A Social Data-Driven System for Identifying Estate-related Events and Topics

Wenchuan Mu[1][0009−0007−2395−9731], Menglin Li[1][0000−0002−7890−7636], and Kwan Hui Lim[1][0000−0002−4569−0901]

Singapore University of Technology and Design
{wenchuan_mu, menglin_li, kwanhui_lim}@sutd.edu.sg

Abstract. Social media platforms such as Twitter and Facebook have become deeply embedded in our everyday life, offering a dynamic stream of localized news and personal experiences. The ubiquity of these platforms position them as valuable resources for identifying estate-related issues, especially in the context of growing urban populations. In this work, we present a language model-based system for the detection and classification of estate-related events from social media content. Our system employs a hierarchical classification framework to first filter relevant posts and then categorize them into actionable estate-related topics. Additionally, for posts lacking explicit geotags, we apply a transformer-based geolocation module to infer posting locations at the point-of-interest level. This integrated approach supports timely, data-driven insights for urban management, operational response and situational awareness.

1 Introduction

Over the past two decades, social media platforms such as Twitter/X and Facebook have undergone unprecedented growth, now reaching approximately 82% of the global online population, with nearly 20% of users' online time spent on these platforms [2]. These platforms have evolved into essential channels for real-time information dissemination and discussion, encompassing topics ranging from entertainment and pop culture to more specialized domains such as politics and human rights. In parallel, organizations increasingly leverage social media as a sensing modality to identify and monitor both large-scale events (e.g., natural disasters, public crises) and localized issues (e.g., infrastructure faults, community disturbances).

Despite the value of social media as an information source, its scale and velocity introduce significant challenges, foremost among them being information overload [5]. This impairs the end users' abilities to efficiently filter, retrieve, and contextualize relevant content. In response, a range of social media analytics systems have emerged. Examples include InfoTrace [1], which tracks the lifecycle of social media campaigns; DISCO [4], a framework for explainable disinformation detection; RAPID [8], which enables real-time mining of streaming social media data; Li et al. [7] that presented a framework implementing clustering and temporal identification for event detection; and Rosa et al. [13] that utilizes user

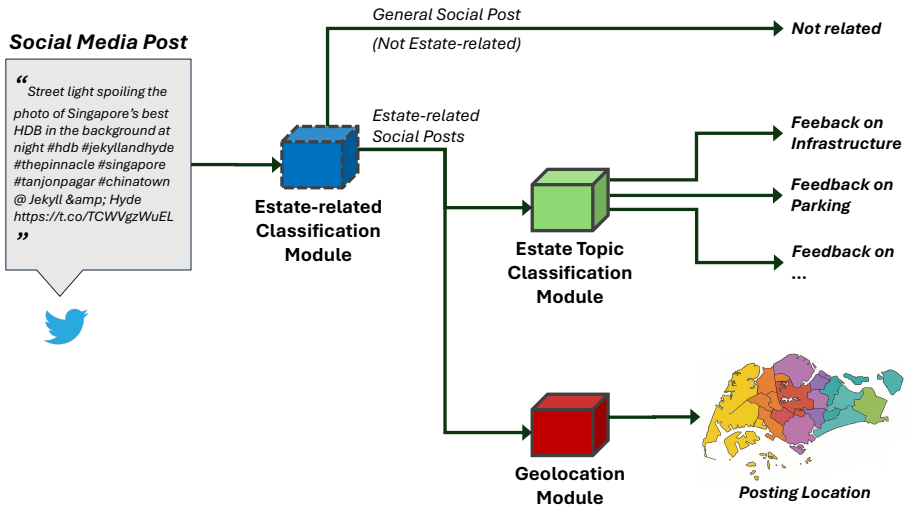

Fig. 1. Model Architecture of Our Proposed System

behaviour changes over time to detect pandemic-related events on social media. While these systems address important facets of social media analysis, a critical gap remains: the automated detection of estate-related events, such as facility breakdowns, noise complaints, or parking violations, within both historical and real-time social media data streams.

To address this gap, we propose a novel system for detecting estate-related events and associated discussion topics from both archival and live social media data. This system is part of the broader Estate-IQ initiative, which aims to automate estate operations and maintenance through AI-driven event detection and decision support.

2 Problem Definition

The ubiquity of smartphones and mobile connectivity has transformed social media platforms into pervasive channels for both real-time news dissemination and the sharing of daily personal experiences. However, the high velocity and volume of user-generated content create significant challenges in filtering and identifying estate-related information amidst the background noise of general posts.

