IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 May 2024, accepted 18 June 2024, date of publication 15 July 2024, date of current version 26 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3427760

APPLIED RESEARCH

Using Twitter Dataset for Social Listening in Singapore

QIONGQIONG WANG[1], (Member, IEEE), HARDIK B. SAILOR[1], KONG AIK LEE[2], (Senior Member, IEEE), KAI MA[3], (Member, IEEE), KIM HUAT GOH[3], (Member, IEEE), AND WAI FONG BOH[3], (Member, IEEE)

[1]Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), Singapore 138632
[2]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong
[3]IT and Operations Management, Nanyang Business School (NBS), Nanyang Technological University (NTU), Singapore 639798

Corresponding author: Qiongqiong Wang (Wang_Qiongqiong@i2r.a-star.edu.sg)

This work was supported by the National Research Foundation, Singapore, and Ministry of National Development, Singapore, under its Cities of Tomorrow Research and Development Program under Award COT-CityScan-2020-1.

ABSTRACT As a highly urbanized nation, Singapore faces unique urban planning challenges due to its geographical attributes and demographics. These include optimizing land and transportation, enhancing quality of life, and preparing for pandemics. Quick responses and understanding of region-specific social voices are essential for effective policy-making and real-time insights into local dynamics. This work delves into analyzing social media data sourced from Twitter within the context of Singapore, forming a crucial component of a broader social listening initiative. Specifically, 96.7 million tweets from 2008 to 2023 were collected using Twitter's free API, providing a decade's worth of social data from Singapore. Alongside the Twitter data, we release a list of 10,357 places and property names with geographic coordinates, mapped to 332 subzones and 55 planning areas in Singapore. In this paper, we further present examples of locating methods that enable region-specific analysis of different urban zones, gathering information reflecting the attitudes of citizens associated with each estate. We showcase the practical application of the dataset through two distinct use cases: sentiment analysis on the prevalent issue of COVID-19 and bursty topic detection during the years 2020 and 2021. Deep learning-based methods are employed for the analysis: sentiment analysis using a zero-shot pretrained model and bursty topic analysis based on the biterm topic model. The experimental analysis demonstrates the efficacy of social listening, providing valuable insights for future city planning in other countries and cities. This work offers invaluable resources and methodologies for the research community, highlighting the potential of social media data in enhancing urban planning and policy-making. The data is realised at https://doi.org/10.21979/N9/PALUID.

INDEX TERMS Social listening, twitter data, sentiment analysis, bursty topic detection, Singapore.

I. INTRODUCTION

Investigating social media has become increasingly valuable due to the exponential growth in social data and user participation. With the surge in telecommunications and the widespread adoption of pervasive systems like mobile phones, coupled with recent advancements in the Internet of Things (IoTs), there exists a remarkable opportunity to amass vast amounts of real-time data encompassing people's behaviors, movements, and land use across diverse geographical regions. We refer to this process as social listening: monitoring digital conversations of target users and online mentions of specific keywords, phrases, brands, or topics to gain insights into public opinion, sentiment, and trends. This wealth of dynamic data, complemented by more static information such as macroeconomic indices, and qualitative insights derived from user-generated content on social media platforms, facilitates a nuanced comprehension of spatio-temporal urban dynamics in unprecedented ways.

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

As a highly urbanized nation, Singapore grapples with distinctive urban planning challenges stemming from its unique geographical attributes and demographics. Spanning a mere 700 square kilometers, Singapore accommodates a remarkably dynamic resident population, driven by the imperatives of globalization. In pursuit of its vision to position Singapore as a world-class city, policymakers have embraced elements of Western planning paradigms in national development endeavors. Central to this vision is the commitment to fostering urban livability and elevating residents' quality of life. This dedication is palpable in the government's manifold projects and initiatives, which strive to bolster social cohesion, cultivate a vibrant cultural scene, and cultivate environmentally sustainable spaces.

Nonetheless, the landscape of sustainable urban planning and development in Singapore is marked by intricacies exacerbated by various factors. The nation contends with an aging population, compounding the challenges of maintaining economic dynamism and social cohesion. Moreover, disruptions wrought by cutting-edge technologies and unforeseen events, such as the COVID-19 pandemic, inject additional layers of complexity into the urban planning milieu. Navigating these multifaceted challenges while striving for sustainable development remains a paramount concern within the national context.

Social media data offer the advantage of region-specific analysis, enabling real-time monitoring and analysis of factors shaping the nation's pulse. By delving into localized data, city planners can better understand the distinctive dynamics of various regions across Singapore. In conjunction with diverse data streams encompassing various aspects of everyday life in Singapore, insights gleaned from social media are instrumental in providing policymakers with a comprehensive grasp of the essence, vitality, and activities within different urban zones. Their insights facilitate answers to complex questions such as optimizing land and transportation usage, enhancing citizens' quality of life within spatial and resource constraints, and preparing for socioeconomic impacts of unexpected events like pandemics, Thereby paving the way for innovative service enhancements tailored to citizens' requirements.

This paper centers on the analysis of social media, particularly focusing on 'X', previously known as 'Twitter' (hereafter referred to as 'Twitter' throughout this paper)., within the context of Singapore. Twitter is a popular platform for individuals to share their thoughts and feelings due to its concise nature, real-time updates, public visibility, and options for anonymity and pseudonymity. We gathered Twitter data spanning a decade in Singapore. Our analysis concentrated on sentiment regarding the 'COVID-19' issue within a subset of data from 2020 to 2021, along with bursty topic detection. We conducted these analyses with the aim of providing illustrative examples showcasing the potential applications of social data in the future. The key innovative aspects of our study include:

- This is the first study that utilizes social media data to develop an information processing system tailored to the Singaporean context.
- We demonstrate the efficacy of leveraging unlabeled data for sentiment analysis, enabling direct search capabilities that provide statistical summaries in response to specific user queries. Our time-wise and location-wise analyses offer insights into how topics of interest vary across different populations over time and geographical locations.
- In our analysis of trending topics, we employ a Bursty Biterm Topic Model (BBTM)-based approach without pre-selecting topics since we do not have labels and utilised entire data. Our results indicate that this approach effectively clusters meaningful topics and identifies trends when analyzed on a monthly basis.
- The entire dataset will be publicly available for further research, serving as a valuable resource for the community, particularly in light of the no more unavailability of freely accessible data from Twitter.

II. PREVIOUS PRELIMINARY WORK/STATE OF CURRENT RESEARCH

Several research initiatives focusing on the concept of 'city pulse' have been reported in the last decade. Published reports in this area come from all across the world, with significant numbers from Europe, USA and Asia in that order. Table 1 lists the details of published research on city pulse and related themes.

