buildings

MDPI

Article

Road-Related Information Mining from Social Media Data: A Joint Relation Extraction and Entity Recognition Approach

Lei Yu [1,2] and Dezhi Li [2,*]

[1] Nanjing Sucheng Real Estate Group, Nanjing 211199, China
[2] School of Civil Engineering, Southeast University, Nanjing 211189, China
* Correspondence: njldz@seu.edu.cn

Abstract: Social media data have been gradually regarded as a prospective social sensor in the transportation domain for capturing road conditions. Most existing social media data-based sensors (SMDbSs) of road conditions, however, rely heavily on lexicon-based methods for information extraction and provide coarse-grained location information. Hence, this work newly devises an SMDbS based on joint relation extraction and entity recognition for sensing road conditions from social media data, which eliminates the reliance on lexicon-based methods and offers finer-grained location information in comparison with existing SMDbSs. This SMDbS development consists of four major steps, including data collection and annotation, data cleansing, two-stage information extraction, and model verification. A tweet dataset in Lexington city is exploited to demonstrate this SMDbS, which shows satisfactory information extraction performance. This study would help facilitate social media data to be an extra information source in the transportation domain.

Keywords: social media data; relation extraction; entity recognition; location granularity

Citation: Yu, L.; Li, D. Road-Related Information Mining from Social Media Data: A Joint Relation Extraction and Entity Recognition Approach. *Buildings* **2023**, *13*, 104. https://doi.org/10.3390/buildings13010104

Academic Editors: Tao Wang, Jian Zuo, Hanliang Fu and Zezhou Wu

Received: 7 December 2022
Revised: 26 December 2022
Accepted: 28 December 2022
Published: 31 December 2022

1. Introduction

Conventional road sensors (e.g., inductive loops, lidars, and video surveillance) are fundamental to intelligent transportation systems (ITS), which facilitate the acquisition of road information. However, these physical sensors are of high cost in installation and maintenance; they are ordinarily placed in a few fixed locations along, under, or above the roads [1–4]. Social media could be complementary to road sensors, as they can provide diverse types of information beyond physical sensors. Social media platforms (e.g., Twitter) also have access to billions of users and offer crowd-sourced data [1–4]. In comparison with conventional road sensors that consume huge costs and are placed at limited locations, social media data hold the advantages of low cost and high coverage and can also be exploited for historical analysis [2,5,6]. With the development of social media, various scholars have exploited social media data to identify traffic congestion [7], detect the occurrence of incidents [8], and recognize accident locations [1,5]. T social media data, thus, are regarded as a viable and prospective information source for sensing road conditions in cities [2,5,6,8].

Although scholars in the transportation domain have invested considerable resources to mine social media data, most existing social media data-based sensors (SMDbSs) of road conditions (1) rely heavily on lexicon-based methods for information extraction and (2) provide coarse-grained location information [1,2,4–6,8–11]. Specifically, existing SMDbSs predefine certain lexicons (e.g., dictionaries of counties, cities, highways, roads, incidents, events, or road statuses) and then exploit these lexicons to extract the desired entities. Such lexicon-based methods suffer from low recall rates [12], as the predefined lexicons cannot deal with out-of-vocabulary (OOV) entities [12]. Another drawback of current SMDbSs stems from locating the captured incidents, congestions, or anomalies at coarse-grained locations of counties, cities, blocks, or roads [2,5,13,14]. The road-level location is the finest

granularity that existing SMDbSs achieve, but one road may have dozens of segments and multiple lanes [10]. The road-level granularity cannot satisfy the requirements of sensing road conditions, as the road segment or lane ordinarily is the basic unit of transportation management [4]. Hence, the existing studies of SMDbS are facing bottlenecks concerning the lexicon-based information extraction methods and coarse-grained locations.

In response to the existing SMDbSs that exploit a lexicon-based information extraction approach and provide coarse-grained location information, a new SMDbS based on joint entity recognition and relation extraction is proposed to extract road-related information including fine-grained locations (i.e., segment- or lane-level locations) from social media data. The devised SMDbS consists of the collection of social media data, data cleansing, a two-stage information extraction, and model verification. Its performance will be demonstrated by a tweet dataset from Lexington city. Compared with existing SMDbSs, the novelties of this newly devised SMDbS are (1) getting rid of the lexicon-based approaches for information extraction and (2) improving the granularity of location information. In addition to these two novelties, the SMDbS designed in this study enables more accurate extraction of the incident-related information from the textual content, which also contributes to facilitating social media data to be an additional information source of road conditions.

The remainder of this article is structured as follows. The literature on social media data mining and location information extraction in the transportation domain is reviewed in Section 2. Then, Section 3 presents the development of an SMDbS based on joint relation extraction and entity recognition for extracting road-related information. Subsequently, Section 4 shows the case study and corresponding results. Finally, the contributions of this work, advantages and disadvantages of social media data sources, reuses of the methodology, and further efforts required by SMDbSs are discussed in Section 5.

2. Literature Review

2.1. Social Media Data-Based Sensors of Road Conditions

Social media data have been viewed as a social sensor of road conditions, which attracts excellent efforts and attention from academia and industry [1,4,6,8–11,15–17]. A wide range of SMDbSs has been developed, as shown in Table 1. The extracted entities from social media data include road names, road statuses, road types, road directions, landmarks, and traffic events (Table 1). These road-related entities are recognized by predefined lexicons corresponding to different types of entities (Table 1). Most lexicon-based SMDbSs struggle to deal with the OOV entities and have low transferability to different tasks, which lead to SMDbSs' low recall rates [4,6,9,15]. Most entities can be extracted from the textual content directly, but the implicit entity relations (e.g., which road segment is closed?) need to be inferred from the text. Many of the existing SMDbSs (Table 1) establish the relations between two entities by default when they co-occur in the same piece of social media data [3,18,19]; this may erroneously build relations between non-associative entities.

Different from the lexicon-based information extraction methods (Table 1), the deep learning-based entity recognition and relation extraction approaches could eliminate the reliance on human-made lexicons [20], which can effectively extract entities and entity relations. Although the joint relation extraction and entity recognition approach has not been applied in SMDbSs, it has been widely used in processing the texts in medical, education, and other fields [20], which hold great potential to resolve the challenge of existing SMDbSs (Table 1).

Table 1. Lexicon-based information extraction method adopted by existing SMDbSs.

No.	Reference	Source	Exploited Methods	Extracted Information or Entities
1	[2]	Weibo	• Dictionaries of roads and others	• Accidents • Traffic-related Activities
2	[14]	Tweet	• Dictionaries of desired entities	• Name, type, and direction of roads
3	[13]	Tweet	• Gazetteer	• Road • Landmark
4	[6]	Tweet	• Dictionary of locations	• Location name
5	[10]	Tweet	• Road dictionary	• Road
6	[21]	Weibo	• City name dictionary, • Traffic-related word dictionary	• Road-level locations • Traffic event
7	[16]	Tweet	• List of road names	• Road
8	[15]	Tweet	• Highway lexicon	• Highway
9	[6]	Tweet	• Dictionary of street names	• Street
10	[9]	Tweet	• Lexicon of roads • Lexicon of hazard	• Hazard and Road
11	[11]	Tweet	• The combination of regular expressions and dictionaries,	• Road name • Lane status • Direction of road
12	[20]	Tweet	• Toponyms and geographical names	• Location
13	[8]	Tweet	• List of cities and streets	• City and Street

