

SENSING AND DETECTING SMALL-SCALE EVENTS
USING GEOSOCIAL MEDIA DATA

by

Shishuo Xu

B.Sc., China University of Mining and Technology, Xuzhou, China, 2013

A dissertation
presented to Ryerson University

in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the program of
Civil Engineering

Toronto, Ontario, Canada, 2020

© Shishuo Xu 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Sensing and Detecting Small-scale Events using Geosocial Media Data

Doctor of Philosophy 2020

Shishuo Xu

Civil Engineering

Ryerson University

Abstract

Small-scale events involve interactive human movement in limited space and time. Social media platforms possibly generate large amount of geospatially-referenced information related to small-scale events. It benefits individuals, management departments, and urban systems if small-scale events can be timely detected from social media platforms, where measuring the abnormal patterns of human movement to discover events and analyzing associated texts to interpret the reasons behind abnormal movement are two keys. Through investigating how people move as different events occur and measuring the patterns on social media platforms, small-scale events can be generally classified into two types, namely type I events with abrupt patterns and type II events with random occurrence of key factors, where social events and traffic events are representative correspondingly.

Despite many studies have been conducted to detect social events and traffic events using geosocial media data, there still are some un-answered questions requiring further research. Most existing studies did not identify occurring events from a full coverage of spatial, temporal, and semantic perspectives. Studies concerning social event detection lack efficient semantic analysis

summarizing event content to infer the reasons driving the abnormal movement. The typical classification-based method regarding traffic event detection lacks investigation on how the spatiotemporal distribution of traffic relevant posts associate with the occurring traffic events, and simply assigns the detected events with predefined categories, missing events that indicate traffic anomalies but go beyond the predetermined categories.

In this thesis, spatial-temporal-semantic approaches are proposed to measure spatiotemporal patterns of posts and users of social media platforms to capture abnormal human movement, and analyze the content of associated posts to mine the reasons driving the movement. A variety of techniques including machine learning, natural language processing, and spatiotemporal analysis are adopted to realize effective detection. Based on one-year Twitter data collected in Toronto, 2014 Toronto International Film Festival and traffic anomaly detection are selected as two case studies to evaluate the performance of proposed approaches. Through comparing with the ground truth data, the result reveals that more than 80% of the detected events do refer to real-world events, which illustrates the feasibility and efficiency of proposed approaches.

Keywords: Small-scale event, Event detection, Geosocial media data, Traffic event, Social event, Twitter, Spatiotemporal clustering

Acknowledgements

Over the last 2-1/2 years of PhD study at Ryerson, completing this thesis is a big milestone that I would never have achieved without the support and encouragement from many people. While it would be impossible to acknowledge all of them, it is my great pleasure to express my special thanks to the individuals mentioned below.

I have known Dr. Songnian Li since I was a third-year undergraduate student in the Surveying Engineering program in China University of Mining and Technology, Xuzhou, China. Dr. Li had a close cooperation with the advisor of my undergraduate thesis project, and gave me guidance. I was determined to start my PhD with Dr. Li since he is an extremely nice and decent person with whom to share and work with. During my pursuing a PhD at Ryerson, he not only supervised me at a high level to keep my research on the right track but also discussed research problems in detail to train my critical thinking. Despite that, Dr. Li shared everything that he found useful to my research progress and personal development. I deeply appreciate his physical and spiritual support particularly providing me with a very free research environment and supervising me with great patience and encouragement. I would also extend my appreciation to my other two supervisory committee members, Dr. Ahmed Shaker and Dr. Cherie Ding, for their constructive comments on both my PhD research proposal and this dissertation. I appreciate Dr. Ahmed El-Rabbany organizing graduate seminars for us to exchange research ideas per week.

I owe a number of thanks to Dr. Wei Huang and Dr. Wai Yeung Yan who selflessly lent me a hand in life abroad and my academic career. They are more like mentoring teachers who taught me fundamental and necessary skills of doing research (e.g., discussing research ideas, plotting figures,

and using Latex). The experiences we shared exploring the best restaurants together made Toronto life much more colorful. I would like to thank Richard Wen for his efforts in revising my papers regarding language proficiency and paper content. Thanks Shahram Sattar, Siyuan Liu, Aleta Mawuenyegah, Pengpeng Huo, and Danya Qutaishat for their contribution to creating an energetic atmosphere in our lab.

Last but not least, I would like to thank my parents for their continuous support. They create a loving and open environment with thoughtful understanding where I can fully engage myself in the research. I would extend my sincere thanks to my brother for taking care of our parents while I was not at home.

Shishuo Xu

January 22th, 2020

Table of Contents

Abstract	iii
Acknowledgements	v
List of Tables	x
List of Figures	xi
List of Appendices	xiv
Chapter 1 Introduction	1
1.1 Problem statement and motivation	1
1.2 Research objectives	3
1.3 Assumptions, scope, and limitations	4
1.4 Research methodology	7
1.5 Main contributions	8
1.6 Organization of dissertation	11
Chapter 2 Literature review and related work	12
2.1 Geosocial media data	12
2.2 Small-scale events	14
2.3 Detecting social events using geosocial media data	17
2.3.1 Spatiotemporal burst-based method	17
2.3.2 Semantic burst-based method	20
2.4 Detecting traffic events using geosocial media data	22
2.4.1 Query for traffic related tweets	23
2.4.2 Preprocessing queried tweets	26
2.4.3 Identifying tweets relevant to real-world traffic events	27
2.4.4 Location inference and visualization	31
2.5 Discussion	33
2.5.1 Social event detection	33
2.5.2 Traffic event detection	35
Chapter 3 Methodology	37
3.1 The general workflow	37
3.2 A spatial-temporal-semantic approach for detecting social events	39

3.2.1	Preliminary and framework	39
3.2.2	Data acquisition	40
3.2.3	Spatiotemporal outlier detection	42
3.2.4	Content analysis of outlier tweets	46
3.2.5	Clustering outliers for social event detection.....	48
3.3	A classification-based approach for detecting traffic events	51
3.3.1	The overall workflow.....	51
3.3.2	Data collection and preprocessing	52
3.3.3	Mining association rules	53
3.3.4	Classifying traffic events into different categories	55
3.4	A clustering-based approach for detecting traffic events.....	58
3.4.1	The overall workflow.....	59
3.4.2	Spatiotemporal clustering for traffic event identification	60
3.4.3	Event content summarization.....	61
3.5	Discussion	63
3.5.1	Social event detection	63
3.5.2	Traffic event detection	64
Chapter 4	Results and analysis.....	66
4.1	Social event detection based on the spatial-temporal-semantic approach	66
4.1.1	Study area and dataset.....	67
4.1.2	Spatiotemporal outliers	68
4.1.3	Content of outlier tweets	75
4.1.4	Detection of TIFF events	77
4.1.5	Evaluation	81
4.2	Traffic event detection based on the classification methods.....	85
4.2.1	Study area and dataset.....	85
4.2.2	Association rules mined by Apriori algorithm.....	87
4.2.3	Event classification	89
4.2.4	Validation with vehicle travel speed data	92
4.3	Traffic event detection based on the clustering method.....	97
4.3.1	Parameter estimation.....	97

4.3.2	Spatiotemporal distribution of clusters indicating traffic events	99
4.3.3	Content of clusters	101
4.3.4	Validation with vehicle travel speed data	104
4.4	Discussion	106
4.4.1	Social event detection	106
4.4.2	Traffic event detection	110
Chapter 5	Conclusions and future work.....	112
5.1	Conclusions	112
5.2	Future work	114
Appendices.....		116
References		119

List of Tables

Table 2.1. Summary of using the keywords-based or accounts-based methods to query for traffic related tweets	24
Table 2.2. Summary of combining the keywords-based with accounts-based methods to query for traffic related tweets.....	25
Table 2.3. Summary of binary classification or multiple classification on traffic events	28
Table 2.4. Summary of binary classification followed by multiple classification on traffic events	29
Table 2.5. Summary of geocoding techniques.....	31
Table 4.1. Venues of TIFF 2014.....	66
Table 4.2. An example of topic distributions over an outlier in region #417	76
Table 4.3. Word distributions over the top five topics of an outlier in region #417.....	77
Table 4.4. “tiff” distributions over four related topics.....	80
Table 4.5. Word distributions over TIFF related topics.....	80
Table 4.6. Details about evaluation metrics.....	82
Table 4.7. Top ten traffic related keywords ranked by their frequency.....	86
Table 4.8. A list of mined association rules.....	88
Table 4.9. Explanation to evaluation metrics.....	91
Table 4.10. Traffic event classification results of three different methods	92
Table 4.11. Content of some clusters summarized by TextRank model	101
Table 4.12. Distributions of top five topics over the detected outliers	106
Table 4.13. Word distributions over topic #71	107

List of Figures

Figure 1.1. An overview of the research methodology	7
Figure 2.1. A typical framework for social event detection using geosocial media data. The steps enclosed in the yellow solid line frame refer to the spatiotemporal burst-based method. The steps enclosed in the blue dotted line frame refer to the semantic burst-based method.	17
Figure 2.2. A typical framework for traffic event detection using geosocial media data.....	23
Figure 2.3. Pre-processing of a sample tweet using NLP (modified from D’Andrea et al., 2015)	26
Figure 2.4. A flowchart of traffic event classification	28
Figure 2.5. Screenshot of an example traffic event visualization system (TEDS) (Liu et al. 2014)	33
Figure 3.1. The general workflow for small-scale event detection using geosocial media data based on space, time, and semantics	38
Figure 3.2. The overall process framework for social event detection	40
Figure 3.3. Process of Latent Dirichlet Allocation (LDA) model	47
Figure 3.4. Overview of clustering outliers based on ST-DBSCAN method that simultaneously measures spatial adjacency and semantic similarity. Outliers of the same color indicate an outlier cluster.	49
Figure 3.5. An explanation of ST-DBSCAN algorithm for grouping outliers (Birant & Kut, 2007)	50
Figure 3.6. The overall workflow of detecting traffic events from Twitter data based on classification methods	52
Figure 3.7. Preprocessing tweets using NLP tools	52
Figure 3.8. The process of Apriori algorithm	54
Figure 3.9. An example showing the spatiotemporal characteristics of traffic relevant tweets ...	59
Figure 3.10. The overall workflow of detecting traffic events from Twitter data based on clustering method.....	60
Figure 3.11. Sorted 4-dist graph by Ester et al. (1996).....	61
Figure 3.12. A sample graph generated by TextRank algorithm	62

Figure 4.1. The geography of Toronto, Ontario, Canada with Twitter data used for social event detection.....	68
Figure 4.2. Frequency distribution of street blocks area in Toronto.....	70
Figure 4.3. Voronoi representation of k-partitioned regions in Toronto (K=1000).....	70
Figure 4.4. The number of outliers changes with different threshold of Mahalanobis Distance..	72
Figure 4.5. Spatiotemporal distributions of the detected outliers in Toronto	73
Figure 4.6. Detection of outliers (red point) in region #417 on weekday and weekend (Mahalanobis Distance)	74
Figure 4.7. Perplexity distribution used for estimating the number of topics for topic modeling	76
Figure 4.8. Statistic distribution of semantic similarity between pairs of outlier tweets	78
Figure 4.9. Statistics of #outliers before and after clustering by investigating spatial adjacency and semantic similarity	78
Figure 4.10. Histogram of the relevance of outlier clusters to Toronto International Film Festival (TIFF) event	81
Figure 4.11. Evaluation of social event detection results measured by accuracy, precision, recall, and F_1 -score	84
Figure 4.12. Detection of outliers (red point) in region #417 on weekday and weekend (boxplot)	85
Figure 4.13. The geography of Toronto, Ontario, Canada.....	86
Figure 4.14. The number of association rules changing with different support and confidence values	88
Figure 4.15. MTO travel speed data	93
Figure 4.16. The distribution of standardized actual travel speed and typical travel speed based on classification method	95
Figure 4.17. The detection rate changing with different thresholds for significance level based on the classification method.....	97
Figure 4.18. Sorted 2-dist graph	98
Figure 4.19. Spatiotemporal distribution of detected traffic events.....	99
Figure 4.20. Spatial distribution of traffic events	100
Figure 4.21. Temporal distribution of traffic events.....	100

Figure 4.22. The distribution of standardized actual travel speed and typical travel speed based on clustering method.....	104
Figure 4.23. Comparison between classification-based method and clustering-based method..	105

List of Appendices

Appendix I. Topic distributions over an outlier in region #417	116
Appendix II. Word distributions over topics in Appendix I	116

Chapter 1 Introduction

1.1 Problem statement and motivation

Small-scale events involve interactive human movement, where abnormal patterns emerge at a small area during a limited time period rather than a wide range of space (i.e., a city, a region, or a country) and time. For instance, Mother's Day occurs throughout the country, which is a national holiday and not detected as a small-scale event (Abdelhaq, Sengstock, & Gertz, 2013; C. Zhang et al., 2016). As more articulated interactions between humans are introduced in a short enough time, these small-scale events can be observed in real time (Batty, Desyllas, & Duxbury, 2003). The most common human interactions that contribute to small-scale events engage the movement of an increasing number of people, or the concentration of similar occurrences in a small region over short time. It is likely that these events draw people's attention and/or constitute concentration, which presents a pattern different from those patterns without the occurrence of small-scale events around this location and time. Multiple reasons can intervene human movements with different patterns. As this movement phenomenon is an ever-growing feature in cities, identifying the reasons that drive such movement is becoming increasingly important to deal with non-trivial problems of planning, management, and control in urban systems.

Scholars from related fields, such as transportation, urban planning, geography, computer science, and geographical information science (GIS), have recently placed intensive efforts on addressing the two aforementioned issues 1) how people move differently when small-scale events occur from normal days and 2) what reasons lie behind such human movement. Various attempts have been made to discover the small-scale anomalies resulted from football matches, rock concerts, street parades, protest, traffic congestion, disaster evacuation with panic and safety issues, and etc. (Batty et al., 2003). Based on the discovery of small-scale anomalies, responsive plans can be made accordingly and urban systems can be better designed. For instance, intelligent transportation systems can be developed and evolved by expanding the roads that are frequently found congested and/or involved in incidents.

In order to investigate how people move unusually with the occurrence of events, space and time are selected as two essential components to measure these abnormal patterns. The space-time patterns can be shaped from the information with appropriate measurement at spatial and temporal dimensions. An event occurs when people moving or gathering with the same momentum are bound together, which is likely to generate a spatiotemporal pattern that goes beyond the geographical regularities. Accordingly, the location and time of occurring events can be estimated by referring to the associated spatial and temporal information. Crowdsourced data, such as vehicle, GPS trajectory and cell-phone cellular network data, have been explored to measure the spatiotemporal patterns to sense and detect small-scale events (Calabrese, Pereira, Di Lorenzo, Liang, & Ratti, 2010; Kamran & Haas, 2007; H. Park & Haghani, 2016).

Moreover, a more important concern is the reasons driving such human movement. What is happening often attracts people to move to or stop at a certain place during a similar period. If text information is available for semantic analysis, the content of occurring events can be investigated to help interpret the reasons lying behind their movement.

In recent years, with the wide use of smart phones and mobile devices, social media platforms (e.g., Twitter, Weibo, and Instagram) have become a promising alternative approach to feasibly collect data with multi-dimensional information. These platforms allow users post short messages, images, and videos with timestamps and geolocation information (Li et al., 2016), providing a cost-effective way to capture a wide variety of information from continuous data streams at any time and any place. Geosocial media data, the data tagged with specific geolocations (e.g., geographical coordinates), can be obtained from these social media platforms if the location service is turned on. As such, the abundant geosocial media data involving space, time, and semantics not only allows us to measure spatiotemporal patterns to sense event occurrence and estimate event location and time, but also conduct semantic analysis of event content to interpret the reasons behind, which well meet the requirements of capturing small-scale events. If the small-scale events can be detected timely with “when, where, and what” information using geosocial media data, it alerts individuals and management departments who place attention on these events to follow the event evolution and adjust their plans, which may help retain harmony and ultimately promotes sustainable development.

1.2 Research objectives

The overall objective of the thesis research is to develop models and methods to efficiently detect small-scale events using geosocial media data, where abnormal spatiotemporal patterns can be measured for event exploration, event location and time can be estimated, and event content can be summarized in an efficient way. The following three specific objectives have been achieved.

1) To study and develop effective methods to extract useful information from geosocial media data streams for event detection.

Geosocial media data randomly distributes in space and time as a whole instead of being partitioned into appropriate regions and time intervals, and contains information irrelevant to events, which may place a negative effect on the small-scale event detection. Thus, the data needs to be arranged and preprocessed for the event detection from spatial, temporal, and semantic perspectives. From the spatiotemporal perspective, geosocial media data is mapped into certain regions and periods based on its associated location and timestamp in order to be efficiently used for identifying small-scale events. How the data is grouped in space and time lays a solid foundation for identifying spatiotemporal pattern. From the semantic perspective, there may exist geosocial media messages that are negatively relevant to events. These irrelevant messages needs be filtered by effective methods or rules, such as keywords. The remaining geosocial media messages are used to infer the content of events.

2) To develop methods that detect abnormal patterns indicating event occurrences based on spatiotemporal information.

As an event occurs, a different pattern is supposed to appear in space and time comparing with the patterns presented around the location during the same time period. For instance, abnormal patterns can be an abrupt of human movement emerging during a short period within a limited space, or a number of geosocial media messages concerning the similar topic unexpectedly concentrating in space and time. Thus, spatiotemporal patterns can be identified as an effective indicator for the occurrence of small-scale events.

3) To develop methods that analyze event content interpreting the reasons behind abnormal patterns based on semantic information.

What people are talking about in a location at a time can be inferred by the associated semantic information. In other words, after the abnormal spatiotemporal patterns are identified, the geosocial media messages associated with these patterns can be further analyzed to extract the content of occurred events. As such, it presents an effective way to explain the reasons driving people move to compose abnormal patterns.

1.3 Assumptions, scope, and limitations

Based on the systematic review and summary of the existing studies concerning the detection of small-scale events using geosocial media data, this thesis is subject to the following two assumptions.

1) The free-accessed geosocial media data is sufficient for detecting small-scale events.

In this thesis, the data was collected from social media platforms through their provided Application Programming Interfaces (APIs). Due to the restrictions imposed by APIs, only a small portion of data could be obtained free. For instance, Twitter Streaming API only allows 1% of all tweets at most to be crawled. Morstatter, Pfeffer, Liu, and Carley (2013) demonstrated that there was no significant difference between the sample data collected through Twitter Streaming API and the complete firehouse data based on statistical analysis, topical analysis, and geographic measures. It means the freely accessed data is sufficient representation of human activity on Twitter as a whole. In addition, Q. Li, Shah, Thomas, Anderson, and Liu (2016) analyzed entity and metadata distribution, and entity coverage and novelty evolution using Twitter data ranging from 1% (i.e., free-accessed data from Twitter Streaming API) to 10% (i.e., enterprise-accessed data from Decahose). The results showed that the streaming data and Decahose data have the same distributions over the topics talking about life and society, which are major concerns in this research. Therefore, it is assumed that the limited amount of geosocial media data collected without cost is at least effective for detecting small-scale events.

2) Social events and traffic events are two representative types of small-scale events researched in the thesis.

As described in Section 1.1, small-scale events include multiple examples from conventional traffic anomalies to vigilant security events. Essentially, they can be categorized into two types at a higher level according to the different patterns of human movement presented.

Type I: Events often link to an abrupt number of moving humans, which can be measured by for example patterns of people posting on social media platforms, during a short period, which creates a significantly abnormal pattern comparing with the geographical regularities. The unusually crowded social activities that well align with these features include concerts, music festivals, and sport games, which are generally summarized as social events in this thesis. Security incidents including protests and terrorist activities can also be identified in this way. More details of the related work can be found in Section 2.2.

Type II: Events involve random occurrence of key factors rather than significant quantity changes, which may not create an obviously abrupt pattern as the former. As reflected on social media platforms, such event is likely to occur in the real world as one or more relevant posts (but not as many as posts related to social events) appear and/or concentrate in a limited space and time. Among all researched small-scale events in existing studies, traffic events represent a good example of this type. Traffic events refer to traffic anomalies, which mainly include traffic incidents and severe traffic conditions. Traffic incidents randomly occur without expectations, such as traffic collisions, road construction, traffic signal failures, and bad weather conditions. Severe traffic conditions concerning traffic flow are usually caused by daily rush hours, traffic jams and traffic delays (Dabiri & Heaslip, 2019). Traffic events can be detected if posts are identified positively relevant to certain types of traffic anomalies. Moreover, if several traffic relevant posts are found around a road within a certain period, it is assumed that a traffic event is very likely to happen, drawing people's attention.