There are works centered on a whole range of aspects from providing an understanding of the concept to proposing frameworks and architectures that would support gathering, aggregating, and visualizing various indicators of city pulse [1], [2], [3], [4]. Out of these, some aim for conceptual clarification [1] while others focus on semantic models [5] or technical architecture [6]. Researchers have also attempted narrowing down on specific dimensions of city pulse based on easily trackable factors such as human mobility as well as activities including attendance of events [7], biking [8], spatiotemporal activity [2] and geo-located social activity [9].

Additionally, a number of research initiatives have been reported on related themes such as smart cities, livable cities or sustainable cities. Such research has been reviewed due to possible overlaps of parameters of interest in the context of city life as well as some methodological parallels that could be drawn between the themes. Research in this category, once again falls within various categories such as seeking conceptual clarification [10], [11], providing means of deriving various indices indicative of the standards of life within a given city [12], [13] or making systematic comparisons or selections of cities [14] based on established benchmarks. Major research reported in the space of City Pulse and related themes in the last 10 years is consolidated and presented in Table 1 below.

TABLE 1. Details of Published Research on City Pulse and Related Themes.

No.	Location	Theme	Output/Findings
1	Europe	CityPulse: Large Scale Data Analytics Framework for Smart Cities [6]	Proposes and tests a CityPulse framework using a distributed system for semantic discovery, data analytics, and large-scale (near) real-time IoT and social media data. The framework also covers data analytics modules that perform intelligent data aggregation, event detection, quality assessment, contextual filtering, and decision support. Demonstrates how the components of the framework interact to support custom-made applications for citizens.
1	USA	A Multi-Granular, Semantic Signatures-Based Information Observatory for the Interactive Visualization of Big Geosocial Data [1]	A theoretical and technical framework to interactively explore the pulse of a city based on social media.
3	USA	Pulse of the city: Visualizing Urban Dynamics of Special Events [7]	Illustrates how special events influence the normal rhythms of the city, and how crowds move and respond to large-scale public events. Discusses the potential of realizing an interface aimed at informing urban dwellers in real time of the dynamics Of their individual locations
4	NYC,USA	Urban Pulse: Capturing the Rhythm of Cities [2]	Defines "urban pulse" which captures spatio-temporal activity in a city across multiple temporal resolutions. Pulses characterized as a set of beats are obtained and compared. Presents a visual exploration framework that allows users to explore the pulses within and across multiple cities under different conditions.
5	Barcelona, Spain	Sensing and Predicting the Pulse of the City through Shared Bicycling [8]	Digital footprints from an emerging urban infrastructure- shared bicycling systems are used to provide a spatiotemporal analysis of bicycle station usage from Barcelona's shared bicycling system, called Bicing.
6	Munich, Germany	City Pulse: Supporting Going-Out Activities with a Context-Aware Urban Display [3]	Design of City Pulse, an urban public display that helps people find going-out locations of their taste. Using sensors incorporated into mobile phones that gather data on motion, pulse and surrounding noise around City Pulse display shows, how crowded and how loud the locations are, which music is playing, whether people dance or drink.
7	Europe	Semantic Modelling of Smart City Data [5]	Puts forward a semantic description model to describe and help discover, index and query smart city data. Presents examples of data that can be collected from cities and discuss issues around collection and use of this data
8	Europe	CityPulse: Real-Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications	Developing and testing of a distributed framework for semantic discovery and processing of large-scale real-time IoT and relevant social data streams for knowledge extraction in a city environment.
9	Greece	CityPulse: A Platform Prototype for Smart City Social Data Mining [9]	Proposes a modular platform for offering smart city services based on geo-located social data mining.
10	Thailand	Development of a Liveable City Index (LCI) Using Multi Criteria Geospatial Modelling for Medium Class Cities in Developing Countries [12]	Presents a Livable City Index (LCI) based on residents' opinions and experts' recommendations with the integration of Geographic Information System (GIS) techniques
11	Italy	Current trends in Smart City initiatives: Some stylized facts [11]	Provides an understanding of the notion of a Smart City through natural resources and energy, transport and mobility, buildings, living, government, and economy and people. Presents evidence that the evolution patterns of a SC depend on its local contextual factors such as economic development and structural urban variables
12	Turkey	The most livable city selection in Turkey with the grey relational analysis [14]	Proposes and applies an approach to selecting the most livable city
13	Canada	Smart Cities: Definitions, Dimensions, Performance, and Initiatives [10]	Clarifies the concept of "smart" in the context of cities through literature review and relevant materials. Also identifies the main dimensions and elements characterizing a smart city
14	Singapore	Towards the Construction of Smart City Index for Analytics (SM-CIA): Pilot-Testing with Major Cities in China Using Publicly Available Data [13]	Proposes and pilot tests the Smart City Index for Analytics (SMCIA), to objectively to measure the degree of smartness in urban cities. A time trend analysis to detect changes of the cities found a significant increasing trend in smart living, mobility, economy and governance domains.

Much of the published research aims to develop various indicators of what could be termed the city pulse through factors such as those that enable monitoring of human mobility or activity, economic health of the region and social media use. Few have adopted a comprehensive integrated means of assessing the pulse of a nation aggregating a wide range of factors that may be indicative of the vibrancy of city life. Additionally, most, if not all of these studies have been undertaken at a country or city level and present consolidated national/city level indices.

Given the uniqueness of Singapore as a city state, having a city pulse at the city level will not provide sufficient nuances to understand the differences in character and pulses of different planning areas in Singapore. Further granularity of information by drilling down to smaller, specific planning areas will be much more beneficial for targeted urban planning of spaces and facilities. Yet, this has been missing in most of these published sources available currently. This is the specific gap we aim to address through this proposed project.

III. DATA

Social media provides rich, multimedia data that reflects the attitudes of citizens. Combining social media data with advanced statistical and machine learning techniques provides a significant range of opportunities to sense and measure the attitudes and public opinions of citizens associated with different estates.

We collect data from social media outlets where APIs are provided for the legal collection of data from public accounts. Specifically, we collected social media data from Twitter to gather information reflecting the attitudes of citizens associated with each estate. Several factors contribute to why people tend to express their feelings on Twitter more than on other social media platforms, as listed below:

- Conciseness and brevity: Twitter's character limit (previously 140 characters, now 280) encourages users to express their thoughts and feelings concisely. This limitation can make it easier for individuals to quickly share their emotions without having to compose lengthy posts.
- Real-time updates: Twitter's format allows for real-time updates, making it conducive for sharing immediate thoughts and feelings as they occur. Users can quickly post about current events, personal experiences, or reactions to news, fostering a sense of immediacy and connection.

