On Improving Toll Accuracy for COVID-like Epidemics in Underserved Communities Using User-generated Data

Hamada A. Aboubakr[a] Amr Magdy[b,c]
[a] Department of Veterinary Population Medicine, CVM
University of Minnesota - Twin Cities
[b]Department of Computer Science and Engineering [c]Center for Geospatial Sciences
University of California, Riverside
aboub006@umn.edu amr@cs.ucr.edu

ABSTRACT

This paper envisions using user-generated data as a cheap way to improve accuracy of epidemic tolls in underserved communities. The global widespread of COVID-19 pandemic has imposed several unprecedented challenges. One of these challenges is constantly monitoring the unprecedented epidemic widespread at a fine-granular spatial scale, so experts can model, understand, and prevent disease transmission and field personnel can reach and treat infected people. Unfortunately, the limited resources compared to the pandemic widespread has led to a significant number of unreported cases in underserved communities and developing countries, including a large number of severe cases.

We propose in this paper enhancing epidemic case reporting in underserved communities through exploiting the power of data that are posted by people on web. Our vision is building a data analysis pipeline that filters and categories use-generated data objects to provide informal estimates for tolls in unreachable regions and enhance estimates in other regions. The pipeline consist of five stages, that starts with filtering epidemic-specific data to visualize advanced aggregates to end users. We also discuss several technical challenges that face different stages of the pipeline.

CCS CONCEPTS

• **Information systems → Information systems applications**; *Information integration.*

KEYWORDS

COVID, user-generated data, big data, query processing

ACM Reference Format:
Hamada A. Aboubakr[a] Amr Magdy[b,c]. 2020. On Improving Toll Accuracy for COVID-like Epidemics in Underserved Communities Using User-generated Data. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3423459.3430758

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430758

1 INTRODUCTION

The global widespread of COVID-19 pandemic has clearly introduced unprecedented challenges to humanity at different fronts. In the front line of these challenges are the health-related challenges, including reaching out and providing appropriate medical care to infected people. However, this pandemic has a global widespread almost in every country, province, and village worldwide, which makes monitoring it a tremendously difficult task. With the limited resources, the health systems have to prioritize patients for care based on different factors [7, 14–16]. Unfortunately, the underserved communities, e.g., rural areas and slums in developed countries or small cities and villages in developing countries, are highly impacted by the consequences of this pandemic relative to other communities. This is because of their higher exposure to the causes of infection and their limited access to COVID testing and equipped medical care facilities [5, 10–12, 25]. Furthermore, failure to monitor and report cases is a growing concern particularly in developing countries because of the limited public health infrastructure, the weak health systems, insufficient laboratory capacity of diagnostic testing, and the poor surveillance systems for diseases [1, 18]. Therefore, the number of reported infections and deaths in underserved communities does not reflect the actual numbers almost everywhere [3, 23]. This leads to a very high cost in lives. For example, as of September 2020, more than 75% of children who have died of COVID-19 in the U.S. are minorities, though they account for just 41% of the overall youth population [28].

To improve access to underserved communities, we propose to use the power of people to mitigate reporting inaccuracy. The main idea is using user-generated data that flows on web around the clock to extract related information that helps in improving epidemic reporting to health officials. Such mitigation will have a great impact as it will enable reaching currently inaccessible cases. This helps health officials to provide appropriate medical care, surround infection foci, and control the situation faster especially in underserved communities that are highly impacted with limited reporting means and highly infectious environments.

Existing work on coronavirus-related social media data puts a particular focus on controlling spread of misinformation that are related to the pandemic symptoms, transmission modes, and other misleading information that could harm people's health [2, 4, 8, 9, 13, 19, 20, 22, 24]. Although this is a crucially important problem to address, it deals with extracting harmful information from user-generated data to prevent the negative aspects of spreading misinformation. On the contrary, our work deals with user-generated

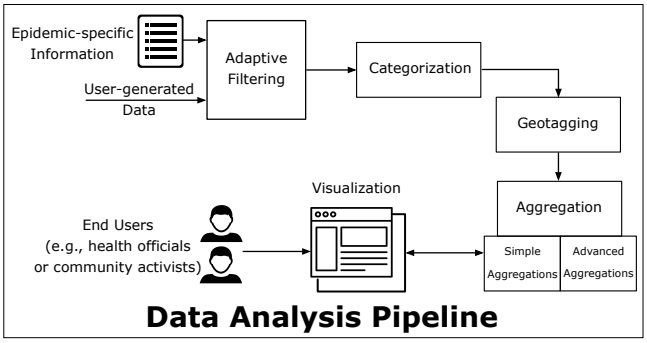

Figure 1: Proposed Pipeline Architecture

data positively as a source of important information that could help health experts. This is also related to orthogonal efforts that deal positively with coronavirus-related user-generated data, including sentiment analysis [27], integrating with IoT data [26], and modelling transmission [21].

