EPIC30M: An Epidemics Corpus of Over 30 Million Relevant Tweets

Junhua Liu*, Trisha Singhal†, Lucienne T.M. Blessing†, Kristin L. Wood§ and Kwan Hui Lim*
*Information Systems Technology and Design Pillar, Singapore University of Technology and Design
†SUTD-MIT International Design Centre, Singapore University of Technology and Design
‡Engineering Product Development Pillar, Singapore University of Technology and Design
§College of Engineering, Design and Computing, University of Colorado Denver
Email: junhua_liu@mymail.sutd.edu.sg, {trisha_singhal, lucienne_blessing, kristinwood, kwanhui_lim}@sutd.edu.sg

***Abstract*—Since the start of COVID-19, there has been several relevant corpora from various sources that were released to support research in this area. While these corpora are valuable in supporting analysis for this specific pandemic, researchers will benefit from additional benchmark corpora that contain other epidemics for better generalizability and to facilitate cross-epidemic pattern recognition and trend analysis tasks. During our research, we discover little disease related corpora in the literature that are sizable and rich enough to support such cross-epidemic analysis tasks. To address this issue, we present EPIC30M, a large-scale epidemic corpus that contains more than 30 million micro-blog posts, i.e., tweets crawled from Twitter, from year 2006 to 2020. EPIC30M contains a subset of 26.2 million tweets related to three general diseases, namely Ebola, Cholera and Swine Flu, and another subset of 4.7 million tweets of six global epidemic outbreaks, including the 2009 H1N1 Swine Flu, 2010 Haiti Cholera, 2012 Middle-East Respiratory Syndrome (MERS), 2013 West African Ebola, 2016 Yemen Cholera and 2018 Kivu Ebola. Furthermore, we explore and discuss the properties of this corpus with statistics of key terms and hashtags and trends analysis for each subset. Finally, we discuss the potential value and impact that EPIC30M could generate through a discussion of multiple use cases of cross-epidemic research topics that attract growing interest in recent years. These use cases span multiple research areas, such as epidemiological modeling, pattern recognition, natural language understanding and economical modeling. The corpus is publicly available at https://www.github.com/junhua/epic.**

I. INTRODUCTION

The Coronavirus disease (COVID-19) has spread around the world since the beginning of the year 2020, affecting around 200 countries and everyone's life. As of 9th November 2020, this highly contagious disease has resulted in over 50 million confirmed cases and cause more than 1.2 million deaths [1]. In times of crisis caused by epidemics, the necessity of rigorous arrangements, quick responses, credible and updated information during the premature phases of such epidemics is especially essential [2].

Social media platforms, such as Twitter, play an important role in disseminating information about the latest epidemic status, via the announcements of public policies in a timely manner. Facilitating the posting of over half a billion tweets daily [3], Twitter emerges as a hub for information exchange among individuals, companies, and governments, especially during times of epidemics where economies are placed in a hibernation mode, and citizens are kept isolated at home. Such platforms help tremendously to raise situational awareness and provide actionable information [4].

Recently, numerous COVID-19 related corpora from various sources are presented that contain millions of data points [5], [6]. While these corpora are valuable in supporting many analyses on this specific pandemic, researchers will benefit from additional benchmark corpora that contain other epidemics to facilitate cross-epidemic pattern recognition and trend analysis tasks. During our other efforts on COVID-19 related work, we discovered little disease related corpora in the literature that are sizable and rich enough to support such cross-epidemic analysis tasks. To address this issue, we curate and make available a dataset spanning numerous epidemics and diseases.

A. Main Contributions

In this paper, we present EPIC30M, a large-scale epidemic corpus that contains more than 30 million micro-blog posts, i.e., tweets crawled from Twitter, from year 2006 to 2020. EPIC30M contains a subset of 26.2 million tweets related to three general diseases, namely Ebola, Cholera and Swine Flu, and another subset of 4.7 million tweets of six global epidemic outbreaks, including the 2009 H1N1 Swine Flu, 2010 Haiti Cholera, 2012 Middle-East Respiratory Syndrome (MERS), 2013 West African Ebola, 2016 Yemen Cholera, and 2018 Kivu Ebola. These epidemics and diseases represent the key outbreaks that have happened in the past 15 years, apart from the recent COVID-19 pandemic that is well covered by various works [5], [6].

In addition to collecting and making available the EPIC30M dataset, we also conduct several exploratory analyses to study the properties of this corpus, such as word cloud visualization and time series trend analysis. Several interesting findings are discovered through these analyses. For instance, we find that a large quantity of topics are related to specific locations; cross-epidemic topics, i.e. one that involves more than one epidemic-related hashtag, appear frequently in several classes; and several hashtags related to non-epidemic events, such as warfare, have relatively high ranks in the list. Furthermore, a time-series analysis also suggests that some of the epidemics,

978-1-7281-6251-5/20/$31.00 © 2020 IEEE

Epidemics	Example Queries	Earliest tweet	Latest tweet	No. of tweets
Outbreak				
2009 H1N1 Swine Flu	h1n1 swine flu panflu	05-01-2009	19-06-2020	2,803,941
2010 Haiti Cholera	haiti cholera	13-01-2010	31-05-2020	359,122
2012 MERS	MERS-CoV #mers	01-09-2012	31-05-2020	265,119
2014 West Africa Ebola	africa ebola	01-12-2013	19-06-2020	1,191,516
2016 Yemen Cholera	yemen cholera	01-01-2017	15-05-2020	102,900
2018 Kivu Ebola	kivu ebola	03-03-2018	31-05-2020	12,063
General				
Cholera	cholera	19-01-2007	19-06-2020	2,321,903
Ebola	ebola	25-12-2006	19-06-2020	20,178,969
Swine Flu	swine flu	31-10-2007	19-06-2020	3,775,217

TABLE I: Statistics of EPIC30M with the *Outbreak* and *General* subsets[1]

i.e., the *2010 Haiti Cholera* and *2018 Kivu Ebola*, show a surge in tweets before the respective start dates of the outbreaks, which signifies the importance of leveraging social media to conduct early signal detection. We also observe that an epidemic outbreak not only leads to rapid discussion of this epidemic, but also triggers exchanges and discussions about other diseases.