To address these challenges, we formally define two key tasks: *Estate-related Post Detection* and *Estate Topic Classification*. Let $S = \{s_1, \ldots, s_n\}$ denote a collection or stream of n social media posts, where s_n represents the most recent entry. Each post s_i is modeled as a sequence of T tokens:

$$s_i = \{s_i^1, \ldots, s_i^T\}$$

Estate-related Post Detection. We define the estate detection task as a binary classification problem:

$$D_E = (S, L^E)$$

where $L^E = \{L_1^E, \ldots, L_n^E\}$ is the set of binary labels with $L_i^E \in \{0, 1\}$. A label of 1 denotes that s_i contains estate-related content (e.g., facility faults, noise complaints), and 0 indicates otherwise. The objective is to learn a classifier:

$$C_D^E : S \to L^E$$

that accurately maps each social media post to its corresponding estate relevance label.

Estate Topic Classification. For posts identified as estate-related, we define a secondary multi-class classification task:

$$D_T = (S, L^T)$$

where $L^T = \{L_1^T, \ldots, L_n^T\}$ and $L_i^T \in \{0, 1, 2, 3\}$. Each label corresponds to a specific estate-related topic: Infrastructure, Parking, Noise and Others.

The corresponding classifier is defined as:

$$C_D^T : S \to L^T$$

which assigns an appropriate estate topic to each post previously identified as relevant by C_D^E.

The overarching goal is to build a pipeline that first filters estate-related content from the general stream of social media data and then categorizes the filtered posts into actionable topic domains. Together, these models form the backbone of an automated estate event detection system, enabling real-time urban situational awareness.

3 System Architecture

Our proposed system is composed of four core components: (i) **Data Repository/Stream**, (ii) **Estate-related Post Classification**, (iii) **Estate Topic Classification** and (iv) **Social Post Geolocation**. Figure 1 provides an overview of the system architecture and its data flow. In the subsections below, we elaborate on each component in greater detail.

3.1 Data Repository/Stream

As introduced in Section 2, our framework operates over a collection of social media posts, which may originate from either a static repository or a live data stream. Each input to the system consists of a single post in textual form, drawn from this larger collection. The system is designed to process posts in real-time or in batch mode, supporting both retrospective and online analyses. For the purpose of applying this work, personal identifiers from the datasets are anonymized to an anonymous ID that still enables us to identify posts by the same user but is otherwise not mappable to the real-life user.

3.2 Estate-related Post Classification

To detect estate-relevant content within social media posts, we utilize the Bidirectional Encoder Representations from Transformers (BERT) model [3]. BERT is a multi-layer bidirectional Transformer encoder trained using self-supervised objectives such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

For our binary classification task, we fine-tune BERT on a curated dataset of annotated social media posts labeled as estate-related or not. The output of this component is a binary prediction indicating whether a given post pertains to estate-related content. Posts identified as relevant are then passed to the next component for fine-grained topic categorization. Experimental results in the later sections will justify our choice of BERT over other baselines methods.

3.3 Estate Topic Classification

Posts classified as estate-related undergo further classification into specific topical categories. This component also uses the BERT model, fine-tuned on a proprietary dataset comprising maintenance reports submitted by estate residents [11]. While the original dataset contains 13 distinct categories, empirical analysis shows that over 94.8% of reports fall into three major categories: *Infrastructure*, *Parking*, and *Noise*.

Accordingly, we consolidate the remaining 10 less frequent categories into a single class labeled *Others*, resulting in a four-class classification task. This setup enables a balance between model simplicity and topic coverage, ensuring robust classification across the most prevalent estate-related concerns.

3.4 Social Post Geolocation

For estate-related posts lacking explicit geotags, we apply an additional geolocation module to infer the likely posting location. We adopt the transformer-based framework, transTagger, proposed in our prior work [6], which builds upon pretrained language models and integrates both textual and non-textual cues for fine-grained location inference at the POI level. The framework categorizes inputs by data modality, applies an optimal feature fusion strategy, and incorporates a hierarchical temporal encoding scheme. A concatenated representation of positional embeddings is used to better capture fine-grained spatiotemporal context. This component enables spatial contextualization of posts, supporting downstream applications like estate monitoring and geospatial event mapping.

4 Experimental Results

We conducted a preliminary evaluation of the core components of our system using a Twitter/X dataset and a proprietary estate incident dataset, both collected from the same geographical region.

Table 1 presents the performance of various models for detecting estate-related posts. We evaluated five methods: Logistic Regression (LogReg), Recurrent Neural Networks with GloVe embeddings (RNN+GloVe)[12], RNN with Word2Vec embeddings (RNN+W2V)[10], BERT, and MPNet [14]. Among these, BERT achieved the highest accuracy and F1-score, establishing it as the most effective architecture for this task. As such, we adopt BERT as the classification backbone in our deployed system.