- Public nature: Twitter is a highly public platform where tweets are often visible to a wide audience, including followers, retweeters, and potentially the broader Twitter community. This public visibility may incentivize users to share their feelings more openly, seeking validation, support, or engagement from others.
- Engagement and interaction: Twitter's design facilitates engagement through features like replies, retweets, and likes. Users can easily interact with others' tweets, fostering conversations and amplifying emotional expression. This interactive nature encourages users to share their feelings and engage with others' emotions.
- Hashtags and trends: Twitter's use of hashtags and trending topics allows users to join ongoing conversations around specific themes or events. This can provide an outlet for individuals to express their feelings in response to shared experiences, cultural moments, or social movements, contributing to a sense of community and solidarity.
- Anonymity and pseudonymity: While Twitter encourages real identities, some users may still maintain a degree of anonymity or pseudonymity, allowing them to express feelings more freely without fear of judgment or repercussions. This anonymity can lead to more candid and uninhibited emotional expression.

A. 16-YEAR PUBLICLY AVAILABLE SOCIAL DATA

1) DATA STATISTICS

We collected and released the social Twitter data specifically in Singapore from 2008 to 2023, including 96,686,894 tweets from 82,324 user accounts. Together with the Twitter data, we also provide a list of 10,357 places and property names obtained from the Google Maps API. This dataset includes geographic coordinates (latitude and longitude) for each place, mapped to one of 332 subzones and 55 planning areas in Singapore [15]. The statistics of the collected data are shown in Table 2.

TABLE 2. Statistics of the collected data.

Attribute	Number	Source
Tweets	96,686,894	Twitter
Users	82,324	
Places	10,357	Google Map API
Sub-zones	332	[15]
Planing areas	55	

B. COLLECTION METHOD

1) TWITTER COLLECTION METHOD

To collect Twitter data, we utilized the Twitter API, specifically *Twitter Search* and *Timeline Search* functionalities, which were free at the time of data collection. Due to the limitations on the number of tweets imposed by the free version of the API, we adopted a two-step method to gather the tweets, as illustrated in Steps 1 and 2 in Figure 1:

- Step 1: One-time Twitter Search. We performed a one-time Twitter Search using the *Geocode* setting in *Timeline Search* (1.346353, 103.807526, 25*km*) indicating the coordinates and radius, to cover the whole of Singapore, as depicted in Figure 2. This search returned approximately 40,000 to 70,000 tweets. From the collected tweets, we extracted and summarized the user IDs, retaining only those with location tags indicating "Singapore".
- Step 2: Tweet search by user IDs. Using the summarized user IDs from Step 1, we employed *Timeline Search* in the Twitter API. This step allows us to retrieve a limited number of tweets from these users, covering the period up to the moment of search.

2) LOCATING METHOD

Twitter offers a *Geotag* function that allows users to attach geographical metadata to their posts or media uploads, thereby enabling the identification of associated locations. However, the proportion of tweets with *Geotags* remains notably low. Consequently, this study aims to employ text analysis techniques to ascertain locations from the tweet contents. We propose two approaches, categorizing the derived locations as either local or global.

We identify local locations by searching for place names within each tweet's text, using place names sourced from the Place dataset. Subsequently, each tweet is augmented with a local location field, which indicates one or multiple subzones and planning areas corresponding to the identified places mentioned in the tweet. For the determination of global locations, we employed a two-step method, as illustrated in Step 3 of the block diagram depicted in Figure 1:

- Step 1: We aggregated users' locations from their tweets across the entire Twitter dataset.
- Step 2: Subsequently, we assigned these locations to each tweet.

The global location method allows for the utilization of a larger volume of tweet data. However, it typically exhibits reduced precision due to users not consistently posting all their tweets from a single location. The same tweets may undergo repeated analysis across multiple locations, leading to potential redundancy. In contrast, the local location approach offers heightened precision by concentrating on the content of individual tweets. Nonetheless, the availability of tweets with local location data may be restricted, potentially leading to a smaller sample size. This reduced sample size can introduce greater uncertainty into subsequent analysis results.

IV. USE CASE

This work is part of the "Assessing Cities: A SCoRE (Societal Comprehensive Reflective Estimate) Methodology"project, supported by the National Research Foundation, Singapore and Ministry of National Development, Singapore under its Cities of Tomorrow R&D Programme (Award No: COT-CityScan-2020-1). The project's goal is to devise a methodology for identifying emerging social trends, addressing gaps in research, and further understanding societal shifts to develop relevant interventions for improving city

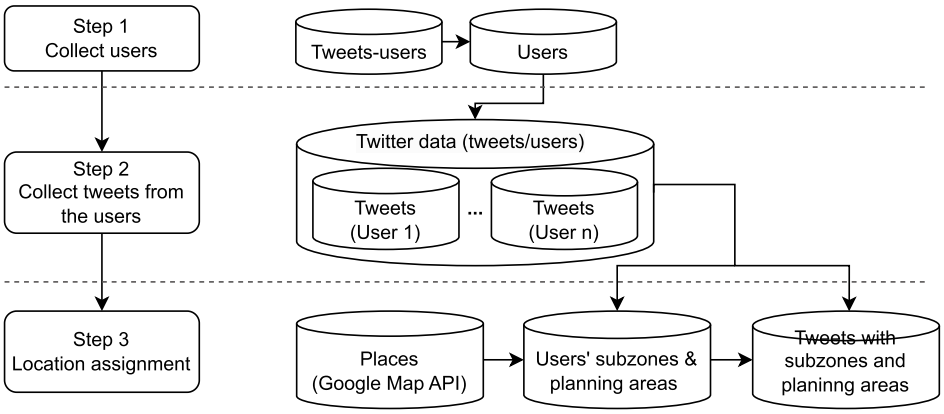

FIGURE 1. *Block diagram to illustrates tweets collection and global location allocation method.*

FIGURE 2. *A Geocode setting as (1.346353, 103.807526, 25km) is used in Twitter API to cover the whole of Singapore.*

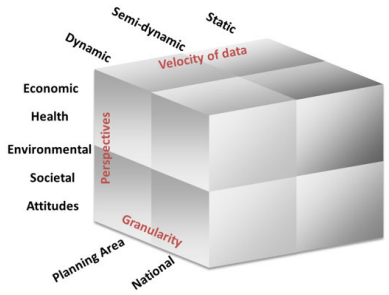

FIGURE 3. *5 dimensions.*

planning and design. The project proposes a multi-perspective framework and system (see Figure 3) to collect and aggregate data on various indicators representing different aspects of life in Singapore, with a focus on planning areas rather than the national level. The framework covers economic, societal, environmental, health, and attitudinal dimensions, with data collected at different temporal scales to reflect changes in each region's pulse. These scales range from static geo-spatial information to highly dynamic behavioral data. As part of the "attitudes" dimension, the social listening component in this work provides real-time access to the target users' pulse, analyzing the attitudes and understand public opinion. This enhances the scope of CityScan by offering valuable insights into community sentiments and engagement.