EPIC30M fills a gap in the literature where little epidemic-related corpora are either unavailable or not sizable enough to support cross-epidemic analysis tasks. Through discussing various potential use cases, we anticipate that EPIC30M will bring value and impact to various fast growing computer science communities, especially in natural language processing, data science and computation social science. We also foresee that EPIC30M is able to contribute partially to cross-disciplinary research topics, such as economic modeling and humanity studies. By including numerous tweets posted throughout the cause of each outbreak in the EPIC30M corpus, we hope that EPIC30M will serve as a suitable form of cross-epidemic benchmark and generate various interesting lines of research.

The rest of this paper is organized and structured as follows. Section II discusses related work in this area, with a focus on datasets based on Twitter and related pandemics and diseases. Next, Section III describes our data collection process. Following which, Section IV discusses the results from our exploratory data analysis based on hashtags visualization, time series analysis and topic modelling. Section V then describes the potential use cases of EPIC30M in different domains. Finally, Section VI summarizes this paper and highlights some future research directions.

II. Related Work

In this section, we discuss the existing Twitter corpora for several domains, such as COVID-19, disasters, and others. These corpora attract a large quantity of interests and enable a large amount of research works in their respective domains, which we believe EPIC30M generalizes to a similar level of impact in the epidemic domain.

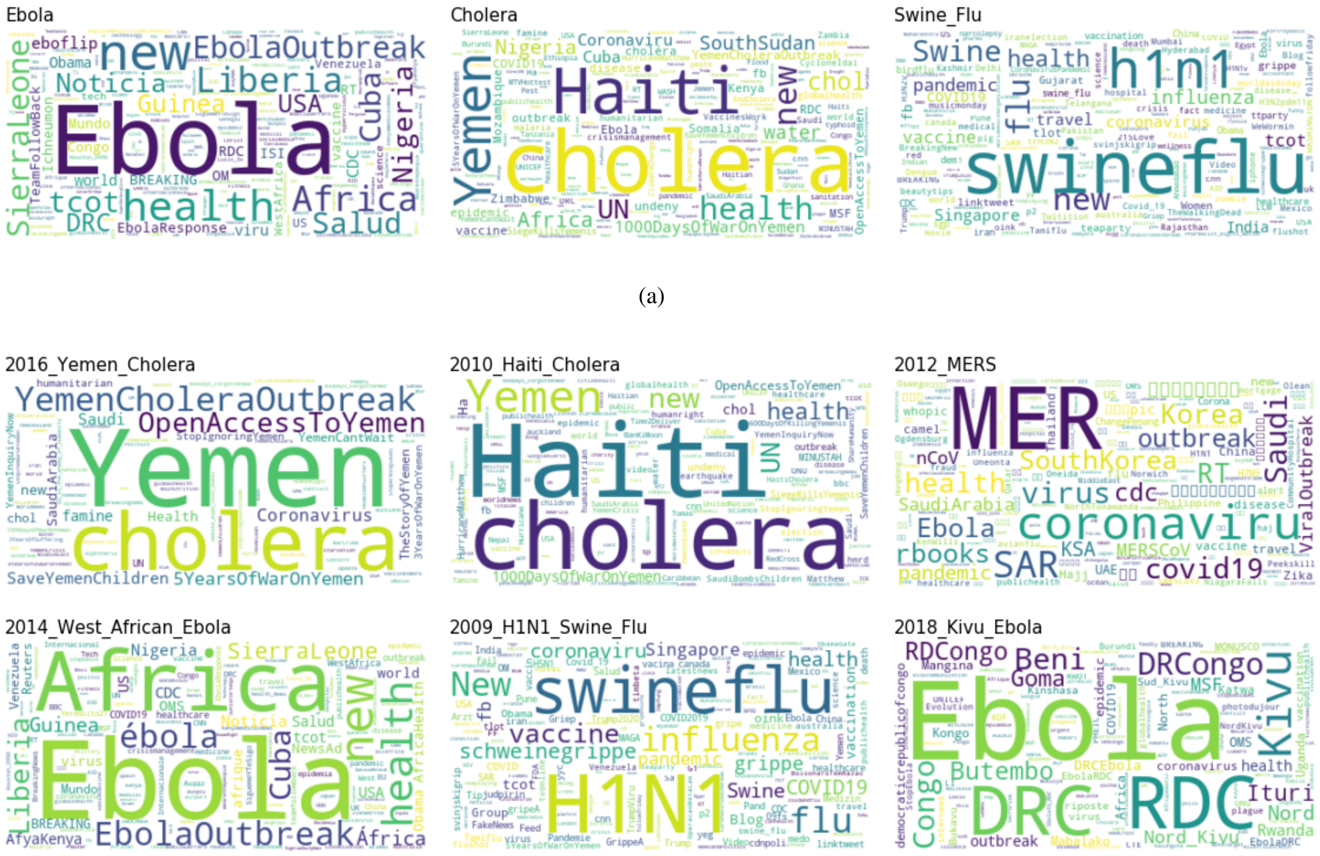
A. Corpora of COVID-19

Recently, the COVID-19 pandemic spread across the world and generated enormous economical and social impact. Throughout the pandemic, numerous related corpora have been released. For instance, [5] released a multi-lingual corpus that consists of 50 million tweets that include tweet IDs and their timestamps, across over 10 languages. Similarly, [8] presented a large-scale COVID-19 chatter corpus that consists of over 152M tweets with retweets and another version of 30 million tweets without retweets. In addition to social media data, others have released COVID-19 dataset relating to mobility patterns [9], [10], [11], government response [12], [13], [14] and medical data [15], [16], among others.

B. English Corpora of Disasters

There are several disaster-related corpora presented in the literature that are utilized for multiple works. CrisisLex [17] consists of 60 thousand tweets that are related to six natural disaster events, queried based on relevant keywords and locations during the crisis periods. The tweets are labelled as *relavant* or *not-relevant* through crowdsourcing. Olteanu et al. [18] conducted a comprehensive study of tweets to analyze 26 crisis events from 2012 to 2013. The paper analyzed about 25k tweets based on crisis and content dimensions, which include hazard type (*natural* or *human-induced*), temporal development (*instantaneous* or *progressive*), and geographic speed (*focalized* or *diffused*). The content dimensions are represented by several features such as informativeness, types and sources. Imran et al. [19] released a collection of over 52 million tweets, out of which 50 thousand come with human-annotated tweets that are related to 19 natural crisis events. The work also presented pre-trained *Word2Vec* embeddings with a set of Out-Of-Vocabulary (OOV) words and their normalizations, contributing in spreading situational awareness and increasing response time for humanitarian efforts during crisis. Others like Phillips [20] released a set of 7 million tweets related to Hurricane Harvey, while Littman [21] published a corpus containing tweet IDs of over 35 million tweets related to Hurricane Irma and Harvey.

