillance about user knowledge, attitudes and behavior, including political activities, criminal activity, health information, and even to detect corruption (Bond et al., 2012; Halpern & Gibbs, 2013; Intravia, Wolff, Paez, & Gibbs, 2017; Jones, Bond, Bakshy, Eckles, & Fowler, 2017;; Mackey, Kalyanam, Katsuki, & Lanckriet, 2017; Patton, Eschmann, & Butler, 2013; Sobkowicz, Kaschesky, & Bouchard, 2012; Stieglitz & Dang-Xuan, 2013).

More purposeful use of online platforms to enable reporting of corruption-related activities and public perception about corruption are taking shape. This includes websites that are used to "crowdsource" information from the public and whistleblowers via dedicated forums. One example is the Indian website ipaidabribe.com (http://www.ipaidabribe.com), which has collected over 80,000 reports of bribery, the global reporting platform BRIBEline, the Romanian initiative Bribe Market, and the Macedonia online reporting platform "Draw a Red Line" (Silveria, 2016; Wickberg, 2013). Though the use of social media technology and dedicated corruption-related crowdsourcing websites shows promise, maintaining on-going engagement, balancing the need for anonymity with source credibility, and methods to verify information reported to these system are key to future design, successful user adoption, and scalability (UNDP, 2011).

Importantly, though these technologies represent significant promise, there are also concerns about potential for misuse. This includes misuse of technology intended for social mobilization and elections by governments themselves, cybersecurity issues associated with vulnerabilities in such systems, and user privacy and confidentiality considerations (Adam & Fazekas, 2018;; Mackey, Kohler, et al., 2017). Further, there are concerns about equity, with communities that may lack access to the Internet or mobile connectivity remaining underrepresented in anti-corruption initiatives or civic engagement opportunities (Wickberg, 2013). Finally, operational issues associated with weak information and communication technology infrastructure, lack of data availability in low and middle-income countries, and cost considerations concerning the deployment of technology, continue to remain challenges (Joudaki et al., 2015c).

1. 3. Corruption in the healthcare setting

Corruption is a serious challenge in the healthcare sector. In fact, it is estimated that some $300 billion US. dollars are lost to corruption and errors annually (equating to approximately 6% of the $7.35 trillion spent on global healthcare services) (Gee & Button, 2015; WHO, 2010). The pressing need to combat corruption, particularly in global health, is now gaining increased international attention, evidenced by a February 2019 joint consultation held by the World Health Organization, the Global Fund, and the UNDP for a proposed Alliance for Anti-Corruption, Transparency and Accountability in Health ("ACTA Alliance") (Mackey, 2019).

Corruption in healthcare is not static, manifesting in different forms (including bribery of healthcare workers, theft and embezzlement of financing for health, collusion in bidding for healthcare goods and services, fraud in healthcare payment systems, unjustified healthcare worker absenteeism, rigged bidding in pharmaceutical procurement, and trade in fake goods and even medicines) and can occur across different parts of the health sector (e.g. health delivery systems, health workforce, health supply chains, global health organizations, and biomedical research) (Mackey & Liang, 2012). Critically, corruption represents a major roadblock to 21st century shared global goals, including ensuring health equity, maintaining integrity of global health financing, and achieving Universal Healthcare Coverage (UHC), all priorities outlined in the SDGs (Mackey, Vian, & Kohler, 2018; Michaud, Kates, & Oum, 2015). In the health sector, corruption not only leads to waste but can also cost lives for those who are directly impacted by diminished access and/or poor-quality healthcare services stemming from corruption (Hanf et al., 2011; Witvliet, Kunst, Arah, & Stronks, 2013).

Targets in the SDG health goal (SDG3) are numerous and include issues such as improving outcomes for maternal child health, infectious diseases, non-communicable diseases (including mental health), substance abuse, road injuries, reproductive health, environmental contamination, tobacco control, access to medicines, healthcare workforce, capacity building for health emergencies, and the overarching goal of achieving UHC. There is no SDG target or indicator that specifically addresses health corruption. However, SDG 3 measures progress towards global health targets that can be directly negatively impacted by corruption (i.e. the presence of corruption can lead to suboptimal outcomes for healthcare/medicines access, higher morbidity and mortality, weaken health workforce capacity, negatively impact health emergency response, and block efforts to achieving UHC) (Mackey, Kohler, et al., 2017; TI, 2006; , n.d.U4).

For example, if programs are put in place to prevent bribery of healthcare workers who may subsequently make suboptimal decisions about patient care (such as overprescribing expensive drugs or being paid to provide healthcare services outside of public clinics), patients can gain better access to essential healthcare services (Target 3.8) and the practice of corruption and bribery in healthcare can be reduced (Gaitonde, Oxman, Okebukola, & Rada, 2016). Or if efforts are put in place to prevent criminal corruption related to introducing falsified and substandard medicines into the global drug supply chain, benefits can come in the form of better access to safe medicines (Target 3.B) and reduction of corruption-related practices in the pharmaceutical sector (Mackey & Liang, 2013). Hence, healthcare represents a critical sector where anti-corruption tools are needed in order to ensure the success of broader human development and population health goals and should be a focal

point for global anti-corruption efforts.

2. Methods

Given the importance of combating global corruption and in order to better explore novel technology-based approaches to detect and characterize corruption in different industrial sectors (including the health sector), we adopted an infoveillance and big data approach to collect, analyze and characterize corruption-related data from one of the world's most popular social media platforms: Twitter. We conducted data mining and analysis on this popular microblogging site that now has approximately 126 million users globally by collecting tweets containing corruption-related keywords.

The study was conducted in three distinct phases: (1) data collection and extraction; (2) data processing; and (3) data analysis. In the data analysis phase, we took steps to further filter data into topic groupings that appeared to be related to user experiences with corruption using an unsupervised machine learning topic model in the family of natural language processing (NLP). We then manually labelled the data extracted from these topic models to confirm if they involved self-reporting of corrupt activities. This approach was undertaken due to the lack of a preexisting training dataset for Twitter messages related to corruption.

The specific aim of this study was to identify Twitter users who self-reported corruption-related experiences and behaviors, not merely conversations about corruption news or personal opinions about corruption events. All data collection, processing and analysis was done in the programming language Python and associated packages. Fig. 1 provides a summary of this methodology and the source code for this project can be found at https://github.com/Mathison/Twitter_corruption_analyze.git.

2.1. Data collection

We collected data from Twitter using the public streaming application programming interface (API) over a period of approximately three months from January 15, 2019 to May 15, 2019 using the python package Tweepy. This allowed us to query a sample of publicly available Twitter messages using our associated Twitter developer account, consumer secret and consumer key, token access, and installing Tweepy for Python. We then created an API object with functions to call the Twitter API by filtering for a list of commonly utilized keywords associated with corruption conversations. This was accomplished using the query parameter and setting the language to "en" (i.e. English characters). Specifically, we chose our keywords on the basis of their association to common forms of corruption and as they were pre-validated in a prior study by authors (Li, Chen, Xu, Shah, & Mackey, 2019). These keywords included: "bribery", "bribe", "corruption", "nepotism", "collusion" and "kickback", all common terms associated with corrupt acts and that