The proposed data analysis pipeline consists of five main stages: adaptive filtering, categorization, geotagging, aggregation, and visualization. Each of these stages has different issues related to either lingual dependency, processing streams, or granularity. The rest of this paper outlines each analysis stage, discussing the technical issues and their implications.

2 DATA ANALYSIS PIPELINE

This section outlines the proposed data analysis pipeline. Figure 1 shows the proposed pipeline architecture. The pipeline consists of five ordered stages, namely, *adaptive filtering*, *categorization*, *geotagging*, *aggregation*, and *visualization*. The stages work in a sequential order, where the output of each stage is an input to the following stage. The first stage takes the input data and epidemic-specific information, while the last stage output visualized aggregates for epidemic cases grouped by spatial locations and temporal intervals. The main functionalities and distinguishing characteristics of each stage are briefly outlined below.

(1) **Adaptive filtering**: This stage takes two inputs: (a) A static dataset or a dynamic data stream of user-generated objects, e.g., tweets, posts, comments, or fusion of them. (b) Epidemic-specific characteristics; a set of seed keywords, optional locations of interest, and optional times of interest. Using the two inputs, an adaptive filter is employed to filter out any data object that does not satisfy the epidemic characteristic. Therefore, any data object that does not contain any of the keywords, lies outside the areas of interest, or posted outside the times of interest will not be considered for further processing. When neither locations nor times are provided, all locations and times are considered relevant, e.g., all locations are relevant for the global COVID-19 pandemic. However, this filter should be adaptive in terms of improving the filtering keywords while the filtration process goes on. To this end, when a relevant data is found based on the seed keywords, the adaptive filter should keep all other words of this data except stop words. Over time, the filter will discover more keywords that identify epidemic-related data adaptively, either by using frequent words or other keyword identification methods. This adaptation should also consider the type of input dataset, as static datasets are easier to discover new keywords compared to dynamic data streams.

(2) **Categorization**: This stage takes the set of relevant data objects, that are output of the first stage, to categorize them based on the epidemic case statuses. For example, for COVID-19 pandemic, three potential case categories are: a death case, a mild infection case, and a severe infection case. Such categorization is epidemic-specific in terms of number of categories and how to identify each category. One way is keyword-based categorization, where each category is defined by a set of keywords and the object is assigned based on the corresponding keywords. This way can be performed jointly with the adaptive filtering stage where the list of filtering keywords are categorized into multiple categories, or separately based on different keyword sets. Another way is using machine learning techniques that have shown effectiveness in document classification. Regardless the categorization method, data in each category will be used in the aggregation stage for improving miscounting accuracy.

(3) **Geotagging**: Another piece of information that is needed in data aggregation is the geographical location of each data object to map the epidemic case to a corresponding city, district, or village. Despite the widespread of mobile devices and mobile users of online platforms, automatic geotagging is still a limitation where majority of data comes either with very coarse spatial granularity or without any spatial information. A main reason is legal privacy concerns, where user-generated data platforms disable automatic geotagging by default to protect personal privacy and avoid legal problems. To overcome this limitation, this stage analyzes the data object's content and metadata to assign a primary relevant location. Geotagging has been studied in the literature for different settings and performance trade-off, including for short posts, long posts, etc. Among the recent work is [17] that uses deep learning to geotag tweets of any language. This type of work is the most relevant for user-generated data of epidemic analysis due to high percentage of short posts and popularity in different languages. This is also related to the cross-lingual issues that will be discussed in Section 3.

(4) **Aggregation**: After processing over the first three stages, the output data objects are ready to be aggregated into corresponding locations and time intervals. This spatio-temporal aggregation stage represents the main counting and analysis stage. Locations could be attached from the original data source or resulted from the geotagging stage. The object timestamp is attached from the original data source in majority of platforms. The aggregation could be either a simple counting aggregation grouped by location and time for all places and times, or advanced aggregation for a specific place or certain time intervals. We outline our vision for both below.