[1] As of 8 Jun 2020



(a)

(b)

Fig. 1: Hashtags analysis with word clouds [7]. Each word cloud contains the top 100 hashtags in their respective class where the sizes represent the frequency of the hash tags. (a) Three general epidemic classes. (b) Six epidemic outbreak classes.

C. Non-English Corpora of Disasters

Numerous non-English crisis corpora are also found in the literature. For instance, Cresci et al. [22] released a corpus of 5.6 thousand Italian tweets from 2009 to 2014 during four different disasters. The features include informativeness (*damage* or, *no damage*) and relevance (*relevant* or *not relevant*). Similarly, Alharbi and Lee [23] compiled a set of 4 thousand Arabic tweets, manually labelled on the relatedness and information-type for four high risk flood events in 2018. Alam et al. [24] released a Twitter corpora composed of manually-annotated 16 thousand tweets and 18 thousand images collected during seven natural disasters (earthquakes, hurricanes, wildfires, and floods) that occurred in 2017. The features of the datasets include Informativeness, Humanitarian categories, and Damage severity categories.

D. Other Twitter Corpora

Apart from epidemic and crisis-related corpora, several Twitter datasets are used for analysis in areas related to politics, news, abusive behaviour and misinformation, Trolls, movie ratings, weather forecasting, etc. For instance, Fraisier et al. [25] proposes a large and complex dataset with over 22 thousand operative Twitter profiles during the 2017 French presidential campaign with their corresponding tweets, tweet IDs, retweets, and mentions. The data was annotated manually based on their political party affiliation, their nature, and gender. There are also numerous Twitter Corpora that are related to other domains, such as politics [26], [27], cyberbullying [28], and misinformation [29], [30], [31].

III. Data Collection

This section describes the data collection process for crawling EPIC30M, which includes the collection of six epidemic outbreaks and three general epidemics, as well as the search queries used.

A. Collection of Epidemic Outbreaks

EPIC30M includes six epidemic outbreaks in the 21st century, recorded by the World Health Organization[2]. These epidemic outbreaks happened after the founding of Twitter in 2006, and represent the main epidemic outbreaks in the past 15 years. These outbreaks include the 2009 H1N1 Swine Flu, the 2010 Haiti Cholera, the 2012 Middle East Respiratory

[2] https://www.who.int/emergencies/diseases/en/

Field	Type	Description
date	datetime	The date and time (in UTC) that the tweet was posted. Example: *4/2/09 17:06*
username	string	Unique username of the user account that posted the tweet. Example: *douance_quebec*
to	string	The twitter account's username that the tweet that was posted to. Example: *CedricFontaine*
replies	integer	The number of replies that the tweet has. A reply is a response to another person's Tweet. You can reply by clicking or tapping the reply icon from a Tweet. Example: *3*
retweets	integer	The number of retweets that the tweet has. A tweet that a user shares publicly with his/her followers is known as a Retweet, which is a conventional way to pass along news and interesting discoveries on Twitter. Example: *3*
favorites	integer	The number of favorites the tweet receives. Favourites are represented by a small heart and are used to show appreciation for a Tweet. Example: *3*
text	string	The content of a tweet that contains up to 280 characters. Example: *H1N1 + H1N5 = Trouble...*
mentions	string	Another account's Twitter username preceded by the "@" symbol. A mention is a Tweet that contains another person's username anywhere in the body of the Tweet. Example: *@cyberlou33*
hashtags	string	The hashtags that the tweet includes. A hashtag is formed by a symbol (#) followed by a relevant keyword. Hashtags are commonly used or phrase in their Tweet to categorize those Tweets and help them show more easily in Twitter search. Example: *#Anger #Ebola #Liberia*
id	string	A unique identifier of a tweet. Example: *1136281607*
permalink	string	The unique URL of a tweet. Whenever you view a Tweet's permanent link, you can see The exact time and date the Tweet was posted and the number of likes and Retweets the Tweet received. Example: *https://twitter.com/douance_quebec/status/1096080744*

TABLE II: Metadata of EPIC30M.

Syndrome (MERS), the 2014 West Africa Ebola, the 2016 Yemen Cholera, and the 2018 Kivu Ebola, as listed in Table I. We intentionally exclude the recent COVID-19 pandemic outbreak to avoid producing redundant work, as there are already numerous COVID-19 datasets released by different parties with millions of data points.

B. Collection of General Epidemics

Beside the set of six specific epidemic outbreaks, we extend EPIC30M by including a subset of three general diseases, namely *Cholera*, *Ebola* and *Swine Flu*. Including these three subsets allow us to identify the main discussions and topics around these three general diseases, in contrast to the earlier six epidemic outbreaks that occur in specific regions. The tweets related to these three general diseases were crawled since their respective first occurrence until 15th May 2020. We expect that the *general epidemic* subset is able to act as an additional benchmarks and contribute substantially to various research topics, such as pattern recognition and trend analysis.

C. Search Queries

For each outbreak, we initialize with a large collection of keywords used as the search queries, with the hypothesis to retrieve most relevant tweets from Twitter. We use a combination of keywords for each outbreak, as listed in Table I, to fetch the related tweets. Two types of keywords were used, namely: (a) general disease-related terms, such as *ebola*, *cholera* and *swine flu*; and (b) specific outbreak-related terms with a combination of location and disease, such as *africa ebola* and *yemen cholera*.