2.2. Location Information Extraction from Social Media Data

Location is one of the most critical pieces of information elicited from social media data (Table 1). The authors specifically review the methods of obtaining location information in existing SMDbSs. Table 2 shows three sources of social media data-provided geoinformation, including user profiles, geotags, and textual contents [22,23]; each of them possesses advantages and disadvantages in terms of granularity and availability [24]. The user-generated profile is a set of structural information (e.g., location field) that is readily obtained, but it only provides macro location information (e.g., county, state, province, and city) [6]. The Geotag of social media data, such as Tweet, provides a rectangle polygon with coordinates of top-left and bottom-right corners instead of a precise location [25], so the locations where users post the microblogs ordinarily are not the place where the incidents happen. It is estimated that the average deviation between the incident locations and the latitude-longitude of geotags is 7.3 miles [6]. Worse is that geotagged posts constitute only a tiny part ranging from 0.42% to 3% of all pieces of social media data [26]. Researchers also try to extract the location information from the social media texts (Tables 1 and 2); they use various predefined dictionaries of nations, states, counties, cities, highways, or roads to recognize the locations [6,12,13,15,16,20,27,28]. This category of lexicon-based methods possesses high precision, but a low recall rate is inflicted when there are OOV locations or any inconsistency (e.g., spelling errors) with predefined terminology lists [29].

Table 2. The granularities of location information in existing SMDbSs.

Sources	Methods	The Granularities of Location Information		
		Nation, State, County, or City	Road	Road Segment or Lane
User profile	Read the location field in the user profile	√	×	×
Geotag	Read the texts of geotags	√	√	×
	Read the coordinates from the tweet API	√	√	×
Textual content	Recognize location entities	√	√	×
	Extract entities and relations in this study	√	√	√

2.3. Research Gaps

Existing SMDbSs (Table 1) rely heavily on lexicon-based methods for mining road-related information including the event locations, and these SMDbSs' performance is significantly hampered by the fixed, entity-specific lexicons. Besides, the location information captured by most current SMDbSs is coarse-grained at the levels of cities, districts, streets, or roads (Table 1). In order to fill deficiencies of existing SMDbSs, it is necessary to get rid of lexicon-based methods and improve the granularity of extracted location information from social media data.

3. Methodology

As shown in Figure 1, an SMDbS is newly devised to mine social media data for extracting road-related information. The collection and annotation of data from social media platforms is conducted first. A data cleansing process is then carried out to remove social media data that is irrelevant to road conditions. Subsequently, a two-stage model of joint relation extraction and entity recognition is developed based on deep learning algorithms (e.g., Bidirectional Encoder Representation from Transformers, Bert). Finally, the developed models are tested by metrics of precision, recall, and F1-score.

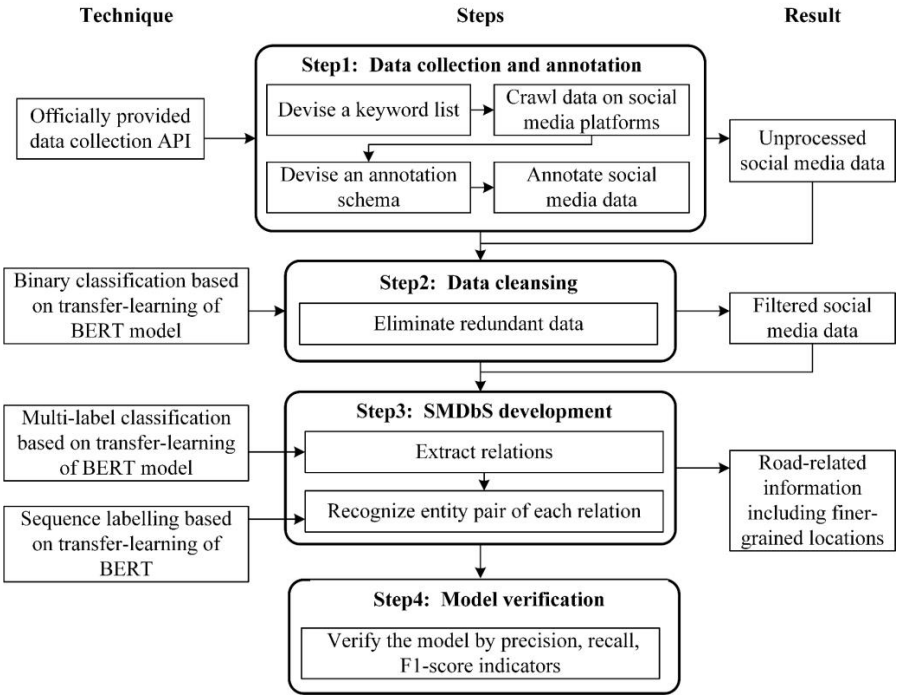

Figure 1. The development of the SMDbS.

3.1. Social Media Data Collection and Annotation

The social media data gathering begins with keyword selection, which refers to the keywords utilized in existing studies (Table 3). The keywords in Table 3 are adopted in this study for retrieving the road-related social media data in a region. Any piece of social media data containing one of these keywords will be automatically collected. After determining the keywords, the API provided by the social media platforms (e.g., Weibo, Twitter, and Facebook) will be adopted to collect the raw data promptly and legally. In this study, a web crawler based on Twitter API is developed to collect data automatically by simulating a browser accessing network resources.

Table 3. The keywords for searching social media data.

Category of Keywords	Keywords	References
Road-related keywords	road, rd, way, street, st, avenue, ave, boulevard, blvd, lane, ln, drive, dr, terrace, ter, place, pl, court, ct, fwy, freeway, alley, aly, boulevard, loop, circle, pass, ramp, pike, pkwy	[1,2,30]
Consequence-related keywords	shutdown, close, incident, accident, crash	[10,15]
Vehicle-related keywords	vehicle, car, bus, vehicular, traffic	[5,11]

After collecting the social media data, ten experienced annotators are invited to do the annotation of the collected data [31]. Each piece of social media data will be labeled with "Y" or "N" for the data cleansing in Section 3.2; the "Y"/"N" means it is relevant or irrelevant to the road conditions.

Besides marking social media data being related or unrelated to the road conditions, an annotation schema (Figure 2) of target information (e.g., name, location, status, and direction of the road) should also be designed to guide the annotation for developing the joint relation extraction and entity recognition model in Section 3.3.