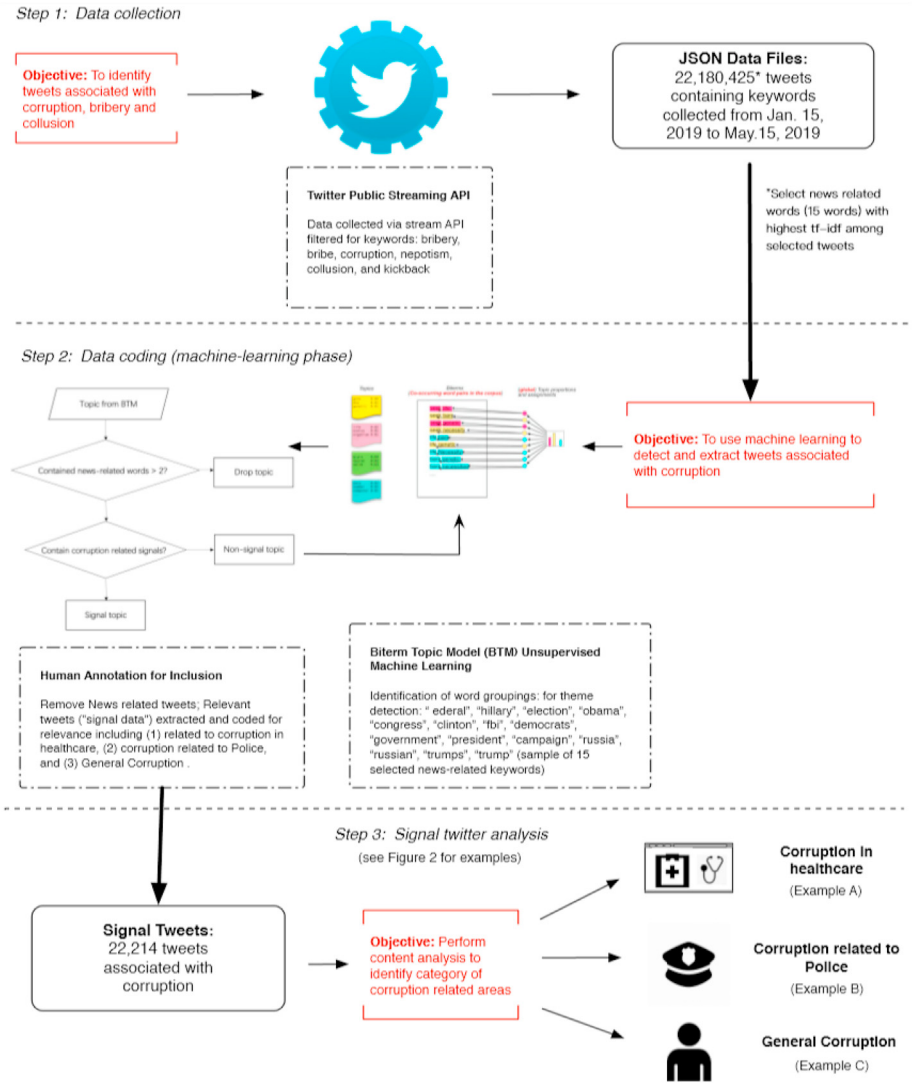

Fig. 1. Methodology using BTM to analyze tweets related to user-generated reports of corruption.

commonly appear in English-language conversations related to corruption on Twitter. Tweets containing at least one of these keywords were captured and processed and the data included the text of the tweet and other metadata associated with the message (geolocation if available, time stamp information, user account/handle, etc.) Data was collected in JSON format returned by the Twitter API.

2.2. Data processing

After collecting data filtered for corruption-related keywords, we extracted the text/message from each Tweet. However, the majority of text in these Tweets contained noise (i.e. messages or content that generally discussed corruption news, attitudes regarding corruption, and opinions about corrupt acts, individuals, or countries, but did not contain self-reported experiences with corruption). For example, the following Tweet expresses opinions about general police corruption, but does not discuss a specific user experience associated with a corrupt act or bribery:

> "[anonymized] Me and my sister we fought hard against #Islam ….we fought for our #lives but we cant fight #Police #corruption and a whole country while having a baby with us … we need help spread the word"

In order to minimize the amount of noise in our keyword filtered data set, we cleaned data for the following attributes:

- Imbedded Hyperlinks: The hyperlink in the text does not provide relevant information as it requires further analyzing the websites that are imbedded in the Tweet. In most of these cases, hyperlinks posted by users relate to sharing news and do not link to self-reporting information.
- Stop words: Stop words (such as the, a, an, in) are commonly used in messages but do not provide much context to the theme or category of the message itself. They are not key words in the text but occupy a high volume of words within social media datasets. The NLTK package was used to filter out stop words in messages collected.
- Special characters and punctuation marks: Special characters like emoji and punctuation marks may communicate a user's sentiment or convey a message. However, we did not prioritize these characters in this study as their interpretation can be subjective and most contextual information is in the form of text.

Following this text cleaning process, the remaining words provide a more representative sample of key concepts/themes of the messages that can then be analyzed using NLP. We also removed all text that consisted of less than three words to further filter messages to focus on those with sufficient thematic context. This additional filtering was based on observations that Twitter discussions about corruption experiences cannot be adequately conveyed or classified if in less than 3 words. Here is the text of the preview tweet example after text processing:

> 'sister', 'fought', 'hard', 'islam', 'fought', 'lives', 'cant', 'fight', 'police', 'corruption', 'whole', 'country', 'baby', 'us', 'need', 'help', 'spread', 'word'

2.3. Data analysis

Due to the large number of Twitter messages we collected, we had to separate the data into 9 parts for the purposes of conducting analysis (also described in "Results" section). For each part, we categorized the data into different clusters based on their themes using an unsupervised machine learning approach. Since this study focused on self-reporting of purported corruption-related activities and first-hand experiences by users, our first priority was to remove tweets that were not generated by individual users. This step focused on removing tweets related to news articles, bots, and other non-individual user accounts. This was particularly important as single issues (e.g. news related to U.S. politics and the Trump administration) dominated discussions around corruption-related topics. This process to remove non-individual user generated content was completed in three steps.

In Step 1, we use the biterm topic model (BTM), an unsupervised topic model used to extract the themes from a group of texts as used in prior studies to detect issues related to opioid behavior and diversion, wildlife trafficking, and other social and public health issues (Han & Kavuluru, 2016; Kalyanam & Mackey, 2017; Mackey & Kalyanam, 2017; Xu, Li, Cai, & Mackey, 2019; Yan, Guo, Lan, & Cheng, 2013). The basic approach of BTM is to combine the distinct words in each text as a biterm, then BTM will learn topics over short texts based on the aggregated biterms to tackle the sparsity problem in a single document. Suppose α and β are the Dirichlet priors. For each topic k, we have a topic specific word distribution $\Phi_T \sim Dir(\beta)$ and the topic distribution $\theta \sim Dir(\alpha)$. For each biterm b in biterm set B, we have two words $w_i w_j$. The joint probability of a biterm $b = (w_i, w_j)$ can be represented as:

$$P(b) = \sum_T P(T)P(w_i|T)P(w_j|T).$$

Since $P(T) = \theta_T$ and $P(w_i|T) = \Phi_{i|T}$ we can also have:

$$P(b) = \sum_T \theta_T \Phi_{i|T} \Phi_{j|T}$$

$$P(B) = \prod_{(i,j)} \sum_T \theta_T \Phi_{i|T} \Phi_{j|T}$$

Also, to infer the topics in a document, we assume that the topic proportions of a document equal to the expectation of the topic proportions of biterms generated from the document, then we can have:

$$P(T|d) = \sum_b P(T|b)P(b|d)$$

Via Bayes' formula based on the parameters established in BTM we can have

$$P(T|b) = \frac{P(T)P(w_{i|T})P(w_j|T)}{\sum_z P(T)(w_{i|T})P(w_j|T)}$$

Since in short text $P(b|d)$ can be represent as a uniform distribution over all biterm in the document d, we can have:

$$P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)}$$

where $n_d(b)$ is the frequency of biterm b in document d.