IV. Exploratory Data Analysis

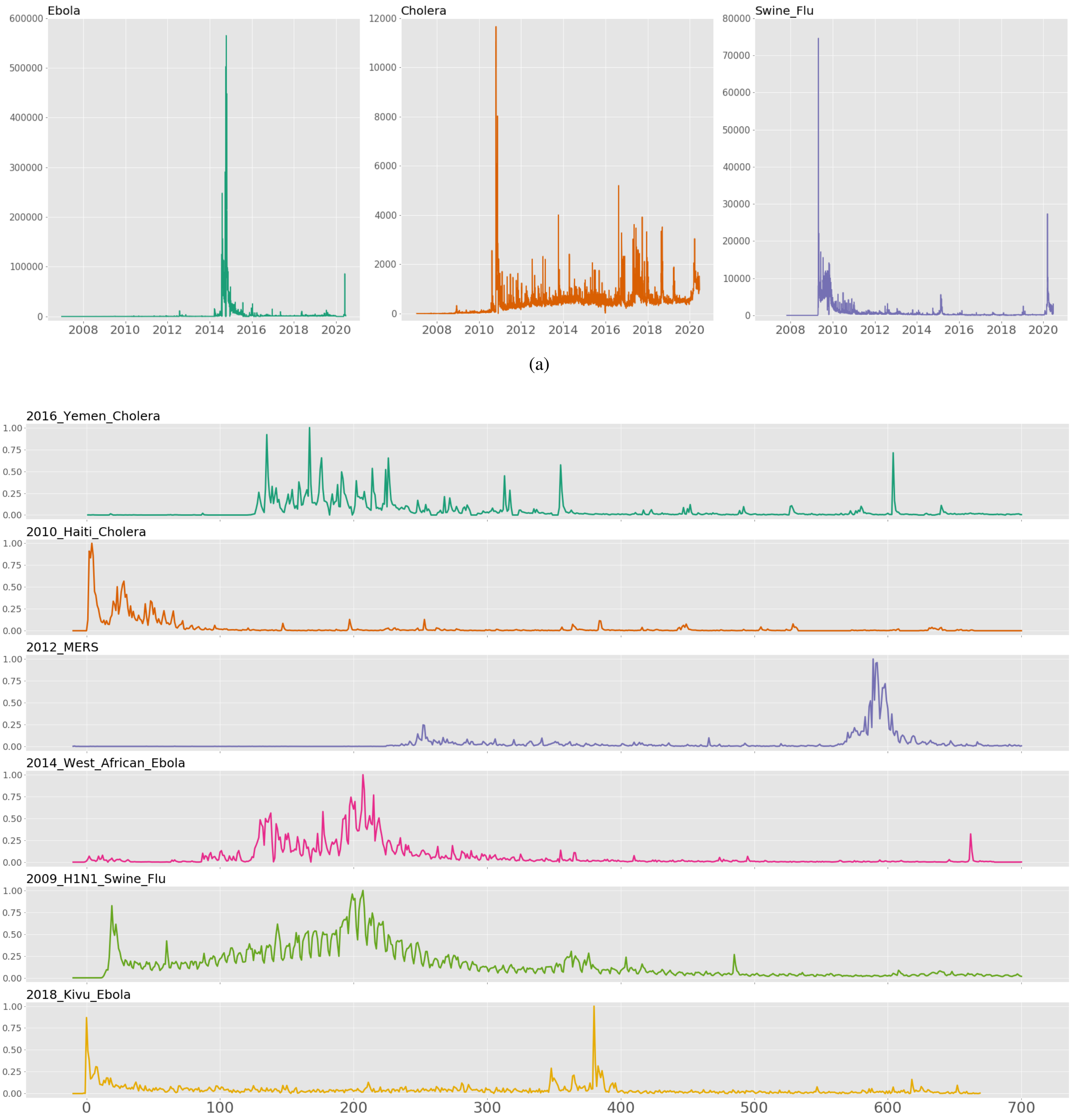
In this section, we describe the exploratory data analysis that we performed on the EPIC30M dataset and discuss some of our main findings from this analysis.

A. Hashtags Visualization Analysis

To provide a general overview of EPIC30M, we first conduct hashtags visualization for each epidemic and plot them on a 3 by 3 grid, as shown in Figure 1. The first row (Figure 1a) represents the three general diseases whereas the second and third rows (Figure 1b) represent the six outbreak classes in chronological order. Each word cloud contains the top 100 hashtags in their respective class, where the font sizes represent their frequencies with more frequent hashtags having larger font sizes.

Through our observation of this visualization, we identify several interesting phenomena, such as:

[3] According to the World Health Organization, https://www.who.int



(a)

(b)

Fig. 2: Time-series trend analysis of the EPIC30M dataset. (a) Each chart represents a class of a general disease, where the x-axis represents the time as yearly dates, and the y-axis represents the corresponding number of tweets. (b) Each chart represents an outbreak class in the corpus. The x-axis represents a range of dates from day -10 to 700, where day 0 represents the respective start date of each outbreak, such as: 2016-09-28 (Yemen Cholera), 2010-10-20 (Haiti Cholera), 2012-09-23 (MERS), 2014-03-23 (West African Ebola), 2009-04-12 (Swine Flu), and 2018-08-01 (Kivu Ebola)[3]. The y-axis represents the number of tweets, normalized to between 0 and 1.

1) Key terms provide semantic indication of the crises, in addition to possible cross-epidemic indicators: such as pandemic, epidemic, healthcare, vaccine, disease, sanitation, and others.
2) Location-related hashtags, such as *#Yemen*, *#Haiti* and *#SierraLeone*, appear in all classes and occupy majority of the key words, which we believe to be the highest concerned feature.
3) Several classes include hashtags of other diseases, i.e., *#COVID19* in the *2016_Yemen_Cholera* class and *#Malaria* in the *Cholera* class, which implies that discussions on cross-epidemic matters are popular
4) Some hashtags refer to non-epidemic related events, such as *#5YearsOfWarOnYemen* and *#earthquake* appearing in the *2016_Yemen_Cholera* and *2010_Haiti_Cholera* sets respectively

B. Time-series Trend Analysis

Subsequently, we conduct a trend analysis with an attempt to identify time-variant patterns from the corpus. For the three general disease classes (Figure 2a), we plot each class into a line chart, where the x-axis represents the time in terms of calendar year and the y-axis represents the corresponding number of tweets. For the six outbreak classes (Figure 2b), the x-axis of each line chart uses the number of days offset from the start date of the outbreak, whereas the y-axis represents the number of tweets normalized to between 0 and 1.

Through the time-series line plots, we observe that some of the epidemics, i.e. *2010 Haiti Cholera* and *2018 Kivu Ebola*, show a surge in tweets before the respective official start dates of the outbreaks, which signifies the importance of leveraging social media to conduct early signal detection. We also observe that an epidemic outbreak not only leads to rapid discussion of its own, but also trigger exchanges of other diseases. Finally, the time-series analyses also show clear dynamic properties or trends with exponential increases (shocks or spikes) in tweet type and a temporal persistence after an initial shock [32]. Other dynamic properties that may be of interest include local cycles and trends. Such dynamic effects, when paired with semantic content (such as healthcare related terms), may provide potential indicators of an onset of a crisis.