Social events and security events hold similar burst characteristics in the detection process. Security events are highly dependent on political situations and have obvious regional features,

while topics related to social activities may be more popular among young generation who are major users of social media platforms all around the world. As such, social events and traffic events are selected as two representatives of small-scale events to be detected from geosocial media data in the thesis. The following work are elaborated based on these two types of events. However, the focus is not on applications *per se* but on introducing a general workflow that might be generalized to the detection of a variety of small-scale events using geosocial media data, which focuses on the text information collected from multiple social media platforms.

Despite that, the research reported in this thesis is subject to two types of possible constraints and limitations listed as follows.

1) Data limitations

- *Twitter data*: The geosocial media data used in this research was collected through the free-cost Twitter Streaming APIs. Only 1% of all Twitter data was obtained.
- *Travel speed data*: The ground truth data, i.e., hourly travel speed data, used for evaluating traffic event detection was obtained from Ministry of Transportation, Ontario (MTO). Due to data ownership, travel speed data at other granularities, such as 30-minute interval and 15-minute interval, are not available.

2) Methodology limitations

- *Workflow process*: The general workflow summarized in Section 3.1 is only suitable for detecting small-scale events from text-based geosocial media data. Other types of geosocial media data, e.g., videos and images, may not be suitable for event detection using this workflow.
- *Model selection*: Due to a relatively large number of models available for certain tasks, it is not feasible to investigate the performance of all these models. For instance, a few most

widely used text classification methods were adopted for traffic event classification. A well-known graph-based model was selected for summarizing traffic event content.

- *Parameter estimation*: Similar to model election, multiple choices can be adopted to estimate parameters for certain models. In this research, the author estimated parameters by taking specific features and requirements of the researched applications into account. For instance, the number K in k-means clustering method was estimated according to the attributes of street blocks. The parameters in ST-DBSCAN used for correlation analysis in social event detection were estimated by plotting frequency distribution of all computed values to obtain the threshold values.

1.4 Research methodology

An overview of the research methodology adopted in this thesis is presented in Figure 1.1, which consists of four major steps together with sub-tasks in each step. It follows the process of theory guiding practice, where theory study was first conducted to lay a solid foundation for the following experiments.

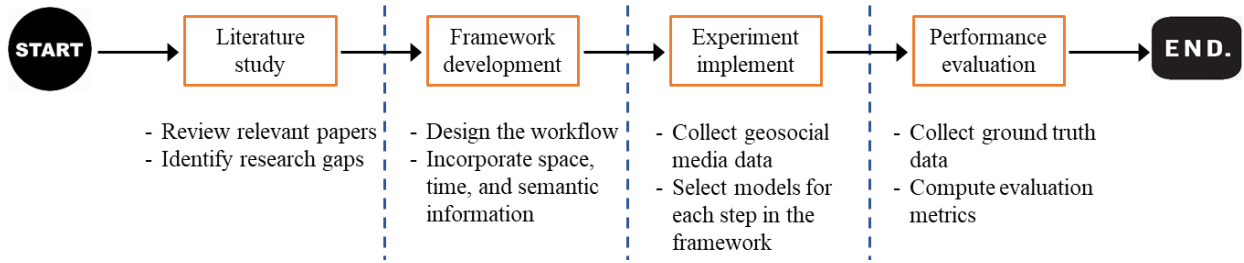


Figure 1.1. An overview of the research methodology

The research started with a literature study, focusing on reviewing relevant papers and identifying research gaps. The author searched research papers concerning detecting social events and traffic events using geosocial media data (e.g., Twitter, Weibo, and Instagram) and organized them according to the methods applied, based on which the author built necessary terminology and knowledge for the research. The general workflow used for social event detection and traffic event

detection was then summarized from the organized papers. Furthermore, the research gaps in existing workflow were identified to guide the author where to make efforts in this research.

In developing framework for capturing social events and traffic events from geosocial media data, the general workflow summarized from the literature study played an essential role. By incorporating space, time, and semantics embedded in geosocial media data, the existing workflow was extended to meet the research objectives declared in Section 1.2 to effectively extract useful information, measure spatiotemporal patterns, and analyze content of posts (as discussed in Section 3.1).

Accordingly, the developed framework was implemented in detecting social events and traffic events using the geosocial media data collected through provided APIs. Appropriate models were then selected and tested for each step based on programming in Python language. Several packages including numpy, nltk, and scikit-learn were applied to conduct spatiotemporal clustering, natural language processing, machine learning classification, text mining, and statistic analysis in this process.

Finally, the performance of the approaches proposed for social event detection and traffic event detection was evaluated using available ground truth data. Different evaluation metrics (see Section 4.1.5 and Section 4.2.3 for detailed description) were computed to quantitatively analyze the performance.

1.5 Main contributions

The thesis presents an efficient way to detect small-scale events using geosocial media data based on spatial, temporal, and semantic dimensions, which provides reference for researchers in the same field or similar fields to carry out further researches. Furthermore, this research benefits the development of urban systems as well. The individuals and management departments are able to sense and track occurring events, and estimate their effect on neighbouring environment or even urban systems. In this manner, responsive plans can be made immediately, and/or relevant policies

can be formulated or improved afterwards. As such, it is possible to prevent urban chaos as similar events occur, which essentially promotes the development of smart cities in the long run.

Through finalizing the above objectives based on the proposed approaches and narrowing the research gaps in existing studies, both the theoretical and methodological contributions have been achieved through this research. The five contributions are elaborated as follows.

Theoretical contributions:

1) A systematic literature review of detecting traffic events using geosocial media data.

The existing literature reviews mainly focused on reviewing detections of all types of events based on geosocial media streams. However, they did not specifically gain insights into traffic domain, or only summarized the main topics in traffic related research using geosocial media data, and analyzed the current collaboration patterns from perspectives of researchers, institutions, and countries, but did not exhaustively discuss the representative methods adopted in detection processes. The research in this thesis considered the characteristics of traffic events, presented a general workflow of extracting traffic events from geosocial media data, and investigated more processing details to complement and extend previous works. To the best of the author's knowledge, this is the first attempt that proposes a comprehensive review from the methodological perspective.

2) An extended framework for traffic event detection with consideration of association rules to effectively extract useful information.

Comparing with the typical framework applied to detect traffic events from geosocial media data, this framework is extended by introducing an extra filtering step to discard negative posts and retain positive posts. This extension aims at dealing with the unbalanced distribution of the results from single keyword query, i.e., negative posts are much more than positive posts, which would generate inaccurate model representation in class-learning tasks by introducing a bias towards the majority class (Ali, Shamsuddin, & Ralescu, 2015). Specifically, traffic related posts were queried using association rules (co-occurrence pattern of two or more words) mined from positive posts by considering the characteristics of geosocial media reported

traffic events. A post is more relevant to a real-world traffic event if it contains two or more words that frequently co-occur in the positive posts. Thus, a higher possibility of returning a positive post can be achieved by queries using a combination of co-occurred words than that of using single keywords. For instance, a post that contains both words “accident” and “injure” is more likely to be a positive post than a post that contains a single word, either “accident” or “injure”. Thus, by adopting the association rules, it can release the negative effect of sample imbalance on training classifiers to identify true traffic events.

Methodological contributions:

- 3) **A new approach of detecting traffic events with exploring the spatiotemporal features of geosocial media data in addition to its abundant semantics.** Most existing studies placed more emphasis on semantic analysis for traffic event identification but less on examining how the spatiotemporal distribution of traffic relevant posts correlated to the occurrence of traffic events. In this research, a space-time-semantic approach was proposed to cluster traffic relevant posts in space and time for traffic event identification as well as automatically summarize the event content. This clustering-based method helps estimate the occurring time and location of detected traffic events.
- 4) **A novel approach to detect social events from spatial, temporal, and semantic perspectives.** The approach efficiently integrates spatiotemporal outlier detection method with short text summarization method for social event detection using geosocial media data. It automatically summarized event content based on a topic modeling approach by investigating a full coverage of outlier messages. It generates comprehensive description of detected events and presents an intuitive view for users to grasp what is going on in a certain region through investigating the topic and word distributions.
- 5) **A new method to incorporate spatial adjacency with semantic similarity to remove event redundancy.** The correlation between abnormal neighbouring regions was examined through measuring their spatial adjacency and semantic similarity. Two or more detected social events were grouped into one event if they were spatially adjacent and

semantically similar to remove event redundancy. In other words, it avoids publishing the same event located in neighbouring regions two or more times during a time period.

1.6 Organization of dissertation

The thesis contains five chapters, starting with this chapter that presents the introduction to the research problems and motivation, objectives and assumptions, overall research methodology, and major contributions of the research.

Chapter 2 presents a literature review of the state-of-the-art availability of social media platforms and related work of small-scale events as well as the research concerning detecting social events, security events, and traffic events using geosocial media data. The research voids in current studies are also discussed in detail so that it paves the way for the following work presented in the subsequent chapters.

Chapter 3 presents the methodology used in this thesis with introducing an overall framework that can be generally used for detecting small-scale events using geosocial media data. Three specific approaches integrating space, time, with semantics are refined in terms of social event detection and traffic event detection, where two distinct approaches are proposed for traffic event detection.

Chapter 4 correspondingly presents three case studies based on the Twitter data collected in Toronto, Ontario, Canada during 2014 and 2015. The experiment results are analyzed and some findings are discussed to acknowledge achievements as well as limitations. Conclusions of the thesis are drawn in Chapter 5, alongside with directions for future work.

Chapter 2 Literature review and related work

This chapter provides necessary background knowledge through reviewing existing studies and related work to help understand proposed approaches and interpret research results in the following chapters. Specifically, the data availability and effectiveness in detecting small-scale events are explained in Section 2.1. Based on the assumption about the two types of small-scale events made in Section 1.3, Section 2.2 mainly reviews existing studies regarding security event detection using geosocial media data, since studies concerning social events and traffic events researched in this thesis will be elaborated in detail in Section 2.3 and Section 2.4, respectively. Similarly, Section 2.3 and Section 2.4 are organized from the methodological perspective, according to the approaches or stages engaged in the detection process. At last, a comprehensive discussion about the applicability and limitations of current researches is presented in Section 2.5, which acts as key for proposing extended and improved approaches in Chapter 3.

2.1 Geosocial media data

Social media platforms provide an easy access for users to express their feelings and/or describe their experiences with limited words tagged with geolocations. It becomes an increasing trend that social media platforms draw more and more people's attention in urban life (Kelley, 2013). Making full use of the free-cost characteristic, users are able to post not only text messages, but also live photos and videos covering a wide spatial and temporal range (Kaplan & Haenlein, 2010). Some popular social media platforms, such as Twitter, Weibo (a Chinese version of Twitter), and Facebook, engage ever expanding user pools, who contribute to composing a huge amount of posts with timestamp and geolocation information (Li et al., 2016). It means that geosocial media posts are likely to be obtained wherever a user is located. As such, these abundant posts seem to be potential sources to extract social events. Further, geosocial media data can be used to not only identify when and where traffic anomalies take place (i.e., traffic pattern), but also explain the reasons behind the traffic anomalies in a real-time manner due to the abundant semantics of geosocial media messages. This provides a significant advantage of geosocial media data over GPS data in detecting traffic events along a road segment (Rashidi, Abbasi, Maghrebi, Hasan, &

Waller, 2017). Therefore, it is likely to be an effective way to extract useful information from these abundant messages to detect traffic events.

Among all of the existing social media platforms, Twitter stands out as it is well-known and accessible in the most part of the world. A number of studies focusing on detecting events have been conducted using Twitter data. Its open APIs including REST API and Streaming API enable developers to collect public tweets with free cost¹. REST API makes it possible to query sample tweets posted in the past seven days with a set of keywords or accounts. Streaming API has the ability to gather tweets in a real-time manner by defining a set of geo-bounding boxes or phrases. As a result, a timestamp, a short message maximum to 140 characters (280 characters have been allowed since November 7, 2017), and a pair of tagged GPS coordinates (i.e., a pair of longitude and latitude) are simultaneously found within a tweet if users' location-based service is enabled. Other popular social media platforms with open APIs, such as Foursquare and Instagram, also provide developers access to collecting data for research purposes. Foursquare allows users to share their experiences or opinions by checking in specific venues. This may result in differences between the exact location of posts and the location of venues, which is likely to have a negative effect on location-sensitive researches. Instagram is picture-sharing platform that draws numerous young people's interest in recent years. Similar to Twitter, it enables users tag location information, timestamps and short texts with their pictures. However, short texts tagged with pictures are usually posted without spaces between words (e.g., "film festival" is represented by "filmfestival"), which poses certain challenges to capture all the spelling patterns for text analysis. As such, Twitter seems to be the most appropriate data source for social event detection since it possesses continuous availability for data collection, geotagged coordinates for location inference, and separate word tokens for event content analysis.

In recent years, many applications that are closely related to our daily lives have been realized bases on Twitter data. Examples include traffic event detection, exploration of human activity patterns, natural disaster detection, crime prediction, monitoring breaking news, social event identification, and prediction of political elections, which have happened in the real world (Gu, Qian, & Chen, 2016; Xu, Li, & Wen, 2018; Huang & Li, 2016; Kryvasheyeu et al., 2016; Zhao,

¹ <https://developer.twitter.com/en/docs/api-reference-index>

Chen, Lu, & Ramakrishnan, 2015; Amer-Yahia et al., 2012; R. Lee & Sumiya, 2010; Metaxas & Mustafaraj, 2012).

2.2 Small-scale events

Although there is no universal definition of “small-scale events” among existing studies, this term is mainly defined in two distinct ways. One is defined as a function of assumed outcomes and event impact (Gratton & Taylor, 2000; Matheson, 2006), and the other one is based on involved resources and event size (Agha & Taks, 2015; Batty et al., 2003). However, a small-sized music festival can have ‘mega’ impact on a small town in terms of tourists and economic benefits (Getz & Page, 2016). As indicated in Section 1.1, this music festival is supposed to be a small-scale event, but it goes beyond the author’s intention given the former definition is applied. As such, in this thesis, small-scale events are defined in the latter way, indicating the events take place in a small area over a short time.

As indicated in Section 1.3, small-scale events can be categorized into type I events and type II events. Type I events mainly include social events concerning leisure and entertainment activities and security events including terrorist activities, riots, and protests. Type II events can be well represented by traffic events including traffic incidents and severe traffic conditions. Since existing studies about social event detection and traffic event detection, which are two representatives researched in the thesis, will be specifically elaborated in Section 2.3 and Section 2.4, this section mainly summarizes the related work of detecting security events using geosocial media data reviewed by Han, Li, Cui, Han, and Song (2019).

Among the current studies concerning analysis of real-world security issues, they mainly focused on detecting security incidents and security situations. Security incidents are detected at a specific place within a short period of time, which indicates the limited space and time enclosing incidents occurrence, such as terrorist activities in city squares and protests along certain streets in a city. Security conditions are sensed at a larger area and a longer period, emphasizing the discussion about the overall security situation of a specific city or country. As such, security incidents are

supposed to be small scale events as defined in the very beginning, of which the relevant studies are summarized as follows.

Security incidents are usually indicated by the burst of relevant topics. Key technologies and approaches targeting security event detection include natural language processing, social network analysis, location inference, image and/or video understanding, and visual analysis. In terms of natural language processing approaches for text content analysis, Hossny and Mitchell (2018) extracted the best word-pairs to represent security events by detecting the spikes within the word-pair signal based on time series analysis, and calculated their Jaccard similarity to describe event occurrence. Those with the highest similarity score were chosen as features to predict if there would be an event or not using machine learning trained classifiers, such as Naïve Bayes, Decision Tree, Support Vector Machine, and Logistic Regression. In addition to that, topic modeling methods and sensitivity analysis were used to uncover the topics and opinions from geosocial media messages. Lin, He, Everson, and Rüger (2012) extended the typical topic model Latent Dirichlet Allocation (LDA) by building an additional layer to analyze users' sentiment, assuming that topics were generated dependent on sentiment distributions and words were generated conditioned on the sentiment-topic pairs. Despite capturing users' emotions from texts, Alfarrarjeh, Agrawal, Kim, and Shahabi (2017) integrated Twitter texts and Flickr images during Hurricane Sandy using Natural Language Toolkit (NLTK) and Google Vision API.

Social network reveals users and their connections on social media platforms through a graph model, where users, communities, and network structure are mainly analyzed in security event detection. Simon, Goldberg, Aharonson-Daniel, Leykin, and Adini (2014) considered Twitter as the crucial channel of communication between the government, emergency responders, and the public during the Westgate Mall terror attack in Kenya. The social networks built among them did great favor for analyzing the main activities and users' patterns. Undertaking the same attack event, Mair (2016) identified al-Shabaab terrorist groups based on the way they communicated on Twitter. The terrorist groups focused on posting the attack scenes to attract the potential audience. As users interact frequently over a certain period, a stable group structure of individuals and social relations is formed, namely communities. Klausen (2015) conducted a community analysis based on snowballing method using Twitter data during Western Foreign Fighters in Syria and Iraq.

Moreover, network structure, which refers to the connections between nodes and edges in a network, helps understand users' roles and the community characteristics of social network and its evolution. Chaurasia and Tiwari (2014) computed the centrality indicators of Lashkar-e-Taiba terrorist networks in the Mumbai attack incident and used the Brokerage method to identify the different roles of terrorists in the network.

Event location inference also benefits from social network analysis. Kong, Liu, and Huang (2014) implemented the SPOT, a large-scale user-location inference system, to estimate users' locations through investigating friends, social closeness, and local social coefficient. Ribeiro and Pappa (2018) improved the accuracy and recall of location inference results by applying the follower network approach. Besides, the location information could be directly uncovered by the geo-tagged coordinates. Those without geotags could be inferred from geosocial media texts using Named Entity Recognition (NER) tool through matching trained templates (Inkpen, Liu, Farzindar, Kazemi, & Ghazi, 2017). Banweer et al. (2018) developed a hybrid filtering model using matrix decomposition method, taking the linkage between user-posted keywords used at different locations into account. Without the help of external place-name database, Ozdakis, Ramampiaro, and Nørvåg (2018) achieved user location inference by mining co-occurrence terms from tweet texts by integrating spatial point patterns with statistical methods.

The visual characteristics of images and short videos also act as crucial features for the analysis of security issues. By interpreting the information associated with these visual data sources, the spatial, temporal, and semantic information related to the visually captured events can be extracted. Up to now, visual analysis has been used to analyze the terrorist organizations recruitment and terrorism propaganda (Conway & McInerney, 2008; Salem, Reid, & Chen, 2008; Sureka, Kumaraguru, Goyal, & Chhabra, 2010), detect violent behaviors (Y. Gao, Liu, Sun, Wang, & Liu, 2016; T. Zhang et al., 2016) and abnormal features (Hu, Wongsuphasawat, & Stasko, 2017), and monitor the trends of security events (Marcus et al., 2011).

2.3 Detecting social events using geosocial media data

With regard to retrieving social events from geosocial media data, there are mainly two distinct types of methods based on the literature review, namely spatiotemporal burst-based method and semantic burst-based method (Figure 2.1). The spatiotemporal burst-based method first measures the geographical regularity pattern of crowds (e.g., the number of tweets and the number of users) as a baseline to detect outliers in a region during a limited time period. The tweet texts of the detected spatiotemporal outliers are further analyzed to identify the event content. The semantic burst-based method mainly focuses on extracting features (e.g., keywords, hashtags, and topics) or documents (e.g., tweets) with increasing frequency and clustering them to determine the social event location. More details about the two types of methods are presented in Section 2.3.1 and Section 2.3.2.