Based on these equations we can get the global topic distribution θ_T and the specific word distribution $\Phi_{w|T}$. In the original BTM paper, the authors decided to use Gibbs sampling to perform approximate inference (Griffiths, T. L., & Steyvers, M. 2004). They first randomly choose the initial state of Markov chain, then calculate the conditional probability based on each biterm b, the detail of the algorithm is shown in the paper by Yan et al., 2013. The result is shown below:

$$\Phi_{w|T} = \frac{n_{w|T} + \beta}{\sum_w n_{w|T} + M\beta}$$

$$\theta_T = \frac{n_T + \alpha}{|B| + K\alpha}$$

Where $n_{w|T}$ is number of times the word w is assigned to topic T, K is the number of topics we chose, |B| is the total number of biterms, α and βis the symmetric Dirichlet priors.

Using BTM, groups of messages/text containing the same word-related themes are categorized into the same clusters, the main themes of those clusters is considered as the topic of the aggregation of the text

(Yan et al., 2013). We set a total number of k different clusters for each data set, if the k value is too large, the text with the same theme could be separated into two clusters, if the k value is too small, the topic(s) with weak signals (i.e. small volume of discussion) could be obscured by topic(s) with stronger signals (i.e. larger number of discussions).

In order to find the appropriate k value for this data set, we use the u-mass coherence score to determine the number (k) of the topics. Coherence score is a value used to measure the performance of a topic model based on the k value, it can help distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference. A k value with higher coherence score means the clusters it categorizes are more identical to each other. We let D(v) be the document frequency of the word type v (i.e., the number of documents containing at least one token of type v) and D (v, v0) be the co-document frequency of word types v and v0 (i.e., the number of documents containing one or more tokens of type v and at least one token of type v0). We define topic coherence as:

$$C\left(t;\ v^{t}\right)=\sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\left(\frac{D\left(v_{t}^{m},v_{t}^{l}\right)+1}{D\left(v_{l}^{t}\right)}\right)$$

Where $V(t)=(v_t^1\ldots,v_t^m)$ is a list of the M most probable words in topic t.a smoothing count of 1 is included to avoid taking the logarithm of zero.

After BTM calculated the correlation values between a topic and word groupings for each cluster, we outputted the top 30 words with the highest correlation. These words can then be used to determine the major theme (i.e. topic) of the clusters. We tried 8 different K values (k = 5,10, 15, 20,25, 30, 35) in each dataset and then calculated the coherence score based on these words. We then selected the clusters with K values that had the highest coherence score; the topics under this k value are highly distinguishable from other topics.

In Step 2 we remove all the clusters that contained signal associated with news, since the news tweets are re-tweeted ("rt") or discussed by users much more frequently than user-generated tweets. This over saturation of news-related messages (e.g. the original news article itself, the sharing of the article by other users/twitter accounts, and user comments and reactions to an article) can cause signal from self-reported user generated tweets to be obscured. If the top 30 words in a cluster contained news-related words, we considered the entire topic to be associated with a news report and then removed clusters with these topics in order to better isolate user-generated non-news tweets.

For each dataset, we also calculated the frequency of the texts after pre-processing and then selected the top 100 tweets which had the highest frequency and then selected tweets that were clearly related to news events. These news tweets represent the news-topic that were most discussed by Twitter users during the period of data collection. We calculated the tf-idf for each of these words, with the words with higher tf-idf representing words with higher importance in these texts. We then selected the top 15 words that had the highest tf-idf. Examples of such words include: 'trump', 'trumps', 'russia', 'russian', 'campaign', 'president', 'government', 'democrats', 'fbi', 'clinton', 'congress', 'obama', 'election', 'hillary', 'federal'. These words represent news topics and clusters Twitter users talked about most frequently during the dataset's time period. Topics containing these words were highly associated with news or reaction to news, primarily regarding alleged corruption in the U.S. Trump administration.

Most of these topics have a clear signal related to news such as:

'rt', 'wait', 'russia', 'collusion', 'trump', 'story', '1', 'seen', 'evidence', 'buzzfeeds', 'like', 'media', '2', 'hasnt', 'entire', 'years', 'reporter', 'co', '3', 'contact', 'predicated', 'buzzfeed', '4', 'campaign', 'two', 'okay', 'team', 'planned', 'cnn', 'contacts'

This cluster was related to news reports about U.S. President Trump and corruption on major media outlets including CNN and BuzzFeed.

However, there were also other clusters that were harder to distinguish such as:

'rt', 'support', 'opposition', 'dharna', 'mamata', 'corruption', 'parties', 'many', 'power', 'west', 'banerjee', 'sit', 'decided', 'received', 'aspire', 'bengal', 'dear', 'co', 'united', 'divided', 'mahagathbandhan', 'region', 'f1kmiwtdft', 'pakistan', 'serious', 'allegations', 'big', 'dont', 'pre', 'convert'

These topics are considered as "non-signal" topics. These clusters may contain multiple topics that do not present a clear theme or indicate that they are non-user generated messages. In this case, all Twitter messages in this cluster are processed in a separate and dedicated round of BTM specific to this cluster in order to filter out more topics in order to detect potential signal in the remaining topic clusters. Using multiple rounds of BTM to iteratively identify signal (e.g. user-generated tweets) and separate out non-signal (e.g. news-related tweets) topic clusters, we removed all clusters that appeared to be associated with news-related topics. We then put the remaining texts with "non-signal" topics into another iteration of step 1 and 2 until all topics are difficult to identify. At this point we have reached thematic saturation, which means that BTM is no longer able to distinguish or filter topics from the remaining messages.

3. Results

3.1. Twitter corruption topics after unsupervised machine learning

Using our keywords to filter and collect data from the Twitter public API stream, we retrieved a total of 22, 180, 425 tweets from Jan 15- May 15, 2019. The data was separated into 9 parts due to the large volume of messages, with each part containing approximately 10 days' worth of data (the last part contained 11 days of data with data time frames of Jan 15- Jan 27, Jan 28 - Feb 9, Feb 10 - Feb 22, Feb 23 - Mar 7, Mar 8 - March 20, March 21 - April 3, April 4 - April 17, April 18 - May 1, May 2 - May 15). After we eliminated all retweets and tweets with duplicate text, the total number of tweets remaining was reduced to 5,249,939.

Table 1 shows what topic clusters were discovered after each iteration or round of BTM. Generally, we used 4 iterations of BTM to filter the dataset before possible signals were detected. In each of the iterations, we found a total of 5 suspected categories of corruption-related signals that could be associated with self-reporting of corruption activities and experiences. These categories included: (1) corruption in law enforcement or police-related activities; (2) corruption associated with education; (3) corruption in professional sports; (4) corruption in the judicial branch or associated with legal issues/proceedings; and (5) corruption in the healthcare sector (one of the aims of this study). These sector specific corruption topics contained keywords that can easily be identified, with many associated with news events of corruption-related scandals, such as the 2019 college admissions scandal involving several U.S. celebrities and educational institutions, and multiple allegations of corruption associated with The Fédération Internationale de Football Association (FIFA), the international governing body for football/soccer. Below we provide a sample of the top 30 keywords for these topics:

Health-related corruption topic clusters:

['corruption', 'health', 'healthcare', 'ndp', 'fraud', 'ahs', 'alberta', 'pharma', 'services', 'staff', 'years', 'scott', 'scandal', 'abuse', 'house', 'scandals', 'pharmaceutical', 'one', 'mental', 'compliance', 'insurance', 'rick', 'corrupt', 'minister', 'bribe', 'solutions', '000', 'audit', 'history', 'times']

Law enforcement and police topic clusters:

['bribe', 'rt', '000', 'co', 'arrested', 'officer', 'police', 'traffic', 'offered', 'three', 'avoid', 'order', 'rs', 'men', 'c', 'taking', 'gauteng',

Table 1
Twitter Message Topic Cluster Result for each Iteration of BTM.