V. ANALYSIS

A. SENTIMENT ANALYSIS FOR DIRECT SEARCH

1) DIRECT SEARCH

A direct search involves searching for tweets related to a specific, user-defined topic, referred to as a direct-search label. Given a direct-search label l and a tweet document i, a confidence score $c_i^l \in [0, 1]$ indicates the relevance of the document to the direct-search label. A score of $c_i^l = 1$ signifies the highest relevance, while$c_i^l = 0$ signifies no relevance. A pre-determined threshold τ is used to select the relevant tweets.

2) SENTIMENT ANALYSIS

Sentiment analysis is a natural language processing (NLP) technique used to determine the sentiment or emotional tone expressed in a piece of text. The goal of sentiment analysis is to understand the opinions, attitudes, or emotions conveyed by the text, whether it is positive, negative, or neutral. Given the posterior probabilities for three classes: positive (pos), negative (neg), neutral (neu) for a tweet i where the sum equals 1

$$P_{i,\text{pos}} + P_{i,\text{neu}} + P_{i,\text{neg}} = 1 \tag{1}$$

we label the tweet as positive, neutral, or negative based on which class's posterior is the largest. Additionally, we define a *positiveness* score for tweet i from its three posterior probabilities, serveing as its sentiment score:

$$s_i = \begin{cases} \log\left(\dfrac{P_{i,\text{pos}}}{P_{i,\text{neg}}}\right) & \text{for positive and negative classes ,} \\ 0 & \text{for neutral class.} \end{cases}$$

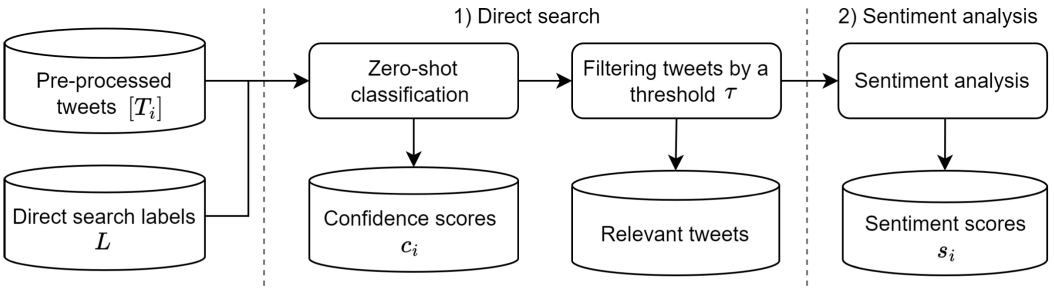

FIGURE 4. Block diagram to illustrate sentiment score calculation for direct search.

3) SENTIMENT SCORES FOR DIRECT SEARCH

For each tweet, a confidence score c^l to a direct-search label l and a positiveness score s can be calculated as mentioned above and illustrated in Figure 4. To analyze the sentiment of a direct-search label l, we propose to use a weighted average sentiment score as follows

$$S^l = \frac{\sum_i^{N_\tau} c_i^l s_i}{\sum_i^{N_\tau} c_i^l}. \tag{2}$$

where N_τ is the number of tweets that have $c_i^l \geq \tau$.

B. BURSTY TOPIC DETECTION

A topic in social media data is termed an emerging or bursty topic if it did not appeared in previous time slice and triggers a significant number of related posts in the current slice [16]. A bursty topic is distinct from frequent topics in social media as it may not necessarily be associated with high frequency. Previously, it was shown that directly applying conventional topic modeling algorithm LDA did not work well due to the short text in twitter data [17]. Therefore, this study employs the Bursty Biterm Topic Model (BBTM) [18], an extension of the earlier work presented in [17], for effective detection of bursty topics. The key idea of our approach is to exploit the burstiness of biterms as prior knowledge to incorporate into BTM for bursty topic modeling. BBTM boasts two significant advantages over preceding techniques. Initially, it adeptly tackles the challenge of data sparsity inherent in topic modeling for short texts, surpassing conventional models. Furthermore, it has the capability to discern and learn bursty topics systematically and efficiently, eliminating the need for heuristic post-processing maneuvers. A more detailed mathematical description of the model is presented in [18].

The block diagram illustrating the training flow is depicted in Figure 5. Here, we take a three month duration as an analysis time slice and apply the model to each month's data. The first step is indexing the words to integers and creating a dictionary. The second step is to calculate biterm statistics considering all three months' data. Intuitively, during the emergence of a trending topic in social media, we often notice a higher occurrence of relevant biterms than usual. For instance, biterms like "Presidential Election"

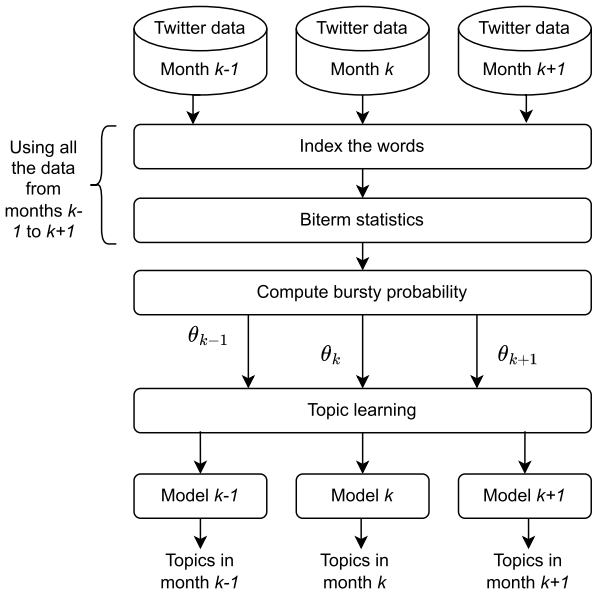

FIGURE 5. Bursty topic model training flow.

experienced a significant surge in popularity on Twitter during the 2023 Singapore elections. Based on the above observation, a probability metric termed "bursty probability" is calculated to assess the burstiness of biterms. This metric can be derived from the temporal frequency patterns of these biterms called as biterm statistics. The bursty probability is calculated for each time slice (here for each month) and is further used by topic model, as shown in Figure 5. Hence, there will be a topic model for each month's data utilizing the corresponding bursty probability.

C. EXPERIMENTAL SETTINGS

We selected a subset of the Twitter data spanning the years 2020 to 2021. The dataset contains a total of 49, 939, 017 tweets. After pre-processing, we retained 40, 659, 542 tweets for the analysis.