C. Topic Modelling

In natural language processing, topic modelling is an unsupervised statistical machine learning model used to identify abstract *topics* that occur in a set of documents. Latent Dirichlet Allocation (LDA) [33] is one of the foundational topic models that is widely used to generate such topics. LDA assumes that each document is a mixture of K topics and each topic has an inherent word frequency distribution. LDA and its variants have frequently been used in works on similar Twitter and social media analytics [34], [35], [36], [37], and thus is suitable for our subsequent analysis.

We considered each tweet in the dataset as a separate document and applied LDA to obtain: (i) topic distribution for each tweet; and (ii) word distribution for each topic. The optimal number of topics $K = 8$ was derived from hyper-parameter tuning. In addition to that, we used hyper-parameter tuning to identify optimal values for: (i) *alpha* which is the symmetric parameter for dirichlet prior on topic distribution; and (ii) *eta* which is the symmetric parameter for dirichlet prior on word distribution for each topic. We identified the optimal values for *alpha* and *eta* to be $50/8$ and 0.1 respectively. All the hyperparameter values can be found in Table III. The overall coherence score obtained was 0.3839.

Hyperparameters	**Values**
K	8
alpha	50/8
eta	0.1
decay	0.5
passes	10

TABLE III: Topic model hyperparameters

Keeping the above in view, the top 10 words in each detected topic were shown in Figure 3 as wordclouds. Besides that, the topic time-series was also plotted from the start of the epidemic to June 2020 based on the weekly frequency of dominant topics in tweets, as shown in Figure 4. Moreover, to evaluate the model, 5-fold cross-validation was performed on the corpus and results were computed for perplexity metric, coherence metric, convergence metric, and difference metric. The model yielded substantially satisfactory results on all defined metrics.

V. Use Cases of EPIC30M Corpus

Twitter has an enormous volume and frequency of information exchange, i.e. over half a billion tweets posted daily, and such rich data potentially exposes information on epidemic events through substantial analysis. In this section, we demonstrate the value and impact that EPIC30M could create by discussing on multiple use cases of cross-epidemic research topics that attract growing interests in recent years. These use cases span multiple research areas, such as epidemiological modeling, pattern recognition, natural language processing and economical modeling. We claim that EPIC30M fills the gap in the literature where little disease related corpora are sizable and rich enough to support such cross-epidemic analysis tasks. Unlike datasets that focus on a single disease or epidemic, EPIC30M offers benchmarks of multiple epidemics to facilitate a wide range of cross-epidemic research topics.

A. Epidemiological Modeling

Epidemiological modeling provides various potential applications to understand the dynamics of Twitter during and after outbreaks, such as compartmental modeling [38] and misinformation detection [39]. To name a few, Jin et al. [40] used Twitter data to detect false rumors and a susceptible-exposed-infected-skeptic (SEIZ) model to group users in four