Table 1. Estate-related Post Classification

Model	Accuracy	F1-score
LogReg	0.810	0.300
RNN+GloVe	0.500	0.375
RNN+W2V	0.300	0.462
BERT	0.950	0.800
MPNet	0.850	0.595

Table 2. Estate Topic Classification

Estate Topic	Accuracy	F1-score
Infrastructure	0.940	0.877
Parking	0.985	0.971
Noise	0.837	0.829
Others	0.188	0.297
Weighted Avg	**0.882**	**0.865**

For the downstream task of estate topic classification, we employed BERT and evaluated performance across four topical categories: Infrastructure, Parking, Noise, and Others. As shown in Table 2, BERT yielded strong performance for the first three categories, with both Accuracy and F1-score exceeding 80%. Although performance for the Others category was comparatively lower, this class constitutes only 5.2% of the dataset and contributes minimally to the overall performance, as reflected in the weighted average scores.

We also evaluated the geolocation module using results from our previous work, transTagger [6]. On the Singapore Twitter/X dataset, transTagger achieved an accuracy of 0.691 and a mean geotagging distance error of 2.21 km. While the accuracy may appear moderate, the geolocation task is inherently challenging due to the presence of 9,666 distinct Points-of-Interest (POIs). To mitigate privacy risks, we employ a variant of transTagger that performs geotagging at the neighbourhood granularity rather than at the POI level.

5 Conclusion and Future Work

In this paper, we presented a social data-driven platform that leverages a pretrained language model to perform hierarchical classification of social media content. Our system first detects estate-related posts and subsequently categorizes them into specific topics of interest. This framework is designed to address the growing challenge of information overload on social media by automatically surfacing relevant, actionable content for estate management and urban operations. By prioritizing such posts, the system facilitates timely interventions and supports more efficient decision-making in densely populated environments. As the system relies on geotagged data and geolocation models, future work will focus on systematically examining the implications of privacy risks and representational bias in both classification and geospatial inference tasks [15, 9].

Acknowledgment. This research/project is supported by the National Research Foundation (NRF), Singapore, and Ministry of National Development (MND), Singapore under its Cities of Tomorrow R&D Programme (CoT Award COT-V2-2020-1). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of NRF and MND. K. H. Lim is also supported by a MOE AcRF Tier 2 (MOE-T2EP20123-0015). The authors would also like to thank Xjera Labs for collaborating on this project.

References

1. Cheng, L.X.W., Chin, D.W.K., Toh, S.S., Mu, W., Ong, J.K.L., Choo, K.T.W., Lee, R.K.W., Lim, K.H.: InfoTrace: A System for Information Campaign Source Tracing and Analysis on Social Media. In: Proceedings of the 2024 ACM/IEEE JCDL'24 (2024)
2. ComScore: It's a social world: Top 10 need-to-knows about social networking and where it's headed. Internet (Dec 2011)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 NAACL (Jun 2019)
4. Fu, D., Ban, Y., Tong, H., Maciejewski, R., He, J.: Disco: Comprehensive and explainable disinformation detection. In: Proceedings of the 31st ACM CIKM (2022)
5. Fu, S., Li, H., Liu, Y., Pirkkalainen, H., Salo, M.: Social media overload, exhaustion, and use discontinuance: Examining the effects of information overload, system feature overload, and social overload. Information Processing & Management **57**(6), 102307 (2020)
6. Li, M., Lim, K.H., Guo, T., Liu, J.: A transformer-based framework for poi-level social post geolocation. In: Proceedings of ECIR. pp. 588–604. Springer (2023)
7. Li, Q., Nourbakhsh, A., Shah, S., Liu, X.: Real-time novel event detection from social media. In: Proceedings of ICDE. pp. 1129–1139 (2017)
8. Lim, K.H., Jayasekara, S., Karunasekera, S., Harwood, A., Falzon, L., Dunn, J., Burgess, G.: RAPID: Real-time Analytics Platform for Interactive Data Mining. In: Proceedings of the 2018 ECML-PKDD (Sep 2018)
9. Malik, M., Lamba, H., Nakos, C., Pfeffer, J.: Population bias in geotagged tweets. In: Proceedings of ICWSM. pp. 18–27 (2015)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. NeurIPS (2013)
11. Mu, W., Lim, K.H.: Label-Free Topic-Focused Summarization Using Query Augmentation. In: Proceedings of the 2024 IJCNN (Jun 2024)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of EMNLP. pp. 1532–1543 (2014)
13. Rosa, R.L., De Silva, M.J., Silva, D.H., Ayub, M.S., Carrillo, D., Nardelli, P.H., Rodriguez, D.Z.: Event detection system based on user behavior changes in online social networks: Case of the covid-19 pandemic. IEEE Access **8**, 158806–158825 (2020)
14. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Proceedings of NeurIPS (2020)
15. Yang, J., Chakrabarti, A., Vorobeychik, Y.: Protecting geolocation privacy of photo collections. In: Proceedings of AAAI. pp. 524–531 (2020)