In this study, we focus on the prevalent issue of the COVID-19 during the years 2020 and 2021, aiming to demonstrate sentiment analysis through targeted direct searches. Specifically, we employ seven distinct phrases,

namely 'corona virus', 'coronavirus', 'covid', 'vaccination', 'pandemic', 'mask', and 'lockdown' (denoted as L) as labels for conducting direct searches. Subsequently, we determine the confidence score for the COVID-19 topic by selecting the maximum among the confidence scores obtained for these seven phrases.

Utilizing Hugging Face's built-in zero-shot classification pipeline [19], [20], we acquire confidence scores. Zero-shot classification represents a machine learning paradigm aimed at understanding and categorizing previously unseen data – data absent from the model's training corpus. In our methodology, we pass all combinations of the pre-processed tweet texts and the predefined seven phrases L as premise/hypothesis pairs to the pre-trained model, specifically, `facebook/bart-large-mnli` [19]. This model corresponds to a specific checkpoint of the BART (Bidirectional and Auto-Regressive Transformers) model [21] after being trained on the MultiNLI (MNLI) dataset [22]. Subsequently, for every tweet text, the model outputs the probability for each phrase independently, serving as the respective confidence scores.

For sentiment analysis, we employ the transformer-based `twitter-roberta-base-sentiment-latest` model, sourced from Hugging Face [23], [24]. This model is employed to compute the posterior probabilities associated with three distinct sentiment classes: positive, negative, and neutral. To conduct location-specific analysis, we leverage a subset comprising 299,752 tweets of the two-year collection that reference specific geographic locations. Subsequently, these tweets are subjected to sentiment analysis. In our examination of sentiments about the issue of COVID-19, a search threshold of 0.5 is established. This implies that only tweets with direct search confidence scores exceeding 0.5 are included in the computation of sentiment scores.

For bursty topic modeling, the experiments are performed considering consecutive three months within a year. The topic models are trained with 30 topics, and results are obtained with the top 10 words for each topic. Differing from prior studies [17], [18], our approach refrains from using specific keywords during the Twitter data crawling process, resulting in the absence of topic labels. However, the topic labels can be inferred from the topic word distribution.

1) TASK-SPECIFIC PROCESSING

The text of the tweets is pre-processed before proceeding to the analysis. For direct search and sentiment analysis, the texts undergo a series of pre-processing steps. First, a removal of usernames, link placeholders, and punctuation is applied. Next, non-English words are filtered out using the nltk English words corpus. Lemmatization is applied before filtering to ensure that English plurals are retained. Only tweets with at least one word remaining after processing will be used for the analysis.

For the bursty topic modeling, in addition to above mentioned text pre-processing, we also use some techniques based on the recommendation from [17]. The additional pre-processing includes: 1. removing duplicate tweets, 2. removing non-Latin characters and stop words and 3. filtering out tweets with a length less than 2.

2) EXPERIMENTAL RESULTS FOR SENTIMENT ANALYSIS FOR DIRECT SEARCH

Histograms of confidence scores related to the 'COVID-19' topic for all tweets in Singapore from January 2020 to December 2021. (a) displays the scores compared to all the 7 phrases in set **L**, while (b) illustrates the maximum scores in comparison to the 7 phrases.

D. EXPERIMENTAL RESULTS

1) DIRECT SEARCH AND SENTIMENT ANALYSIS

To filter the COVID-19 related tweets from the entire two-year tweet collection, we first calculate the confidence score of each tweet for each of the seven labels in L: 'corona virus', 'coronavirus', 'covid', 'vaccination', 'pandemic', 'mask', and 'lockdown'. The number of tweets multiplied by 7 are computed and displayed in Figure 6 (a). It is evident that the number of tweets decreases exponentially as the confidence score increases. To ensure that relevant tweets are included for further analysis, we select the maximum score among those obtained for these seven labels as the confidence score for the tweet for the COVID-19 topic. The distribution the these maximum scores is shown in Figure 6 (b). Utilizing a predetermined search threshold of 0.5 for the confidence scores, we identified 21, 799, 980 tweets, constituting 53.62% of the collected dataset. We refer to it as COVID-19 related tweets. This figure significantly surpasses the ratio obtained when utilizing scores compared to individual labels in set L, which averages at 8.89%.

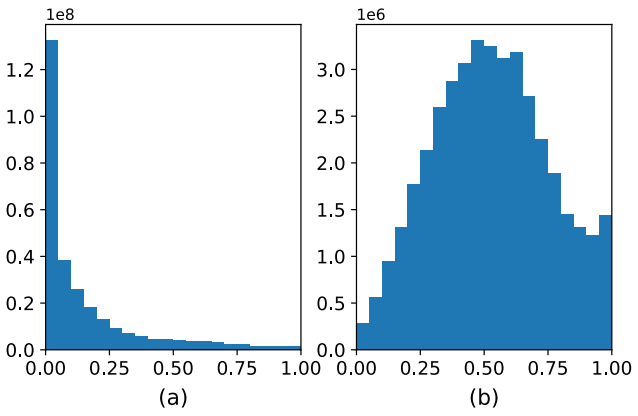

FIGURE 6. Histograms of confidence scores related to the 'COVID-19' topic for all tweets in Singapore from January 2020 to December 2021. (a) displays the scores compared to all the 7 phrases in set L, while (b) illustrates the maximum scores in comparison to the 7 phrases.

For the sentiment analysis, we evaluate two measures: the ratios of positive, neutral, and negative tweets, and the weighted average sentiment scores of exclusively positive and negative tweets. The sentiment scores are measured by positiveness as described in Section V-A2. Using both measures provides a more comprehensive analysis and deeper insights. First, we investigate the overall sentiment changes in tweets over the two-year period, as shown in Figure 7. This analysis is not limited to any specific topics.We observe a drop in the sentiment score of positive tweets in March 2020. Since then, the score of positive tweets has been gradually increasing. A similar drop in the ratio of positive tweets occurred in March 2020, with an additional decline observed in May 2021. The lowest weighted average sentiment score for negative tweets (indicating the most negativity) is seen in June 2020 and May 2021. Correspondingly, the ratio of negative tweets also peaks during these two months. Overall, the absolute values of the sentiment scores for positive tweets are higher than those for negative tweets, and the ratio of positive tweets exceeds that of negative tweets. This indicates that the public sentiment was generally more positive.

FIGURE 7. Monthly sentiment analysis of tweets in Singapore from January 2020 to December 2021. Top: Ratios of positive, neutral, and negative tweets. Middle & Bottom: Weighted average sentiment scores for positive and negative tweets.