	1/15–1/27	1/28–2/9	2/10–2/22	2/23–3/7	3/8–3/20	3/21–4/3	4/4-4/17	4/18–5/1	5/2–5/15
Iteration 1	News: 73.6%	News:65.3%	News:71.3%	News:88.2%	News:74.6%	News:78.5%	News:81.1%	News:70.4%	News:72.6%
Iteration 2	News:53.1%	News:42.6% Police:5.6%	News:52.8% Police:6.7%	News:42.6% Police:4.3%	News:51.2%	News:53.6% Police:8.7%	News:48.5%	News:66.6%	News:57.2% Police:1.9%
Iteration 3	News:38.1% Police:13.5%	News:28.9% Police:23.5% Health:8.9%	News:46.8% Police:19.1% Health:8.2%	News:31.8% Police:21.3%	News:24.6 Health:11.5% College:9.9% Police:6.23%	News:46.3% Police:11.2% College:5.4% Health:4.5%	News:44.2% Police:27.1% Health:3.4%	News:36.2% Police:11.6% Health:5.1%	News:53.7% Police:13.7
Iteration 4	News:62.8% Health:2.1%	News:38.4% Police:4.5%	News:26.7% Police:14.8% Health:11.2%	News: 38.2% Police:6.2% Health:11.6%	News:33.4% Police:7.1%	News:29.3% Police:6.23% Health:1.6%	News:28.2% Police:6.3%	News:41.2% Police:6.7% Law:2.4% Sports:1.2%	News:35.4% Police:15.0% Health:1.6% Law:4.4% Sports:3.1%

'mandla', 'r4000', 'sikhonyane', 'accepting', 'councillor', 'temba', '2', 'received', '1', 'accused', 'caught', 'go', '400']

Education-related topic clusters:

'bribery', 'college', 'corruption', 'scandal', 'bribe', 'president', 'arrested', 'temer', 'admissions', 'former', 'michel', 'brazil', 'ex', 'kids', 'yale', 'via', 'student', '1', 'parents', 'million', 'admission', 'probe', 'investigation', 'lori', 'get', 'coach', 'charges', 'paid', 'scheme', 'loughlin'

Sports-related topic clusters:

'corruption', 'league', 'oecd', 'new', 'bahamas', 'collusion', 'football', 'lakers', 'year', 'gov', 'anti', 'one', 'fifa', 'city', 'go', 'means', '2', 'among', 'prosecution', '2019', 'effective', 'top', 'team', 'nassau', 'liverpool', 'bribe', 'survey', 'players', 'like', 'integrity'

Legal and judicial-related topic clusters:

'500', 'subpoenas', 'collusion', 'warrants', 'search', '2', 'witnesses', 'fbi', '40', 'agents', '19', 'lawyers', '800', 'records', '230', '50', '13', 'foreign', '2800', 'evidence', 'mueller', 'requests', 'orders', '000', 'communication', 'million', 'interviewed', 'phone', '25', 'obstruction'

We found that in the first two iterations of BTM, most topics detected were clearly news-related, with the only topic containing likely user-generated messages in the law enforcement topic cluster. This means that some users were likely discussing potential experiences with "police corruption or bribery" during the period of data collection. We note that in the context of bribery-related activities, Transparency International reports that police institutions have the highest perception of being corrupt (International, 2017). In the third and fourth BTM iterations, other non-news topics begin to be detected, such as further user-generated discussions in the law enforcement, healthcare and education-related topic clusters.

3.2. Police bribery and healthcare-related corruption tweets

Based on the results from BTM (see Table 1), we choose two clusters (e.g. police-related and health-related) that contained the largest volume of non-news messages, with the total number of tweets extracted from these two clusters totaling 22,214 tweets (0.01% of the total dataset). In order to confirm if these tweets actually represented true signals of self-reporting of corruption activities and experiences, we manually coded all of these tweets.

After manual annotation, categories of signal content included: (a) tweets self-reporting first-hand experience with corruption; and (b) tweets that contained content revealing or reporting corruption or bribery behavior engaged by another person or organization but that the user purports having some first or secondhand knowledge. We also assessed if tweets included some form of evidence or more specific information about a corrupt act. Based on this coding scheme, 2383 tweets (10.7% of suspected signal tweets) from 1556 users contained true signal, with 1483 related to police bribery as shown by the example in Fig. 2 (a), 514 associated with corruption in the health sector as depicted in Figs. 2 (b), 265 related to administrative corruption issues as shown in Fig. 2 (c), and the remaining 121 messages containing variations of forms of corruption that were not classified for a specific sector or theme but could be considered as general corruption or corruption in international development in Fig. 2 (d).

Based on the detected true signal tweets, we further sub-categorized messages into three categories including: (A) reports where users discuss their personal experience with corruption; (B) messages where users report some type of evidence, such as documents or video of corruption; and (C) messages reporting a statement without specific evidence or experience but contains information about a corruption-related event not news related. Some tweets contained features for both types A and B. In order to distinguish from these two types, we specified that type B tweets should contain information of the reported subject of corruption, such as the name of the person, company or place. Sample tweets for these categories are shown in Fig. 3.

In Table 2, we report police and health-related tweets stratified based on the above sub-categorizations. Compared to police-related reports, most health-related reports were categorized as personal experience, while the police-related tweets usually had specific information or even a user documenting the corrupt act via a video uploaded to Twitter as shown in Fig. 4. Overall, police-related corruption tweets we detected often involved bribery between a motor driver and a police officer. Due to this close interaction some users posted evidence to report such behavior including specific street names or even a police officer's name. In contrast, though bribery of physicians and other healthcare professionals was reported, most messages we reviewed in this topic related to more systematic forms of corruption, such as inappropriate financial relationships between a physician and pharmaceutical or insurance company, though these reports did not necessarily involve witnessing corruption first-hand by the patient. Specifically, health sector reports included corruption related to alleged physician bribery, violation of privacy laws, bribery associated with trying to influence health policy decisions, and concerns about the impact of physicians receiving payments from industry or otherwise having conflicts of interest that impacted patient care.

3.3. Geolocated police bribery and healthcare tweets

In addition to conducting content analysis of signal tweets, we also collected metadata associated with the posts, including any geolocation or geographic identifiers in the metadata of a user's account information or geotagging of the social media post that indicated the location of the

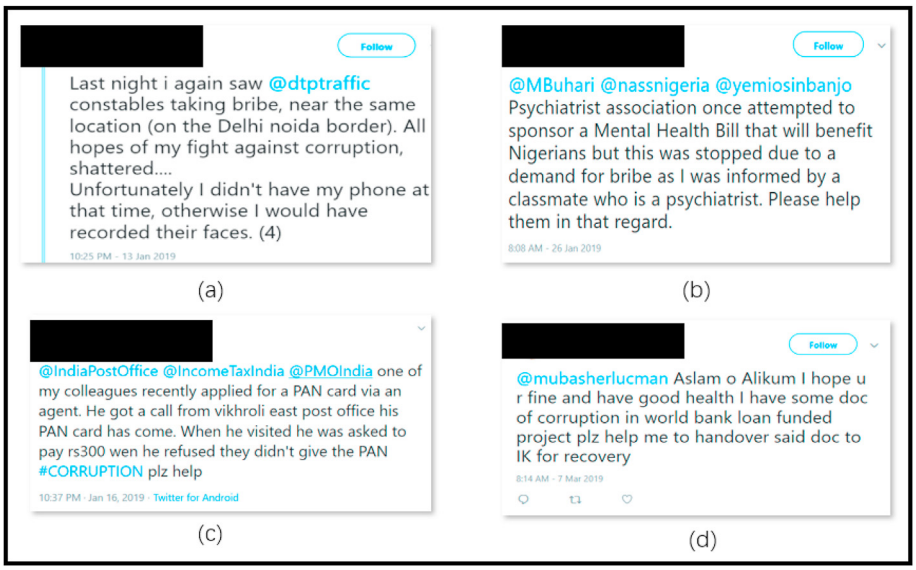

Fig. 2. Different types of corruption-related tweets based on the content of reporting.