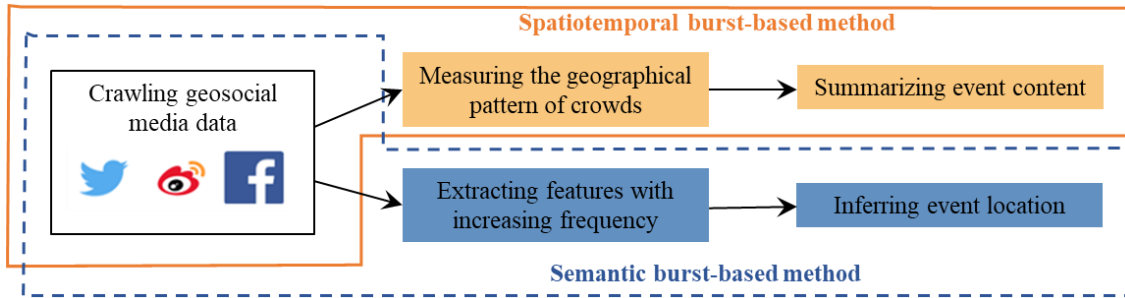


Figure 2.1. A typical framework for social event detection using geosocial media data. The steps enclosed in the yellow solid line frame refer to the spatiotemporal burst-based method. The steps enclosed in the blue dotted line frame refer to the semantic burst-based method.

2.3.1 Spatiotemporal burst-based method

(1) Non-spatial clustering-based method

Non-spatial clustering-based method estimates the geographical regularity of crowds in threshold-based regions, administrative districts, or geometrically partitioned grids. Walther and Kaiser (2013) created clusters if there were more than 3 tweets issued during the previous 30 minutes

within a radius of 200 meters, which were further updated using a threshold-based method. These candidate outlier clusters were evaluated to see whether they constituted real-world events or not based on a binary machine learning classification method. A ranked list of tweets were displayed in a descending order to summarize the identified social events. A handful of tweets were also extracted by Krumm and Horvitz (2015) to summarize social events happening in Hierarchical Triangular Mesh (HTM) partitioned space during 20 minutes, 1 hour, 3 hours, 6 hours, 12 hours, and 24 hours. These events were declared when an abrupt increase in the number tweets was identified in comparison with the amount predicted by a regression model. Gao, Cao, Jin, et al. (2013) modeled social events by aggregating individuals and collective behaviors, namely the number of users and the number of tweets collected within a specific time period in the geographical region. A social event was detected if the aggregated number of users or tweets exceeded the geographical pattern threshold, and tracked by the top k representative tweets. Similarly in Gao, Cao, He, et al. (2013), the threshold-based outlier detection method was adopted to discover unusual administrative districts, where the tweets were clustered based on content similarity. Another threshold was set to determine candidate clusters as real-world social events, of which the content was summarized by highly frequent keywords. Xia et al. (2015) and Xie et al. (2013) integrated Twitter and Instagram data to explore outlier cells that were split by identical grids based on the Gaussian Process Regressor (GPR) time series model. These candidate event signals were then classified into true events or noise. Both keywords of tweets and captions of photos related to true events were extracted for event description. This method is not flexible since the occurrences of social events do not usually follow the predefined spatial patterns.

(2) Spatial clustering-based method

Fujisaka, Lee, and Sumiya (2010) partitioned the study area based on k-means clustering method. The hourly crowd movement pattern within each rectangular space was extracted from Twitter data using aggregation and dispersion models. Crowds of which the total moving distance was high were attached with high activity scores, which were regarded as indicators of social events occurrence. Furthermore, R. Lee and Sumiya (2010) and R. Lee et al. (2011) manually checked tweet's texts to investigate the focus of the event taking place in k-means clustered geographical regions during morning (6am-12pm), afternoon (12pm-6pm), evening (6pm-12am), and night

(12am-6am), after it was identified as spatiotemporal outlier by comparing with geographical regularities (i.e., number of tweets, number of users, and number of moving users).

Khalifa, Díaz Redondo, Vilas, and Rodríguez (2017) clustered Twitter users of New York City into a number of groups using DBSCAN method. The crowd's behavior on a normal day was represented by the number of tweets posted in each cluster at different time slots, which was obtained as daily reference pattern to detect burst urban crowds involved in certain event through boxplot method. Moreover, Domínguez, Díaz Redondo, Vilas, and Khalifa (2017) improved this mechanism to better obtain the pulse of city using Instagram data covering several months rather than a single day. This avoided accidental errors caused by improperly selecting a random day as the normal day. However, the content of the posts (e.g., tweet texts and image tags) are not analyzed in the aforementioned studies to identify what the detected event refers. Costas, Vilas, Vicente, and Díaz Redondo (2018) extended this work by adopting content aggregation models to generate representative contents every 30 minutes in order to give users a timely sense of what was going on in the specified region. Driven by the spatial distribution of data points, the spatial clustering-based method is adjustable to detect different types of social events.

(3) Space-time scan statistics-based method

Instead of grouping geosocial media data in a fixed time dimension (e.g., daily, hourly, equally divided time intervals), space-time scan statistic method was used to adapt the spatiotemporal variation of geosocial media data, which was realized by a cylindrical window that extruded a circular geographic base with a height corresponding to time. The cylinder with the maximum Poisson generalized likelihood ratio (GLR) was measured as the candidate for a true outbreak (Kulldorff, Athas, Feuer, Miller, & Key, 1998; Kulldorff, Heffernan, Hartman, Assunção, & Mostashari, 2005). Cheng and Wicks (2014) adopted this method to spatiotemporally cluster Twitter data through cylindrical windows regardless of tweets content. A window was identified as a cluster where tweets count emerged. The significance of each cluster was then indicated by p-value. In order to measure the relationship of each cluster to space-time events, Latent Dirichlet Allocation (LDA) method was used to discover the topics engaged in cluster's tweets. It was detected as a space-time event if at least half of the topics were attributable to real world. However,

different spatiotemporal clusters are likely to be composed if two or more variables (e.g., number of tweets and number of users) were selected as indicators to measure outbreaks, which poses certain challenges to identify the actual time and location for social events. In contrast, Andrienko et al. (2013) investigated regular or repetitive spatiotemporal pattern using Twitter data through a space-time visualization approach. It clustered tweets with considering the spatiotemporal distances between tweets. By tagging word terms with spatial and temporal information and classify them into multiple categories based on meaning, different thematic patterns represented by the most prominent terms could be spatiotemporally visualized on the map.

2.3.2 Semantic burst-based method

Documents and features are mainly selected as signals to detect burst in a specific geographical area during a given time period. Documents refer to the raw tweets, while the features are terms that are extracted from tweets, which include hashtags, keywords, and topics.

(1) Hashtag-based method

Cordeiro (2012) retrieved hashtags from tweets and organized them every five minutes. A wavelet analysis method was used to identify the significant increases in the volume of specific hashtags in a given time interval. The peak was detected as events, of which the topics was estimated by applying LDA topic model among tweets that were related with the burst hashtags. However, this work neglected the location inference of the detected events, which could not help people make effective plans to take part in or avoid these events with spatial reference. Feng et al. (2015) organized tweets according to the spatial and temporal hierarchy. It enabled users to explore hashtag clusters, which were created by a single-pass hashtag clustering algorithm, in different space and time granularity. Given a particular region and time frame, event ranking considering popularity, burstiness, and localness as three indicators was done to find localized events based on a cube model. By incorporating Twitter data and Instagram data that were only posted during the past hour, DBSCAN method were adopted twice by Ranneries et al. (2016) . The first was used to estimate the location of events within a candidate hashtag cluster, which was made up by measuring the posts similarity. The second was applied to group candidates located nearby with a

certain distance based on their estimated locations. This method heavily depends on prior knowledge and has limitations of detecting other types of social events.

(2) Keyword-based method

With regard to keyword-based method, Liang, Caverlee, and Cao (2015) modeled keywords as signals from spatial, temporal, and semantic perspectives using Twitter data. Watanabe, Ochi, Okabe, and Onai (2011) identified a social event by investigating the co-occurrence of keywords and popular places within a short time and a small geographic area. The geotags observed frequently in tweet texts during a specific period were defined as popular places based on Geohash algorithm. The words appearing three times or more in tweet texts associated with the popular place were selected as keywords. The place with one or more extracted keywords that was detected as a social event, which was demonstrated by reading these tweets. In this way, tweets posted in the specific geographic area but without preselected place mentions may also refer to true social events, which are discarded as noise. This method was compared with the work of Boettcher and Lee (2012), where some possible sets of keywords were preselected and spatially clustered for potential event identification, which were further filtered through a binary logistic regression. It revealed that the precision was improved and events could be detected without the limitation of a list of popular places. Abdelhaq, Gertz, and Sengstock (2013) identified burst keywords for candidate events descriptions in terms of their spatiotemporal characteristics using a sliding window approach. A novel graph-based regularization was proposed to deal with the noise and sparsity of geo-tagged tweets. Abdelhaq, Sengstock, et al. (2013) further clustered burst keywords based on spatial similarity to extract localized events. A scoring scheme was introduced to determine the most important events in a time scheme. Considering the possibility that people far away from the location of an event could also post tweets containing relevant keywords, Abdelhaq, Gertz, and Armiti (2017) filtered such spatial outliers by investigating the correlation between the keywords and place names, namely some event-related keywords often co-occurred with the mentions of places. The central location of keywords was estimated for geo-locating the detected social event. Moreover, both visual and textual information of Twitter were analyzed by Kaneko and Yanai (2016) through adding image clustering with keywords burst detection. The detected social events were shown with the representative photos on the map. Another visual analytics

method was conducted by Thom, Bosch, Koch, Worner, and Ertl (2012) based on clustering approach.

(3) Topic-based method

Pozdnoukhov and Kaiser (2011) first labelled the geo-referenced tweets with the most probable topic using space-time LDA topic modeling method. The prevailing topic was obtained with spatial kernel density and comparative analysis. Markov-modulated time varying Poisson process model was then trained to quantify a volume of abnormal tweets or users within the topic to indicate the social event. While Chae et al. (2012) adopted seasonal trend decomposition for abnormality estimation of the preselected topic. Flickr and YouTube data were collected for cross validation. This supervised method poses limitations on monitoring other topics rather than the preselected topic.

(4) Document-based method

In spite of using derived features as social event signals, tweets are also directly used for events monitoring. S. Zhang, Cheng, and Ke (2017) defined a set of tweets that were spatially close and semantically coherent as a geo-topic cluster. The spatiotemporal burstiness of candidate geo-topic clusters were ranked and updated to identify real-time social events. A web-based system named Event-Radar was further developed to demonstrate the real-time social events for public interests with continuously mapping keywords and tweets into the same low-dimension vector space to better capture the semantic of short Twitter messages, which was also considered by C. Zhang et al. (2017).

2.4 Detecting traffic events using geosocial media data

With respect to characteristics of traffic events defined in the beginning, a typical framework is presented in Figure 2.2, which illustrates a number of components to deal with different stages of operation in detecting traffic events from geosocial media data. The workflow consists of querying for traffic related geosocial media data, pre-processing geosocial media data, identifying geosocial

media data relevant to real-world traffic events, extracting location information, and summarizing and notifying the detected traffic events to users.

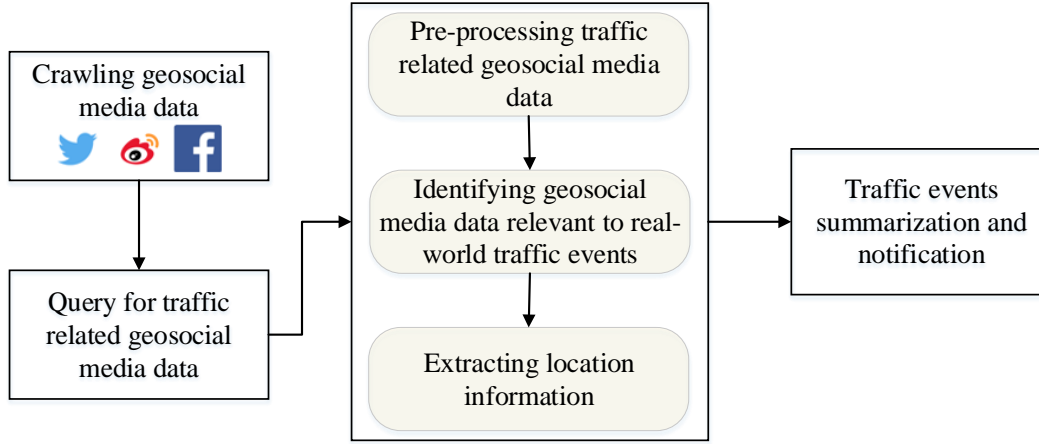


Figure 2.2. A typical framework for traffic event detection using geosocial media data

2.4.1 Query for traffic related tweets

Despite the huge amount of raw data obtained from social media platforms, the information specific to the traffic domain only draws the author’s interest. The accurate extraction of traffic related tweets is very important, as this is the entry point to the following classification process.

There are two main approaches to extract traffic related information from raw tweets: keywords-based approach and accounts-based approach. When certain traffic related keywords (e.g., accident, traffic, and crash) are detected in the tweet, it indicates that a potential traffic event is observed by the user who posts it. Besides, there are also many organization accounts operated by traffic authorities whose main function is to disseminate real-world traffic events information. In this way, traffic events information is often posted after the traffic management officials are already notified. Personal tweets are more likely to disseminate a traffic event in a more timely manner than organizational accounts (Yazici, Mudigonda, & Kamga, 2017). Table 2.1 shows some examples of querying for traffic related tweets using keywords-based or accounts-based methods separately, while Table 2.2 summarizes the studies that combine keywords-based with accounts-based methods to make queries.

Table 2.1. Summary of using the keywords-based or accounts-based methods to query for traffic related tweets

Articles	Query methods	Details
Mai and Hranac (2013)	Predefining traffic related keywords	“accident”, “crash”, “traffic”, “road”, “freeway”, “highway”
D’Andrea, Ducange, Lazzerini, and Marcelloni (2015)		“traffic”, “crash”, “queue”
Nguyen et al. (2016)		“accident”, “crash”, “delay”, “traffic”
Kuflik et al. (2017)		Traffic domain experts review the keywords list ranked by their frequency in a corpus of traffic related documents.
Yazici et al. (2017)		“accident”, “crash”, “traffic”, “road”, “freeway”, “highway”, “lane”, “wreck”, “car”, “cars”, “delay”, “NB”, “northbound”, “SB”, etc.
R. Li et al. (2012)	Extending traffic related keywords	The iteratively refined rules are used to retrieve more tweets based on seed keywords.
Schulz, Ristoski, and Paulheim (2013)		WordNet is used to extend the seed keywords like “incident”, “injury”, “police”, “vehicle”, “accident”, “road”.
S. Zhang et al. (2015)		Topic modeling combined with a hierarchical clustering algorithm are adopted to filter out the irrelevant tweets.
Endarnoto, Pradipta, Nugroho, and Purnama (2011)		@TMCPoldaMetro
Ribeiro Jr. et al. (2012)	Preselecting influential traffic accounts	@TransitoBH, @Transito98FM, @waytaxi

Daly et al. (2013)	@LiveDrive, @AARoadwatch, @GardaTraffic
Kurniawan, Wibirama, and Setiawan (2016)	@ATCS_DIY, @atcs_kotasmrg, @atcs_kotatgr, @atcs_pekalongan, @ntmclantaspolri, etc.

Table 2.2. Summary of combining the keywords-based with accounts-based methods to query for traffic related tweets

Articles	Query methods	Details
S. Wang, He, Stenneth, Yu, & Li (2015)	Preselecting influential traffic accounts and predefining traffic related keywords	@ChicagoDrives, @ChiTraTracker, @roadnowChicago, @traffic Chicago, etc. “stuck”, “congestion”, “jam”, “crowded”, “pedestrian”, “driver”, “accident”, “crash”, etc.
Fu, Lu, Nune, & Tao (2015)	Preselecting influential traffic accounts and keywords query expansion	Tf-idf method is used to rank the importance of each word based on tweets from seed users (@WTOPTraffic, @VaDOT, @drgridlock, @DCPoliceDep). Top words are then selected as keywords to acquire more traffic related tweets.
Gu, Qian, & Chen (2016)	Preselecting influential traffic accounts and extending traffic related keywords	Influential traffic accounts are collected manually by Google searching, and an adaptive keywords-based method is also used to acquire more traffic related tweets.
Aziz, Prihatmanto, Henriyan, & Wljaya (2015)	Preselecting specific hashtag and account	#lalinbdg, @ lalinbdg

2.4.2 Preprocessing queried tweets

Due to the limited length of Twitter messages, there exist problems of mining text information, such as language ambiguity, uncertainty, and abbreviation. It is necessary to clean the tweet texts for the feature extraction to identify real-world traffic events. Pre-processing tweets is mainly based on basic natural language processing (NLP) techniques, including tokenization, normalizing all words to lowercase, removing non-English and duplicate posts (i.e., retweets), removing stop words, links and mentions to other Twitter accounts, correcting spelling errors, slang replacement, stemming, lemmatizing, and POS tagging. A sample tweet pre-processing is shown in Figure 2.3.

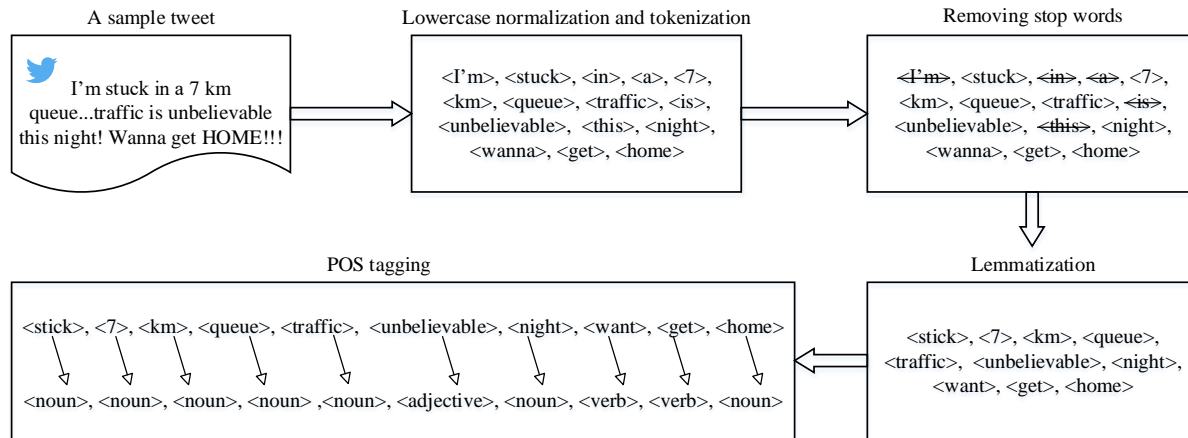


Figure 2.3. Pre-processing of a sample tweet using NLP (modified from D’Andrea et al., 2015)

Tokenization aims to break the short tweet text into separate tokens (Aziz et al., 2015; D’Andrea et al., 2015), and all the letters are normalized to lowercase (Aziz et al., 2015; Kurniawan et al., 2016; Yazici et al., 2017) at the same time. Non-English and duplicate posts (Kurniawan et al., 2016), accent marks (Ribeiro Jr. et al., 2012), and links and mentions to other Twitter accounts (Kurniawan et al., 2016; Ribeiro Jr. et al., 2012) are further removed. Filtering stop words, such as punctuations (Yazici et al., 2017) and non-alphanumeric (alphabets and numbers) characters (Kurniawan et al., 2016), are also an important step in the cleaning process (Aziz et al., 2015; D’Andrea et al., 2015; Kumar, Jiang, & Fang, 2014; Nguyen et al., 2016; Schulz & Ristoski, 2013;

Schulz et al., 2013; Yazici et al., 2017). Moreover, as users may post a tweet with spelling errors and slangs, a replacement is required to make corrections (Schulz et al., 2013).

Stemming is the process of reducing each word to its stem or root form by removing its suffix or prefix. D’Andrea et al. (2015) and Kumar et al. (2014) used the Porter algorithm to reduce inflected words and transform variants of words into a single stem. Endarnoto et al. (2011), Schulz and Ristoski (2013), Schulz et al. (2013), and Nguyen et al. (2016) made use of the Stanford lemmatization function (Manning et al., 2014) to normalize words, and then applied the Stanford POS tagger (Manning et al., 2014) to filter meaningless categories of words. Furthermore, Gutiérrez et al. (2015) used a POS tagger for analyzing the time expressions and verbal forms to extract temporal information from the tweet texts.