Next, we investigate public sentiment towards the COVID-19 topic using 21 million COVID-19 related tweets. The monthly sentiment analysis is shown in Figure 8. The two measures of sentiment analysis in COVID-19 related tweets (Figure 8) show a correlation with those of all tweets (Figure 7), but with more pronounced changes. This indicates that the public has stronger opinions on the COVID-19

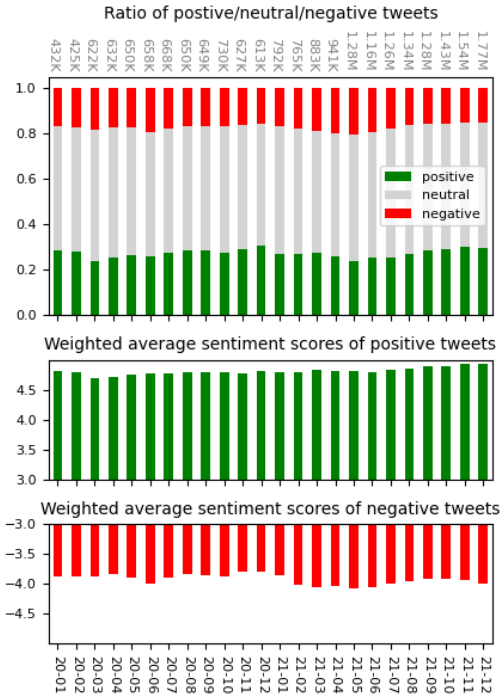

FIGURE 8. Monthly sentiment analysis of the 'COVID-19' related tweets. Top: Ratios of positive, neutral, and negative tweets. Middle & Bottom: Weighted average sentiment scores for positive and negative tweets.

specific topic compared to the average of all topics. Notably, there is a similar drop in the ratio of negative tweets and the sentiment score of negative tweets in June 2020. The period from February 2021 to August 2021 shows the lowest overall sentiment scores and the highest ratios of negative tweets.

We also investigate the difference in public sentiment towards COVID-19 between regions. Figure 9 presents sentiment analysis across 55 planning areas using the 21 million COVID-19 related tweets. The bars are arranged in descending order based on the ratio of negative tweets, with the 'SG' bar serving as a reference benchmark for Singapore nationwide statistics. Notably, the bars corresponding to planning areas Mandai, Bukit Panjang, Bukit Batok, and Clementi are positioned to the left of the 'SG' bar, indicating higher ratios of negative tweets compared to the national average. It is essential to highlight that the statistics for 'SG' encompass tweets lacking location specifications, leading to a significantly larger total tweet count for 'SG' compared to the sum of tweets across all planning areas.

2) BURSTY TOPICS DETECTION

Example of bursty topics discovered by the model is shown in the Table 3. The first topic is background topic with a large topic probability, while others are bursty ones. The high values of topic probability indicate that the topics are extensively discussed within a particular time slice and are not considered as bursty topics.

Since we do not have topic labels, we examined the topics discovered by the model and selected two topics for

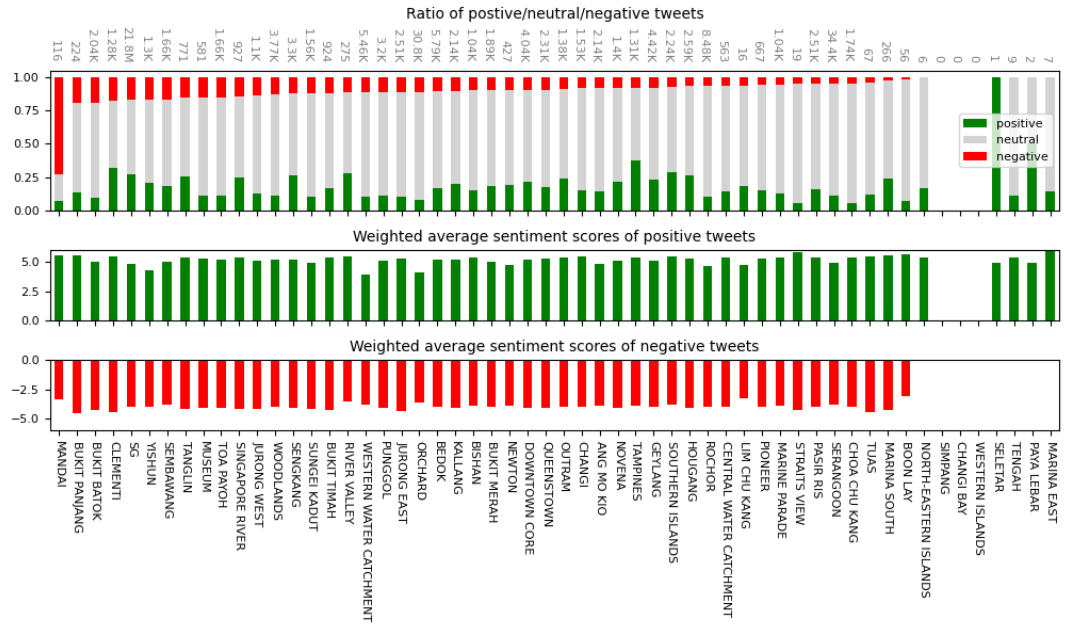

FIGURE 9. Planning area-based sentiment analysis of COVID-19 related tweets in Singapore (2020-2021). Top: Ratios of positive, neutral, and negative tweets. Middle & Bottom: Weighted average sentiment scores in positiveness for positive and negative tweets. Bars are ordered by the ratio of negative tweets, with 'SG' as a nationwide benchmark.

TABLE 3. Example of burty topics with top 10 words.

Topic ID	P(z)	Top 10 most probable words
1	0.923575	im time people day love happy life world singapore feel
9	0.003312	#btc #eth eth buy join chain los spot trading unfi
10	0.002599	bir ve #ace ile gibi kadar ben olarak deil mi
13	0.00235	united liverpool league premier manchester bruno #mufc chelsea club nah
14	0.002341	jan january covid covid19 vaccine test positive day #covid19 singapore
15	0.00229	id backup lvl battle awards golden congratulations xiao music disc
16	0.001922	rubina boss lady #food #bb14 #foodporn #foodie akl #technology bdt
17	0.001914	#master #forex weather #gold cold #fx #eurusd tattoo sl #masterfilm
18	0.001661	hr #hr hrsingapore india test #training series #hrsingapore #singapore hrlaw
19	0.001654	got7 album katy perry #got7 wa #uknow #bts youngjae #bobby
20	0.001396	#nct #nct2020 #nctdream air #nct127 #sggo #eggsandskittlesgos kelly #nctu rhian
21	0.001392	digimon south beat #beauty africa gang gucci #aishwaryarai #aishwaryaraibachchan #bollywood
22	0.001387	pain vha #handmade #polymerclay #photography selling #kids #artandcraft #singapore ndi
24	0.001269	mix #stockmarket #trading #stocks #intraday #equity #nifty nie #priceaction #trade
26	0.001198	#adoptmetrades #adoptmetrading #royalehightrades #royalehightrading #adoptmetrader #royalehigh #adoptme #royalehightrader #royalehighselling #rh
29	0.001042	culinary #yco21virtual hope #forwardtogether battle mentor worlds biggest virtual #youngchef
30	0.001012	#ellasukasabella #kitajagakita #sabellasentiasa #mosscrepe #jagajarak #igsg #sabella insecure #bandai #sgig