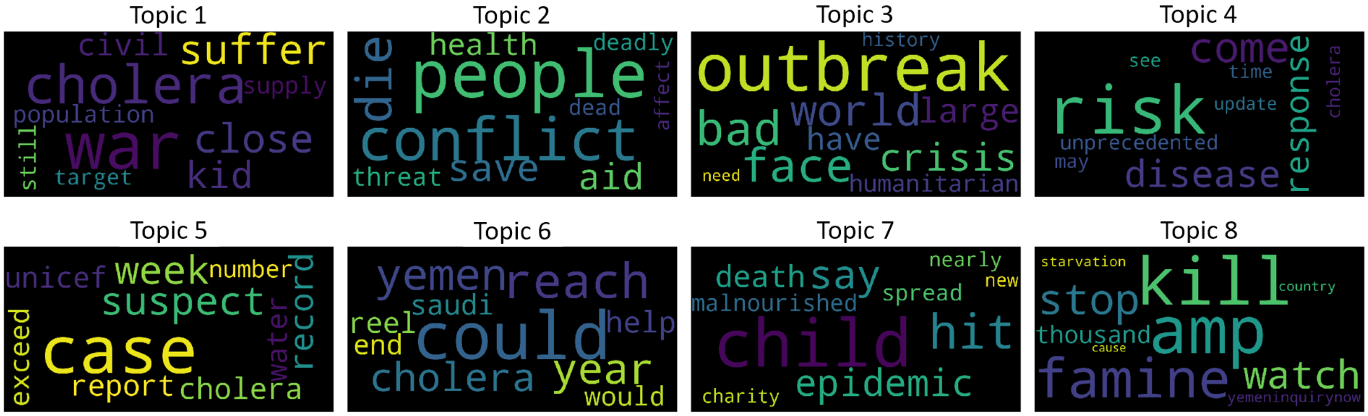


Fig. 3: Top 10 words in detected topics

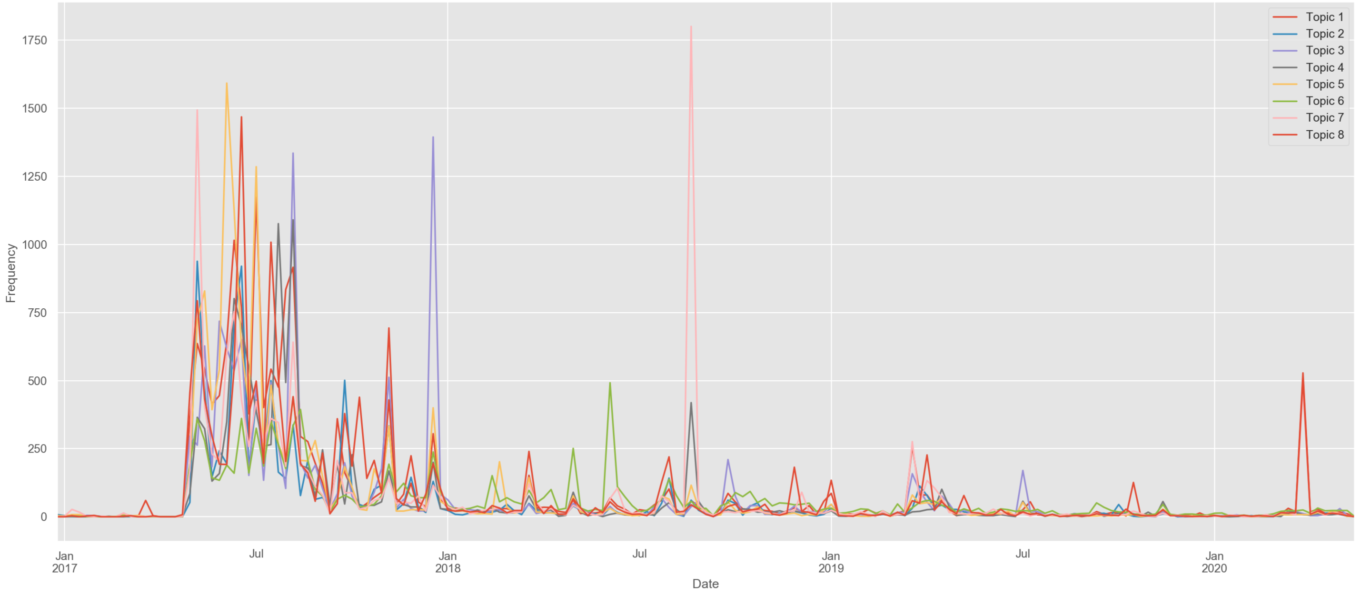


Fig. 4: Topics time-series in terms of the weekly frequency of each topic. Refer to Figure 3 for the set of representative keywords for each topic.

compartments. Skaza and Blais [41] used the susceptible-infectious-recovered (SIR) epidemic models on Twitter hashtags to compute the infectiousness of a trending topic. During the recent event of COVID-19, these models are repeatedly applied to predict discrete questions, such as the proposal by Chen et al. to use a time-dependent SIR model for estimating the total number of infected persons and the eventual outcomes, i.e., recovery or death.

B. Trend Analysis and Pattern Recognition

Extensive prior works leverage social media data to perform trend analysis and pattern recognition tasks [42], [43], [44], [45]. For instance, Kostkova et al. [46] studied the 2009 swine-flu outbreak and demonstrated the potential of Twitter to act as an early warning system up to a period of two or three weeks. Similarly, Joshi et al. [47] predicted early warnings of Western Africa Ebola epidemic, three months earlier than the official announcement. Given how early detection and warning systems for crisis events may reduce overall damage and negative impacts [3], EPIC30M provides a source of high volume and timely information that facilitate trend analysis and pattern recognition tasks for similar epidemic events.

C. Sentiment and Opinion Mining

The observation of social sentiments and public opinions plays an important part in benchmarking the effect of proposing new initiatives or making public policy amendments. Several prior works leverage sentimental analysis and opinion mining to extract the contextual meaning of social media content. For instance, Beigi et al. [48] provided an overview of the relationship among social media, disaster relief and situational awareness in crisis time, and Neppalli et al. [49] performs location-based sentimental analysis on tweets for Hurricane Sandy in 2012. Others like Kwan and Lim [36] used Twitter to understand public sentiments, opinions and controversy surrounding COVID-19 related issues.

D. Topic Detection

Topic detection or modeling may enable authorities in anticipating crises and thus take the appropriate preventive actions to mitigate the effects of such crises. These techniques help in recognizing hidden patterns, understanding semantic and syntactic relations, annotating, analyzing, organizing, and summarizing the huge collections of textual information. Considering the same, several researchers have implemented similar approaches on crises datasets to detect and categorize the potential topics. Chen et al. [50] suggested two topic modeling prototypes to ameliorate trends estimation by seizing the underlying states of a user from a sequence of tweets and aggregating them in a geographical area. In [51], researchers perform optimized topic modeling using community detection methods on three crises datasets [17], [18], [30] to identify the discussion topics.

E. Natural Language Processing

Several works leverage Twitter datasets to conduct Natural Language Processing (NLP) tasks. As a challenging downstream task of NLP, Automatic Text Summarization techniques extract latent information from text documents where the models generate a brief, precise, and coherent summary from lengthy documents. Text summarization is applicable in various real-would activities during crisis, such as generating news headlines, delivering compact instructions for rescue operations and identifying affected locations. Prior works have demonstrated the application of such techniques during crisis time. For instance, [52] and [53] proposed two relevant methods that classify and summarize tweets fragments to derive situational information. More recently, Sharma et al. [54] proposed a system that produces highly accurate summaries from the Twitter content during man-made disasters. Several other works focused on other NLP tasks on social media data, such as information retrieval [55], [56] and text classification [57], [58].

F. Disease Classification

Applications of Machine Learning and Deep Learning in the healthcare sector have gathered growing interests in recent years. For instance, Krieck et al. [59] analyzed the relevance of Twitter content for disease surveillance and activities tracking, which help alert health official regarding public health threats. Lee et al. [60] conducted text mining on Twitter data and deployed a real-time disease tracking system for flu and cancer using spatial, temporal information. Ashok et al. [61] developed a disease surveillance system to cluster and visualise disease-related tweets.

G. Crisis-time Economic Modeling

Estimating the economical impact of crises, such as epidemic outbreaks, is a crucial task for policy makers and business leaders to adjust operational strategies [62] and make the appropriate decisions for their organizations in time of crises. Several papers have conducted surveys and proposed approaches for this application domain. For instance, Okuyama [63] provided an overview and a critical analysis of the methodologies used for estimating the economic impact of disaster; Avelino and Hewings [64] proposed the Generalized Dynamic Input-Output framework (GDIO) to dynamically model higher-order economic impacts of disruptive events. Such studies correlate disaster events and economy impact, which rely on disaster-related data and financial market data, respectively. We believe that EPIC30M is able to contribute to future economic modeling studies for epidemic events.

H. Health Informatics

Compared to the cases above, a more general use case area is healthcare informatics, i.e., "the integration of healthcare sciences, computer science, information science, and cognitive science to assist in the management of healthcare information" [65], [66], [67]. While social media and online sources are used to connect with patients and provide reliable educational content in health informatics, there is growing interest in using Twitter and other real-time feeds to study and understand indicators for health trends or particular behaviors or diseases. For example, Nambisan et al. [68] utilized Twitter content to study the behavior of depression and Kwan and Lim [36] studied how emotions and sentiments change due to the COVID-19 pandemic. EPIC30M contains behavioral information across various diseases via the social media messages that people post and this information serves as an indicator of how the populace behaves with the onset and persistence of the diseases. Multiple disease cases will provide the research opportunity to correlate behavioral information across different instances.

I. News and Fake News

With the proliferation of news content through internet and virtual media, there is a growing interest in developing an understanding of the science of news and fake news [69]. Data mining algorithms are advancing to study news content [70]. EPIC30M contains real news content that grows over time from social lay-person terminology to technical and professionally based information and opinion. Likewise, EPIC30M includes fact-based information as well as distorted or fake content. Through multiple cases over time, the field will have a rich source to study news content, especially when correlating with reliable news sources for particular snapshots of time.