2.4.3 Identifying tweets relevant to real-world traffic events

As described in Section 1.5, some tweets containing traffic related keywords do not actually refer to a real-world traffic event. The main task of this stage is to identify and remove these noisy tweets, in other words, classifying the processed tweets into real traffic events and non-traffic events. To this end, a summarized process including feature extraction and classification is presented in Figure 2.4, which illustrates the process of identifying real traffic events. Specifically, the techniques applied in each stage are summarized in Table 2.3 and Table 2.4.

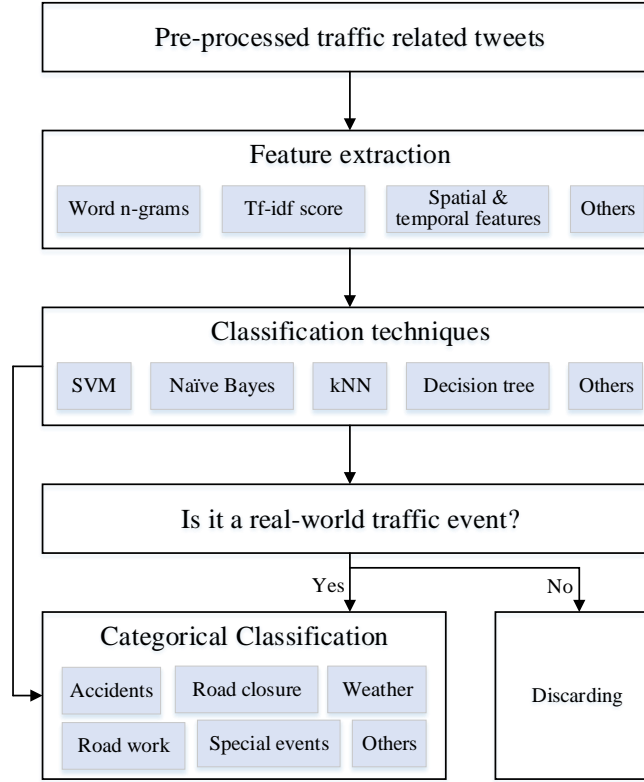


Figure 2.4. A flowchart of traffic event classification

Table 2.3. Summary of binary classification or multiple classification on traffic events

Articles	Features	Classification	
		Classification on traffic events/non-traffic events	Classification on traffic events categories
Kurniawan et al. (2016)	Words and their appearance account in a tweet	Naïve Bayes, SVM, Decision tree	N/A
Yazici et al. (2017)	Word n-grams, tf-idf score	Naïve Bayes	N/A
Z. Zhang et al. (2016)	Correlated words with high coefficient	Maximum likelihood estimation model (MLE)	N/A

	in traffic related tweets		
Schulz et al. (2013)	Word n-grams, char n-grams, tf-idf score, syntactic features, spatial & temporal features, FeGeLOD features	Naïve Bayes, Ripper rule learner (JRip), SVM	N/A
Sakaki, Matsuo, Yanagihara, Chandrasiri, and Nawa (2012)	Dependency features, context features, position features, time expression features, word features	N/A	SVM; Heavy traffic, traffic restrictions, police checkpoints, rain, mist
Ribeiro Jr. et al. (2012)	N/A	N/A	Manually creating rules to identify traffic conditions (e.g., slow) and events (e.g., accident)

Table 2.4. Summary of binary classification followed by multiple classification on traffic events

Articles	Features	Classification	
		Classification on traffic events/non-traffic events	Classification on traffic events categories
Gu, Qian, and Chen (2016)	Single words and combinations of some words that are positively correlated	Semi Naïve Bayes	Latent Dirichlet Allocation (LDA); Accidents, road work, hazards & weather, special

	with being a traffic related tweet		events, and obstacle vehicles
Cui, Fu, Dong, and Zhang (2014)	Word n-grams	Bayesian classifier	Natural language processing based structural labelling; Traffic flow, traffic accident and traffic control SVM;
Kuflik et al. (2017)		SVM	Expression of an opinion, transport need, reporting an event
Gutiérrez et al. (2015)	Word n-grams, tf-idf score	SVM	Constructing a list of corresponding synonyms of representative keywords for each class; Traffic jam, road work, freight traffic, road closure, ice, wind & snow, traffic accident, and others
D'Andrea et al. (2015)	Relevant stems that are retrieved from traffic related tweets	SVM, Naïve Bayes, C4.5 decision tree, kNN, PART	SVM, Naïve Bayes, C4.5 decision tree, kNN, PART; Traffic due to external event, traffic congestion or crash, and non-traffic
Nguyen et al. (2016)	Bag of words, lemma, POS and chunk features, pattern recognizer, bag of tags	kNN, Bayesian Network, SVM, C4.5 decision tree	Conditional Random Fields (CRFs) labelling; Queue, accident, breakdown, police activities, road work

2.4.4 Location inference and visualization

After the identification of real-world traffic events, some studies further extracted their location information and geocoded them with different tools. The geographic location information carried by tweets is rich but may be very noisy and may not be explicitly available. There are generally three types of location information: tagged GPS coordinates, user profiles, and tweet texts (e.g., geographic names and street names). The use of the location information in traffic event detection are summarized in Table 2.5.

Table 2.5. Summary of geocoding techniques

Articles	GPS coordinates	User profiles	Tweet texts
Wanichayapong, Pruthipunyaskul, Pattara-Atikom, & Chaovalit (2011)	×	×	√
R. Li et al. (2012)	√	√	×
Sakaki et al. (2012)	√	×	√
Ribeiro Jr. et al. (2012)	×	×	√
Daly et al. (2013)	×	×	√
Schulz et al. (2013)	×	×	√
Chen, Chen, & Qian (2014)	√	×	√
Cui et al. (2014)	√	×	√
Kumar et al. (2014)	√	×	×
Gutiérrez et al. (2015)	×	×	√
S. Zhang et al. (2015)	√	×	×
S. Wang et al. (2015)	√	×	√
Tejaswin, Kumar, & Gupta (2015)	×	×	√
Gu, Qian, & Chen (2016)	×	×	√
Nguyen et al. (2016)	√	×	√

Some tweets carry a pair of coordinates, namely longitude and latitude, when they are tweeted from smart phones that have location-based service enabled. These coordinates correspond to the more precise locations where users post tweets. They are usually regarded as where the traffic events take place if the tweets are relevant to real-world traffic events (Kumar et al., 2014; Zhang et al., 2015; Zhang et al., 2016). Some tweets are posted by accounts whose profiles are shared with the public, such as city, country, and sometimes finer-grained business names and street address of the business. Mining the text of tweets is also able to infer the event locations at various granularity levels (e.g., city, district, street name or block number) depending on the event type and can provide more than one location (Ozdikis et al., 2017).

Once a traffic event is detected from the social media platforms, it needs to be presented to the user in a meaningful and informative way, as the tweets themselves do not necessarily provide a good summary of an event. The detected traffic events are visualized in some proposed systems to give the user an intuitive view of the events. An example screenshot of these visualization systems is illustrated in Figure 2.5. TEDS (Liu, Fu, Lu, Chen, & Wang, 2014) summarized event content using a long twitter text based on the multi-document method. As shown in Figure 2.5, the list of detected events is shown on the left, and a traffic information summary is illustrated on the right.

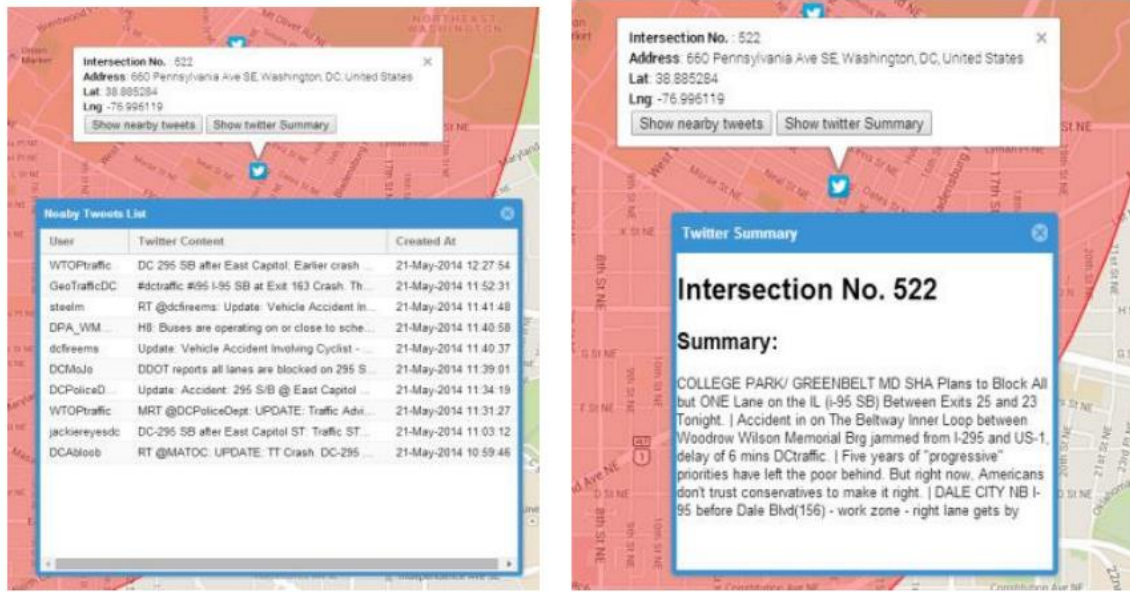


Figure 2.5. Screenshot of an example traffic event visualization system (TEDS) (Liu et al. 2014)

2.5 Discussion

2.5.1 Social event detection

Both the spatiotemporal burst-based method and the semantic burst-based method discussed in Section 2.3 can detect spatiotemporal social events using geosocial media data. Given an event has drawn great attention before it really takes place, such as a famous concert or sport game, a certain number of related features (e.g., keywords, hashtags, and topics) or documents (e.g., tweets) are likely to have been posted around the location. As the event occurs during the specific period, the semantic burst-based method may not be able to find the significant burstiness of features or documents by monitoring the changes over time. However, a social event not only results in a hot discussion but also attracts a number of people coming from other places, which can be successfully captured by the spatiotemporal burst-based method since it can measure the abnormal increase of the number of geosocial media messages in addition to the number of users. Therefore, in this study, the spatiotemporal-based method was used to detect social events from geosocial media data.

In the case of spatiotemporal-based method, there is no semantic analysis of the detected events (Domínguez et al., 2017; Khalifa et al., 2017), or the event content is inferred by manually checking tweets text (R. Lee & Sumiya, 2010; R. Lee et al., 2011), which is time consuming and labor cost. Furthermore, a list of representative tweets or keywords were also selected to describe the social events (Gao, Cao, He, et al., 2013; Gao, Cao, Jin, et al., 2013; Krumm & Horvitz, 2015; Walther & Kaisser, 2013; Xia et al., 2015; Xie et al., 2013). Due to the dispersion of social media messages, the keywords/tweets-based summarization method may not comprehensively describe the detected event in an efficient way. For example, representative keywords/tweets selected for event summarization may indicate different topics, which will confuse users to identify the event content.

Cheng and Wicks (2014) described the space-time events using a set of topics with word distributions, but they did not analyze the correlation between neighboring outlier regions. In other words, the semantic similarity among space-time events were not considered to avoid event redundancy. It is possible that two or more (but not all) neighboring outlier regions indicate the same social event since spatiotemporal social event is found to be spatially adjacent and semantically similar. As such, more text analysis need to be conducted so that users can get intuitive information of the events at first sight.

In summary, two major problems remain unsolved. First, the meaning of real-world events that outlier tweets refer to may not be accurately captured through simply looking into the limited amount of keywords/tweets. Second, previous studies ignored the cases that two or more neighbouring outliers may indicate the same social event if they contain similar Twitter messages. In this thesis, a hybrid approach was proposed to detect the spatiotemporal outliers by comparing with the regular patterns of normal days. Further, the corresponding geosocial media messages within the outliers were analyzed to summarize the social events in an efficient way with investigating the correlation between neighboring outlier regions.

2.5.2 Traffic event detection

As indicated in Section 2.4.1, the traffic related information disseminated by organization accounts does not seem to be as timely as the events reported by individual users. In order to include the tweets from both organization accounts and individual users, a collection of traffic related keywords were initially defined to obtain a large number of tweets in this study. However, among all the tweets queried by the list of single keywords, there exist a much larger number of negative tweets in comparison to the number of positive tweets, which creates a bias towards the negative tweets and places a negative effect on training classifiers to identify real-world traffic events.

Despite the data noise caused by the single keyword based query, two more aspects may limit the performance of this classification-based method. First, most of the existing studies identified the content of detected events based on supervised classification methods, where a predetermined category was assigned to the detected event. In this manner, those tweets that are relevant to traffic events but go beyond the predetermined categories may be missed by false. Second, existing studies focused on detecting traffic events from semantic perspective, but did not investigate the correlation between the spatiotemporal distribution of traffic relevant tweets and the occurrence of traffic events. Due to GPS deviations on mobile devices and the spatiotemporal influence of traffic events, certain spatial and temporal difference exist between the mapped event point and its actual location and time. Any posts captured within that certain area during that period will count for detecting the same traffic event. In other words, a concentration of several similar posts in space and time may indicate a traffic event.

In this thesis, a hybrid approach was proposed to discard negative tweets returned by a single-keyword query, while retaining positive tweets by adopting the automatically mined word combinations (association rules), where the length and the order of the words were not limited. The association rule-filtered tweets were further classified into different types of traffic events based on different machine learning-based classification methods. In addition, a comparative experiment was conducted by taking the spatial and temporal characteristics of the association rule-filtered tweets into account to help estimate the event location and time. The content of

detected events was automatically generated based on short text summarization method rather than assigning them with predetermined categories.

Chapter 3 Methodology

In order to address the limitations of existing studies summarized in the previous chapter, this chapter illustrates the further efforts being made in methodology. A general workflow is first presented to outline the general process of detecting small-scale events from geosocial media data with a full consideration of spatial, temporal, and semantic perspectives. This general process is subsequently mapped to detect social events and traffic events with introducing detailed methods and technologies from Section 3.2 to Section 3.4, where techniques such as machine learning, natural language processing, and spatiotemporal analysis are adopted to realize effective detection. Section 3.5 discusses the factors that are likely to influence the performance of proposed approaches, and provides potential solutions to achieve better performance.

3.1 The general workflow

Based on summarizing the event detection process by the engaged stages, a general workflow was outlined in Figure 3.1, which can be applied to detect multiple types of small-scale events using geosocial media data. As geosocial media data contains texts, images, and videos, this workflow mainly takes the geotagged text information with timestamps posted to social media platforms into account. Images and videos may be not flexible for event detection using this workflow. Researchers first determine which type of events will be detected, and provide a clear explanation to the targeted events. Normally, the social media platforms allow developers to crawl data with different choices. For instance, the Standard APIs provide free access for developers to collect data with certain limitations (e.g., a limited proportion of all data), while the Enterprise APIs assign developers higher authority comparing with the free access, and charge them depending on the selected spatial area and temporal duration.

The collected geosocial media data are then preprocessed and organized by extracting useful spatial, temporal, and semantic information to be prepared for the following spatiotemporal measurement and semantic analysis. As indicated in Section 1.3, small-scale events can be mainly categorized into two types (i.e., type I events and type II events) by referring to their characteristics of occurrence. In order to measure the abrupt patterns as type I event occurs, it is required to split

the research area and period into a number of sub-regions and time intervals according to the requirements of different applications. Grid division and clustering methods can be used for spatiotemporal partition. With regard to type II events, it is critical to efficiently extract relevant posts in this process. Text search technique and natural language processing tools can be adopted for this purpose.

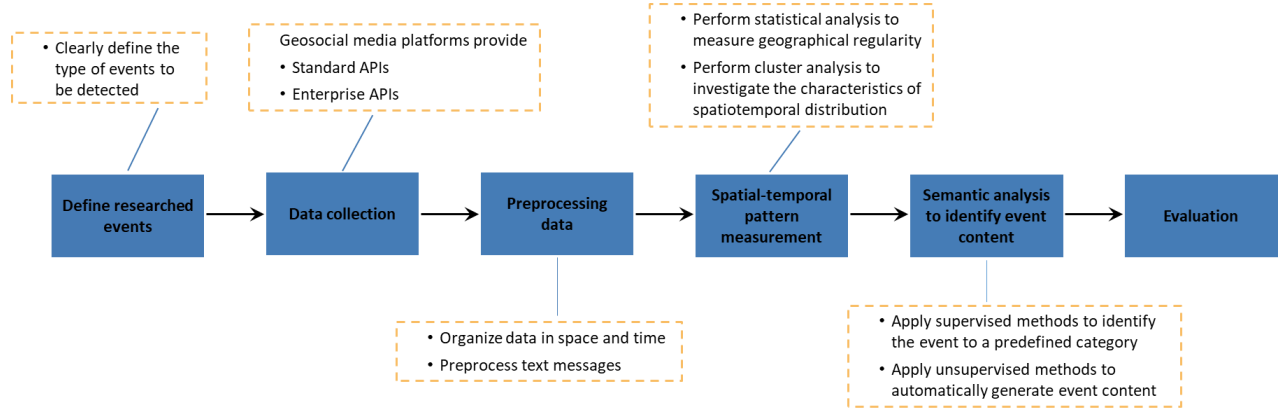


Figure 3.1. The general workflow for small-scale event detection using geosocial media data based on space, time, and semantics

An event is usually detected when an abnormal pattern occurs. Type I events investigate the spatial-temporal pattern in a sub-region during a certain interval through measuring the geographical regularity that is often shaped by the number of posts and the number of involved users. The abrupt in the number of posts and the number of users is detected as outlier by performing statistical analysis. An abnormal pattern refers to an outlier that is beyond the range of measured regularities. Based on the relevant posts obtained in the previous step (i.e., preprocessing data), type II events explore how these posts distribute in space and time by conducting cluster analysis. A spatiotemporal cluster grouping the posts that potentially indicate the same event represents an abnormal pattern in this case.

Subsequently, the content of the geosocial media messages that correspond to the abnormal patterns is analyzed through supervised methods or unsupervised methods, which can be both applied in type I events and type II events. In the case of supervised learning, the event content is

identified as a predetermined category or topic based on the training-testing mechanism with prior knowledge. In contrast, the unsupervised methods automatically summarize the event content with abstractive or extractive terms without prior knowledge. Finally, the detection results are evaluated with the ground truth data to examine the performance of proposed approaches.

3.2 A spatial-temporal-semantic approach for detecting social events

In this study, a method was proposed to detect social events using Twitter data from spatial, temporal, and semantic perspectives. The burst in the number of tweets and the number of users appearing in a specific geographical region (e.g., people participating in a music festival tend to share their excitement with the public via Twitter) within a certain period were detected as spatiotemporal outliers by comparing with the regular pattern estimated in normal days. A machine learning-based method was used to generate the content of the outliers automatically by making full use of topic modeling techniques. In addition, it was assumed that geographically close outliers are likely to indicate the same social event if they are semantically coherent. The correlation between the detected outliers was investigated by examining their spatial adjacency and semantic similarity. The correlated outliers were then regrouped into spatiotemporal outlier clusters for social event identification.

3.2.1 Preliminary and framework

Definition 1. Geosocial media data is the geotagged texts with timestamps posted to social media platforms.

Definition 2. Geographical pattern here refers to the number of tweets and number of users that are measured in a certain geographic region within a specific time period.

Definition 3. Spatiotemporal outlier is an abnormal geographical pattern, where an abrupt burst in the number of tweets and the number of users is detected, comparing with the regular pattern of the same region during a similar day and hour (e.g., weekday morning). It can be regarded as candidate social event.

Definition 4. Outlier tweets are text messages that are posted within spatiotemporal outliers, which are analyzed to summarize the content of candidate event.

Figure 3.2 presents the overall process, which consists of data acquisition, spatiotemporal outlier detection, outlier content analysis, and clustering outliers that indicate the same event for social event detection in the real world. The four stages will be further described in the following sections.