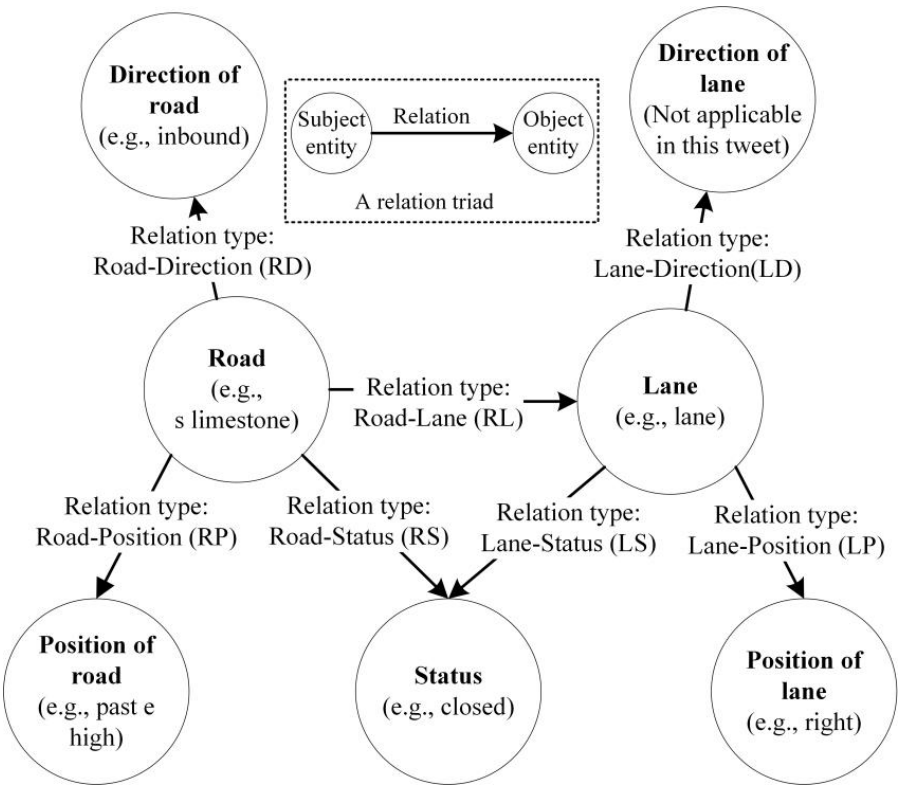

Figure 2. Annotation schema of desired information extracted from social media data.

The annotation schema consists of multiple "subject entity-relation-object entity" triads, which are the derivatives of commonly used "subject-predicate-object" triads. The items involved in the annotation schema hinge on desired information. For instance, the desired information (e.g., status of road, direction of road, and location of lane) in this study is linked and organized through seven categories of triads (Figure 2). Not every piece of social media data contains all seven triads at the same time (Figure 2). For example, the decomposed tweet in Figure 2 contains six triads without the information of lane direction; one of the triads (Figure 2) is [subject entity: s limestone, relation type: RP, object entity: past e high].

3.2. Social Media Data Cleansing

Due to the fact that irrelevant social media data may be brought by the keywords-based retrieval, a data cleansing process is performed to eliminate redundant data and improve data relevance (Figure 1). In this section, a binary classification model is developed based on the transfer-learning of Bert to filter raw social media data. The data cleansing model is comprised of two parts, as shown in Figure 3. The transferred Bert embedding layer is the first part. In this part, the raw social media data without all emojis, links, and extra spaces are used as input. The numerical representation of one piece of social media data is achieved by overlaying three layers of embedding information on the input word-token sequence: token embedding, positional embedding, and segmental embedding (Figure 3). The input embedding layer is then processed by the 12 transformer layers and refined into the hidden layer [32].

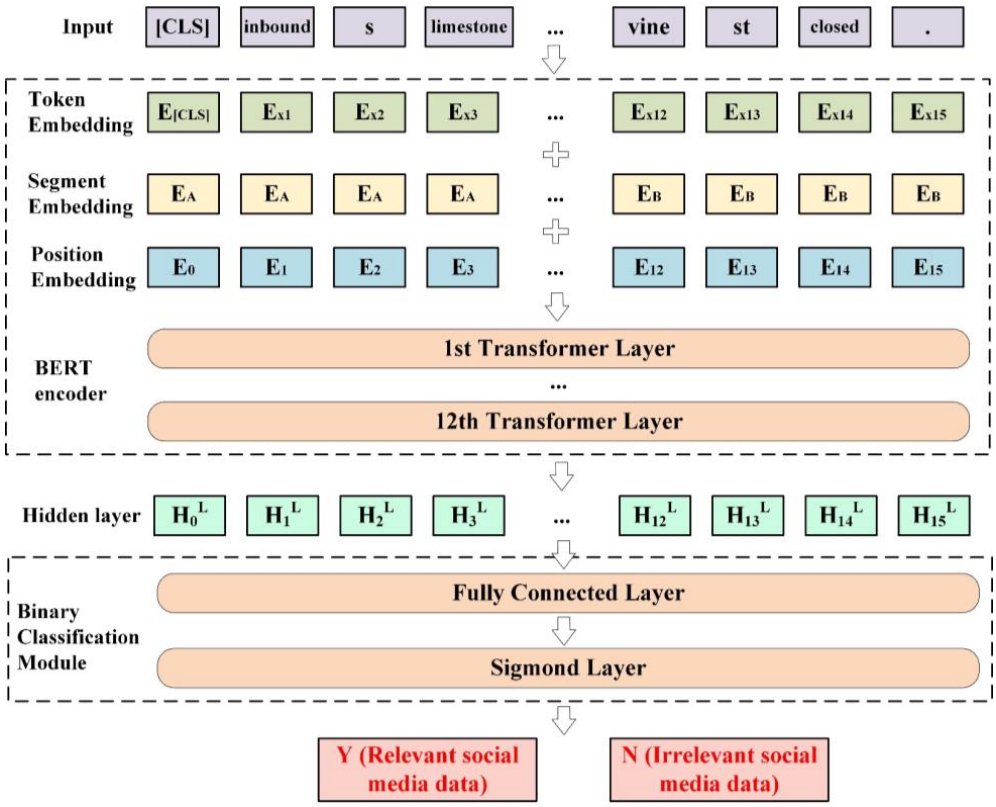

Figure 3. Binary classification model for cleansing social media data.

The second part (Figure 3) is the module of binary classification, which consists of a full-connected and a sigmoid layer. In this part, the output from the hidden layer is transformed into a real-valued vector through the fully connected layer, whose length is the same as the number of text class labels. The activation function of the sigmoid layer is then used to predict the final label [33]. The final label is either "Y" or "N" (Figure 3), which

implies that the piece of social data is relevant or irrelevant to the road conditions. This study develops this classification model by integrating TensorFlow, Keras, and Natural Language Toolkit (NLTK) in Python language. The codes of all algorithms developed in this study are available in the supplemental materials.

3.3. Joint Relation Extraction and Entity Recognition Model for Extracting Road-Related Information

The two-stage model (Figure 4) is constructed to identify the relation types in the social media data and recognize the subject entity and object entity corresponding to each relation. Transfer learning is also adopted for developing the two-stage model, and the base model (i.e., Bert) is extensively used in academia and industry [33,34]. The pre-trained Bert can be reutilized in a range of different downstream text mining tasks, and it merely needs to be fine-tuned according to the particular task, reducing training time and resource consumption considerably [32]. The relation extraction and entity recognition sub-models in the two stages share a similar transfer-learning and fine-tuning process. They both (1) inherit the encoder part of pre-trained Bert, (2) add the tailor-made multi-label classification or sequence labeling modules, (3) split the annotated data into the training, validation, and test sets in the ratio of 8:1:1, (4) train the sub-models, and (5) test their performance.

3.3.1. Relation Extraction in Social Media Data

Both the relation extraction sub-model (Figure 4) and the data cleansing model (Figure 3) are essentially classification tasks. The difference is that the data cleansing model (Figure 3) is a binary classification task, while relation extraction (Figure 4) is a multi-label classification task. Therefore, the same structural design is used for both models in the classification module after the Bert encoder (Figure 4). The Bert-based relation extraction model is also composed of two parts. In the first part, the input is a piece of social media data related to road conditions, where the first token of each text sequence is always a special classification embedding ([CLS]), and each remaining token represents a word. The input embedding layer converts each word in the text into a vector representation [32]. In the second part, the multi-label classification module is developed to infer the different types of relations that each piece of social media data contains. For example, a total of six relations (i.e., "RD", "RP", "RS", "RL", "LD", and "LS") are identified by the relation extraction model in the example tweet (Figure 4).