All in all, we believe that EPIC30M provides a set of rich benchmarks and is able to facilitate extensions of the above-mentioned works on a higher order, e.g., in cross-epidemic settings. As a result, the research findings derived from these works will potentially be more robust and closer to real-world scenarios.

VI. Conclusion and Future Work

A. Conclusion

During our other efforts on COVID-19 related work, we discovered little disease related corpora in the literature that are sizable and rich enough to support such cross-epidemic analysis tasks. In this paper, we present EPIC30M, a large-scale epidemic corpus that contains more than 30 million tweets from 2006 to 2020. The corpus includes a subset of tweets related to three general diseases and another subset related to six epidemic outbreaks. We conduct an exploratory analysis to study the properties of the corpus and identify several phenomena, such as a strong correlation between epidemics and locations, frequent cross-epidemic topics, and surge of discussion before the occurrence of these outbreaks. Finally, we discuss a wide range of use cases that EPIC30M can potentially facilitate. We anticipate that EPIC30M will be able to generate substantial value and impact to both fast growing computer science communities, such as natural language processing, data science and computation social science, and multi-disciplinary areas, such as economic modeling, health informatics and the science of news and fake news.

B. Future work

For certain epidemic outbreaks, such as the *2009 H1N1 Swine Flu* and *2014 West Africa Ebola*, EPIC30M includes relevant tweets posted throughout the respective duration of the epidemics. We expect the data of these few classes could serve as a suitable source of cross-epidemic and cross-disease benchmarks. On the other hand, several epidemics, such as the *2018 Kivu Ebola* and *2016 Yemen Cholera*, are still ongoing and EPIC30M will benefit from the inclusion of more recent tweets. We intend to extend the corpus by actively or periodically crawling tweets and update the current version of EPIC30M with these additional tweets. Furthermore, we plan to further develop the corpus by collecting additional epidemic outbreak that occurred more recently, such as the *2019 multi-national Measles outbreaks* in the DR Congo, New Zealand, Philippines and Malaysia, the *2019 Dengue fever epidemic* in Asia-Pacific and Latin America, and the *2018 Kerala Nipah virus outbreak*. Lastly, we also intend to develop an active crawling web service that automatically update EPIC30M, and migrate to cloud-based relational database services to ensure its availability and accessibility.

VII. Acknowledgement

We thank the anonymous reviewers for their useful comments. This research is funded in part by the Singapore University of Technology and Design under grant SRG-ISTD-2018-140.

References

[1] WHO, "Who coronavirus disease (covid-19) dashboard," https://covid19.who.int/, 2020, [Online; accessed 09-Nov-2020].

[2] W. H. Organization, *Managing epidemics: key facts about major deadly diseases*. World Health Organization, 2018.

[3] J. Liu, T. Singhal, L. T. Blessing, K. L. Wood, and K. H. Lim, "Crisisbert: a robust transformer for crisis classification and contextual crisis embedding," *arXiv preprint arXiv:2005.06627*, 2020.

[4] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster-related messages in social media." in *Iscram*, 2013.

[5] E. Chen, K. Lerman, and E. Ferrara, "Covid-19: The first public coronavirus twitter dataset," *arXiv preprint arXiv:2003.07372*, 2020.

[6] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset," *arXiv preprint arXiv:2003.10359*, 2020.

[7] B. Camburn, R. Arlitt, D. Anderson, R. Sanaei, S. Raviselam, D. Jensen, and K. L. Wood, "Computer-aided mind map generation via crowdsourcing and machine learning," *Research in Engineering Design*, pp. 1–27.

[8] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, and G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration," *arXiv preprint arXiv:2004.03688*, 2020.

[9] E. Pepe, P. Bajardi, L. Gauvin, F. Privitera, B. Lake, C. Cattuto, and M. Tizzoni, "Covid-19 outbreak response, a dataset to assess mobility changes in italy following national lockdown," *Scientific data*, vol. 7, no. 1, pp. 1–7, 2020.

[10] Y. Kang, S. Gao, Y. Liang, M. Li, J. Rao, and J. Kruse, "Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic," *arXiv preprint arXiv:2008.12238*, 2020.

[11] Y. Feng and W. Zhou, "Is working from home the new norm? an observational study based on a large geo-tagged covid-19 twitter dataset," *arXiv preprint arXiv:2006.08581*, 2020.

[12] C. Cheng, J. Barceló, A. S. Hartnett, R. Kubinec, and L. Messerschmidt, "Covid-19 government response event dataset (coronanet v. 1.0)," *Nature human behaviour*, vol. 4, no. 7, pp. 756–768, 2020.

[13] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies *et al.*, "A county-level dataset for informing the united states' response to covid-19," *arXiv preprint arXiv:2004.00756*, 2020.

[14] C. Cheng, J. Barcelo, A. Hartnett, R. Kubinec, and L. Messerschmidt, "Coronanet: A dyadic dataset of government responses to the covid-19 pandemic," 2020.

[15] J. Zhao, Y. Zhang, X. He, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," *arXiv preprint arXiv:2003.13865*, 2020.

[16] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv preprint arXiv:2006.11988*, 2020.

[17] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[18] A. Olteanu, S. Vieweg, and C. Castillo, "What to expect when the unexpected happens: Social media communications across crises," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 994–1009.

[19] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages," *arXiv preprint arXiv:1605.05894*, 2016.

[20] M. E. Phillips, "Hurricane harvey twitter dataset," 2017.

[21] J. Littman, "Hurricanes harvey and irma tweet ids," Published to Harvard Dataverse by GWU Libraries Dataverse, 2017, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QRKIBW.

[22] S. Cresci, M. Tesconi, A. Cimino, and F. Dell'Orletta, "A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1195–1200.

[23] A. Alharbi and M. Lee, "Crisis detection from arabic tweets," in *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 2019, pp. 72–79.

[24] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[25] O. Fraisier, G. Cabanac, Y. Pitarch, R. Besancon, and M. Boughanem, "# élysée2017fr: The 2017 french presidential campaign on twitter," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[26] L. Wrubel, J. Littman, and D. Kerchner, "2018 u.s. congressional election tweet ids," Published to Harvard Dataverse by GWU Libraries Dataverse, 2019, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AEZPLU.

[27] I. Brigadir, D. Greene, and P. Cunningham, "Analyzing discourse communities with distributional semantic models," in *Proceedings of the ACM Web Science Conference*, 2015, pp. 1–10.