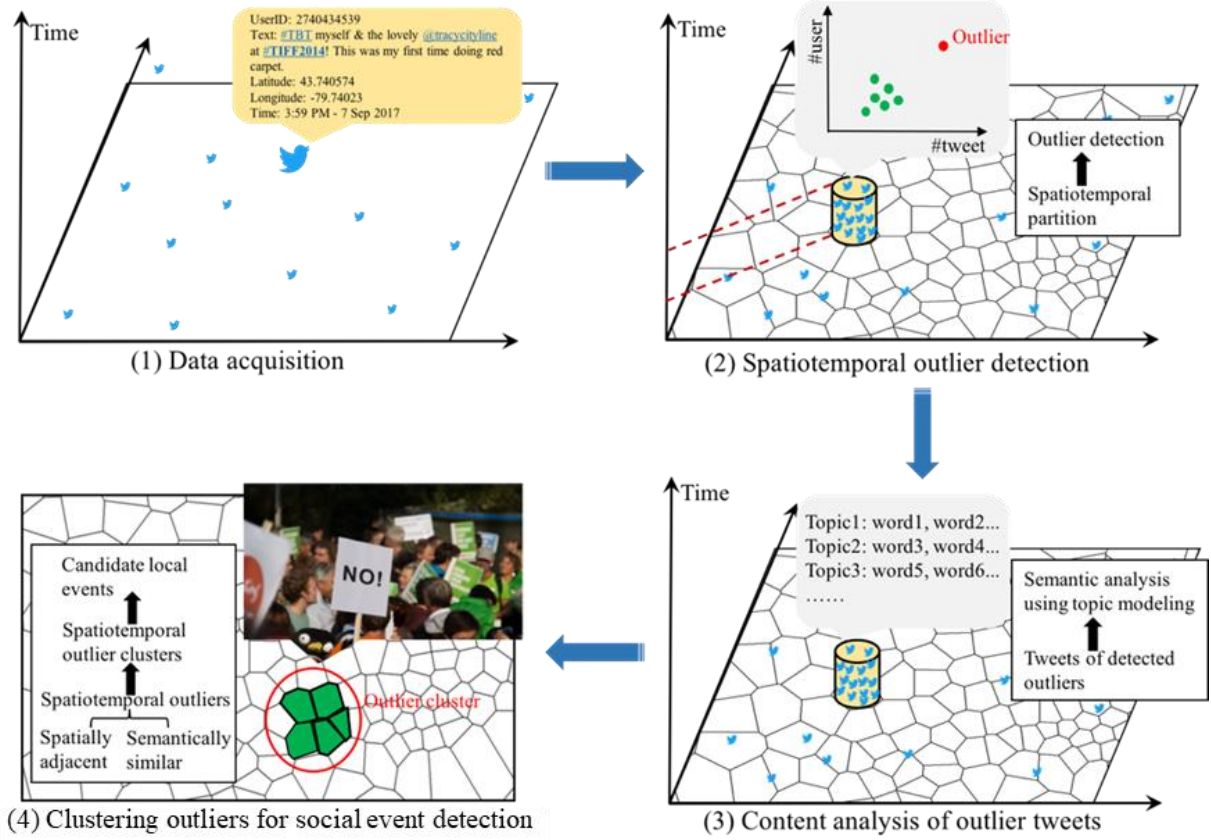


Figure 3.2. The overall process framework for social event detection

3.2.2 Data acquisition

Data in this study was accumulatively collected using Twitter Streaming API², which allows users to retrieve at most 1% of all the real-time tweets that meet user defined parameters for free (Huang, Fan, & Zipf, 2017; Morstatter et al., 2013). These parameters consist of 1) follow: up to 5,000 comma-separated user IDs, 2) track: up to 400 comma-separated keywords, and 3) locations: up to 25 0.1-360 degree bounding location boxes, which can be combined with an “OR” operator.

² <https://developer.twitter.com/en/docs/api-reference-index>

The returned Tweet objects³ were encoded using JavaScript Object Notation (JSON) that are composed of a pair of named attributes and associated values. Each Tweet object can be encapsulated with more than 150 attributes covering tweets and users.

In this study, Twitter streams were obtained by defining the “locations” parameter with a bounding box enclosing Toronto, Canada. It enables developers to crawl tweets that are posted within the area predefined by a geo-bounding box in a real-time manner. With this spatial restriction, only the tweets geo-located within the bounding box were collected. A tweet is allowed to be geotagged with a pair of GPS coordinates or a Twitter place indicating a point of interest (POI) (e.g., Dundas Square, Toronto, Canada) if the location service is turned on while posting. GPS coordinates refer to a pair of longitude and latitude, while a Twitter place is defined as an enclosing box surrounding this place⁴. With respect to testing whether a tweet matches the location query based on valid geotags, a tweet is included if its pair of longitude and latitude falls in the bounding box or the geographical region corresponding to the tagged place has intersection with the bounding box. In other words, a tweet that is not geotagged with GPS coordinates or a Twitter place is discarded by this query.

The tagged GPS coordinates are with relatively high location accuracy, especially for mobile phones, and the estimated median horizontal error ranges from 5m to 8.5m (Zandbergen & Barbeau, 2011). However, Giridhar, Abdelzaher, George, and Kaplan (2015) found that that only 2%-3% of all tweets were geo-located, which poses certain challenges for researches adopting Twitter as valid data source for experiments due to data sparsity. In fact, the way Twitter data is collected acts as a major indicator for the geotagged percentage. It illustrates that a geo-bounding box-based query method contributes to a higher percentage since crawled tweets meet the geospatial requirement for GPS coordinates (Ozdikis et al., 2017). Among all geotagged tweets collected through Twitter Streaming API with drawing a geospatial bounding box, as many as 30.11% and 27.49% of them originated from Asia and North America, respectively (Morstatter et al., 2013). This indicates that Asia and North America have higher chance of retrieving geotagged tweets from sampling Twitter streams. In this study, only the tweets tagged with GPS coordinates

³ <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html>

⁴ <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects>

were taken into consideration for event detection, since the location information indicated by the attached Twitter places that can be freely determined by users seems to be lack of reliability (Fujisaka et al., 2010).

In order to extract abundant information for detecting traffic events, a set of attributes serving as spatial, temporal, and semantic basis were selected. They include “user” (unique identifier of user), “created_at” (UTC timestamp), “coordinates” (a pair of coordinates indicating the exact location), “place” (four pairs of vertex coordinates of the polygon enclosing the Twitter place), and “text” (short message with limited characters).

3.2.3 Spatiotemporal outlier detection

(1) Spatiotemporal partition

In order to retrieve social events from a large geographical region, spatial partition is first conducted to identify the abnormal geographical pattern of crowds in each sub-region. There are multiple partition methods, which generally include grid-based method, administrative district-based method, and cluster-based method. This study aims to capture groups of points that have the potential to compose spatial clusters based on their geolocations in the target region. Both the grid-based method and the administrative district-based method simply split the target region into equal cells and administrative districts (e.g., municipal election wards), respectively, but neglect the unbalanced spatial distribution of tweets indicating crowd activity patterns. Thus, these two methods are not suitable for accurately extracting active regions enclosing social events for event location estimation.

The cluster-based method can not only reflect the geographical distribution of points, but also deal with heterogeneous regions flexibly, which is an appropriate choice for spatial partition in the social event detection process. Due to its simplicity and flexibility, the k-means clustering method was applied in this study (Kanungo et al., 2002; Lloyd, 1982). This unsupervised learning method is adjustable to group similar data points into one cluster based on the geographical distributions of geotagged Twitter dataset. Only the number of grouped clusters K needs to be adequately set

for performing k-means method. The way how the number K is determined highly depends on specific applications. Given any initial set of K points as the centroids of clusters, this method assigns each data point to the cluster that owns the least squared Euclidean distance between the data point and the centroids. After one iteration, the K different centroids emerge and the distance are recalculated. The process repeats until K clusters are composed, i.e., when the algorithm becomes convergent and the assigned data points to each cluster becomes stable. As a result, K sub-regions are then generated using Voronoi diagram based on the centers of the k-means results (i.e., a pair of longitude and latitude). With respect to the temporal perspective, both weekdays and weekends were split into four equal time intervals to monitor geographical pattern of crowds in this research, which were 1) Morning: 6am – 12pm, 2) Afternoon: 12pm – 6pm, 3) Evening: 6pm – 12am, and 4) Night: 12am – 6am.

(2) Outlier detection

After being able to organize Twitter data into spatiotemporal groups, it is necessary to determine how to extract outliers from these groups by investigating their geographical pattern. An outlier is regarded as an abnormal observation that significantly distinguishes from other observations, which causes the assumption that it was created in a different manner (Hawkins, 1980). As defined in Section 3.1, a spatiotemporal outlier is detected as an abnormal geographical pattern in this case. The number of tweets and the number of users are considered as two necessary variables to generate the geographical pattern. In the case without examining the number of users, an abnormal burst in the number of tweets may not refer to a spatiotemporal outlier that indicates a potential social event if only a few users post more tweets regarding personal matters, while the number of users keeps stable. Given that some users adapt to posting a significant number of tweets every day, others only post few on national holidays (e.g., Christmas) that do not belong to social events. It is possible that the overlap of these two types of users' posting time results in a burst in the number of users but no significant change in the number of tweets. This case can be wrongly identified as a social event if the number of users is selected as unique variable for geographical pattern measurement. Therefore, investigating both the number of tweets and number of users for outlier detection performs as a better strategy to capture social events.

R. Lee and Sumiya (2010) and R. Lee et al. (2011) adopted boxplot method, which is a univariate outlier detection method, to detect outliers of the two variables. A spatiotemporal outlier was identified if two variables were simultaneously detected as outliers. However, Acuna and Rodriguez (2004) indicated that an instance containing several variables did appear to be an outlier but it was not an outlier in each variable. Thus, a multivariate and non-parametric technique is in need for the purpose of outlier detection in this study since the author does not have prior knowledge about the distribution of the two variables.

The approach that uses robust estimation of the Mahalanobis Distance is a good solution meeting the author's requirements, which has been proven to work well for identifying scattered outliers (Rocke & Woodruff, 1996). It takes the correlation between multiple variables into account as it is computed based on the inverse of the variance-covariance matrix (De Maesschalck, Jouan-Rimbaud, & Massart, 2000). Therefore, to be able to compute the Mahalanobis Distance, the variance-covariance matrix of p variables is first constructed as:

$$C_x = \frac{1}{n-1} (X_c)^T (X_c), \quad (1)$$

where X_c is the column-centered matrix ($X - \bar{X}$). X is the matrix containing n instances in the rows measured by p variables, and \bar{X} is the matrix with the means of each variable as columns. In the case of p equals to 2, namely two distinct variables, x_1 and x_2 , the variance-covariance matrix C_x is calculated as

$$C_x = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad (2)$$

where $\rho_{12}\sigma_1\sigma_2$ is the covariance between x_1 and x_2 , and σ_1^2 and σ_2^2 are the variances of the values of x_1 and x_2 , respectively. Accordingly, the inverse of C_x is

$$C_x^{-1} = \begin{bmatrix} \sigma_2^2 / \det(C_x) & -\rho_{12}\sigma_1\sigma_2 / \det(C_x) \\ -\rho_{12}\sigma_1\sigma_2 / \det(C_x) & \sigma_1^2 / \det(C_x) \end{bmatrix} \quad (3)$$

where the determinant of the variance-covariance matrix $\det(C_x) = \sigma_1^2 \sigma_2^2 (1 - \rho_{12}^2)$.

The Mahalanobis Distance for each instance x_i , which is measured by two variables, x_1 and x_2 , is then

$$MD_i = \sqrt{(x_i - \bar{x})C_x^{-1}(x_i - \bar{x})^T}$$

$$= \sqrt{\left(\frac{x_{i1} - \bar{x}_1}{\sigma_1}\right)^2 + \left[\left\{\left(\frac{x_{i2} - \bar{x}_2}{\sigma_2}\right) - \rho_{12}\left(\frac{x_{i1} - \bar{x}_1}{\sigma_1}\right)\right\} \frac{1}{\sqrt{1 - \rho_{12}^2}}\right]^2}, \quad (4)$$

where x_{i1} and x_{i2} are the values of the two variables of the instance x_i , respectively. More details about the Mahalanobis Distance algorithm can be found in De Maesschalck et al. (2000). An instance is identified as an outlier if its Mahalanobis Distance is abnormally large.

As described above, both the number of tweets and the number of users were selected as two necessary indicators (i.e., variables) for spatiotemporal outlier detection, where an outlier is extracted if its Mahalanobis Distance to the center (i.e., mean values of variables) is larger than a certain threshold λ . This threshold can be defined by a function of z-scores, if variables follow a normal distribution, or a statistical technique specifying a percentile (Warren, Smith, & Cybenko, 2011). In this study, the latter method was adopted to define a percentile for outlier detection without knowing the distinct distributions of variables in all active regions during all time intervals. Based on the concept of hypothesis test in statistics, it makes the null hypothesis that a new instance and the reference samples come from populations with equal means of Mahalanobis Distance. The new instance is determined as an outlier if the null hypothesis is rejected. In this manner, the possibility of rejecting the null hypothesis is the percentile determined for outlier detection. An instance is identified to be an outlier if the probability of its occurrence is less than a given percentile and it falls in the low probability area (S. Chen, Wang, & van Zuylen, 2010; E. S. Park, Turner, & Spiegelman, 2007; Y. Zhang, Meratnia, & Havinga, 2010).

3.2.4 Content analysis of outlier tweets

The detected spatiotemporal outliers are further analyzed from semantic perspective in order to automatically describe the candidate event in an efficient way. This work aims at enabling people to have a general understanding of the event at first sight. Summarizing the outlier tweets via several representative topics seems to be an available way. Topic modeling is widely used for text mining in machine learning-based applications, which can infer the topics of a text corpora in a cost-effective way (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2002). Topic modeling essentially represents a document with a set of topics, and each topic is indicated by a mixture of words. It is generally a process of distributing probabilities, where a document is summarized by a probability distribution over topics, and each topic is explained by a probability distribution over words. In this thesis, outlier tweets correspond to documents in training a topic model. The candidate event is represented by the distributions of topics over documents. Furthermore, the content of candidate events is identified by investigating the distribution of topics over documents and the distribution of words over each topic at the same time.

Multiple techniques intend to generate probability distributions, among which LDA is a typical topic model. As such, LDA is used for content analysis of outlier tweets in this study. The procedure of distribution estimation by LDA mainly includes two steps. **Step 1:** estimating the probability distributions of topics (t) over document (i.e., outlier tweets reflecting candidate event), $P(t)$, and **Step 2:** estimating the probability distributions of words (w) over each topics (t) that is involved in the document, $P(w/t)$. The probability of the k^{th} word in a candidate event can be measured as:

$$P(w_k|e) = \sum_{m=1}^T P(w_k|t_k = m)P(t_k = m|e), \quad (5)$$

where t_k refers to the topic where the k^{th} word is included, $P(w_k|t_k = j)$ represents the possibility of the k^{th} word in topic m , $P(t_k = m|e)$ is the possibility of the k^{th} topic sampled for the k^{th} word token for candidate event e .

The whole process of the LDA model, together with the explanation to the corresponding parameters, is illustrated in Figure 3.3. The shaded w means that word tokens can be observed directly, while other parameters such as t , $\varphi(t)$, and $\delta(e)$ are latent and need to be estimated. The direction of arrows indicates the dependency relationship between two parameters. Gibbs sampling, a Markov chain Monte Carlo (MCMC)-based technique, is able to generate representative topics for text corpora of large size with high efficiency (Steiyvers & Griffiths, 2007). Therefore, this technique is chosen to do inference for LDA model using Twitter data in this study. A detailed introduction to the mechanism of Gibbs sampling was presented by Griffiths and Steiyvers (2004).

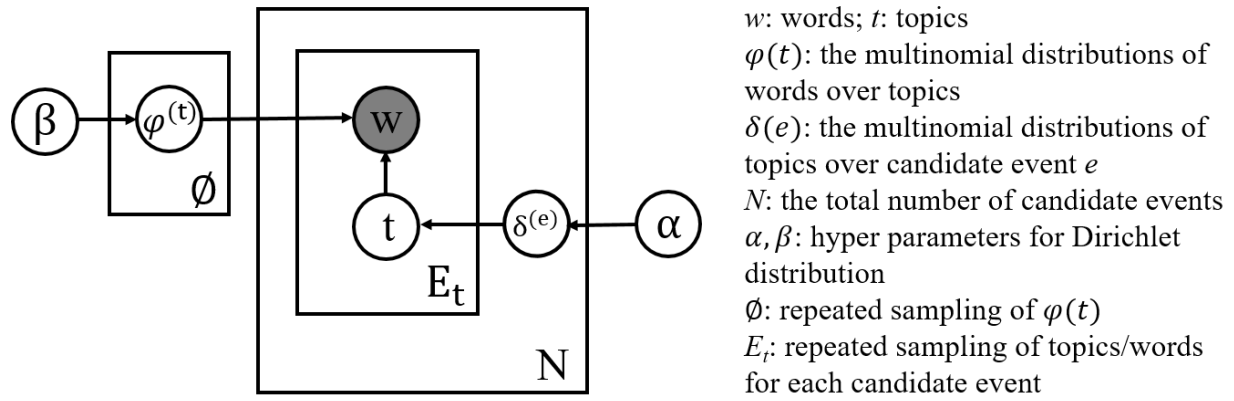


Figure 3.3. Process of Latent Dirichlet Allocation (LDA) model

In order to train a topic model, the number of topics needs to be predefined. The perplexity is selected as an indicator for this predefinition since it is a standard metric measuring the generation performance within the domain of machine learning (Blei et al., 2003). The main idea of perplexity is that the likelihood of the dataset exponentially decrease in the training process. A lower perplexity value reveals a better performance of the machine-learning model. In this study, the perplexity is defined in Equation (6) based on a training dataset of E groups of outlier tweets.

$$\text{Perplexity} = \exp \left\{ -\frac{\sum_{e=1}^E \log P(w_e | E)}{\sum_{e=1}^E E_t} \right\}, \quad (6)$$

where E_i represents the number of word tokens in each group of outlier tweets, w_e refers to the words of outlier tweets of group e , and $P(w_e|E)$ is computed by Equation (5).

3.2.5 Clustering outliers for social event detection

As indicated in Section 1.5, tweets indicating the same social event are geographically close and semantically coherent. Thus, there are cases of multiple outliers for the same event within a certain geographical distance. An approach integrating spatial clustering method with text similarity measurement is required to handle these occurrences heuristically. If detected outliers are spatially adjacent and semantically similar, they will be clustered as a social event.

Suppose that spatially adjacent outliers are first grouped into spatial clusters based on spatial clustering method (e.g., DBSCAN and k-means), within each spatial cluster, outliers with similar semantics are further separated to form new clusters to represent potential social events. It is possible that outliers in a new cluster are not spatially adjacent if the new cluster has a large spatial coverage or strip shape, which does not meet the requirement of spatial proximity for social event detection. To avoid the above false case, a clustering method incorporating spatial and non-spatial dimensions is adopted instead to examine spatial distance and text similarity at the meantime. An overview of this clustering method is shown in Figure 3.4.

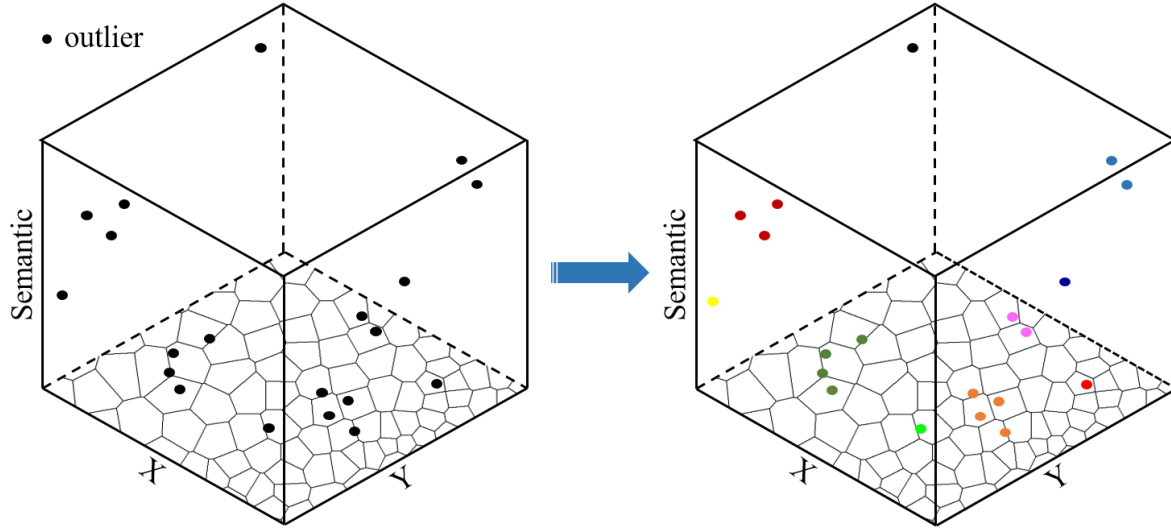


Figure 3.4. Overview of clustering outliers based on ST-DBSCAN method that simultaneously measures spatial adjacency and semantic similarity. Outliers of the same color indicate an outlier cluster.