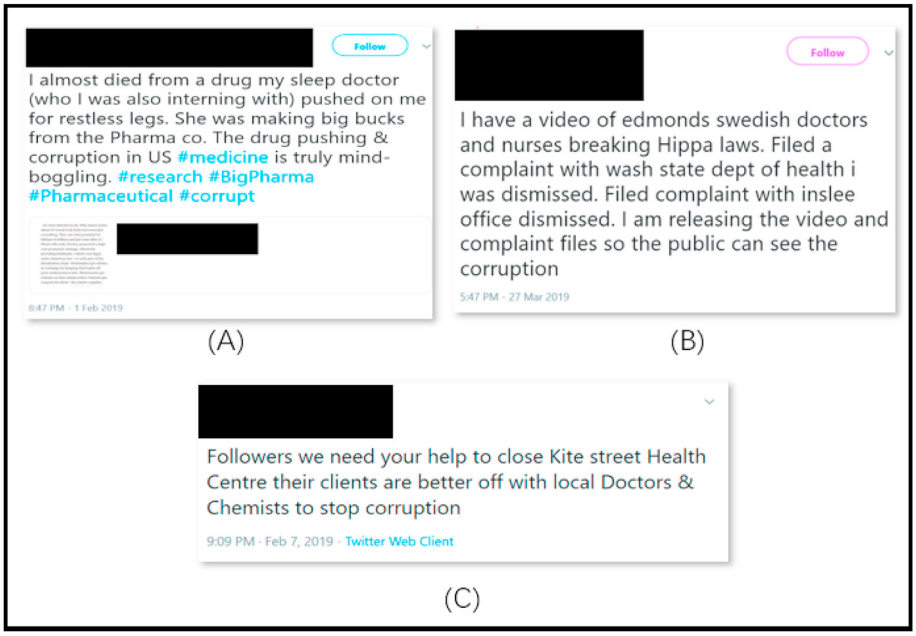

Fig. 3. Example of Corruption-related Tweets based on type of reporting.

Table 2
Number of tweets for different types of corruption reports.

	# of type A tweets (personal experience)	# of type B tweets (tweets with evidence/ documentation)	# of type C tweets (statement about corruption event not news related)
Health-related	317 (13.3%)	32 (1.3%)	165 (6.9%)
Police-related	630 (26.4%)	752 (31.5%)	101 (4.2%)
Total	947 (39.7%)	784 (32.8%)	266 (11.2%)

Legend: (A) reports where users discuss their personal experience; (B) messages where users report some type of evidence, such as a documents or video; and (C) messages reporting a statement without specific evidence or experience but contains information about a corruption-related event not news related.

user (e.g. via geocoordinates, or mention of street, city, county or country name) at the time of the post. However, only 1031 (43.2% of true signals) in the police related tweets and 419 (17.6%) in the health-related tweets had geocoded information. Fig. 5 shows the distribution of geocoded self-reported tweets related to health corruption and Fig. 6 shows the distribution of tweets related to police corruption. Based on these maps, it is

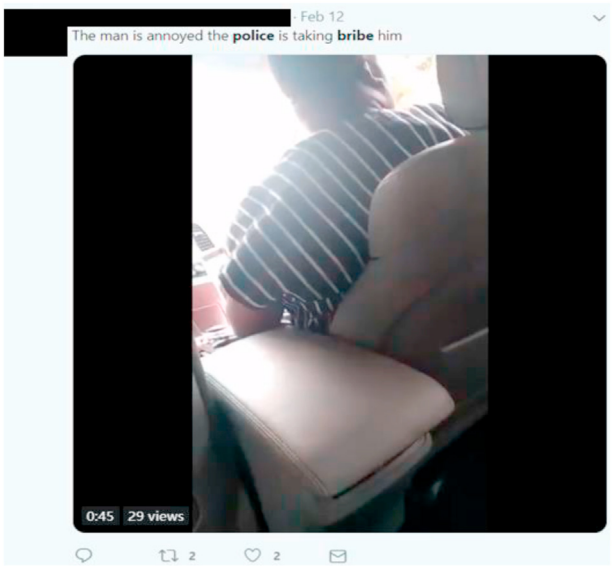

Fig. 4. Tweet of police bribery accompanied by video.

clear that the two groups have similar geographic representation, with the top 3 countries of health-related corruption detected as Nigeria (n = 104), the USA (n = 69) and Kenya (n = 52). Similarly, for police corruption the top 3 countries were Kenya (n = 298), the USA (n = 175), and Nigeria (n = 154). Actual police bribery in the USA is likely low (based on data indicating that only 214 cases involved bribery sentencing in 2018), bringing into question the veracity of some of these Twitter user-generated reports (USSC). In contrast, both Kenya and Nigeria rank low in the Transparency International Corruption Perception Index (tied for rank #144 out of 180 countries respectively) indicating they might be countries at heightened risk for these types of corruption.

3.3.1. Limitations

This study has certain limitations. First our study was limited to the Twitter platform, posts that contained English characters, and focused on specific corruption-related keywords. Specifically, the high presence of tweets located in the United States and Canada detected in our study (Figs. 5 and 6) likely reflects the fact that our dataset was filtered for English-characters and that the highest proportion of Twitter users are located in the United States. This is also impacted by the general skewed geographic global distribution of all tweets, as has been explored in other studies that found a small number of countries (primarily the United States) account for the largest share of the total Twitter user population (Kulshrestha, Kooti, Nikravesh, & Gummadi, 2012). The implications of filtering for English-character tweets and the existing skewed Twitter demographic means that it is difficult to infer broader geospatial trends for tweets on specific corruption-related topics for smaller and non-English speaking countries unless data collection is more targeted. Hence, our results are not generalizable to broader trends in global corruption and bribery. Future studies should conduct multilingual searches on Twitter (in more languages than English and possibly in languages targeted for countries with higher levels of corruption such as those detected in this study) to better estimate the overall volume of corruption-related user generated messages occurring on Twitter globally.

Future studies should also examine other popular social media platforms such as Facebook, YouTube or Instagram or structured online

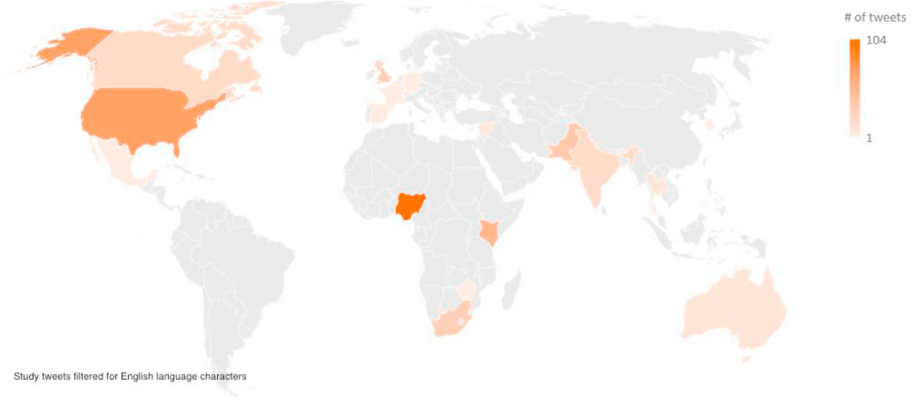

Fig. 5. Study distribution of geocoded health-related corruption tweets.