3.3.2. Entity Recognition for Each Extracted Relation

Recognizing the entity pair (i.e., subject and object entities) of each relation essentially could be viewed as a sequence labeling task. Sequence labeling is the task of marking each token in a one-dimensional linear input sequence. In English text, the word is the smallest unit, and each word will be marked with a one-to-one label. These labels include B_SE, I_SE, B_OE, I_OE, O, and [relation], where "B" implies that the word is the beginning word of the subject entity (SE) or object entity (OE), "I" means that the word is the intermediate word of the entity; and "O" refers to the idea that the word is not in the entity.

In the first part of the sub-model (Figure 4), the input is a sequence of words consisting of the raw data and one extracted relation (e.g., RS in Figure 4). The raw textual content is segmented from the [relation] by a [SEP] symbol, and two different text vectors are appended for differentiation. In the second part of the sub-model, the sequence labeling module of the entity pair recognition model consists of the fully connected and softmax layers (Figure 4). In the end, the entity recognition model (Figure 4) outputs the relation and corresponding entity pair. For example, by inputting a sequence of the tweet and "RP" relation, the model recognizes the subject "limestone" as the subject entity, identifies the "past e high" as the object entity, and obtains the required triad [s limestone, RP, past e high].

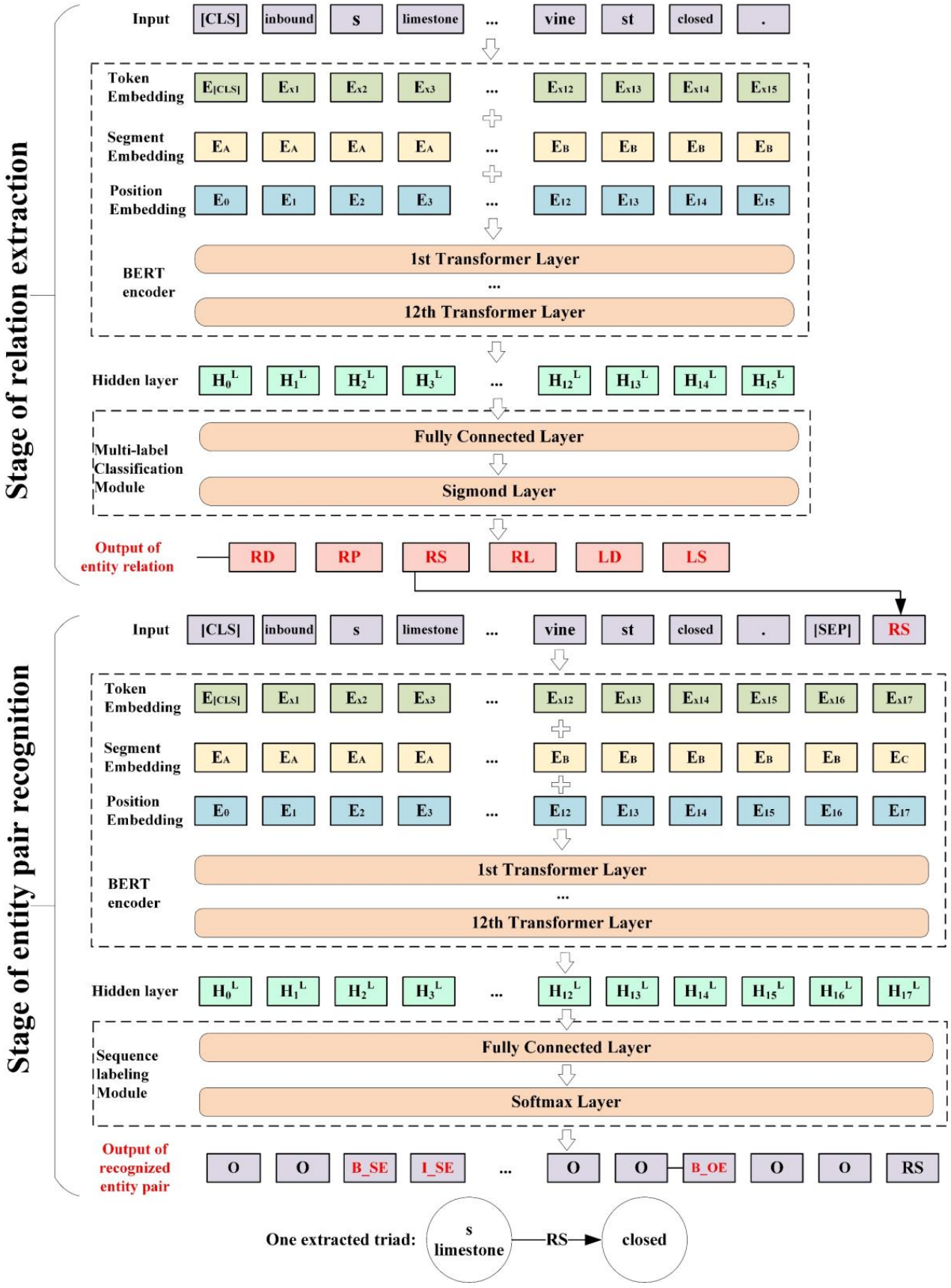

Figure 4. Joint relation extraction and entity recognition model developed in the SMDbS.

3.4. Model Verification

In this paper, precision, recall, and F1-score are exploited to judge the effectiveness of all developed models (Figures 3 and 4). Recall is the percentage of correctly predicted results against the number of results that should return (Equation (1)). Precision is the percentage of the number of correctly predicted results against the number of all returned results (Equation (2)). F1-score is calculated by the harmonic average of the precision and recall (Equation (3)). If the results of the developed model are unsatisfactory, the quality of the data annotation should be checked, and the model structures (Figures 3 and 4) should be adjusted.

$$Recall = \frac{TP\ (True\ Positive)}{TP\ (True\ Positive) + FN\ (False\ Negative)} \quad (1)$$

$$Precision = \frac{TP\ (True\ Positive)}{TP\ (True\ Positive) + FP(False\ Negative)} \quad (2)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

There is no one-size-fits-all standard for determining the base algorithm of transfer learning. In addition to Bert (Figures 3 and 4), the transfer-learning models based on optional FastText, TextCNN, TextRNN, and TextRCNN algorithms are also developed, tested, and compared. The codes of all algorithms utilized in this study are available in the supplemental materials.

4. A Demonstrative Case

To demonstrate the devised methodology, 11,066 pieces of road-related tweets ranging from January 2014 to January 2021 in Lexington, Kentucky were collected and cleansed from Twitter. The collected tweet data are annotated in the manner of Figure 2, and they are available in Table S1. With the process illustrated in Section 3.3, the devised joint relation extraction and entity recognition model (Figure 4) in the SMDbS are trained and tested for extracting road-related information, including the event locations from tweets.

4.1. Performance of the Joint Relation Extraction and Entity Recognition Model

Figure 5 shows the satisfactory performance of the developed data cleansing (Figure 5a), relation extraction (Figure 5b), and entity recognition models (Figure 5c) through precision, recall, and F1-score indicators. The transfer-learning-based SMDbS (Figures 3 and 4) devised in this work show the best performance. The F1-scores of all sub-models based on the transfer-learning of Bert (red columns in Figure 5a–c) are greater than 90%. The selected Bert algorithm also performs better than other widely used algorithms of FastText, TextCNN, TextRNN, and TextRCNN. These optional algorithms are well-performed in data cleansing (Figure 5a) and relation extraction (Figure 5b), while their performance significantly drops down in entity pair recognition (Figure 5c). The codes of all algorithms are available in the supplemental materials.