Simple aggregations. The simple spatio-temporal aggregation stage sums up data objects counts based on user-defined hierarchies for both spatial and temporal dimensions. For the spatial dimension, end users, e.g., health officials or activists, might, for example, define a hierarchy of <*city, county, state*> to count different categories of epidemic cases for each provided city, county, and state within the USA. Users should be able to control defining this hierarchy based on the needs and the different administrative region divisions around the world. Also, the provided regions are not necessarily

to be of predefined borders, but could be arbitrary, e.g., output of a regionalization algorithm, to enable exploring areas based different attributes, e.g., economic level, population density, or environmental factors. In all cases, the attached location information to data objects affects the count accuracy for this user-defined spatial hierarchy. For example, if the attached information provides city-level locations but not district-level, any hierarchy that includes districts will suffer form low counting accuracy. This is discussed among the technical issues in Section 3.

Unlike the spatial dimension, the temporal dimension is more deterministic and has clearer aggregation options, and in turn less issues. Users can still define temporal hierarchy for aggregations. For example, the user can define <*day,week,month*> hierarchy to count cases for each day, week, and month in each city, county or state. Unlike spatial information, attached temporal information are provided in fine granularity, e.g., second-level granularity, in majority of platforms. This makes temporal aggregation easier and of much better accuracy. By default, each level of the temporal hierarchy assumes disjoint time intervals, e.g., disjoint days, for ease of use due to popularity of this temporal aggregation model. However, users should be also able to define temporal hierarchies of overlapping time intervals. For example, if the analysis is performed on a continuous data stream, a sliding window of three days will be of interest for health officials to monitor in different places, which is by definition a set of overlapping time intervals. This could be also applicable to static datasets in certain analysis scenarios. So, allowing overlapping time intervals will be a useful aggregation feature to support.

Advanced aggregations. Beyond the simple count aggregations for all levels of spatial and temporal hierarchies, end users will be interested in more advanced aggregations that better show the situation in specific places and at certain times. For example, when Southern California appears as a region with high number of cases on the epidemic map, health officials will be interested in producing advanced aggregates for Southern California counties that show the absolute and relative increase in number of cases over the past seven days. Another example is finding areas that have the highest rate of increase over the past three days to mitigate most vulnerable regions. We can discuss endless examples that combine spatial, temporal, and counts in an advanced way to show a different information or insight. It is important to identify the most important blocks that are used in such advanced analysis based on the need of domain experts, making use of existing analysis frameworks as data analysis infrastructures.

(5) **Visualization**: The last stage is visualizing both simple and advanced spatio-temporal aggregates to end users, e.g., health officials or leading community activists, to enable them making use of these counts effectively. This stage should make use of the existing rich literature of visualization frameworks, such as UCI Cloudberry [6], to provide low-effort and effective visualization. Obviously, a geographical map will be an essential element in such visualization. Domain experts should be involved in collecting requirements for all needed visualization features, so they are effective for them as end users. For this context, fundamental visualization elements that should be supported are heatmaps that are either based on administrative borders or cross-borders, hover display boxes that shows cases counts in each spatial entity, filters that allow fragmenting the data based on location and time, and filters that allow fragmenting based on other important attributes such as case category, e.g., either death, mild infection, or severe infection of COVID-19. In addition, the traditional pan, zoom, linking, and brushing features of interactive geovisualization should be supported to enable effective display and exploration for both simple and advanced aggregates.

3 DISCUSSIONS

This section discusses some technical issues that should be addressed while developing the proposed data analysis pipeline. We discuss issues of *language dependency*, *real-time streams*, *granularity*, and *multi-locatable objects*.

Language dependency. One of the main challenges in supporting underserved communities for epidemic data applications is the language issue. Obviously, the language and its usage is highly variant from one underserved community to another, depending on the country and even the locality within that country. Orthogonal from differences in languages among countries, it is known that dialects could be very different within different parts of the same country. This issue affects the first three stages of our pipeline, adaptive filtering, categorization, and geotagging. Addressing this issue could take one of two forms. The first way is tailoring the developed pipeline for a certain underserved community, and hence use its specific language and dialects as input to process. This means tailoring the filtering and categorization keywords and using language-specific geotagging tool, e.g., place ontology. The alternative way is training machine learning models that uses blended datasets of several languages to adapt for a multi-lingual setting. This approach is used in the literature for different tasks. For example, the work in [17] uses this approach for cross-lingual geotagging.