The ST-DBSCAN method, which is a modified DBSCAN algorithm, was used to discover clusters from both spatial and non-spatial dimensions (Birant & Kut, 2007). Instead of using only one parameter *eps* to define distance between spatial data in DBSCAN method, ST-DBSCAN adopts two distance metrics, *eps1* and *eps2*, to measure similarity of data points by examining two density features. *Eps1* is the geographical distance of data points computed by spatial values, and *eps2* measures the similarity of non-spatial values. Thus, three parameters, *eps1*, *eps2*, and *MinPts* (i.e., the minimum number of data points to be contained to compose a cluster), are necessary to be defined for using ST-DBSCAN method to perform spatial and non-spatial clustering. The process of ST-DBSCAN is expressed in Figure 3.5. In this study, the spatial adjacency and semantic similarity between outliers were measured to compose clusters for social event identification.

```

Function ST-DBSCAN (D, eps1, eps2, MinPts)                                /* D is a dataset containing n data points */
    C ← 0                                                                    /* Initial cluster C */
    for unvisited P ∈ D
        P ← visited
        NeighborPts = regionQuery (P, eps1, eps2)    /* Return all points within P's neighborhood */
        if sizeof (NeighborPts) < MinPts
            P ← noise
        else
            P ← next cluster
            expandCluster (P, NeighborPts, C, eps1, eps2, MinPts) /* Expand cluster C based on center point P */

Function expandCluster (P, NeighborPts, C, eps1, eps2, MinPts)
    C ← P
    for P' ∈ C
        if P' ≠ visited
            P' ← visited
            NeighborPts' = regionQuery (P', eps1, eps2) /* Return all points within P''s neighborhood */
            if sizeof (NeighborPts') ≥ MinPts
                NeighborPts ← NeighborPts ∪ NeighborPts'
            if P' ∈ none cluster
                C ← P'

```

Figure 3.5. An explanation of ST-DBSCAN algorithm for grouping outliers (Birant & Kut, 2007)

To be specific, the centroid of each outlier region is extracted as a data point. *MinPts* is set to 1 since any point (outlier) is not considered as noise. In other words, those spatially close and semantically similar outliers are grouped to an outlier cluster representing a potential social event, otherwise each outlier stays separately to represent a potential social event. The geographical distance *eps1* is calculated using the coordinates (*X, Y*) of the centroids. *Eps2* refers to the semantic similarity between outlier tweets, which is measured by the cosine similarity method. The topic distributions of outlier tweets are expressed as vectors, so the similarity between two sets of outlier tweets can be calculated by the cosine value of the angle between the two vectors (i.e., corresponding topic distributions). Therefore, the semantic similarity between two sets of outlier tweets is calculated as follows:

$$\cos\theta = \frac{\sum_1^n (p \times q)}{\sqrt{\sum_1^n p^2} \times \sqrt{\sum_1^n q^2}}, \theta \in [0, 90] \quad (7)$$

where n is the number of topic set to represent the outlier tweets, p and q are corresponding topic distributions of the two sets of outlier tweets, and θ refers to the angle between the two vectors represented by topic distributions.

The value of $\cos\theta$ ranges from 0 to 1, where the larger it is, the more similar the two vectors are. In this study, the two sets of outlier tweets, between which $\cos\theta$ is larger than a certain threshold, are identified to be semantically similar. As such, ST-DBSCAN method will group outliers, which are located nearby within a distance of $eps1$, and meanwhile semantic similarity is larger than $eps2$, to an outlier cluster. Eventually, a spatiotemporal social event is represented by an outlier cluster where the spatial adjacency and the semantic similarity between outliers are simultaneously investigated. As shown in Figure 3.4, outliers with the same color represent a social event.

3.3 A classification-based approach for detecting traffic events

In this thesis research, the framework typically used in current studies was extended by adding association rules mining to extract positive tweets while discarding negative tweets after a list of keywords were used separately to extract potential tweets relevant to traffic. These association rule-filtered tweets were then classified into different types of traffic events based on the Naïve Bayes (NB), Support Vector Machine (SVM), and Logistic Regression (LR) methods. The location and time of detected events were inferred by referring to the tagged GPS coordinates and timestamps.

3.3.1 The overall workflow

As illustrated in Figure 3.6, the overall workflow of the classification-based method mainly includes collecting Twitter data, querying the raw tweets based on single keywords, preprocessing the queried tweets, mining the association rules that are embedded in positive tweets, conducting a query using a list of mined association rules, classifying the potentially traffic related tweets into different categories, and geocoding them to real world locations. These steps are discussed in detail in the following sections.

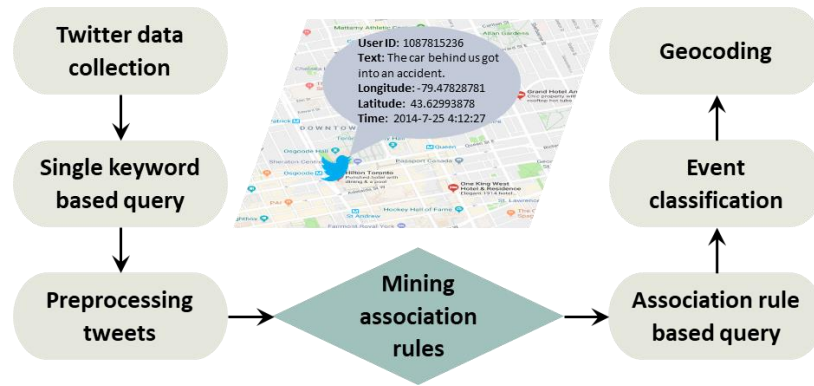


Figure 3.6. The overall workflow of detecting traffic events from Twitter data based on classification methods

3.3.2 Data collection and preprocessing

As explained in Section 3.2.2, Twitter data used in this study was also collected through Streaming API by defining a geo-bounding box. The collected Twitter data in JSON format was stored in MongoDB⁵, which is an open-source document database and leading NoSQL program. It arranges data with different fields according to the mentioned spatial, temporal, and semantic attributes, and supports full text search⁶ through matching stemmed keywords.

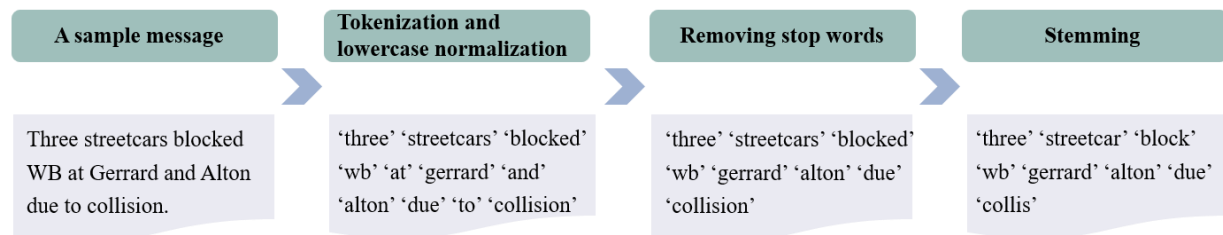


Figure 3.7. Preprocessing tweets using NLP tools

Tweets returned by the search of stemmed keywords were then processed using NLP tools. The NLP process is presented by providing a sample tweet in Figure 3.7. The tweet are tokenized into

⁵ <https://www.mongodb.com/>

⁶ <https://docs.mongodb.com/manual/reference/operator/query/text/>

a number of word tokens and normalized to lowercase. Stop words are then removed and the remaining words are transformed into stemmed format. These text preprocessing methods improves efficiency for exploring association rules since it filters out stop words that are not very useful, and removes the redundancy produced by variant words coming from the same stem.

3.3.3 Mining association rules

As declared in Section 1.5, association rules refer to the co-occurrence pattern of two or more words embedded in positive tweets. Apriori algorithm (Agrawal & Srikant, 1994), a representative algorithm for mining frequent itemsets to establish Boolean association rules in a large database of sales transactions, was applied in this study. It includes three parameters, namely support, confidence, and lift. Frequent itemsets refer to the sets of items of which the transaction support is above the minimum support. The number of items within one itemset ranges from two to k (i.e., an integer that is not less than two). Thus, an association rule can be established by k -itemset.

In this study, a word refers to an item, and a wordset (i.e., k words included in one association rule) refers to an itemset. When k equals to two, support, confidence, and lift are defined in Equation (8) – (11), where w_1 and w_2 are two different words. Support (w_1) and Support (w_2) indicate how frequently w_1 and w_2 appear in all of the tweets, respectively. Confidence ($w_1 \rightarrow w_2$) reveals how often the association rule that when w_1 appears, w_2 co-occurs is found to be true. Lift is a measure combining support and confidence. If lift is larger than 1, w_1 and w_2 are positively associated. Otherwise, w_1 and w_2 are not associated (lift = 1) or negatively associated (lift < 1).

$$\text{Support } (w_1) = \frac{\# \text{ tweets containing } w_1}{\text{total number of tweets}} \quad (8)$$

$$\text{Support } (w_2) = \frac{\# \text{ tweets containing } w_2}{\text{total number of tweets}} \quad (9)$$

$$\text{Confidence } (w_1 \rightarrow w_2) = \frac{\# \text{ tweets containing both } w_1 \text{ and } w_2}{\# \text{ tweets containing } w_1} \quad (10)$$

$$\text{Lift}(w_1 \rightarrow w_2) = \frac{\text{Confidence}(w_1 \rightarrow w_2)}{\text{Support}(w_2)} \quad (11)$$

The working process of Apriori algorithm is elaborated in Figure 3.8. The association rules A_k are generated by a collection of k -wordsets where the frequency of k words co-occurring is larger than the predefined minimum support. Two major steps are involved in this procedure.

```

 $A_1 \leftarrow \{1\text{-wordset}\}$ 
 $k \leftarrow 2$ 
while  $A_{k-1} \neq \emptyset$ 
     $CD_k = \text{apriori-gen}(A_{k-1})$  /* Candidate  $k$ -wordsets */
    for tweet  $t \in T$  /*  $T$  is the collection of all tweets */
         $CD_t \leftarrow \{cd \in CD_k \mid cd \subseteq t\}$  /* Candidates  $cd$  contained in tweet  $t$  */
        for candidates  $cd \in CD_t$ 
             $cd.\text{count} \leftarrow cd.\text{count} + 1$ 
     $A_k \leftarrow \{cd \in CD_k \mid \frac{cd.\text{count}}{\text{Num}(T)} \geq \text{minSupport}\}$  /*  $\text{Num}(T)$  is the total number of all tweets  $T$  */
     $k \leftarrow k + 1$ 
FinalResult =  $\bigcup_k A_k$ 

```

Figure 3.8. The process of Apriori algorithm

Step 1: Generating candidate k -wordsets (CD_k) by investigating the association rules mined in the previous pass (A_{k-1}) based on the apriori-gen function, which includes the join operation and the prune operation. First, in the join process, each $(k-1)$ -wordset is extended to k -wordset by adding one more words that are external to this $(k-1)$ -wordset, but internal to all tweets T . Next, in the prune process, the previously generated k -wordsets are deleted if their $(k-1)$ -subsets are not in A_{k-1} . CD_k is finally formed by the remaining k -wordsets.

Step 2: Counting how many times the candidates cd ($cd \in CD_k$) appear through scanning all the tweets T , and composing the association rules at k level A_k by adding cd if their support is higher than the minimum support. More details of the Apriori algorithm can be found in Agrawal and Srikant (1994).

3.3.4 Classifying traffic events into different categories

Identifying the event type described by tweets engages machine learning techniques of feature extraction and text classification, which are respectively elaborated as follows.

(1) Feature extraction

Term frequency-inverse document frequency (tf-idf) method, one of the most term-weighting schemes, was used to extract representative features to train classification models. Tf-idf is essentially determined by integrating the relative frequency of word w in a specific tweet S_t (term frequency) with the inverse proportion of w over the entire collection of tweets E_t (inverse document frequency) (see Equation 12). Specifically, term frequency $tf(w, S_t)$ and inverse document frequency $idf(w, E_t)$ are represented in Equation 13 and Equation 14, where $f(w, S_t)$ refers to the number of times w appear in S_t ($S_t \in E_t$), $f(w, E_t)$ refers to the number of tweets where w appears, and $len(E_t)$ is the total number of tweets in E_t . As such, the words with lower tf-idf values indicate that they are prevalent throughout the entire tweets collection and are not significant enough to be selected as features.

$$tfidf(w, S_t, E_t) = tf(w, S_t) \times idf(w, E_t) \quad (12)$$

$$tf(w, S_t) = f(w, S_t) \quad (13)$$

$$idf(w, E_t) = \log \frac{len(E_t)}{f(w, E_t)} \quad (14)$$

(2) Classifying traffic events into different categories

Three supervised text classification methods, which are NB (Manning, Prabhakar, & Schütze, 2010), SVM (Chang & Lin, 2011; Cortes & Vapnik, 1995), and LR (Fan, Chang, Hsieh, & Lin, 2008), were adopted to categorize tweets into class c_1, c_2, \dots, c_m , where m is the number of event types. Intuitively, all tweets are split into training dataset and test dataset, where the classification

models produced by the training dataset is used to predict the class label of unseen instances in the test dataset. In practice, based on the extracted features, a tweet S_t is accordingly represented by a numerical feature vector with the conditional probability of features $f_1, f_2, f_3, \dots, f_n$, where n is the dimension of feature space.

The NB method stems from the famous Bayes' theorem in probability. It assumes that the position of word w in tweet S_t does not matter, and words are independent in a given class c . The probability of a tweet S_t belongs to class c is computed as

$$P(S_t|c) \propto P(c) \prod_{1}^n P(f_n|c), \quad (15)$$

where \propto is the proportionality relationship, $P(c)$ is the prior probability of a tweet S_t appearing in class c , and $P(f_k|c)$ is the conditional probability of the k^{th} ($1 \leq k \leq n$) feature f_k appearing in a tweet of class c . Both $P(c)$ and $P(f_k|c)$ are estimated from training data based on Equation (16) and (17). N_{ct} is the number of tweets in class c , and N_t is the total number of tweets. $P(f_k|c)$ is measured by the relative frequency of f_k in the tweets that belong to class c . NT_{ct} is the number of occurrences of f_k in the training dataset regarding class c . Therefore, by computing the probability of a tweet in each class using Equation 15, this tweet is classified into the class C_t that has the highest probability (Equation 18).

$$P(c) = \frac{N_{ct}}{N_t} \quad (16)$$

$$P(f_k|c) = \frac{NT_{ct}}{\sum_{1}^n NT_{ct}} \quad (17)$$

$$C_t = \operatorname{argmax} P(S_t|c) \quad (18)$$

The SVM method is based on the structural risk minimization principle (Vapnik, 2013) in the computational learning domain. It aims at minimizing the possibility of true errors through generating an optimal hyperplane, which refers to the maximum marginal hyperplane best dividing

the dataset into different classes. In practice, this process is implemented by adopting the kernel function to transform the input data into a higher dimensional space, which is formulated by Equation (19). Thus, the problem that low-dimensional space data is difficult for classification is likely to be solved by adding more favorable dimensions.

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ \text{subject to} \quad & C^T \alpha = 0 \end{aligned} \tag{19}$$

$$0 \leq \alpha_i \leq R, i \in [1, L]$$

$$Q_{ij} = C_i C_j K(TV_i, TV_j)$$

In terms of the L -dimensional training vectors TV , e is a vector of all ones, C is a vector indicating class labels, Q is a L by L positive semidefinite matrix computed by the kernel function K , and R is the regulation parameter used for the determination of error tolerance. Consequently, the class of a tweet C_t' can be identified by the decision function

$$C_t' = \text{sign}(\sum_{i=1}^L C_i \alpha_i K(TV_i, TV) + b), \tag{20}$$

where b is a bias term represented by a real number.

The LR method estimates the probability of a tweet S_t in class (i.e., $P'(S_t|c)$) by measuring the weighted relationship among categorical dependent features and independent features. This is directly related to the logistic/sigmoid function $\sigma(f, \lambda)$ (Equation (21), λ_k refers to weight of the k^{th} feature f_k), which is different from NB method where the probability is computed by the integration of a likelihood and a prior probability as shown in Equation (15). The tweet is assigned to the class with the highest probability (Equation (18)).

$$P'(S_t|c) = \sigma(f, \lambda) = \frac{1}{1 + e^{-(f_1\lambda_1 + f_2\lambda_2 + f_3\lambda_3 + \dots + f_n\lambda_n)}} \quad (21)$$

3.4 A clustering-based approach for detecting traffic events

The tweets queried by the association rules were regarded as potential traffic events. If a certain number of association rule-filtered tweets concentrate around a location within a certain time period, it is likely that a real-world traffic event is occurring nearby since this actual event involves several people and/or draws their attention. For instance, the account “@680NEWSTraffic” is an organization account that reports the up-to-the-minute traffic updates in Toronto and Great Toronto Area (GTA). As shown in the capture in the upper left corner of Figure 3.9, it announced a road closure between Finch Avenue and Drewry Avenue along Yonge Street due to a traffic collision investigation at 4:52 PM on August 7, 2014. This closure covered the road with a length of 850 meters, where five relevant tweets were found around and posted about one or two hours away the event. It makes sense since it takes time for police to clear this traffic event. As such, an assumption was made that an actual traffic event occurs if a group of traffic related tweets were found spatiotemporally concentrated. A spatiotemporal clustering method was adopted to group these similar tweets into one cluster. In this manner, a traffic event was represented by a spatiotemporal cluster, where the event location and time were estimated by the average values of the tweets included. The event content was further automatically generated with representative terms.

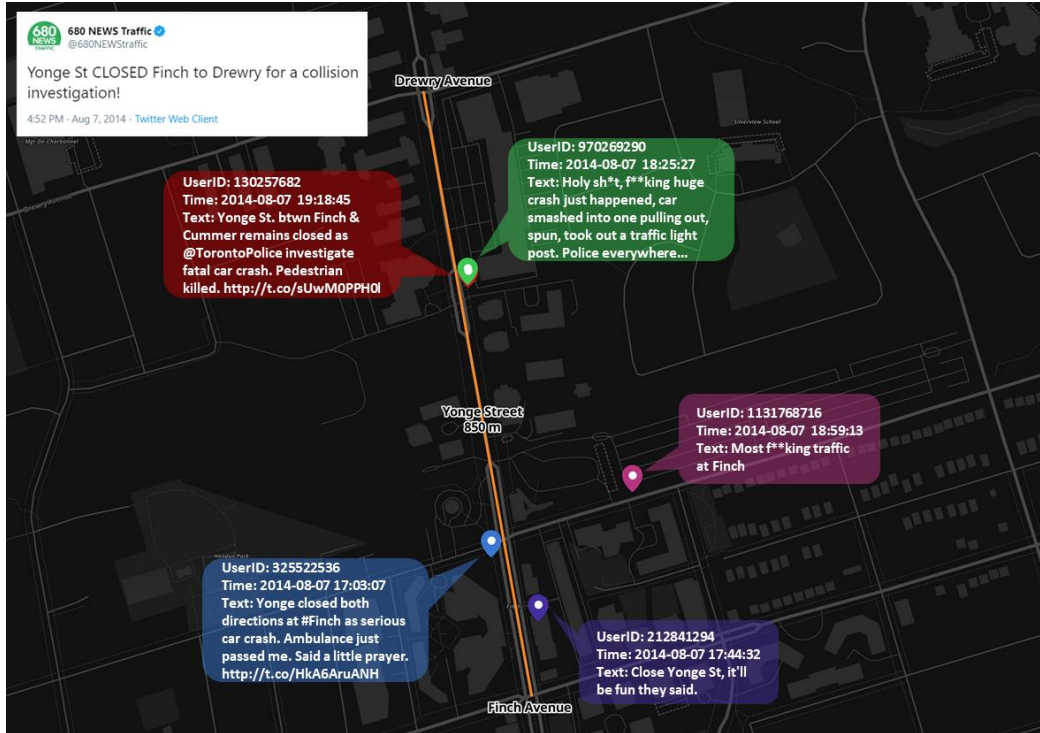


Figure 3.9. An example showing the spatiotemporal characteristics of traffic relevant tweets

3.4.1 The overall workflow

Based on the above mentioned assumption, the association rule-filtered tweets that were potentially relevant to traffic matters were spatiotemporally grouped into different clusters for traffic event identification. Instead of inferring the event content to a predefined category as the classification-based method, the clustering-based method automatically generated what the detected event referred to in the real world with a number of representative terms through a graph-based model. This detection process engaging spatiotemporal clustering as well as semantic analysis is shown in Figure 3.10.