4.2. Granularity Improvement of the Location Information Extracted by SMDbS

The comparison of the location granularities between existing and newly developed SMDbSs is displayed in Figure 6, which shows the captured and processed tweets in downtown blocks of Lexington. Once a road-related event (Figure 7) is extracted, most existing SMDbSs (Table 1) could only locate the events at the road level (The left subfigures in Figure 6a–c), while the SMDbSs developed in this study could locate the same events at the segment or lane (The right-side subfigures in Figure 6a–c). All road segments and lanes in one road (e.g., "s upper st" in Figure 6a, "e high st" in Figure 6b, and "w wine st" in Figure 6c) may be reckoned to be affected by existing SMDbSs, while an event ordinarily affects one road segment or lane (Figure 7). Existing SMDbSs entail the risks of exaggerating the number of incidents that occur on one segment or lane, and certain examples are shown

in Table 4. The newly developed SMDbS significantly increase the granularity of extracted location information (Figure 6).

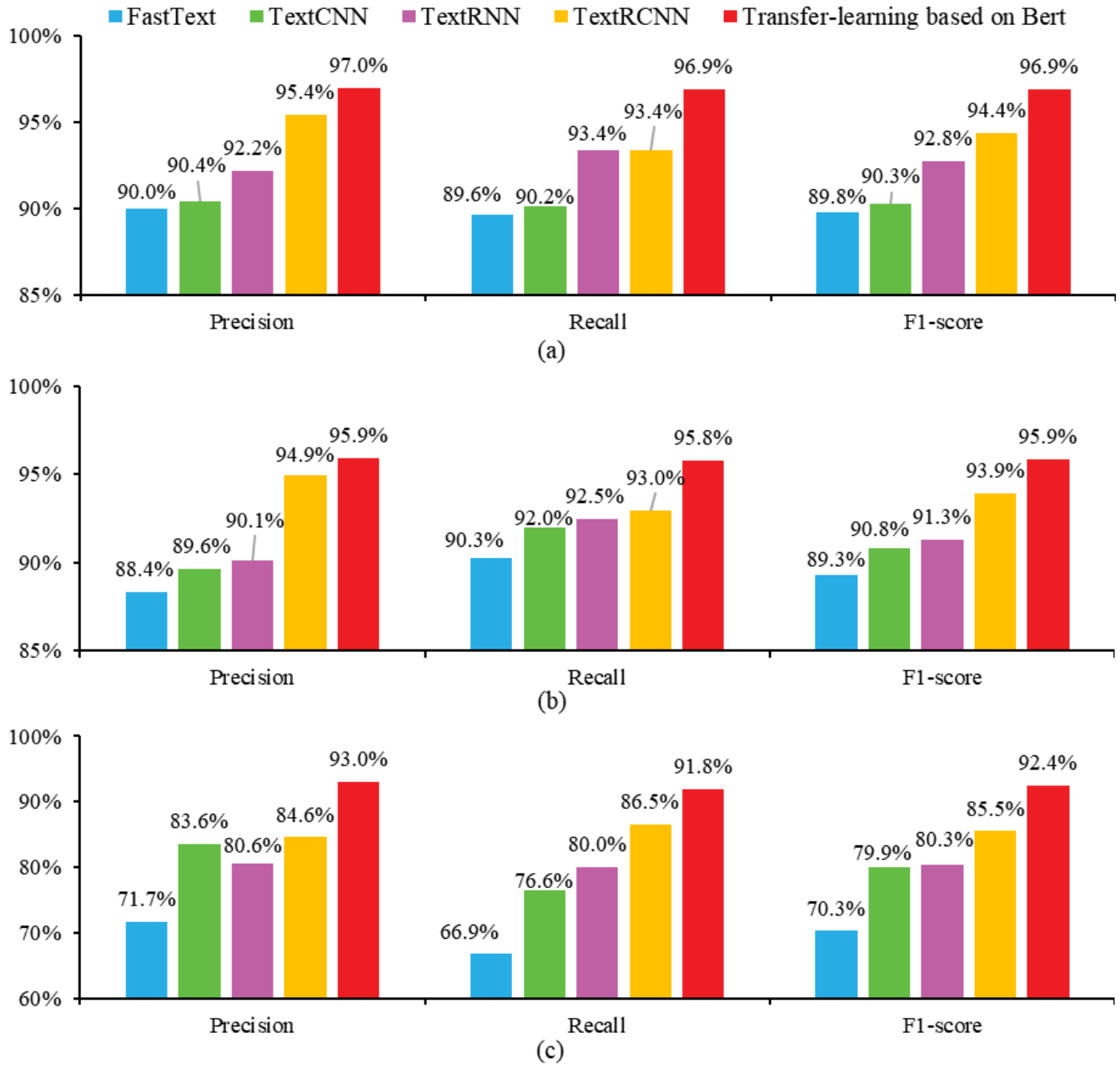

Figure 5. Precision, recall, and F1-score of (**a**) data cleansing model, (**b**) relation extraction, and (**c**) entity recognition model.