Real-time streams. When data analysis is performed on a dynamic data stream that continuously receive data objects around the clock, different aspects of data analysis change including data storage, processing schemes, and query models. This has triggered the whole literature of streaming data management that is active for a couple of decades. For our proposed data analysis pipeline, analyzing streaming data will have impact on all stages. The least affected stage is geotagging as many of existing geotagging methods depend solely on the data object's content and metadata, without much dependency on previous or upcoming objects. A main problem for this stage will be geotagging efficiency in real time, however, using fast machine learning models could solve this problem [17]. For other stages, the impact is clearer. The filtering phase will be a driver stage as it will help to significantly reduce the streaming data size and output only relevant data objects, so the number of objects to be processed by the following stages are much smaller in size. This will eliminate one of the major overhead in stream processing, which is excessive data size. The other major overhead, which is incremental data processing, will clearly affect the other three stages, categorization, aggregation, and visualization. In categorization, incremental document classification has to be incorporated. If keyword-based categorization is employed, incremental processing will be straightforward to incorporate. Machine learning based categorization will be more challenging to handle.

Although several existing techniques handle this setting, the result accuracy is expected to be lower compared to static datasets. The incremental aggregation will be easier and less impacted by the streaming nature. The reason is that all our aggregations depend on counting, which is easy to maintain incrementally. The visualization will consume the aggregation results as is. However, incremental results updates will need to be visualized incrementally to end users. However, in case of epidemics, even hourly updates are considered fast enough for most of the cases, and this can be adequately served by existing visualization platforms.

Granularity. At different stage of the proposed pipeline, granularity plays a role in trading off usability and processing overhead of analyzing user-generated data. For example, adaptive filtering could classify objects as *relevant* or *irrelevant* and output one type of relevant objects. It could also filter at a finer granularity and further classifies relevant objects into further types to distinguish epidemic-specific cases. This is clearer in the categorization stage that can provide coarse-granular or fine-granular categories with a wide variety of options. Finer granularity levels will provide better accuracy and more information, but it will come with further processing requirements. Granularity is also a trade off for geotagging, where accurate point geotagging consumes much larger processing overhead, while city-level or province-level geotagging is much faster. The granularity of aggregation over space and time will also introduce the same trade off, but it will add a storage trade off as well to decide how much data to store. In general, granularity is a cross-stage issue to consider while designing and developing the proposed pipeline, and it should consider the trade off between available computing resources and required functionality.

Multi-locatable objects. Some data objects might be attachable to multiple locations. Examples for sources of such phenomenon are location ambiguity, e.g., Alexandria is a city name in different countries, mentioning multiple locations either within the content or in both content and metadata, e.g., the user profile shows a city in USA and the post is about a city in India. Regardless the source of multiple locations, this represents a challenge as we cannot assume the case is replicated in multiple physical places unless the locations are nested, e.g., California and USA. However, for the general case where the attachable locations are different, it is essential to promote one of them as the primary location to be used in further analysis. Location selection could be rule-based or certainty-based. Rule-based location selection will apply some heuristic rules to promote the most probable location, e.g., favoring the content words over the user profile location or favoring the earliest mentioned location. Certainty-based location selection will depend on assigning a probability to each potential location, either based on a probabilistic model or a multi-class classifier. Then, locations can be considered or neglected based on these probabilistic values. This certainty-based model opens the door to consider more than one location in the aggregation by introducing uncertain query processing. However, we believe that might be confusing for non-expert end users. Another option is to consider all uncertain locations to contribute partially while distinguishing them from certain locations in both aggregation and visualization stages.

REFERENCES

[1] H. Aboubakr and S. Goyal. Involvement of Egyptian Foods in Foodborne Viral Illnesses: The Burden on Public Health and Related Environmental Risk Factors: An Overview. *Food and Environmental Virology*, 11(4):315–339, Sept. 2019.

[2] Social Media Firms Fail to Act on Covid-19 Fake News. https://www.bbc.com/news/technology-52903680, 2020.

[3] Two Charts Estimate the True Scope of the US's Coronavirus Infections and Deaths. https://www.businessinsider.com/us-coronavirus-cases-deaths-real-scale-estimates-charts-2020-7, 2020.

[4] J. S. Brennen, F. Simon, P. N. Howard, and R. K. Nielsen. Types, Sources, and Claims of COVID-19 Misinformation. *Reuters Institute*, 7:1–13, 2020.

