Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology

1st Jiawei Li
UCSD Extension
University of California, San Diego
San Diego, USA
jil206@ucsd.edu

2nd Wen-Hao Chen
Department of Computer Science and Engineering
University of California, San Diego
San Diego, USA
w2chen@ucsd.edu

3rd Qing Xu
UCSD Extension
University of California, San Diego
San Diego, USA
qix036@ucsd.edu

4th Neal Shah
UCSD Extension
University of California, San Diego
San Diego, USA
neal.shah417@gmail.com

5th Timothy Mackey
Department of Anesthesiology, School of medicine
University of California, San Diego
San Diego, USA
tmackey@ucsd.edu

***Abstract*—Corruption is a serious impediment to global goals of ensuring sustainable development and is now a threat specifically recognized in the UN Sustainable Development Goals under Target 16.5. Though corruption remains challenging to identify, measure, and combat, technology advances provide new opportunities to advance humanitarian goals, including the detection of corruption reported by the public. In this study, we address this challenge by developing a method using an unsupervised machine learning model to detect reports of corruption-related activity on the micro-blogging platform Twitter. In total, we collected over 6 million tweets containing keywords related to corruption between January and February 2019. We use the Biterm Topic Model to then isolate tweets from users who report corruption and found that most topics focus on police bribery and corruption in health-care. Though preliminary, these results shave the potential of identifying the scope and prevalence of corruption in society and also advance shared goals of combating corruption and advancing sustainable development in the 21st century. Index TermsCorruption, Machine Learning, Natural Language Processing, Topic modeling**

***Index Terms*—Corruption, Machine Learning, Natural Language Processing, Topic modeling**

I. Introduction

Corruption is a key challenge in achieving shared humanitarian goals of equity, economic progress, strong institutions, human rights, upholding the rule of law, and ensuring sustainable development. Recognizing the corrosive impact of corruption on sustainable global development, the United Nations Sustainable Development Goals (SDGs) calls for reducing corruption and bribery in all their forms as part of Goal 16 that promotes accountable and inclusive institutions and access to justice for global society [1]. However, corruption is notoriously hard to detect and quantify, given its criminal nature and the fact that corruption thrives in environments with lack of transparency and accountability and weak governance [2]. However, without knowing the true scope and spread of corruption, it is difficult to develop and implement evidence-based anti-corruption interventions and policies.

Importantly, due to the opaqueness of corruption-related activities, directly measuring corruption can be difficult. In response, the anti-corruption community is actively exploring proxy indicators as a means to measure in-direct data points that may point to the presence of corruption. Further, advances in technology, including big data and machine learning approaches, are creating new opportunities to analyze large datasets to detect and characterize corruption (e.g. such as use of data mining and machine learning to detect fraudulent healthcare reimbursement claims [3]). Websites such as www.ipaidabribe.com are also leveraging technology to directly source reports about corruption from the public [1]. Finally, the use of social media globally is growing, with the Pew Research Center finding [4] that across 39 countries, approximately 53% use an online social network site (such as Twitter or Facebook). Conversations on social media platforms vary, including news, promotional content, bot traffic, and user-generated content, that cover a wide range of topics. This includes conversations about corruption, which includes news reports about corruption events (e.g. bribery of public officials, prosecutions of people for corrupt activities, corruption in politics, etc.), perceptions and attitudes related to corruption, and actual self-reporting of corruption by the public that have been detected through manual searches.

These developments create new opportunities to conduct big data surveillance of social media to detect corruption trends and activities. For example, on average around 6,000 tweets are [2] posted on Twitter per second. From these conversations, we are interested in identifying tweets that self-report corruption activities which may not be reported otherwise. This could help uncover new data points about the presence of corruption, types of corruption being reported, and communities and users that are being impacted and could

[1] https://www.ipaidabribe.com/#gsc.tab=0
[2] http://www.internetlivestats.com/twitter-statistics/

aid in making progress towards SDG 16.5. Hence, in this study we describe the use of big data and machine learning approaches to detect conversations about corruption activities via the popular microblogging social media platform Twitter. Specifically, in this pilot study we are trying to identify how likely people are to report corruption issues and what kind of corruption activities they report for the purposes of testing a proof-of-concept data collection and classification system as a potential corruption surveillance and reporting mechanism.. However, report-based (or identication-based) tweets have a relatively small volume compare to discussion-based tweets (i.e. tweets about attitudes towards corruption, etc.), which makes detection of these tweets more difficult.