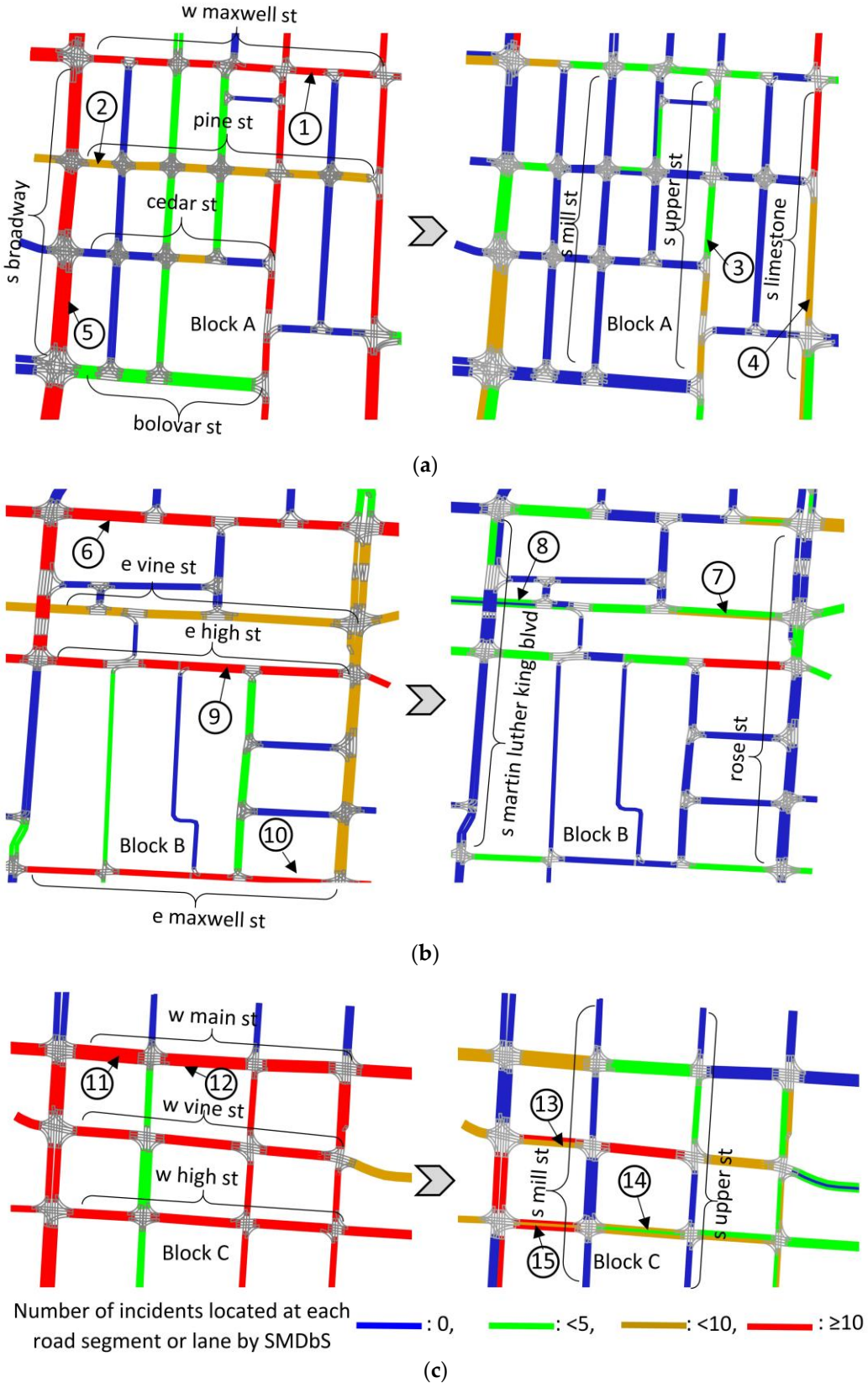

Figure 6. Granularity comparison between existing SMDbSs and the newly proposed SMDbS in three downtown blocks.

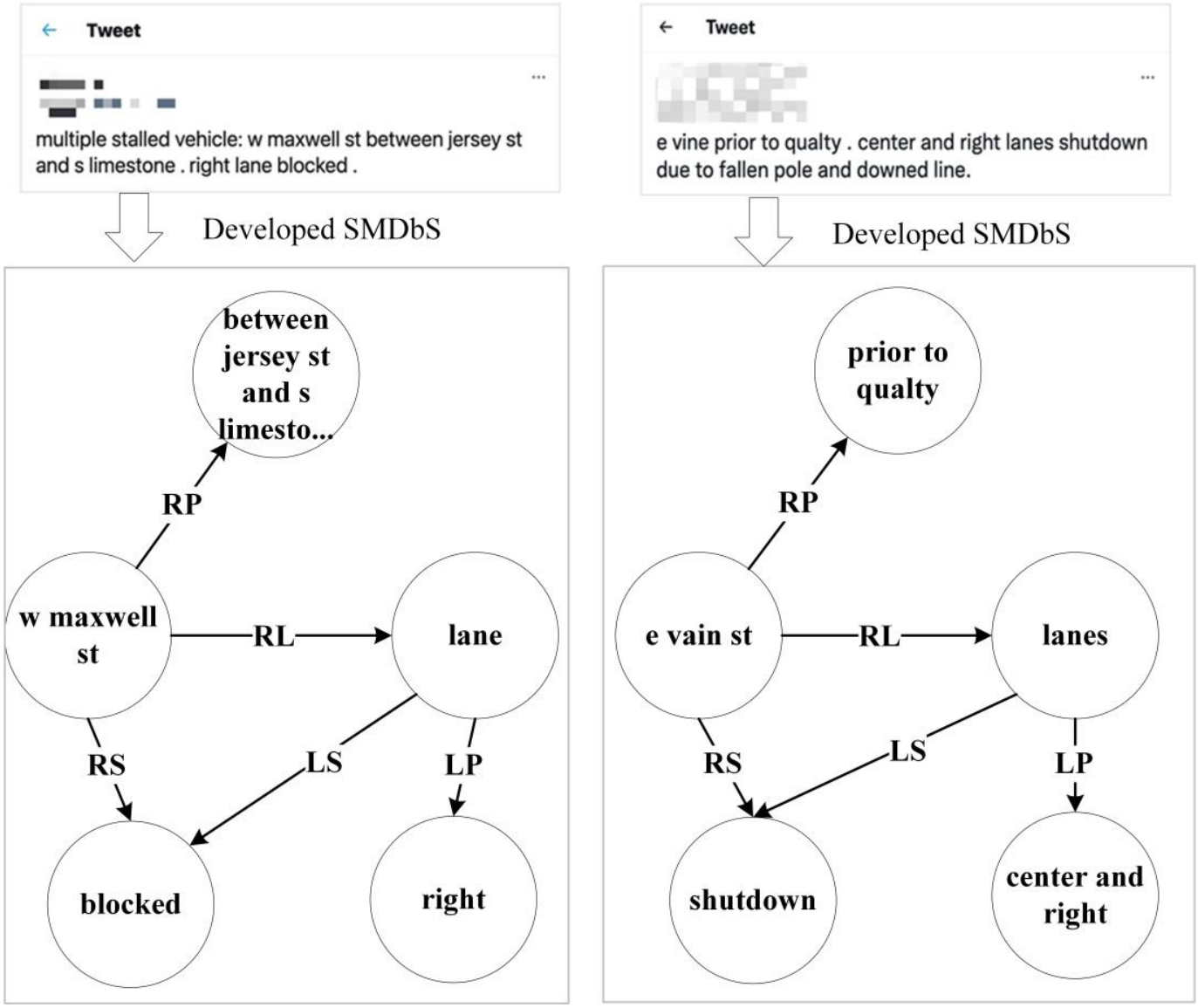

Figure 7. Examples of processed tweets by the SMDbS.

All case tweets in the downtown blocks (Figure 6) are available in Table S2, and two specific examples show the process from raw tweets to extracted information (Figure 7). There are no lexicons recording these entities (e.g., "w maxwell street", "blocked", and "center and right" in Figure 7), while the developed SMDbS possesses the capacity to extract the desired information intelligently. It is recognized that many events are unreported by social media data, so Figure 6 just aims to show the devised SMDbS's capacity of extracting finer-grained location instead of concluding the spatial distribution of traffic incidents in Lexington City.

Table 4. Comparison of the number of incidents extracted by the existing SMDbS and the newly devised SMDbS.

Codes in Figure 6	Examples of Segment or Lane *	Number of Incidents that Occur on this Segment or Lane	
		Existing SMDbS (Table 1)	The Newly Devised SMDbS
No.1	segment of w maxwell st between s upper st and jersey st	13	2
No.2	segment of pine st between s broadway and plunkett st	7	2
No.3	segment of s upper st between pine st and cedar st	14	3
No.4	segment of s limestone between pine st and winslow st	16	6
No.5	segment of broadway between pine st and cedar st	19	7
No.6	segment of e main st between esplanade and s martin luther king road	11	3
No.7	central lane of e vine st between quality st and rose st	9	1
No.8	leftmost lane of e vine st between s martin luther king road and beck alley	9	2
No.9	segment of e high st between hagerman ct and stone ave	15	4
No.10	segment of e maxwell st between stone ave and rose st	12	4
No.11	segment of w main st between s broadway and s mill st	13	8
No.12	segment of w main st between s mill st and s upper st	13	3
No.13	right lane of w vine st between s broadway and s mill st	15	5
No.14	central lane of w high st between s mill st and s upper st	26	3
No.15	central lane of w high st between s broadway and s mill st	26	6

* More tweets showing incidents on different segments or lanes are provided in Table S2.

5. Discussion

5.1. Contributions

This work contributes to devising a new SMDbS based on joint relation extraction and entity recognition for extracting road-related information from social media data, which facilitates social media data to be an additional sensor in the transportation domain for capturing road conditions.