[5] Health Equity Considerations and Racial and Ethnic Minority Groups. https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html, 2020.

[6] Cloudberry Big Data Visualization. http://cloudberry.ics.uci.edu/, 2020.

[7] P. D. N. et. al. Multi-Criteria Decision Analysis to Prioritize Hospital Admission of Patients Affected by COVID-19 in Low-resource Settings with Hospital-bed Shortage. *International Journal of Infectious Diseases*, 98:494–500, Sept. 2020.

[8] SOCIAL MEDIA STRUGGLES WITH CORONAVIRUS MISINFORMATION. https://www.rpc.senate.gov/policy-papers/social-media-struggles-with-coronavirus-misinformation-, 2020.

[9] Battling the pandemic of misinformation. https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods/, 2020.

[10] Coronavirus infection by race: What's behind the health disparities? https://www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-infection-by-race/faq-20488802, 2020.

[11] HHS Initiatives to Address the Disparate Impact of COVID-19 on Ethnic Minorities. https://www.hhs.gov/sites/default/files/hhs-fact-sheet-addressing-disparities-in-covid-19-impact-on-minorities.pdf, 2020.

[12] The Impact on Underserved Communities in Times of Crisis. https://www.himss.org/resources/impact-underserved-communities-times-crisis, 2020.

[13] A. N. Islam, S. Laato, S. Talukder, and E. Sutinen. Misinformation Sharing and Social Media Fatigue During COVID-19: An Affordance and Cognitive Load Perspective. *Technological Forecasting and Social Change*, 159:120201, Oct. 2020.

[14] California Sets Guidelines on Which Patients Are Prioritized if Hospitals Overwhelmed by Coronavirus. https://www.latimes.com/california/story/2020-04-21/california-healthcare-guidelines-shortages-coronavirus-treatment, 2020.

[15] Coronavirus: Italy Doctors Forced to Prioritise ICU Care for Patients with Best Chance of Survival. https://www.euronews.com/2020/03/12/coronavirus-italy-doctors-forced-to-prioritise-icu-care-for-patients-with-best-chance-of-s, 2020.

[16] The Heart-wrenching Choice of Who Lives and Dies. https://www.bbc.com/future/article/20200428-coronavirus-how-doctors-choose-who-lives-and-dies, 2020.

[17] M. Izbicki, V. Papalexakis, and V. J. Tsotras. Geolocating Tweets in any Language at any Location. In *CIKM*, pages 89–98, 2019.

[18] T. P. Mashamba-Thompson and E. D. Crayton. Blockchain and Artificial Intelligence Technology for Novel Coronavirus Disease-19 Self-Testing. *Diagnostics*, 10(4):198, Apr. 2020.

[19] Quantifying the COVID-19 Misinformation Epidemic on Twitter. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7152572/, 2020.

[20] Hoaxes Are Making Doctors' Jobs Harder. https://www.nytimes.com/2020/08/28/opinion/sunday/coronavirus-misinformation-faceboook.html, 2020.

[21] Z. Peng, R. Wang, L. Liu, and H. Wu. Exploring Urban Spatial Features of COVID-19 Transmission in Wuhan Based on Social Media Data. *ISPRS International Journal of Geo Information*, 9(6):402, 2020.

[22] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, June 2020.

[23] Actual Covid-19 Case Count Could be 6 to 24 Times Higher than Official Estimates. https://www.statnews.com/2020/07/21/cdc-study-actual-covid-19-cases/, 2020.

[24] S. Tasnim, M. M. Hossain, and H. Mazumder. Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, 53(3):171–174, May 2020.

[25] COVID-19 Risks Amplified for underserved Communities. https://www.mobihealthnews.com/news/covid-19-risks-amplified-underserved-communities, 2020.

[26] B. Wang, Y. Sun, T. Q. Duong, L. D. Nguyen, and L. Hanzo. Risk-Aware Identification of Highly Suspected COVID-19 Cases in Social IoT: A Joint Graph Theory and Reinforcement Learning Approach. *IEEE Access*, 8:115655–115661, 2020.

[27] T. Wang, K. Lu, K. Chow, and Q. Zhu. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access*, 8, 2020.

[28] Coronavirus Kills Far More Hispanic and Black Children Than White Youths. https://www.washingtonpost.com/health/2020/09/15/covid-deaths-hispanic-black-children/, 2020.