[28] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[29] P.-M. Hui, C. Shao, A. Flammini, F. Menczer, and G. L. Ciampaglia, "The hoaxy misinformation and fact-checking diffusion network," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[30] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PloS one*, vol. 11, no. 3, 2016.

[31] O. Roeder, "Why we're sharing 3 million russian troll tweets," *FiveThirtyEight, July*, vol. 31, 2018.

[32] J. D. Hamilton, *Time series analysis*. Princeton New Jersey, 1994, vol. 2.

[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[34] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European conference on information retrieval*. Springer, 2011, pp. 338–349.

[35] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 889–892.

[36] J. S.-L. Kwan and K. H. Lim, "Understanding Public Sentiments, Opinions and Topics about COVID-19 using Twitter," in *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'20)*, 2020.

[37] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the covid-19 pandemic: infoveillance study," *Journal of medical Internet research*, vol. 22, no. 4, p. e19016, 2020.

[38] D. H. Anderson, *Compartmental modeling and tracer kinetics*. Springer Science & Business Media, 2013, vol. 50.

[39] L. Wu, F. Morstatter, X. Hu, and H. Liu, "Mining misinformation in social media," in *Big Data in Complex and Social Networks*. Chapman and Hall/CRC, 2016, pp. 135–162.

[40] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Proceedings of the 7th workshop on social network mining and analysis*, 2013, pp. 1–9.

[41] J. Skaza and B. Blais, "Modeling the infectiousness of twitter hashtags," *Physica A: Statistical Mechanics and its Applications*, vol. 465, pp. 289–296, 2017.

[42] C. Comito, A. Forestiero, and C. Pizzuti, "Bursty event detection in twitter streams," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 4, pp. 1–28, 2019.

[43] K. H. Lim, S. Jayasekara, S. Karunasekera, A. Harwood, L. Falzon, J. Dunn, and G. Burgess, "RAPID: Real-time Analytics Platform for Interactive Data Mining," in *Proceedings of the 2018 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'18)*, 2018, pp. 649–653.

[44] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on twitter," *Computers in Human Behavior*, vol. 89, pp. 258–268, 2018.

[45] S. Madisetty, "Event recommendation using social media," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 2106–2110.

[46] P. Kostkova, M. Szomszor, and C. St. Louis, "# swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic," *ACM Transactions on Management Information Systems (TMIS)*, vol. 5, no. 2, pp. 1–25, 2014.

[47] A. Joshi, R. Sparks, S. Karimi, S.-L. J. Yan, A. A. Chughtai, C. Paris, and C. R. MacIntyre, "Automated monitoring of tweets for early detection of the 2014 ebola epidemic," *PloS one*, vol. 15, no. 3, p. e0230322, 2020.

[48] G. Beigi, X. Hu, R. Maciejewski, and H. Liu, "An overview of sentiment analysis in social media and its applications in disaster relief," in *Sentiment analysis and ontology engineering*. Springer, 2016, pp. 313–340.

[49] V. K. Neppalli, C. Caragea, A. Squicciarini, A. Tapia, and S. Stehle, "Sentiment analysis during hurricane sandy in emergency response," *International journal of disaster risk reduction*, vol. 21, pp. 213–222, 2017.

[50] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Syndromic surveillance of flu on twitter using weakly supervised temporal topic models," *Data mining and knowledge discovery*, vol. 30, no. 3, pp. 681–710, 2016.

[51] K. H. Lim, S. Karunasekera, and A. Harwood, "Clustop: A clustering-based topic modelling algorithm for twitter using word networks," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 2009–2018.

[52] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events: a classification-summarization approach," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 583–592.

[53] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra, "Summarizing situational tweets in crisis scenario," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 2016, pp. 137–147.

[54] A. Sharma, K. Rudra, and N. Ganguly, "Going beyond content richness: Verified information aware summarization of crisis-related microblogs," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 921–930.

[55] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "Aidr: Artificial intelligence for disaster response," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 159–162.

[56] J. Liu, L. Loh, E. Ng, Y. Chen, K. L. Wood, and K. H. Lim, "Self-Evolving Adaptive Learning for Personalized Education," in *Proceedings of the 2020 ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW'20)*, 2020.

[57] B. E. Parilla-Ferrer, P. Fernandez, and J. Ballena, "Automatic classification of disaster-related tweets," in *Proc. International conference on Innovative Engineering Technologies (ICIET)*, vol. 62, 2014.

[58] J. Liu, Y. C. Ng, K. L. Wood, and K. H. Lim, "IPOD: A Large-scale Industrial and Professional Occupation Dataset," in *Proceedings of the 2020 ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW'20)*, 2020.

[59] M. Krieck, J. Dreesman, L. Otrusina, and K. Denecke, "A new age of public health: Identifying disease outbreaks by analyzing tweets," in *Proceedings of health web-science workshop, ACM Web Science Conference*, 2011, pp. 10–15.

[60] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1474–1477.

[61] A. Ashok, M. Guruprasad, C. Prakash, and S. Shylaja, "A machine learning approach for disease surveillance and visualization using twitter data," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE, 2019, pp. 1–6.

[62] J. Liu, K. L. Wood, and K. H. Lim, "Strategic and Crowd-Aware Itinerary Recommendation," in *Proceedings of the 2020 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'20)*, 2020.

[63] Y. Okuyama, "Critical review of methodologies on disaster impact estimation," *Background paper for EDRR report*, 2008.

[64] A. F. Avelino and G. J. Hewings, "The challenge of estimating the impact of disasters: many approaches, many limitations and a compromise," in *Advances in Spatial and Economic Modeling of Disaster Impacts*. Springer, 2019, pp. 163–189.

[65] K. McCormick and V. K. Saba, *Essentials of nursing informatics*. McGraw-Hill, 2015.

[66] G. Avinash, R. Liu, and S. Roehm, "System and method for integrated learning and understanding of healthcare informatics," May 24 2007, uS Patent App. 11/284,855.

[67] K. Siau and Z. Shen, "Mobile healthcare informatics," *Medical informatics and the Internet in medicine*, vol. 31, no. 2, pp. 89–99, 2006.

[68] P. Nambisan, Z. Luo, A. Kapoor, T. B. Patrick, and R. A. Cisler, "Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter," in *2015 48th Hawaii International Conference on System Sciences*. IEEE, 2015, pp. 2906–2913.

[69] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[70] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.