Compared with the current SMDbSs enumerated in Table 1, this devised SMDbS (1) gets rid of lexicon-based methods (Figure 4) and (2) improves the granularity of extracted location information (Figure 6). The majority of existing SMDbSs mine road information from social media data by predefined dictionaries [1,4,6,8–11,15,16]; however, such manipulations often fail to adapt to the diversity of user-generated social media data [4,35]. The newly devised SMDbS exploits a two-stage model consisting of relation extraction and entity recognition (Figure 4) for information extraction, as well as affirming the efficacy of Bert-based transfer learning (Figure 5). As the positions and directions of specific road segments and lanes can be extracted more accurately and efficiently (Figure 7), this SMDbS can achieve finer-grained extraction of location information than existing SMDbSs (Table 1). In addition, the authors provide an open standard hand-annotated Twitter dataset (Table S1) for the SMDbS research area, which may help promote social media data-related research in the transportation domain [4,36].

5.2. Advantages and Disadvantages of Social Media Data

In this paper, social media data are adopted to extract road information. Compared with road sensors in intelligent transportation systems, social media data hold the advantage of reflecting more types of information and being easily accessible. As a complement to road sensors in intelligent transportation systems, social media data can also be adapted for various analyses [4,35]. For example, social media data holds the potential to be exploited for historical analysis (e.g., hot spots of traffic incidents), since it covers transportation data during a certain period of time in the past [1,3]. In addition, the situations of multiple cities located in different countries can be compared based on social media data, as they are accessible through social media APIs, which are easier to access than conventional physical sensors (e.g., inductive loops) in ITS.

Both social media data and commercial maps (e.g., Google Maps) provide road conditions, but the two types of data sources still have differences. (1) Social media data provide diverse types of road-related information in Figure 2; they also hold the potential to provide more information (e.g., the types of occurred crashes, the affected stakeholders, and the reasons for incidents) [11]. The data provided by Google Maps focus on a limited number of types of traffic information (e.g., congestion). (2) Compared with the traffic data acquired from Google Maps, social media data is easier to access, and the analysis process (e.g., the SMDbS in this work) is also freely available to the public. The traffic information from Google Maps has certain permission restrictions, and the data analysis process of commercial maps is a black box, which is not conducive to scholars doing further research related to traffic data and road conditions analysis.

However, social media data also have some disadvantages, which make them only a supplemental data source for the transportation domain. It is difficult to achieve real-time monitoring of road conditions, as the time of posting the microblogs may not be the time of accidents [5]. Also, the locations where users post the microblogs may not be the place where the incidents happen [4,23]; and the users also may lack the capacity to accurately and concisely present the incident locations [4]. What's more, it is admitted that many events currently are unreported by social media data. Therefore, the purpose of this paper is to propose an approach to extract road conditions from social media data possessing the desired information, which could promote social media data to be a source of data rather than achieving dynamic monitoring of road conditions based on social media data.

5.3. Reuses of the Methodology

The proposed methodology (Figure 1) is designed to be modular and transferable; it may be reused to extract more types of road-related information from user-generated social media data in different cities. As there is no model fitting any social media data across disparate scenarios and tasks [33], the reuse of the methodology requires some modifications regarding gathering social media data in particular cities, designing a task-specific annotation schema (Figure 2), and training the two-stage joint relation extraction and entity recognition model anew (Figure 4). When reusing the methodology proposed in this paper to different cities or tasks, it is important to collect city-specific raw data for making the models learn local language characteristics [12]. If the user-generated social media data involve more desired information items, more triads could be added to the annotation schema, such as [event, event-cause, cause], [road, road-user, user], and [road, road-type, road type]. Although the transfer-learning based on Bert shows satisfactory performance (Figure 5), the advanced algorithms of relation extraction and entity recognition in the future could still be exploited in the two-stage model.

5.4. Further Efforts

This study still has limitations and requires more improvements. Firstly, due to the word limit of many social media platforms, one piece of social media data is unable to offer comprehensive information concerning road conditions [37]. Although more individuals are willing to share their observations through social media, the majority of them might

not have the ability to accurately and succinctly characterize essential elements of observed road conditions. Second, if the suggested SMDbS could be applied to other tasks, situations, and cities, further tests on a larger number of cases would help more thoroughly validate this devised SMDbS. Last but not least, with the advancement of text-mining techniques and technologies, the base algorithms adopted by each sub-model (Figures 3 and 4) in the SMDbS are replaceable.

6. Conclusions

This work proposes an SMDbS based on joint relation extraction and entity recognition to help perceive road conditions from social media data. The development of this SMDbS consists of four major steps, including (1) data collection and annotation, (2) data cleansing, (3) joint relation extraction and entity recognition model development, and (4) model verification. The devised SMDbS has been preliminarily demonstrated by the tweet dataset in Lexington city. Based on the results (Figures 5 and 6), it is affirmed that social media data could be one of the supplemental sources of transportation information. The superiority of the transfer-learning of Bert for SMDbS has also been concluded through the comparison results of different algorithms' performance (Figure 5), which outperformed multiple conventional algorithms.

Compared with existing SMDbSs (Table 1), this newly devised SMDbS relieves the heavy reliance on lexicon-based methods and offers finer-grained location information. This study would help promote social media data to be an additional information source for perceiving real-world road conditions. In terms of information extraction, the SMDbS in this study develops a Bert-based two-stage method for entity recognition and relation extraction, relieving the problems of low information recall and generality of traditional lexicon-based information extraction methods. In terms of information granularity, the location information obtained by most of SMDbSs is coarse-grained (e.g., city-, district-, street- or road-level locations). The SMDbS developed in this study is able to locate traffic events on segments or lanes, improving the granularity of location information extracted by conventional SMDbSs.

The future research directions of SMDbS include different application cities and scenarios, diverse types of information extracted from social media data, and more advanced relation extraction and entity recognition algorithms. (1) The SMDbS can be applied to road information extraction in different cities by collecting and annotating local social media data. (2) In addition to road information in Figure 2, the method can be extended to more information extraction by adapting the annotation schema to the specific application scenarios. (3) Although the transfer-learning based on Bert shows satisfactory performance, future advanced text mining algorithms can be used to improve the performance of the SMDbS.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings13010104/s1, Table S1: Tweet dataset; Table S2: Processed tweets in the downtown.

Author Contributions: Conceptualization, D.L.; methodology, L.Y.; writing—original draft preparation, L.Y. and D.L.; writing—review and editing, L.Y. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Social Science Fund of China (No.19BGL281).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in the Supplementary Materials.