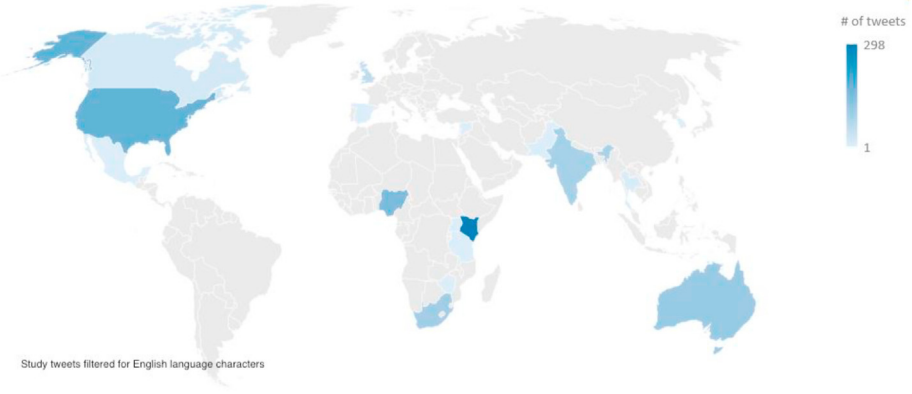

Fig. 6. Study distribution of geocoded police-related corruption tweets.

forums (such as ipaidabribe.com) to better understand the characteristics of user-generated reports about corruption. We also did not specifically examine private messages between users (e.g. direct messages, group chats, etc.) due to privacy considerations and the difficulty of collecting data. Further, we were not able to verify if reported corruption activities actually occurred or had further factual basis to allegations. Additionally, there were specific limitations associated with the methods including that some true positive messages of users self-reporting corruption did not exhibit significant contextual difference from false positives when applying BTM. For instance, the only difference between a true positive and false positive message could be the target/subject of the corrupt act (e.g. a tweet mentioning bribing an official for document processing versus another tweet describing a parent's approach to bribing a child with sweets to take medicine). Future studies will need to further contextualize the difference in meaning and intent of this similarly structured language/conversations that may include corruption-related keywords but may not convey the same meaning.

4. Discussion

We began this study by collecting over 22 million tweets filtered for common corruption-related keywords. Through our use of an unsupervised machine learning approach, we were able to identify clusters of messages that we suspected were associated with self-reporting of corruption-related activities. Though we were able to detect a number of corruption-related topic clusters (including education, sports, judicial, etc.), we decided to focus on two key sectors that we felt were important: police bribery and healthcare.

Bribery is a key issue of anti-corruption primacy, with the SDGs specifically prioritizing bribery as a unit of measure in assessing whether global goals are being met to fight corruption. Bribery topics also appeared to be the primary topic area detected from user self-reporting, particularly as bribery of police may be experienced more broadly than other forms of corruption in the general population, particularly in low- and middle-income countries. Healthcare was important for us to examine in the context of what has been stated before; namely, the dual burden of this form of corruption as it negatively impacts individual and population health.

After applying our machine learning approach and manually annotating for verification, we were only able to detect 2383 total posts (0.0001% of the total dataset) self-reporting corruption experiences. This low volume of true signal reflects the fact that most corruption-related Twitter posts appear to be associated with news or media information, and users reacting or opining to this information. In particular, single events or individuals, such as news and user conversations regarding the Trump Administration in the United States, appeared to predominate our dataset and had the potential to crowd out self-reporting. Hence, actual self-reporting of corruption experiences appears to be low relative to the total volume of tweets analyzed, though it is clear that some individuals are actively reporting their experiences.

Within the tweets of corruption self-reporting, the majority of reports in both the police-related and health-related categories included messages about a personal experience with corruption (39.7%, $n = 947$) or included some type of evidence associated with the corrupt act (32.8%, $n = 784$). This demonstrates that though Twitter is not purposefully used as a platform to collect structured data about corruption, it could nevertheless act as a tool for the public to report and share this information to other social media users or for the purposes of disseminating it for raising awareness and social mobilization. In fact, some posts we identified in the police bribery category seemed to warn other users that a specific location should be avoided due to the need to pay a bribe.

Importantly, this study can also inform progress towards SDG Goal 16, Target 16.5, and specifically indicators 16.5.1 and 16.5.2 that examine the proportion of persons and businesses that have paid or were asked to pay a bribe to a public official. Currently, the SDG Goal 16 progress and info section for 2018 and 2017 only list that one in five firms reported receiving a bribe request for regulatory or utility transactions and in 2015, an estimated 18% of firms worldwide reported receiving at least one bribery payment request. Hence, it appears that most of the focus on measuring SDG progress is on identifying self-reporting by private firms based on their own activities involving bribery or requests for bribery from public sector officials. While informative, this may miss other reporting focusing on individuals with first-hand corruption-related experiences or may suffer from underreporting by firms who do not wish to disclose bribery-related activities that they may or may not be part of. Hence, approaches such as the one conducted in this study will be important to gain a more holistic picture and better estimate of global corruption activities, while also providing more nuance to the types of corruption that occur in specific sectors and geographies.

Mobilizing these technology-based approaches to detecting and characterizing corruption will require global partnerships seeking to leverage shared goals of the SDGs. This includes the need to build a coalition of like-minded stakeholders from UN specialized agencies (i.e. the World Health Organization, UN Office of Drugs and Crime, and UN Development Programme), civil society, and national governments. These partnerships should also assess if existing international treaty instruments (e.g. the UN Convention against Corruption) or other domestic law (e.g. the Foreign Corrupt Practices Act or UK Anti-bribery Act) can act as the basis for coherent global policy action and data collection that is evidence-based and can help with prevention, enforcement, and prosecution (Mackey et al., 2016; 2017b; Mackey & Liang, 2012). These efforts are starting to take tangible shape, as the recent consultation on the ACTA Alliance drive global dialogue, exchange anti-corruption solutions and experiences, and forge a plan of action to operationalize the network (Mackey, 2019).

Despite these developments, if past is prologue, corruption is unlikely to be eradicated even with renewed attention and innovation, as corruption is an endemic problem to society. Fortunately, there is widespread consensus from world leaders regarding the serious consequences of corruption, catalyzed by the 1996 speech by former WB President James Wolfensohn's, calling for global action against the "cancer of corruption." These sentiments were echoed by former World Bank President Jim Yong Kim, who denounced corruption as "public enemy number one," and also former U.S. Secretary of State John Kerry, who characterized corruption as a "pandemic." Reflecting this increased attention, in May 2016, world leaders from over 40 countries (along with civil society and private sector representatives) joined for an anti-corruption summit hosted by UK Prime Minister David Cameron culminating with a Global Declaration to "expose," "pursue and punish" and "substantially reduce corruption and bribery in all their forms."

However, global condemnation of corruption will not be enough. Though steps have been taken to raise the alarm about corruption, it is clear that focused and reinvigorated political will and advocacy is needed and should be supported by more data. Herein lies the potential of the SDGs to act as a "shared" governance space for global society to coalesce and act to solve these challenges in a measurable and systematic way. Specifically, if technology-based solutions or infrastructure (such as e-government, data mining and machine learning for corruption detection) can be integrated into anti-corruption tools through global partnerships, simultaneous progress towards SDGs 16, 3, 9, and 17 can be achieved (see Table 3).

Clearly technology will not act as a "silver bullet" to entirely address the multifaceted and complex nature of global corruption, but this study demonstrates the potential utility of a big data and machine

Table 3
UN SDGs that align with combating corruption from a health, technology and global partnership standpoint.