The remainder of this paper is structured as follows. In section 2 we describe our data collection process and the corpus of data we collected, as well as the model used to analyze data. In section 3, we describe the output of our model and what specic corruption related messages we identified In section 4, finally we discuss how these results can help catalyze progress towards shared anti-corruption SDG goals, targets, and indicators.

II. METHODOLOGY

Our methods consist of three distinct phases: data collection, data processing and data analysis. In data analysis, we have two parts, part1: filtering data using an unsupervised machine learning model, and part2: labeling the data from part 1 and using it as a training set for a supervised machine learning model. Primarily, this study focuses on exploration of Twitter messages without a preexisting training data set and is focused on text analysis using an unsupervised machine learning approach. The overall goal of the study is to identify, classify, and report types of corruption-related activities as self-reported by Twitter users.

A. Data Collecting

We collected data from Twitter using the public streaming application programming interface (API) over approximately 30 days, from January 15 2019 to February 13 2019. The data includes the text of the tweet and other metadata associated with the message (geocoded data, time stamp data, user account information, etc.). In order to curate the dataset specic to potential topics and conversations associated with corruption, we chose a list of corruption-related keywords to lter the API for including: bribery, bribe, corruption, nepotism, collusion and kickback.

B. Data Processing

After we collected data, we extract the text from each Tweet. However, the majority of text in these Tweets contained a lot of noise (i.e. messages or content in a message that are irrelevant to the study aims) unrelated to core issues related to corrupt behaviors, attitudes, or events. For example the following Tweet expresses opinions about international development, but is not specific to corruption:

> @ejwwest The UN did just that asked almost 10m people across the world for their top priorities. The world wants education, health, jobs, no corruption and nutrition Climate at the *very* bottom. Now we've consulted with them - will you respect their wishes? https://t.co/cCG9MWhlyv

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

- Imbedded Hyperlinks: The hyperlink in the text does not provide much information unless we further analyze the website that the link resolves to. Twitter has its own way to represent these kind of links using shortened URLs (beginning with https://t.co/-)
- 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 our data. The NLTK package has a list of all stop words, which we use to filter out those words in texts.
- Special characters and punctuation marks: Special characters like emoji and punctuation marks could exhibit certain meaning or sentiment. However, in this study, we do not prioritize these characters as they can be subjective in interpretation in relation to text.

The following is an example after applying the above steps to the original text example:

> un asked almost 10m people across world top priorities world wants education health jobs corruption nutrition climate bottom consulted respect

The remaining words represent key concepts/themes of the text. We then remove all text that consist of less than three words to further filter messages to focus on thematic meaning.

C. Data Analysis

In the first part of our data analysis we need to clean the data (get training data) using an unsupervised machine learning model. Since this study is focusing on self-reporting of purported corruption-related activities and behavior, we need to remove tweets that are not user-generated. This includes removing tweets related to: (1) news articles or reports (particularly since news-related messages regarding U.S. politics and the Trump administration dominate our ltered dataset for corruption keywords); (2) promotional or commercial content; and (3) bot trafc. This process of removing non user generated content was completed in three steps.

In step 1, we calculate the frequency of each word and extract the first 200 words as s1. These are the words very likely to be used in this data set. Then we randomly select 500 news related tweets, calculating the frequency of each word and extracting the first 200 words from them as s2. We then intersect s1 and s2 which gives us 15 words: trump, trumps, russia, russian, campaign, president, government, democrats, fbi, clinton, congress, obama, election, hillary, federal. These words are highly related to news events and articles, including conversations from users, though are not the focus of this study. We store these words in a new word set s3.

In step 2, we use the biterm topic model (BTM) [5] which has been adopted in some studies for thematic detection [6]–[8]. BTM is an unsupervised topic model which learns the distribution of biterms. Biterm is defined as a pair of words occurring in a piece of text. Texts with similar biterm distribution can then be clustered to the same topic. For each topic, BTM outputs the correlation values between a topic and word groupings. We output the top 20 words with highest correlation, with these word groupings representing highly correlated clusters of words that can be interpreted as topic buckets.

Since the raw data contains a lot of noise, most news-related topics we get contain words 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'

Other words that are still news-related but appear less frequently include:

- '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'

Another reason for the high saturation of news tweets is that they are re-tweeted("rt") or discussed by users much more frequently than user-generated tweets. This oversaturation of news can cause signal from self-reported user generated tweets to be obscured. To locate this type of signal data, we need to remove the topics that contain noise.