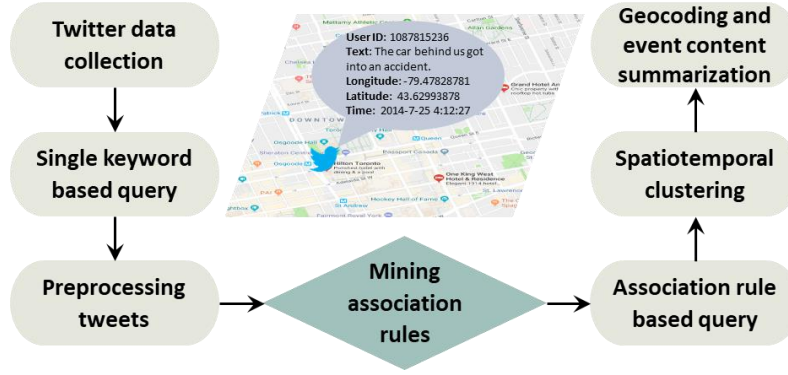


Figure 3.10. The overall workflow of detecting traffic events from Twitter data based on clustering method

3.4.2 Spatiotemporal clustering for traffic event identification

Similar to Section 3.2.5, the ST-DBSCAN method was also applied to conduct the spatiotemporal clustering for traffic event detection. A major difference existed that $eps2$ referred to temporal distance instead of semantic similarity in this study. Thus, a cluster was composed by a minimum number of points (i.e., $MinPts$) that are located nearby within a distance of $eps1$, and meanwhile temporal distance is less than $eps2$.

In this study, $eps1$ and $eps2$ were estimated by referring to the sorted k -dist graph, where k equals to $MinPts$. Specifically, with respect to each data point (i.e. a tweet), the distance to its k^{th} nearest neighbour is computed. These k -distance values are sorted in a descending order to compose the sorted k -dist graph, where the value indicating the first “valley” is selected as the threshold (i.e., $eps1$ or $eps2$) to separate noise data from cluster data (Ester, Kriegel, Sander, & Xu, 1996). As demonstrated in Figure 3.11, a small number of data points with higher 4-dist values on the left side of the threshold point are considered noise points, while those points associated with lower 4-dist values on the right side of the threshold point are used for clustering.

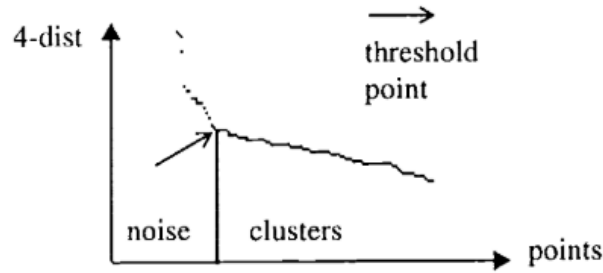


Figure 3.11. Sorted 4-dist graph by Ester et al. (1996)

3.4.3 Event content summarization

With regard to each cluster referring to a traffic event, the content of the cluster was automatically summarized with a set of representative terms that best described the tweets in this cluster. The set of terms were extracted by an unsupervised graph-based ranking model named TextRank (Mihalcea & Tarau, 2004), which is derived from the well-known Google's PageRank (Brin & Page, 1998). TextRank holds a competitive performance for keyword extraction and short text summarization in natural language applications, since it does not require deep linguistic knowledge nor domain or language specific annotated corpora, which makes it highly portable to other domains, genres, or languages (Mihalcea & Tarau, 2004).

First, the tweet texts in each cluster are tokenized into word tokens, which are further annotated with part of speech tags. These tagged words are used as nodes, and an edge is added between the tagged words if they co-occur within a window of maximum N words to compose a graph. N can be set anywhere from 2 to 10 words. As the length of the collected tweets in this research was limited to 140 characters, N was set as 10 in order to cover the co-occurrence relation between words within each tweet.

TextRank essentially determines the importance of a node within a graph based on the global information recursively retrieved from the entire paragraph instead of the local information specified to a node. Given one node connects to another one through an edge, it is basically casting a vote for that other node. Taking the global graph into account, the importance of a node not only

relies on the votes casting for it, but also the nodes casting these votes. Thus, the score indicating the importance of a node N_i is computed by

$$S(N_i) = (1 - d) + d * \sum_{j \in In(N_i)} \frac{1}{|Out(N_j)|} S(N_j), \quad (22)$$

where d is a damping factor that is usually set as 0.85 (Brin & Page, 1998), $In(N_i)$ refers to the nodes that point to the node N_i (i.e., predecessors), and $Out(N_j)$ refers to the nodes that N_j points to (i.e., successors). Within the undirected and unweighted graph, the score of each node is initially set as 1. Equation (22) keeps running on the graph for several iterations (usually 20 to 30 iterations) until the result becomes convergent.

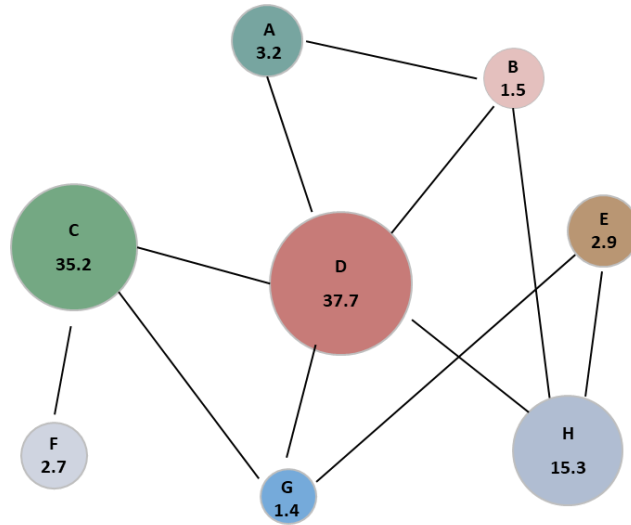


Figure 3.12. A sample graph generated by TextRank algorithm

By ranking the scores of all nodes, the K nodes (i.e., words) with the top K scores are obtained as potential keywords for post-processing, where the constant K is usually set to be from 5 to 20. During post-processing, all the K keywords are marked in the tweet texts, and multi-word keywords are composed by the combination of sequences of adjacent keywords. For example, in the tweet “*Traffic accident involving a charter bus and multiple vehicles on WB Hwy 118 at Madera Rd*”, if the adjacent words “*traffic*” and “*accident*” are selected as potential keywords by

TextRank model, they will be combined into one single keyword “*traffic accident*”. Finally, a list of representative keywords including single-word keywords and multi-word keywords within in the graph are generated to summarize the content of detected clusters. A sample graph built from a cluster is illustrated in Figure 3.12, where keywords *A* to *H* are used to represent the content of this cluster.

3.5 Discussion

3.5.1 Social event detection

To prepare data basis for efficiently retrieving social events from geosocial media data, the research area was split into sub-regions, where a spatial clustering method was conducted based on the data itself. This data-driven method fully measures the spatial distribution characteristics of geosocial media data. That a day was equally split into four time intervals may result in certain time offset comparing with the actual occurring time.

In the process of outlier detection by selecting a threshold for Mahalanobis Distance method, selection of a lower percentile may result in more false negative detections (i.e., the events detected as non-real-world events but they actually occur), where a portion of real-world social events are missed by the small number of outliers. Selection of a higher percentile may involve more false positive detections (i.e., the events detected as real-world events but they actually do not occur), where some detected outliers are not significant enough to indicate the occurrence of social events in the real world. Thus, a sensitivity analysis can be conducted to examine how the number of outliers changes with different settings of percentile values for Mahalanobis Distance method.

As summarizing event content based on topic modeling method, determining the number of topics by referring to the perplexity distribution avoids the deficiency that may exist if the number was empirically selected. As indicated in Section 3.2.5, an outlier cluster was composed if the semantic similarity between outliers was larger than *eps2* within a geographical distance of *eps1*. The minimum number of outlier grouping an outlier cluster *MinPts* in ST-DBSCAN was set as 1, which may group all outliers into one big cluster given the threshold for geographical distance *eps1* was

set as a larger value and the threshold for semantic similarity eps_2 was set as a smaller value. To avoid this false case, these two parameters can be estimated by reviewing all computed values and taking the delimitation value as reference distinguishing similar outliers from all outliers.

3.5.2 Traffic event detection

In the thesis, two approaches were proposed for extracting traffic events from geosocial media data, namely the classification-based approach and the clustering-based approach. The Apriori algorithm applied in association rules mining were both used in these two approaches. The key to the performance of this algorithm is the selection of thresholds for the minimum support and the minimum confidence. A higher threshold for the minimum support only taking those words with extremely high frequency into account misses some words for mining association rules, while a lower threshold results in redundant words and increases processing time. In correspondence, a higher threshold for the minimum confidence only returns word combinations with extremely significant correlation and some useful information is missed, while a lower threshold returns some word combinations that are not significant enough to be efficiently used. The selection of these two thresholds highly depends on the requirements of researched applications.

To conduct the classification-based method, without a clear conclusion about which method performs best in text classification, a potential solution may be testing several widely used method for investigation. Due to a number of classification methods, it is not feasible to check all existing classification models to review their performance. Instead, a few representative methods that were most frequently used in researches regarding traffic event detection, i.e., SVM, NB, and LR (refer to Table 2.3), were selected to illustrate how the classification-based method works. The results obtained by the one that holds the best performance can be used for validation with the ground truth data to represent the performance of the classification-based method.

Based on the clustering method integrating spatial dimension and non-spatial dimension used for spatial-semantic clustering in the social event detection study, another way to estimate parameters was adopted through k -dist graph regarding spatial and temporal perspectives in the clustering-based method for traffic event. The graph-based TextRank model was selected to generate top-

scored words or phrases to summarize event content with consideration of global information retrieved from the entire graph and nodes rather than the frequency-based method, e.g., tf-idf method, which simply returns the most frequently used keywords by sorting their frequency of appearing in the clustered tweets indicating traffic events. Other short text summarization techniques, such as adjusted topic modeling for tweets analysis (Hong & Davison, 2010), can be further tested.

Essentially, it is possible that traffic events occurring during grand festivals (e.g., Christmas) be detected through social event detection method. For instance, many people in Toronto tend to drive to Niagara Falls to celebrate Christmas Eve, resulting traffic anomalies on connected highways with a higher number of tweets and users than usual. This burst in the number of tweets and number of users can be detected as outliers by social event detection method. The content of outliers is further analyzed to infer what is happening, based on which traffic anomalies can be identified. After all, such types of traffic events can be detected by either social event detection method or traffic event detection method. The event location and time estimated by two different methods can be further investigated to see if there exists any significant difference.

Chapter 4 Results and analysis

This chapter presents the results of implementing and testing the approaches proposed in Chapter 4 using one-year Twitter data collected in Toronto, Ontario, Canada from April 1, 2014 to March 31, 2015. The 2014 Toronto International Film Festival and traffic anomaly detection in Toronto area were selected as case studies to evaluate the performance of the proposed approaches. Through comparing with the ground truth data, the detection results were quantitatively measured by certain evaluation metrics, followed by a discussion on the achievements obtained and limitations identified.

4.1 Social event detection based on the spatial-temporal-semantic approach

The proposed method was tested in detecting the events of 2014 Toronto International Film Festival (TIFF), which was held in Toronto, Ontario, Canada from September 4 to 14, 2014⁷. It is an annual festival that started from 1976 with the goal of collecting the best films all over the world and sharing them with interested audiences in Toronto. TIFF has become one of the most prestigious film festival in the world, attracting over 480,000 people annually. TIFF Bell Lightbox is a core destination that not only runs opening ceremony, but also provides live film events, interactive gallery and workshops⁸.

Table 4.1. Venues of TIFF 2014

Venue name	Address	Longitude	Latitude
Roy Thompson Hall	60 Simcoe St	-79.386436	43.64664
Princess of Wales Theatre	300 King St W	-79.389165	43.64687
Elgin and Winter Garden Theatre Centre	189 Yonge St	-79.379483	43.653124
Ryerson Theatre	43 Gerrard St E	-79.379948	43.659572
Scotiabank Theatre	259 Richmond St W	-79.391235	43.648949
Bloor Hot Docs Cinema	506 Bloor St W	-79.410575	43.665575

⁷ https://en.wikipedia.org/wiki/2014_Toronto_International_Film_Festival

⁸ https://en.wikipedia.org/wiki/Toronto_International_Film_Festival

Isabel Bader Theatre	93 Charles St W	-79.392385	43.667357
Jackman Hall (AGO)	317 Dundas St W	-79.392679	43.653652
TIFF Bell Lightbox	350 King St W	-79.390481	43.646517
Glen Gould Studio	250 Front St W	-79.388449	43.64479

Totally ten venues located in downtown Toronto were selected for film screening during TIFF 2014. Their geographical information is summarized in Table 4.1. The performance of the proposed method was tested through investigating how many detected social events were true TIFF events during TIFF 2014 by comparing with the schedule of each venue, which included information about start time, end time, film title, and director and so on.

4.1.1 Study area and dataset

Twitter data posted in Toronto, Canada was collected for study based on a geo-bounding box through Twitter Streaming API. As shown in Figure 4.1, Toronto is located in Southern Ontario on the northwestern shore of Lake Ontario. It ranks first and forth in Canada and North America, respectively, when sorted by population⁹. 6,782,651 tweets tagged with GPS coordinates or places were obtained from April 1, 2014 to September 14, 2014, where data in June was missing due to server problems. 1,743,533 tweets tagged with GPS coordinates, namely around 25.7% of the obtained tweets, were used for analysis in this study. Considering there existed users (e.g., advertising robots) who kept posting tweets at an identical location, such spams were further filtered and 1,433,213 tweets were left. In this way, around 18% of GPS tagged tweets were discarded as noise.

⁹ https://en.wikipedia.org/wiki/List_of_North_American_cities_by_population

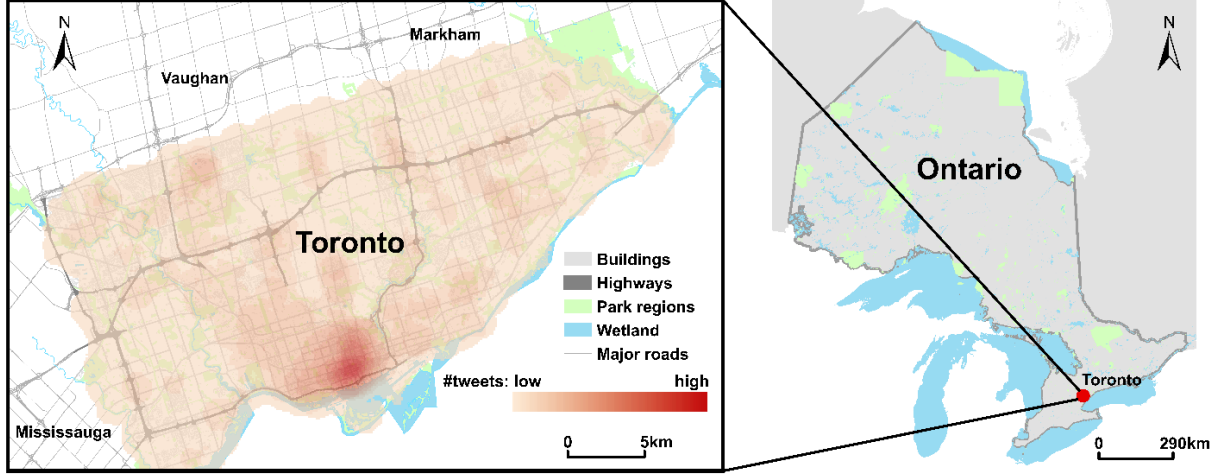


Figure 4.1. The geography of Toronto, Ontario, Canada with Twitter data used for social event detection

4.1.2 Spatiotemporal outliers

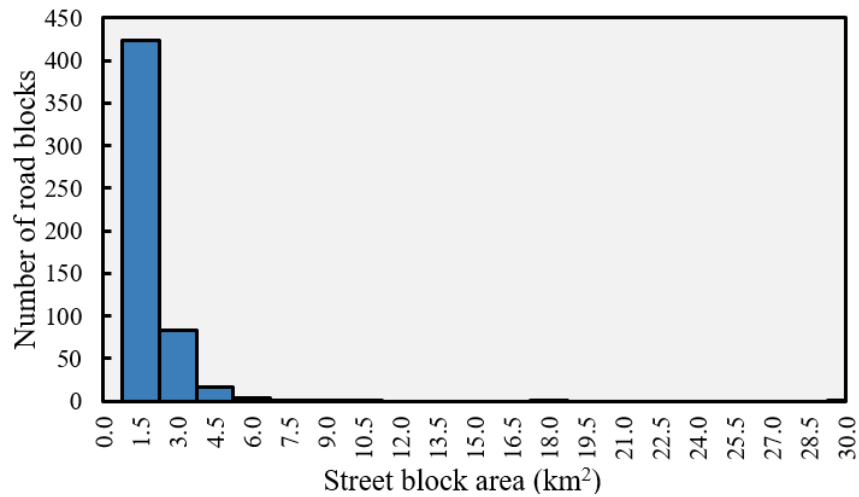
The research area was split into a set of geographical regions for spatiotemporal outlier detection using the collected Twitter data. As explained in Section 3.2.3, the research area was partitioned based on the k-means clustering method. The number of partitioned regions K was inferred by referring to the attributes of street blocks since the pattern people move or gather is closely linked to street blocks coverage, who are participants composing social events. K was estimated by

$$K = \frac{\sum S_G}{\bar{s}_b}, \quad (23)$$

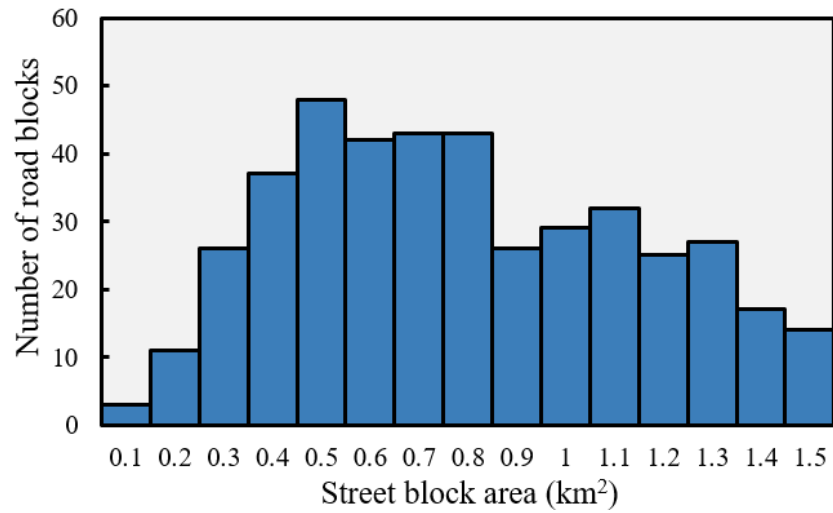
where $\sum S_G$ is the total geographical area of research area, and \bar{s}_b is the average area of investigated street blocks.