SDG Goal	SDG Target(s) (relevant language bolded)	Rationale for Inclusion
Goal 3: Ensure healthy lives and promote well-being for all at all ages	3.8: **Achieve universal health coverage, including financial risk protection, access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all** 3.B: Support the research and development of vaccines and medicines for the communicable and non-communicable diseases that primarily affect developing countries, **provide access to affordable essential medicines and vaccines**, in accordance with the Doha Declaration on the TRIPS Agreement and Public Health, which affirms the right of developing countries to use to the full the provisions in the Agreement on Trade-Related Aspects of Intellectual Property Rights regarding flexibilities to protect public health, and, in particular, provide access to medicines for all	3.8: Several Goal 3 targets benefit from improved medicines procurement. However, this target specifically calls for access to safe, effective, quality and affordable medicines, an objective directly undermined by corruption in medicines procurement. 3.B: Target also calls for access to essential medicines and vaccines.
Goal 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels	16.5: **Substantially reduce corruption and bribery in all their forms** 16.6: **Develop effective, accountable and transparent institutions at all levels**	16.5: Includes language that calls for reducing corruption and bribery, which is inclusive of corruption-related activities as they relate to medicines procurement. 16.6: Target generally calls for greater transparency and accountability for institutions at all levels, which is inclusive of efforts to improve transparency and accountability in medicines procurement, health system financing, and using technology to enhance transparency and access to data that may relate to corrupt practices.
Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	9.5: Enhance scientific research, **upgrade the technological capabilities of industrial sectors in all countries,** in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers per 1 million people and public and private research and development spending 9.B: **Support domestic technology development**, research and innovation in developing countries, including by ensuring a conducive policy environment for, inter alia, industrial diversification **and value addition to commodities**	9.5: Includes language that calls for upgrades of technological capabilities within the context of "resilient infrastructure". This includes the health and pharmaceutical sector where improved medicines procurement could lead to sustainable human development. 9.B: Domestic technology development can include technology to strengthen health supply chains which can also bring additional value to health commodities by ensuring they are safe and effective.
Goal 17: Strengthen the means of implementation and revitalize the global partnership for sustainable development	17.6: Enhance North-South, South-South and triangular regional and international cooperation on and access to science, technology and innovation and enhance knowledge sharing on mutually agreed terms, including through improved coordination among existing mechanisms, in particular at the United Nations level, and through a global technology facilitation mechanism people and public and private research and development spending 17.16: Enhance the **global partnership for sustainable development, complemented by multi-stakeholder partnerships** that mobilize and share knowledge, expertise, technology and financial resources, **to support the achievement of the sustainable development goals** in all countries, in particular developing countries	17.6.1: Number of science and/or technology cooperation agreements and programmes between countries, by type of cooperation 17.16.1: Number of countries reporting progress in multi-stakeholder development effectiveness monitoring frameworks that support the achievement of the sustainable development goals .

learning approaches to shed additional light on corruption with the hopes of enabling next generation anti-corruption solutions. Hopefully technology can be used as a force for good to address corruption, which continues to act as a threat to human development, health and prosperity.

CRediT authorship contribution statement

Jiawei Li: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Wen-Hao Chen:** Data curation, Formal analysis, Investigation, Methodology. **Qing Xu:** Formal analysis, Investigation, Project administration. **Neal Shah:** Formal analysis, Investigation, Writing - original draft. **Jillian C. Kohler:** Funding acquisition, Supervision, Writing - original draft. **Tim K. Mackey:** Conceptualization,

Formal analysis, Investigation, Methodology, Visualization, Supervision, Funding acquisition, Project administration, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Author has received funding from the World Health Organization to attend international meetings and workshops associated with addressing health corruption and also received funding to attend and participate in the February 2019 joint consultation held by the World Health Organization (WHO), the Global Fund, and the UNDP for a proposed Global Network on Anti-Corruption, Transparency and Accountability. Authors received support from the WHO Collaborating Centre for Governance, Transparency and Accountability in the Pharmaceutical Sector, University of Toronto to carry out this study. Authors report no other conflict of interest associated with this manuscript.

References

Adam, I., & Fazekas, M. (2018, December). Are emerging technologies helping win the fight against corruption in developing countries? *Govtransparency.Eu*. Retrieved March 4, 2020, from http://www.govtransparency.eu/wp-content/uploads/2019/02/ICT-corruption-24Feb19_FINAL.pdf.

Aldaz-Carroll, E., & Aldaz-Caroll, E. (2018, February 1). Can cryptocurrencies and blockchain help fight corruption? *Brookings.Edu*. Retrieved March 4, 2020, from https://www.brookings.edu/blog/future-development/2018/02/01/can-cryptocurrencies-and-blockchain-help-fight-corruption/.

Banning-Lover, R. (2016, May 26). Nine ways to use technology to reduce corruption. *Theguardian.com*. Retrieved March 4, 2020, from http://www.theguardian.com/global-development-professionals-network/2016/may/26/nine-ways-to-use-technology-to-reduce-corruption.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., et al. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295–298. https://doi.org/10.1038/nature11421

Corruption costs developing countries $1.26 trillion every year - yet half of Emea think it's acceptable. Corruption costs developing countries $1.26 trillion every year - yet half of EMEA think it's acceptable (n.d.). Weforum.org. Retrieved February 8, 2020, from https://www.weforum.org/agenda/2019/12/corruption-global-problem-statistics-cost/.

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research, 11*(1), e11. https://doi.org/10.2196/jmir.1157

Gaitonde, R., Oxman, A. D., Okebukola, P. O., & Rada, G. (2016). Interventions to reduce corruption in the health sector. *Cochrane Database of Systematic Reviews, 135*(8), 1. https://doi.org/10.1002/14651858.CD008856.pub2

Gee, J., & Button, M. (2015, September). *The financial cost of healthcare fraud 2015: What data from around the world shows*. Pianoo.Nl. Retrieved February 13, 2019, from: https://www.pianoo.nl/sites/default/files/documents/documents/thefinancialcostofhealthcarefraud-september2015.pdf.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Gupta, S., Davoodi, H., & Tiongson, E. (2000, June). Corruption and the provision of health care and education services. *Imf.org*. Retrieved June 13, 2013, from http://www.imf.org/external/pubs/ft/wp/2000/wp00116.pdf.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior, 29*(3), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

Hanf, M., Van-Melle, A., Fraisse, F., Roger, A., Carme, B., & Nacher, M. (2011). Corruption kills: Estimating the global impact of corruption on children deaths. *PLoS One, 6*(11), Article e26990. https://doi.org/10.1371/journal.pone.0026990

Han, S., & Kavuluru, R. (2016). Exploratory analysis of marketing and non-marketing E-cigarette themes on twitter. In *Social informatics : 8th international conference, SocInfo 2016, bellevue, WA, USA, november 11-14, 2016, proceedings. Part II. SocInfo (conference) (8th : 2016 : Bellevue, Wash.), 10047* pp. 307–322). https://doi.org/10.1007/978-3-319-47874-6_22, 2.

Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2018). Approaches for identifying U.S. medicare fraud in provider claims data. *Health Care Management Science, 25*(6), 1603–1618. https://doi.org/10.1007/s10729-018-9460-8

Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018). Big Data fraud detection using multiple medicare data sources. *Journal of Big Data, 5*(1), 1–21. https://doi.org/10.1186/s40537-018-0138-3

Holeman, I., Cookson, T. P., & Pagliari, C. (2016). Digital technology for health sector governance in low and middle income countries: A scoping review. *Journal of Global Health, 6*(2), Article 020408. https://doi.org/10.7189/jogh.06.020408

International, T. (2017, November 14). Global corruption barometer: Citizens' voices from around the world. *Transparency.org*. Retrieved April 9, 2020, from https://www.transparency.org/news/feature/global_corruption_barometer_citizens_voices_from_around_the_world.

Intravia, J., Wolff, K. T., Paez, R., & Gibbs, B. R. (2017). Investigating the relationship between social media consumption and fear of crime: A partial analysis of mostly young adults. *Computers in Human Behavior, 77*, 158–168. https://doi.org/10.1016/j.chb.2017.08.047

Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., & Fowler, J. H. (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. *PLoS ONE, 12*(4), Article e0173851. https://doi.org/10.1371/journal.pone.0173851

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., et al. (2015a). Improving fraud and abuse detection in general physician claims: A data mining study. *International Journal of Health Policy and Management, 5*(3), 165–172. https://doi.org/10.15171/ijhpm.2015.196

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., et al. (2015b). Improving fraud and abuse detection in general physician claims: A data mining study. *International Journal of Health Policy and Management, 5*(3), 165–172. https://doi.org/10.15171/ijhpm.2015.196

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., et al. (2015c). Using data mining to detect health care fraud and abuse: A review of literature. *Global Journal of Health Science, 7*(1), 194–202. https://doi.org/10.5539/gjhs.v7n1p194

Kalyanam, J., & Mackey, T. (2017, December 1). *Detection and characterization of illegal marketing and promotion of prescription drugs on twitter*. arXiv.org.