In step 3, we compare the top 20 words for each topic with the words in s3. If the top 20 words contain at least one word in s3, we consider this topic to be associated with a news report. We then remove the clusters with these topics. We also keep the topics that contain non-news signal such as:

- 'drug', 'corruption', 'pharma', 'bribe', 'police', 'money', 'doctors', 'drugs', 'kickback', 'get', 'scheme', 'said', 'big', 'force', 'pharmaceutical', 'home', 'alleged', 'charged', 'anti', 'bribery', 'immigration', 'fentanyl', 'million', 'patients', '000', 'corrupt', 'conspiracy', 'companies', 'ato', 'allegations'

The remaining topics will be those that generally do not contain news related signal and also do not show strong signal association with other topics. We will than repeat step3 on the remaining tweets in the data set. Using this method, we further filter out the tweets that contain weak signal and isolate user-generated tweets that contain corruption-related keywords and themes. Fig 1 shows the basic structure of our methods.

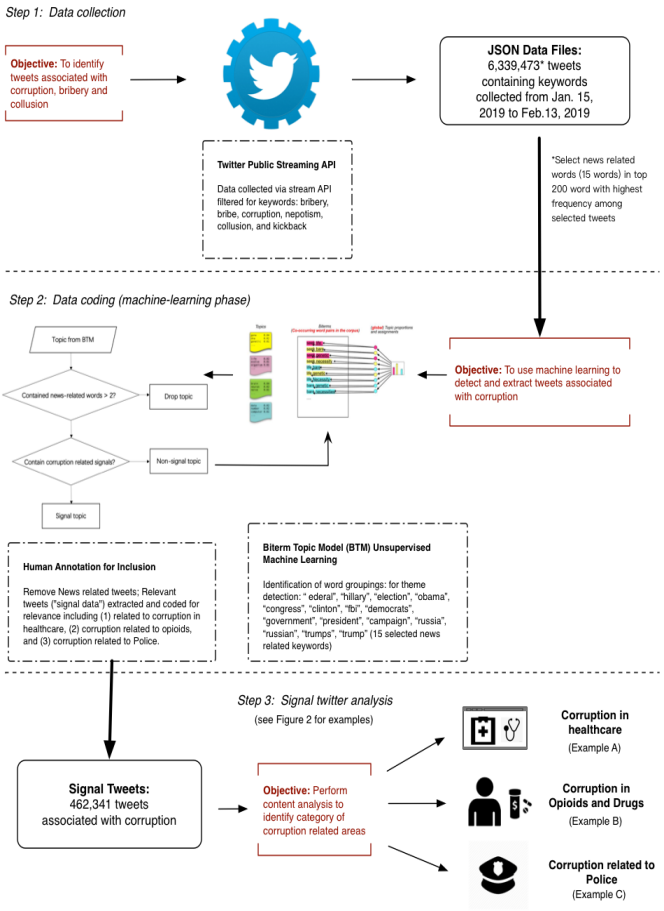

Fig. 1. Method of analyzing twitter data for corruption related tweets

In part 2, we labeled the data we get from part 1, we filter out the tweets that are most likely related to the news, then we retrieve the tweets that are user reports by manually coding, which includes messages related to different corruption topics as we report in the Results section. However, due to the large volume of the data, its difficult to manually annotate the whole dataset, therefore, we only use part of the data as a training set. We use Support Vector Machine (SVM) [9] to learn the feature of the texts, and use a trained model to nd similar tweets from the rest of the data (the data that doesnt contain our labeling data). The model is trained using labeled text after we have applied data processing.

III. RESULTS

We collected a total of 6339473 tweets from Jan 15 to February 13 2019, and we ran three iterations on step 3 to lter for the signal data we were trying to detect. The results are shown in table 1. The rst row is the round of the iteration. The second row shows the topic we found and the percentage between the number of tweets in the selected topics compared to the total number of tweets in that iteration. In the rst two iterations, most topics are news-related. The other topics show very weak signal that cannot be easily distinguished. In the third and fourth iteration, other topics relevant to corruption in particular sectors of society begin to show up, such as law enforcement, health care and pharmaceuticals education.