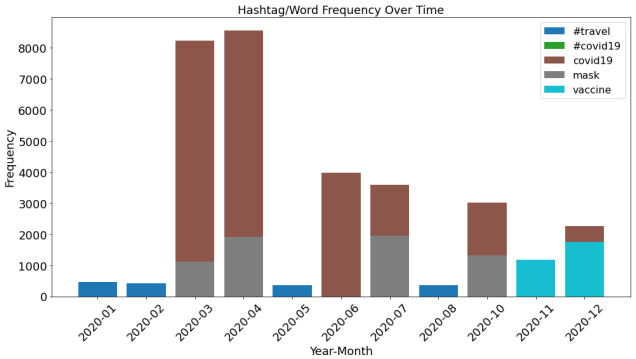

FIGURE 10. Selected words/hashtag distribution over time obtained from topics in a year 2020.

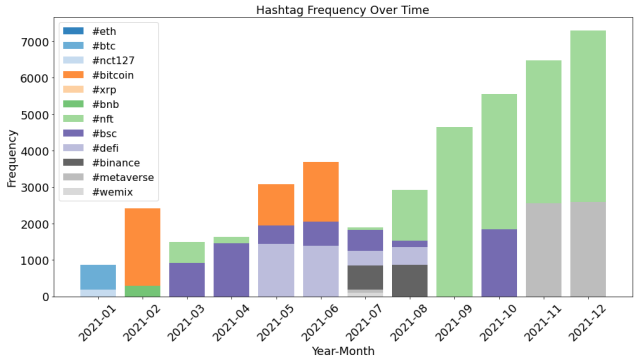

FIGURE 11. Selected words/hashtag distribution over time obtained from topics in a year 2021.

visualization. The topics related to COVID-19 and crypto currency were chosen based on manual inspection and word distribution is plotted for all the months in year 2020 and 2021, respectively. Figure 10 shows words and hashtags in topic related to COVID-19. We can observe that 'covid19' keyword and mask related tweets are significantly higher

in the months of March and April, gradually decreasing in subsequent months. During the last two months of 2020, there were also tweets regarding vaccination that align well with vaccination programs in Singapore. Similarly, the plot for word distribution for crypto currency topic is shown in Figure 11. Here, we can observe that different tweets regarding different crypto-related topics were discussed throughout the year 2021. Interestingly, several clusters of hashtags can be observed during different time slices, such as #eth, #btc, #bitcoin, and #bsc during first few months, while #bsc and #defi were discussed from March to October 2021. The special case of #binance (related to Binance company) is also observed in months July and August. This monthly analysis of specific keywords for each topic shows how the topic is discussed throughout the year or in a given timeline.

VI. CONCLUSION

In conclusion, this paper has presented a study of social media data obtained from Twitter, as a part of a research project study. We have detailed our data collection methodology, including the estimation of location information, offering transparency and insight into our process. Moreover, we have showcased the practical application of our dataset through two distinct use cases: sentiment analysis and bursty topic detection. Our examination of sentiment regarding the prominent issue of COVID-19 across 55 planning areas in Singapore, alongside temporal variations, underscores the diverse spectrum of opinions and attitudes prevalent in the discourse. Additionally, our bursty topic modeling approach has illustrated the capacity to extract meaningful topic clusters from monthly data, even without supervision, demonstrating the robustness and utility of our analytical framework. Overall, our findings contribute valuable insights into the dynamics of social media discourse and highlight the potential for leveraging such data for nuanced analysis and understanding of societal trends.

ACKNOWLEDGMENT

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore, and the Ministry of National Development, Singapore.

REFERENCES

[1] G. McKenzie, K. Janowicz, S. Gao, J.-A. Yang, and Y. Hu, "POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data," *Cartographica, Int. J. Geographic Inf. Geovisualization*, vol. 50, no. 2, pp. 71–85, Jun. 2015.

[2] F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Gonçalves, L. Wilson, M. Hsieh, and C. T. Silva, "Urban pulse: Capturing the rhythm of cities," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 791–800, Jan. 2017.

[3] M. Obaid, E. Kurdyukova, and E. Andre, "City pulse: Supporting going-out activities with a context-aware urban display," in *Proc. Adv. Comput. Entertainment, 9th Int. Conf.*, Nov. 2012, pp. 529–532.

[4] D. Puiu, P. Barnaghi, R. Toenjes, D. Kumper, M. I. Ali, A. Mileo, J. Xavier Parreira, M. Fischer, S. Kolozali, N. Farajidavar, F. Gao, T. Iggena, T.-L. Pham, C.-S. Nechifor, D. Puschmann, and J. Fernandes, "CityPulse: Real-time IoT stream processing and large-scale data analytics for smart city applications," *IEEE Access J.*, vol. 4, pp. 1086–1108, 2016, doi: 10.1109/ACCESS.2016.2541999.

[5] S. Bischof, A. Karapantelakis, C.-S. Nechifor, A. Sheth, A. Mileo, and P. Barnaghi. (Feb. 2014). *Semantic Modelling of Smart City Data*. [Online]. Available: https://corescholar.libraries.wright.edu/knoesis/572

[6] D. Puiu, P. Barnaghi, R. Tönjes, D. Kümper, M. I. Ali, A. Mileo, J. X. Parreira, M. Fischer, S. Kolozali, N. Farajidavar, F. Gao, T. Iggena, T.-L. Pham, C.-S. Nechifor, D. Puschmann, and J. Fernandes, "CityPulse: Large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.

[7] A. Vaccari, M. Martino, F. Rojas, and C. Ratti, "Pulse of the city: Visualizing urban dynamics of special events," in *Proc. 20th Int. Conf. Comput. Graphis Vis.*, Jul. 2013, pp. 1–10.

[8] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and predicting the pulse of the city through shared bicycling," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, Jan. 2009, pp. 1420–1426.

[9] M. Giatsoglou, D. Chatzakou, V. Gkatziaki, A. Vakali, and L. Anthopoulos, "CityPulse: A platform prototype for smart city social data mining," *J. Knowl. Economy*, vol. 7, no. 2, pp. 344–372, Jun. 2016.

[10] V. Albino, U. Berardi, and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," *J. Urban Technol.*, vol. 22, no. 1, pp. 3–21, Jan. 2015.

[11] P. Neirotti, A. De Marco, A. C. Cagliano, G. Mangano, and F. Scorrano, "Current trends in smart city initiatives: Some stylised facts," *Cities*, vol. 38, pp. 25–36, Jun. 2014.