Acknowledgments: The authors would like to acknowledge and thank Ling Mao, Hui Wang, Yongheng Zhao, and Wentao Wang for providing support, assistance, and suggestions regarding coding and model development.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zia, M.; Furle, J.; Ludwig, C.; Lautenbach, S.; Gumbrich, S.; Zipf, A. SocialMedia2Traffic: Derivation of Traffic Information from Social Media Data. *ISPRS Int. Geo-Inf.* **2022**, *11*, 20. [CrossRef]
2. Wang, Y.; He, Z.; Hu, J. Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1132–1144. [CrossRef]
3. Park, J.Y.; Mistur, E.; Kim, D.; Mo, Y.; Hoefer, R. Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs. *Sustain. Cities Soc.* **2022**, *76*, 17. [CrossRef]
4. Xu, S.; Li, S.; Wen, R. Sensing and detecting traffic events using geosocial media data: A review. *Comput. Environ. Urban Syst.* **2018**, *72*, 146–160. [CrossRef]
5. Zayet, T.M.A.; Ismail, M.A.; Varathan, K.D.; Noor, R.M.D.; Chua, H.N.; Lee, A.; Low, Y.C.; Singh, S.K.J. Investigating transportation research based on social media analysis: A systematic mapping review. *Scientometrics* **2021**, *126*, 6383–6421. [CrossRef]
6. Khan, S.M.; Chowdhury, M.; Ngo, L.B.; Apon, A. Multi-class twitter data categorization and geocoding with a novel computing framework. *Cities* **2020**, *96*, 102410. [CrossRef]
7. Chang, H.L.; Li, L.S.; Huang, J.X.; Zhang, Q.P.; Chin, K.S. Tracking traffic congestion and accidents using social media data: A case study of Shanghai. *Accid. Anal. Prev.* **2022**, *169*, 17. [CrossRef]
8. Agarwal, A.; Toshniwal, D. Face off: Travel Habits, Road Conditions and Traffic City Characteristics Bared Using Twitter. *IEEE Access* **2019**, *7*, 66536–66552. [CrossRef]
9. Alhumoud, S. Twitter Analysis for Intelligent Transportation. *Comput. J.* **2019**, *62*, 1547–1556. [CrossRef]
10. Paule, J.D.G.; Sun, Y.; Moshfeghi, Y. On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Inf. Process. Manag.* **2019**, *56*, 1119–1132. [CrossRef]
11. Yao, W.; Qian, S. From Twitter to traffic predictor: Next-day morning traffic prediction using social media data. *Transp. Res. Part C Emerg. Technol.* **2021**, *124*, 102938. [CrossRef]
12. Li, J.; Sun, A.X.; Han, J.L.; Li, C.L. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
13. Milusheva, S.; Marty, R.; Bedoya, G.; Williams, S.; Resor, E.; Legovini, A. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PLoS ONE* **2021**, *16*, e0244317. [CrossRef] [PubMed]
14. Yuan, F.; Liu, R.; Mao, L.; Li, M. Internet of people enabled framework for evaluating performance loss and resilience of urban critical infrastructures. *Saf. Sci.* **2021**, *134*, 105079. [CrossRef]
15. Chen, Y.; Wang, Q.; Ji, W. Rapid Assessment of Disaster Impacts on Highways Using Social Media. *J. Manag. Eng.* **2020**, *36*, 04020068. [CrossRef]
16. Essien, A.; Petrounias, I.; Sampaio, P.; Sampaio, S. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web* **2021**, *24*, 1345–1368. [CrossRef]
17. Chaniotakis, E.; Antoniou, C.; Pereira, F. Mapping Social Media for Transportation Studies. *IEEE Intell. Syst.* **2016**, *31*, 64–70. [CrossRef]
18. Chen, J.V.; Nguyen, T.; Oncheunjit, M. Understanding continuance intention in traffic-related social media Comparing a multi-channel information community and a community-based application. *Internet Res.* **2020**, *30*, 539–573. [CrossRef]
19. Berube, M.; Tang, T.U.; Fortin, F.; Ozalp, S.; Williams, M.L.; Burnap, P. Social media forensics applied to assessment of post-critical incident social reaction: The case of the 2017 Manchester Arena terrorist attack. *Forensic Sci. Int.* **2020**, *313*, 9. [CrossRef] [PubMed]
20. Martínez, N.J.F.; Pascual, C.P. Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets. *Rev. Electron. De Linguist. Apl.* **2020**, *19*, 136–163.
21. Li, X.; Zhou, J.; Pedrycz, W. Linking granular computing, big data and decision making: A case study in urban path planning. *Soft Comput.* **2020**, *24*, 7435–7450. [CrossRef]
22. Rettore, P.H.L.; Santos, B.P.; Lopes, R.R.F.; Maia, G.; Villas, L.A.; Loureiro, A.A.F. Road Data Enrichment Framework Based on Heterogeneous Data Fusion for ITS. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1751–1766. [CrossRef]
23. Tang, J.; Tang, X.Y.; Yuan, J.S. Traffic-optimized Data Placement for Social Media. *IEEE Trans. Multimed.* **2018**, *20*, 1008–1023. [CrossRef]
24. Shen, D.Y.; Zhang, L.F.; Cao, J.P.; Wang, S.Z. Forecasting Citywide Traffic Congestion Based on Social Media. *Wirel. Pers. Commun.* **2018**, *103*, 1037–1057. [CrossRef]
25. Twitter. Filtering Tweets by location. Available online: https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location (accessed on 25 June 2015).
26. Mahmud, J.; Nichols, J.; Drews, C. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol. (TIST)* **2014**, *5*, 1–21. [CrossRef]
27. Salazar-carrillo, J.; Torres-ruiz, M.; Davis, C.A., Jr.; Quintero, R.; Moreno-ibarra, M.; Guzmán, G. Traffic congestion analysis based on a web-gis and data mining of traffic events from twitter. *Sensors* **2021**, *21*, 2964. [CrossRef]
28. Yang, T.; Xie, J.; Li, G.; Mou, N.; Chen, C.; Zhao, J.; Liu, Z.; Lin, Z. Traffic Impact Area Detection and Spatiotemporal Influence Assessment for Disaster Reduction Based on Social Media: A Case Study of the 2018 Beijing Rainstorm. *ISPRS Int. Geo-Inf.* **2020**, *9*, 136. [CrossRef]

29. Goyal, A.; Gupta, V.; Kumar, M. Recent named entity recognition and classification techniques: A systematic review. *Comput. Sci. Rev.* **2018**, *29*, 21–43. [CrossRef]
30. Wang, D.; Al-Rubaie, A.; Clarke, S.S.; Davies, J. Real-Time Traffic Event Detection From Social Media. *ACM Trans. Internet Technol.* **2017**, *18*, 23. [CrossRef]
31. Lu, H.; Zhu, Y.F.; Shi, K.Z.; Lv, Y.S.; Shi, P.F.; Niu, Z.D. Using Adverse Weather Data in Social Media to Assist with City-Level Traffic Situation Awareness and Alerting. *Appl. Sci.* **2018**, *8*, 18. [CrossRef]
32. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
33. Zhang, Z.H.; He, Q.; Gao, J.; Ni, M. A deep learning approach for detecting traffic accidents from social media data. *Transp. Res. Pt. C-Emerg. Technol.* **2018**, *86*, 580–596. [CrossRef]
34. Li, D.W.; Zhang, Y.J.; Li, C. Mining Public Opinion on Transportation Systems Based on Social Media Data. *Sustainability* **2019**, *11*, 15. [CrossRef]
35. Rashid, M.T.; Zhang, D.; Wang, D. DASC: Towards a road Damage-Aware Social-media-driven Car sensing framework for disaster response applications. *Pervasive Mob. Comput.* **2020**, *67*, 23. [CrossRef]
36. Ghandour, A.J.; Hammoud, H.; Dimassi, M.; Krayem, H.; Haydar, J.; Issa, A. Allometric scaling of road accidents using social media crowd-sourced data. *Phys. A Stat. Mech. Its Appl.* **2020**, *545*, 8. [CrossRef]
37. Lock, O.; Pettit, C. Social media as passive geo-participation in transportation planning-how effective are topic modeling & sentiment analysis in comparison with citizen surveys? *Geo-Spatial Inf. Sci.* **2020**, *23*, 275–292.