TABLE I
TOPIC RESULT FOR EACH ITERATION

	Iteration 1	Iteration 2	Iteration 3	Iteration 4
Topic	News: 81 %	News: 31.6%	News: 22.2% Health: 15.9% Drug: 12.3% Police:6.23%	News: 8.2% Health: 11.6% Drug: 12.3% Police:6.23% Education:2.1%

We then specifically analyzed the topics we extracted from iteration 3 and 4, which contain 462341 tweets. We randomly selected 3000 tweets, and then we manually annotated this sample of tweets to assess whether any directly reported corruption activities. Importantly, comments to tweets were not included in this analysis. After manual annotation, we found 235 users tweet and re-tweets reporting corruption related activities. One hundred and twenty-three were associated with reporting police bribery, 37 reported corruption related to trafc and motor violations involving bribery, 18 reported purported corruption in the pharmaceutical sector (e.g. corruption related to physician-industry relationships and conicts of interest), 18 were related to corruption in health care (e.g. bribery related to health care policy, access and infrastructure projects), 9 were related to bribery associated with issuance of government documents/passports/identity documents, with the remaining 30 related to nancial and company bribery or topics not easily categorized.

In the second part of data analysis, these data are used as a training set for a machine learning model. After processing the data, we use SVM to learn the word vector from each text of the 3000 tweets which we have already labeled, then we apply the model to the rest of the data (the original 6339473 tweets). The model detected 2051 tweets with 1245 tweets related to user reports which comprised of 903 Twitter users. One thousand and fty two of them are related to police bribery, 45 of them are related to health care and pharmacy, 31 are related to law and government issues (non-health care), the remaining 117 belonged to different types of reports such as education/ticket selling or types that are not easy to identify.

Figure 2 shows one example of police bribery (including a post where alleged bribery is taking place and a video of the bribery act is recorded) and an example of corruption related to pharmaceutical prescribing (where a patient alleges that their physician has an inappropriate relationship with industry).

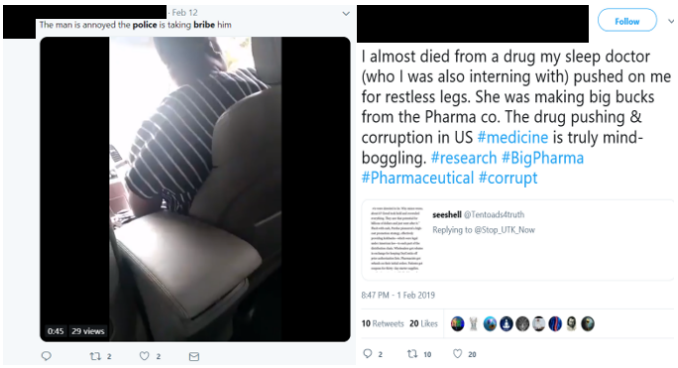

Fig. 2. Twitter about police bribery and pharmacy corruption

We also build a map to show the distribution of the manually annotated corruption Tweets that have geographic data based in different countries in Fig 3, unfortunately, not many of the result contain identical geo-coded information. We find that Australia, India and the USA are the top three countries that have corruption tweets, but there are only 2 people reporting in Australia, while there were 34 for India and 23 for the USA.

In this study, one limitation is that user corruption reporting tweets and unrelated tweets do not exhibit significantly contextual difference after we

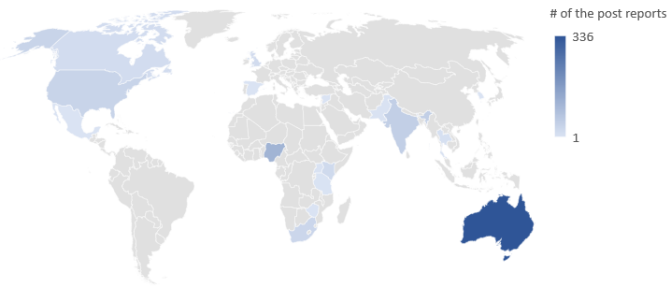

Fig. 3. The distribution of the corruption related Tweets

apply the Data Processing as described in the previous section. For instance, the first tweet below is a user reported tweet, the second one is an unrelated tweet:

- Today I refused to give bribe at Jkia. Official demanded a tip for giving me a standard single visa which is usually 3 months.
- Last night I tried to bribe Emerson with an amoxicillin and when she wouldnt take the bait I threatened to hit her in the head with a can of biscuits. Where do we go from here??