[12] W. Onnom, N. Tripathi, V. Nitivattananon, and S. Ninsawat, "Development of a liveable city index (LCI) using multi criteria geospatial modelling for medium class cities in developing countries," *Sustainability*, vol. 10, no. 2, p. 520, Feb. 2018.

[13] Y.-L. Theng, X. Xu, and W. Kanokkorn, "Towards the construction of smart city index for analytics (SM-CIA): Pilot-testing with major cities in China using publicly available data," in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 2964–2973.

[14] E. Kose, D. Vural, and G. Canbulut, "The most livable city selection in Turkey with the grey relational analysis," *Grey Syst., Theory Appl.*, vol. 10, no. 4, pp. 529–544, Jun. 2020.

[15] (2020). *2019 City Population—Statistics, Maps and Charts | Singapore: Regions*. [Online]. Available: https://www.citypopulation.de/en/singapore/cities/

[16] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1427–1445, Mar. 2022.

[17] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 1445–1456.

[18] X. Yan, J. Guo, Y. Lan, J. Xu, and X. Cheng, "A probabilistic model for bursty topic discovery in microblogs," in *Proc. AAAI Conf. Artif. Intell.*, 2015, vol. 29, no. 1, pp. 1–7.

[19] *Bart-Large-Mnli*. Accessed: Dec. 31, 2023. [Online]. Available: https://huggingface.co/facebook/bart-large-mnli

[20] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3914–3923.

[21] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[22] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 1112–1122.

[23] *Twitter-ROBERTa-base for Sentiment Analysis—UPDATED (2022)*. Accessed: Dec. 31, 2023. [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[24] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "TimeLMs: Diachronic language models from Twitter," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2022, pp. 251–260.

QIONGQIONG WANG (Member, IEEE) received the B.E. degree from the Undergraduate School of Physics, Shanghai Jiao Tong University, China, in 2011, and the M.E. degree in computer science from Tokyo Institute of Technology, Japan, in 2013. She was a Researcher with the Biometrics Research Laboratories, NEC Corporation, Japan, from 2013 to 2021. She is currently a Lead Research Engineer with the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore. Her research interests include multi-modality modeling, foundation models, paralinguistic AI, speech anti-spoofing, and speech enhancement.

HARDIK B. SAILOR received the B.E. degree from the Government Engineering College (GEC), Surat, in 2010, and the M.Tech. and Ph.D. degrees from the Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, in 2013 and 2019, respectively. From 2019 to 2020, he was a Postdoctoral Researcher with The University of Sheffield, U.K. From 2020 to 2022, he was a Chief Engineer with Samsung Research Institute Bangalore (SRIB). He is currently a Senior Scientist with the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include multi-modality modeling, foundation models, paralinguistic modeling, speech recognition, and audio classification.

KONG AIK LEE (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. After this, he joined the Institute for Infocomm Research, Singapore, as a Research Scientist and then a Strategic Planning Manager (concurrent appointment). He is currently an Associate Professor with The Hong Kong Polytechnic University (PolyU), Hong Kong. Before joining PolyU, he was an Associate Professor with Singapore Institute of Technology, Singapore, while holding a concurrent appointment as a Principal Scientist and a Group Leader with the Agency for Science, Technology and Research (A*STAR), Singapore. From 2018 to 2020, he was a Senior Principal Researcher with Data Science Research Laboratories, NEC Corporation, Tokyo, Japan. His research interests include the automatic and para-linguistic analysis of speaker characteristics, ranging from speaker recognition, language and accent recognition, voice biometrics, spoofing, and countermeasures. He is an elected member of the IEEE Speech and Language Processing Technical Committee. Since 2016, he has been an Editorial Board Member of *Computer Speech and Language* (Elsevier). He was a recipient of Singapore IES Prestigious Engineering Achievement Award, in 2013, and the Outstanding Service Award from IEEE ICME 2020. He was the General Chair of the Speaker Odyssey 2020 Workshop. From 2017 to 2021, he was an Associate Editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

KAI MA (Member, IEEE) received the B.E. degree from the College of Computer Science and Technology, Beijing University of Technology, China, in 2008, and the M.E. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2011. He was a Research Engineer with SMU, Singapore, and a Senior Applications Developer with NIE, Nanyang Technological University. He is currently a Research Associate with the Division of Information Technology and Operations Management, Nanyang Business School, Nanyang Technological University. His research interests include data pre-processing, data manipulations, system architecture and development, and large-scale web crawling.

KIM HUAT GOH (Member, IEEE) received the Ph.D. degree in business administration with a specialization in economics and information systems from the Carlson School of Management, University of Minnesota, Twin Cities. He is currently a Professor of information systems with the Nanyang Business School (NBS), Nanyang Technological University, Singapore. He is also the Associate Dean (Graduate Studies) of the Nanyang Business School. He was with various organizations, such as the Federal Deposit Insurance Corporation (USA), Khoo Teck Puat Hospital, Ng Teng Fong General Hospital, Tan Tock Seng Hospital, Rockwell Automation, Telenor (Norway), and Singapore Exchange in executive training, research, and business analytics related projects. Since 2018, he has been the lead Principal Investigator for various competitive grants (MOE Tier 2, Social Science Research Council, National Research Foundation) with a total amount exceeding U.S. $3.3 million. He has published in top scientific and Financial Times-ranked journals, such as *Nature Communications*, *MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*, *Journal of Consumer Psychology*, and *Human Resource Management*. His research interests include using analytics and economic theories to model human behavior in healthcare and technology-mediated settings and evaluating health systems implementation. He was formerly the Senior Editor of *Electronic Commerce Research Applications* and the Co-Editor of *Journal of Management Information Systems* (Special Issue). He is a Senior Editor of the *Journal of the Association for Information Systems*.

WAI FONG BOH (Member, IEEE) received the Ph.D. degree from the Tepper School of Business, Carnegie Mellon University. She is currently the President's Chair Professor of information systems with Nanyang Technological University (NTU), Singapore. She is also the Vice President of Lifelong Learning and Alumni Engagement with NTU and an Interim Dean of the Nanyang Business School (NBS). She is also the Director of the Information Management Research Centre, NBS. She is also the Co-Director of the NTU Centre in Computational Technologies for Finance (CCTF). She has published in leading journals, including *Management Science*, *MIS Quarterly*, and *Academy of Management Journal*. She published a book on *Identifying Business Opportunities Through Innovation*. She is a Senior Editor of *MIS Quarterly*. Her research interests include innovation and technology management. She was recently awarded Singapore's Public Administration Medal (Silver). She was an Associate Editor of *Management Science* and ISR. She is currently on or had been previously on the editorial board of multiple leading *Information Systems*.