Kohler, J. C., Chang Pico, T., Vian, T., & Mackey, T. K. (2018). The global wicked problem of corruption and its risks for access to HIV/AIDS medicines. *Clinical Pharmacology & Therapeutics, 104*(6), 1054–1056. https://doi.org/10.1002/cpt.1172

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. P. (2012). *Geographic dissection of the twitter network*. Sixth International AAAI Conference on Weblogs and Social Media.

Lalountas, D. A., Manolas, G. A., & Vavouras, I. S. (2011). Corruption, globalization and development: How are these three phenomena related? *Journal of Policy Modeling, 33*(4), 636–648.

Li, J., Chen, W. H., Xu, Q., Shah, N., & Mackey, T. K. (2019). Leveraging big data to identify corruption as an SDG goal 16 humanitarian technology. In *2019 IEEE global humanitarian technology conference (GHTC), seattle, WA, USA* (pp. 1–4). https://doi.org/10.1109/GHTC46095.2019.9033129, 2019.

Mackey, T. K. (2019). Opening the policy window to mobilize action against corruption in the health sector comment on "we need to talk about corruption in health systems. *International Journal of Health Policy and Management, 8*(11), 668–671. https://doi.org/10.15171/ijhpm.2019.65

Mackey, T. K., & Kalyanam, J. (2017). *Detection of illicit online sales of fentanyls via Twitter, 6* p. 1937). https://doi.org/10.12688/f1000research.12914.1. F1000Research.

Mackey, T. K., Kalyanam, J., Katsuki, T., & Lanckriet, G. (2017). Machine learning to detect prescription opioid abuse promotion and access via twitter. *American Journal of Public Health, 107*(12), e1–e6. https://doi.org/10.2105/AJPH.2017.303994

Mackey, T. K., Kohler, J., Lewis, M., & Vian, T. (2017). Combating corruption in global health. *Science Translational Medicine, 9*(402), Article eaaf9547. https://doi.org/10.1126/scitranslmed.aaf9547

Mackey, T. K., Kohler, J. C., Savedoff, W. D., Vogl, F., Lewis, M., Sale, J., et al. (2016). The disease of corruption: Views on how to fight corruption to advance 21(st) century global health goals. *BMC Medicine, 14*(1), 149. https://doi.org/10.1186/s12916-016-0696-1

Mackey, T. K., & Liang, B. A. (2012). Combating healthcare corruption and fraud with improved global health governance. *BMC International Health and Human Rights, 12*(1), 23. https://doi.org/10.1186/1472-698X-12-23

Mackey, T. K., & Liang, B. A. (2013). Improving global health governance to combat counterfeit medicines: A proposal for a UNODC-WHO-interpol trilateral mechanism. *BMC Medicine, 11*, 233. https://doi.org/10.1186/1741-7015-11-233

Mackey, T. K., & Nayyar, G. (2017). A review of existing and emerging digital technologies to combat the global trade in fake medicines. *Expert Opinion on Drug Safety, 16*(5), 587–602. https://doi.org/10.1080/14740338.2017.1313227

Mackey, T. K., Vian, T., & Kohler, J. (2018). The sustainable development goals as a framework to combat health-sector corruption. *Bulletin of the World Health Organization, 96*(9), 634–643. https://doi.org/10.2471/BLT.18.209502

Michaud, J., Kates, J., & Oum, S. (2015, May 8). *Corruption and global health: Summary of a policy roundtable. Kff.org*. The World Bank. Retrieved January 30, 2018, from.

Nishtar, S. (2010). Corruption in health systems. *Lancet, 376*(9744), 874. https://doi.org/10.1016/S0140-6736(10)61413-4

Patton, D. U., Eschmann, R. D., & Butler, D. A. (2013). Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Computers in Human Behavior, 29*(5), A54–A59.

Promise and peril: blockchain, Bitcoin and the fight against corruption. (2018, January 31). Promise and peril: Blockchain, Bitcoin and the fight against corruption. Transparency.org. Retrieved March 4, 2020, from https://www.transparency.org/news/feature/blockchain_bitcoin_and_the_fight_against_corruption.

Reynolds, L., & McKee, M. (2010). Organised crime and the efforts to combat it: A concern for public health. *Globalization and Health, 6*, 21. https://doi.org/10.1186/1744-8603-6-21

Shim, D. C., & Eom, T. H. (2008). E-government and anti-corruption: Empirical analysis of international data. *Intl Journal of Public Administration, 31*(3), 298–316. https://doi.org/10.1080/01900690701590553

Silveria, L. (2016, April 18). *4 technologies helping us to fight corruption*. Weforum.org. Retrieved March 4, 2020, from https://www.weforum.org/agenda/2016/04/4-technologies-helping-us-to-fight-corruption/.

Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly, 29*(4), 470–479. https://doi.org/10.1016/j.giq.2012.06.005

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291. https://doi.org/10.1007/s13278-012-0079-3

Technology against corruption. (2013, May 8). Technology against corruption. Transparency.org. Retrieved March 4, 2020, from https://www.transparency.org/news/feature/technology_against_corruption.

Transparency International. (2006). *Global corruption report 2006: Corruption and health*. Transparency.org. Retrieved July 17, 2017, from https://www.transparency.org/whatwedo/publication/global_corruption_report_2006_corruption_and_health.

Transparency International, & Global Corruption Report 2006. (2005). *Global corruption report 2006*. London: Transparency International.

U4. Corruption in the health sector. U4 anti-corruption resource centre (n.d.). Retrieved October 27, 2011, from http://www.cmi.no/publications/file/3208-corruption-in-the-health-sector.pdf.

UNDP. (2011, October). *Fighting corruption in the health sector: Methods, tools and good practices*. Undp.org.Tt. Retrieved May 7, 2012, from http://www.undp.org.tt/News/UNODC/Anticorruption%20Methods%20and%20Tools%20in%20Health%20Lo%20Res%20final.pdf.

USSC. Quick facts: Bribery offenses (n.d.). Ussc.Gov. Retrieved April 9, 2020, from https://www.ussc.gov/sites/default/files/pdf/research-and-publications/quick-facts/Bribery_FY18.pdf.

Vian, T. (2008). Review of corruption in the health sector: Theory, methods and interventions. *Health Policy and Planning, 23*(2), 83–94. https://doi.org/10.1093/heapol/czm048

WHO. (2010). The world health report: Health systems financing - the path to universal coverage. *Who.Int.* Retrieved February 13, 2019, from https://www.who.int/whr/2010/10_chap04_en.pdf.

Wickberg, S. (2013, March 28). *Technological innovations to identify and reduce corruption*. Retrieved March 4, 2020, from https://www.u4.no/publications/technological-innovations-to-identify-and-reduce-corruption/.

Witvliet, M. I., Kunst, A. E., Arah, O. A., & Stronks, K. (2013). Sick regimes and sick people: A multilevel investigation of the population health consequences of perceived national corruption. *British Journal of Clinical Pharmacology, 18*(10), 1240–1247. https://doi.org/10.1111/tmi.12177

Xu, Q., Li, J., Cai, M., & Mackey, T. K. (2019). Use of machine learning to detect wildlife product promotion and sales on twitter. *Frontiers in Big Data, 2*. https://doi.org/10.3389/fdata.2019.00028, 1989.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). *A biterm topic model for short texts. the 22nd international conference*. New York, New York, USA: ACM. https://doi.org/10.1145/2488388.2488514