After applying the data processing for both of the tweets the word bag is as below:

- 'today', 'refused', 'give', 'bribe', 'jkia', 'official', 'demanded', 'tip', 'giving', 'standard', 'single', 'visa', 'usually', '3', 'months'
- 'last', 'night', 'tried', 'bribe', 'emerson', 'amoxicillin', 'wouldnt', 'take', 'bait', 'threatened', 'hit', 'head', 'biscuits', 'go'

The former two tweets are all talking about bribery, the only difference is that the target of the "bribery" in the first text is different from the second one, and that makes the result significantly different. Future studies will need to develop better models to further contextualize the difference between these similar conversations which have distinct differences in meaning and intent.

IV. FUTURE WORK

Our study for detecting self-reporting of corruption conversations and activities via Twitter using big data and machine learning is still at an early stage. The preliminary results indicate that certain Twitter users actively self-report forms of corruption on social media, generally with the hopes of sharing their frustration or experience with other users, though the overall volume of this content appears to be low. The most common topics they appear to report are related to police bribery and the pharmaceutical and health system (importantly ensuring healthy lives and promote[ing] well-being for all at all ages is Goal 3 of the SDGs, which includes several population health targets that are directly negatively impacted by the presence of corruption [3]). Since this is an exploration study, we relied on unsupervised machine learning with no sample tweets or a training data set that could tell us what tweets to look for during the first phase of this project. We then utilized a traditional machine learning model(SVM) to nd more signal data, though future study should use other models like Decision Tree [10] or LSTM [11] and compare the performance of detecting corruption-related tweets between these models.

As for the challenge mentioned in the Results section regarding contextual difference between user reports and unrelated tweets, the main problem is the difference in the target, usually these targets are the significant words in a sentence, therefore our future approach will include using Tf-idf [12], to calculate the significant words in each sentence, then use K nearest neighbor(KNN) [13] to categorize them into different clusters, therefore, we can make sure in a different cluster, they are focusing on a similar target (based on keywords like "bribe", "corruption").

Though preliminary, the results from this study provide early evidence of the use of social media by users to report corruption that can be used to help measure the proportion of persons who have paid a bribe, a data point specifically required by SDG16.5s indicator. It is important to note that news reports (such as a high volume of tweets concerning corruption and the U.S. Trump administration), can drown out potential user-generated signal about corrupt acts. Future studies will need to take this into account, while also developing more targeted models to detect and classify corruption user generated messages, and also develop structured approaches and educational outreach to encourage the public to use social media to crowd source corruption reporting.

This study also provides an early framework for a social media-based surveillance and reporting tool to detect corruption, report it to authorities, and can also be used in the future to develop targeted anti-corruption interventions. For example, future uses could include more structured approaches and outreach so users more purposefully report corruption on social media channels that could act as a future deterrent, developing social bots that can provide education to those impacted by corruption on how to report to local authorities and take action, and acting as a data source for evidence for future law enforcement or prosecution activity against those identified and even tagged in social media corruption posts. Technology alone will not end corruption, but transparency, data, and evidence about its harms generated through big data and machine learning has the real potential to curtail it.

[3]https://www.transparency.org/whatwedo/publication/global_corruption_report_2006_corruption_and_health

REFERENCES

[1] T. K. Mackey, T. Vian, and J. Kohler, "The sustainable development goals as a framework to combat health-sector corruption," in *Bulletin of the World Health Organization*, 2018.

[2] T. K. Mackey, J. Kohler, M. Lewis, and T. Vian, "Combating corruption in global health," *Science Translational Medicine*, vol. 9, 2017.

[3] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, "Using data mining to detect health care fraud and abuse: A review of literature," in *Global journal of health science*, 2014.

[4] J. Poushter, C. Bishop, and H. Chwe, "Social media use continues to rise in developing countries but plateaus across developed ones," *Pew Research Center*, vol. 22, 2018.

[5] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: ACM, 2013, pp. 1445–1456. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488514

[6] X. Wang, B. Lafreniere, and T. Grossman, "Leveraging community-generated videos and command logs to classify and recommend software workflows," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018, pp. 285:1–285:13. [Online]. Available: http://doi.acm.org/10.1145/3173574.3173859

[7] Asaki and S. Ono, "Query expansion for microblog retrieval focusing on an ensemble of features," 2018.

[8] V. W. Chu, R. K. Wong, S. J. Fong, and C.-H. Chi, "Emerging service orchestration discovery and monitoring," *IEEE Transactions on Services Computing*, vol. 10, pp. 889–901, 2017.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[10] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *icml*, vol. 99, 1999, pp. 124–133.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[12] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." Carnegie-mellon univ pittsburgh pa dept of computer science, Tech. Rep., 1996.

[13] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.