In Toronto area, there are totally 531 street blocks split by major roads (shown in Figure 4.1), among which the maximum and minimum area are 0.013 km^2 and 29.02 km^2 , respectively. Specifically, the frequency distribution of all street blocks area is shown in Figure 4.2 (a). About 80% of the street blocks area are less than 1.5 km^2 , which can be used to effectively represent the

overall distribution of street block area in Toronto by discarding the other 20% as noise. The filtered street blocks (i.e., around 80%) were further investigated with frequency distribution in Figure 4.2 (b). It illustrates that the number of street blocks whose area are below 1.5 km^2 distributes relatively evenly. The average area of street blocks in Toronto was then computed as 0.653 km^2 based on these street blocks involved in Figure 4.2 (b). The geographical area of Toronto is 630.2 km^2 . Therefore, the number of partitioned regions K was estimated by $630.2 / 0.653 \approx 1000$ in this experiment. The spatial partition result is presented in Figure 4.3. Obviously, downtown Toronto represented by the red polygon is with higher density of partitioned geographical regions, meaning that more tweets were posted in downtown area than other areas.



(a) Frequency distribution of all street blocks area



(b) Frequency distribution of street blocks area ≤ 1.5 km²

Figure 4.2. Frequency distribution of street blocks area in Toronto

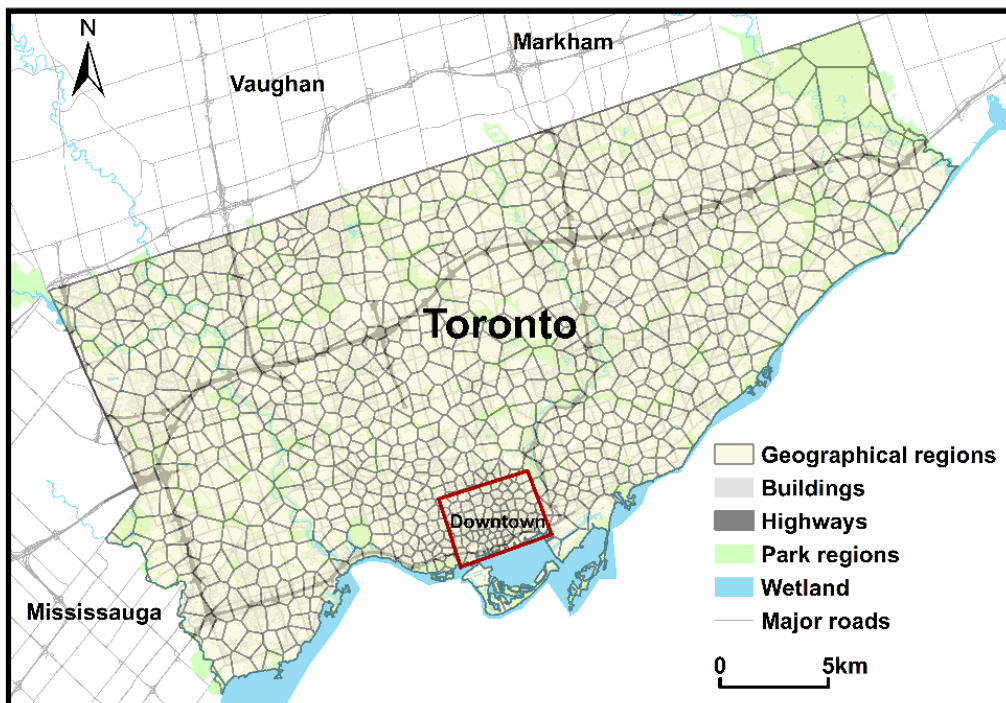


Figure 4.3. Voronoi representation of k-partitioned regions in Toronto (K=1000)

In order to improve processing efficiency, the inactive regions where the number of days with the number of the posted tweets below the first quartile (i.e., the 25th percentile) of all regions were filtered. In other words, the author discarded regions where people seldom posted tweets, since these regions were hardly involved in social events in the real world. Consequently, 259 “inactive” regions were discarded as noise, and 741 regions were left for study.

Within each “active” region, spatiotemporal outliers were detected by comparing with the geographical regularities of normal days in the morning, afternoon, evening, and night. The Twitter data collected from April 1, 2014 to August 31, 2014 (June data were missing due to server problems) were used as training data to separately estimate the geographical regularities on weekdays and weekends in term of four time intervals (i.e., morning, afternoon, evening, and night). The data collected during TIFF event (September 4-14, 2014) were used for testing outliers based on the Mahalanobis Distance method, where the number of tweets and the number of users were selected as indicators (i.e., variables) for spatiotemporal outlier detection.

As explained in Section 3.5.1, the selection of percentile value directly affects the performance of Mahalanobis Distance method for outlier detection. Therefore, instead of empirically setting a percentile to separate normal instances and outliers as Park et al. (2007), a sensitivity analysis was done to examine how the percentile value ranging from 0.01 to 0.1 influences the number of detected outliers. As shown in Figure 4.4, the larger the percentile, the more outliers were detected. The number of outliers increases sharply when the percentile is less than 0.04. A short convergence appears as the percentile is between 0.04 and 0.05 comparing with the following near linear increase. In order to avoid false negative detection and false positive detection as much as possible, 5% of the instances that fell in low probability regions were identified as outliers to conduct spatiotemporal outlier detection in this study. As a result, 774 outliers were detected in these “active” regions during TIFF events. In other words, among four time intervals, a total of 774 regions were engaged in abnormal patterns indicating potential TIFF events, which would be further identified by investigating the content of tweets posed in associated time intervals and regions.

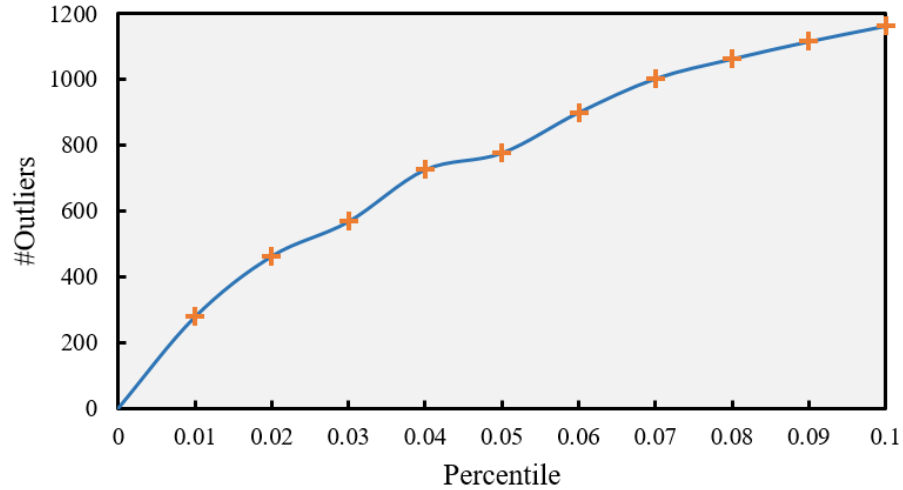


Figure 4.4. The number of outliers changes with different threshold of Mahalanobis Distance

A space-time map illustrating spatiotemporal distributions of the detected outliers is shown in Figure 4.5. TIFF event lasting 11 days consists of 7 weekdays and 4 weekends, thus the outlier distributions of the four time intervals were represented by 8 separate figures. The frequency of a region being detected as an outlier was counted during each time interval on weekdays and weekends, which were visualized using circles of different colors and sizes. The darker and bigger a circle is, the more times this region is detected as an outlier. It reveals that the detected outlier regions of high frequency are intensive in downtown Toronto, exactly where the TIFF events occurred. In spite of that, two more regions located in the northwest and northeast of Toronto, which corresponds to Humber College and Seneca College, respectively, were frequently detected as outliers on the weekday morning and afternoon. This phenomenon does make sense since a large number of students are back to school in early September after a long summer vacation. More tweets can be found expressing their excitement or complains about classes.

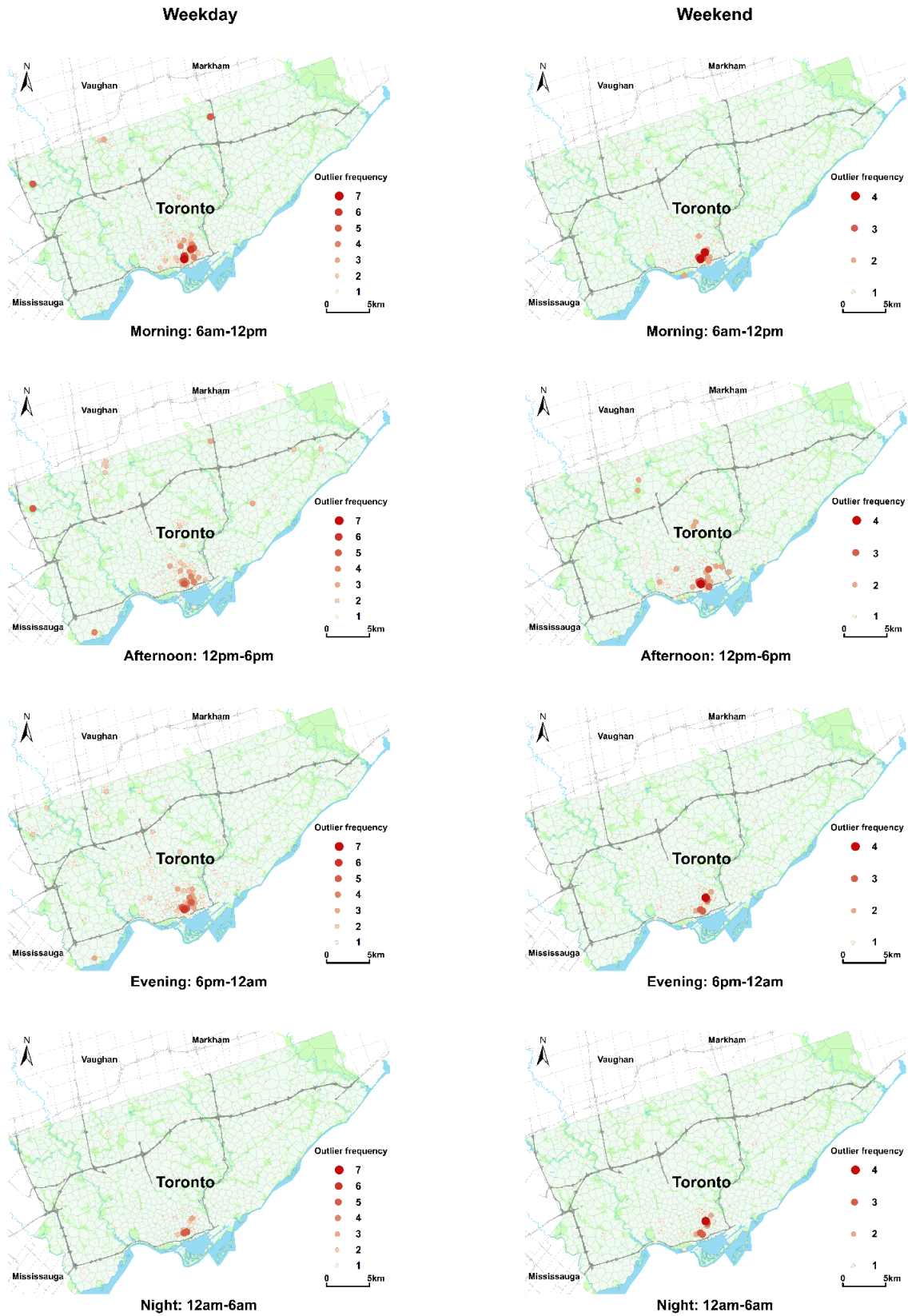


Figure 4.5. Spatiotemporal distributions of the detected outliers in Toronto

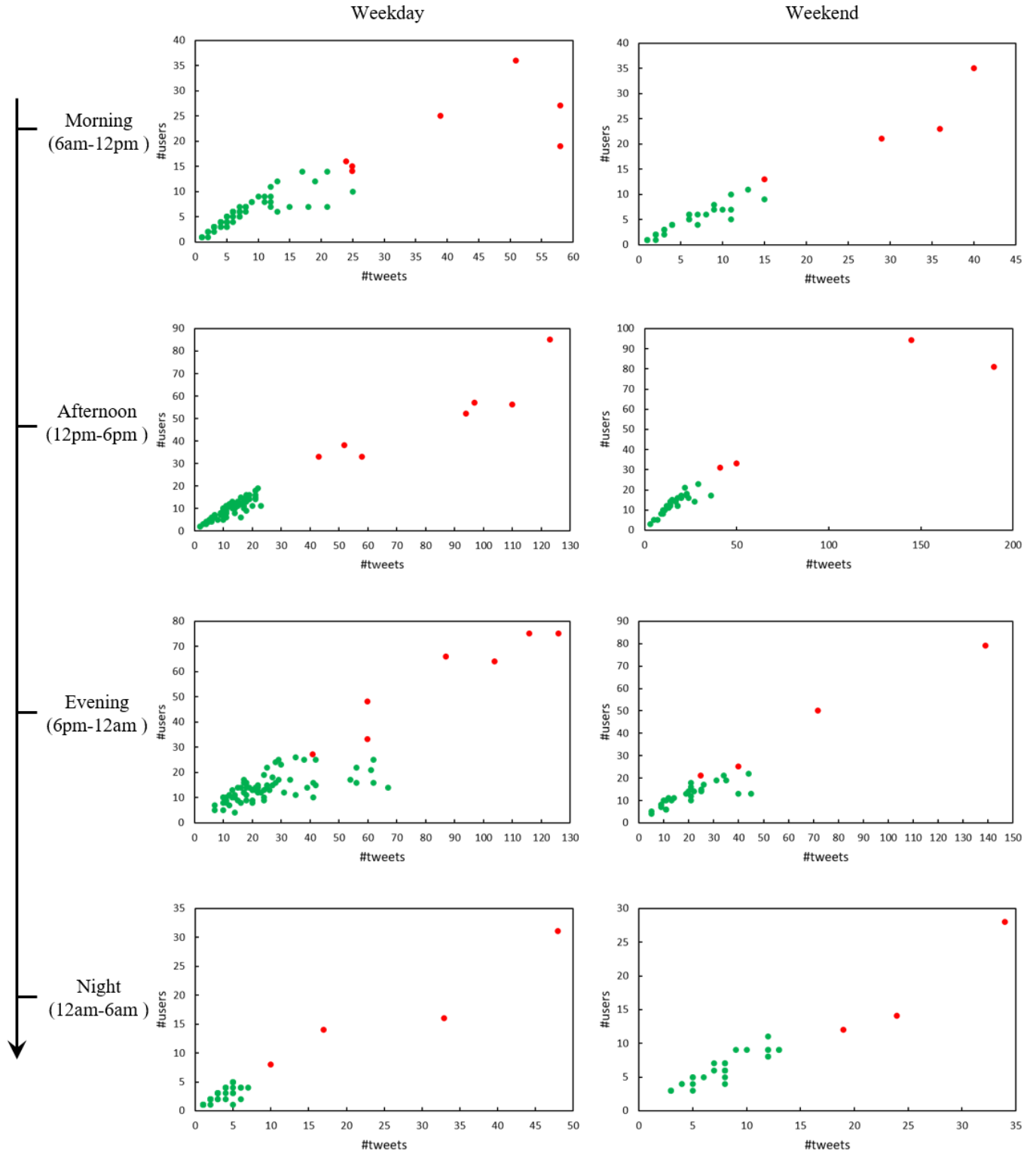


Figure 4.6. Detection of outliers (red point) in region #417 on weekday and weekend (Mahalanobis Distance)

Figure 4.6 presents an example of outlier detection result in region #417 in the four time intervals of weekday and weekend, namely 40 outliers were identified. Region #417 is a hot pot for TIFF events where two TIFF venues, namely TIFF Bell Lightbox and Princess of Wales Theatre, are located. The geographical regularities were represented as green points, while outliers were marked as red points. The detected outliers indicate a burst in the number of tweets and the number of users simultaneously (i.e., an abnormal geographical pattern), which is consistent with what was described in Section 3.2.3. For instance, in weekend evenings, two instances with slightly burst patterns were detected as outliers by measuring the correlation between the number of tweets and number of users, which may be missed by a univariate outlier detection method separately measuring the significant burst in the number of tweets and number of users. As such, it is obvious that the Mahalanobis Distance method works well for two-variable outlier detection.

4.1.3 Content of outlier tweets

Topic modeling was then implemented based on a Java library – Mallet (Mccallum, 2002) to infer the content of outlier tweets by investigating the topic distributions over outlier tweets and word distributions over topics. The number of topics set for training LDA model was estimated by calculating perplexity (Equation (6)), where 80% of outlier tweets were used for training and the rest were used for evaluating (Huang et al., 2017). The perplexity distribution over the number of topics is shown in Figure 4.7. It can be found that the perplexity becomes convergent (or decreases flatly) from 100 topics. However, a relatively sharp decrease exists between 190 and 220 topics, and it becomes stable after 220 topics. It is difficult to determine the optional number of topics to infer the topics involved in outlier tweets. Subsequently, both 100 topics and 220 topics were tried to do topic inferring. The word distributions over inferred topics were manually investigated, and it showed that more topics were with similar meanings by adopting 220 topics than 100 topics. To reduce the redundancy and highlight the distinction of inferred topics, the number of topics was defined to be 100 to conduct LDA topic model. According to Steyvers and Griffiths (2007), the hyper-parameters of Dirichlet prior α was set as 0.01, and β was calculated by $50/N_t = 0.5$, where N_t is the number of topics.

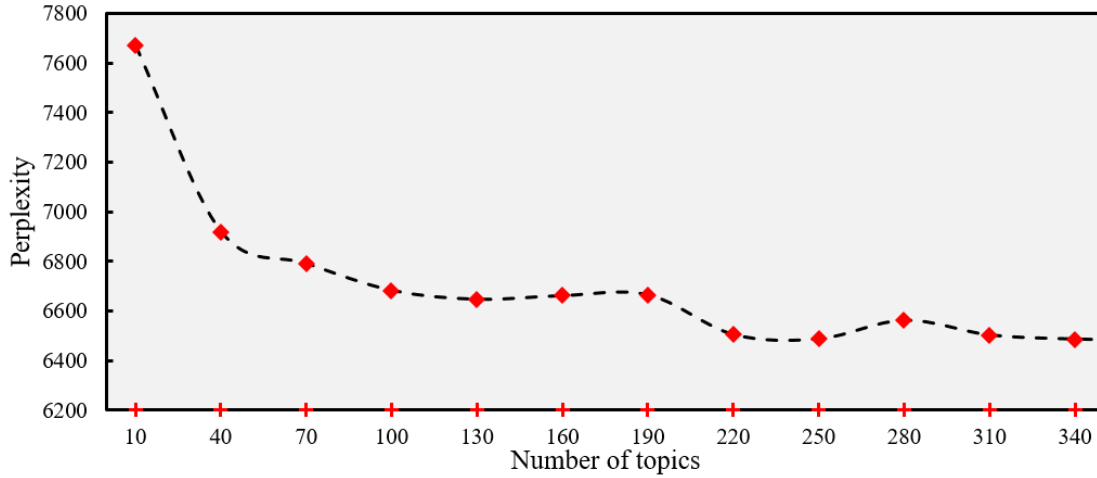


Figure 4.7. Perplexity distribution used for estimating the number of topics for topic modeling

The topic distributions over outlier tweets depict the attribute of the candidate event, which can be used to measure the semantic similarity between two sets of outlier tweets. Further, the event content can be precisely inferred by word distributions over each involved topic. The author randomly selected an outlier detected in region #417 and illustrated the distributions over its top five topics in Table 4.2. The words distribution over the top five topics are specifically presented in Table 4.3. Topic #92 (with a distribution of 0.463) and topic #44 (with a distribution of 0.124), containing words such as “tiff”, “film”, “festival”, “movie” and “theatre” and so on, are both closely related to TIFF events, while topic #34 (with a distribution of 0.19) describing a positive feeling about Toronto is less clearly relevant to TIFF events. Therefore, based on looking into the topic distributions in Table 4.2 and corresponding word distributions in Table 4.3 (a full list of topic distributions and word distributions are summarized in appendix), it is much likely that this outlier is linked to a TIFF event. The relevance of an outlier to a TIFF event will be quantitatively identified in the following section.

Table 4.2. An example of topic distributions over an outlier in region #417

Topic ID	P(t)
92	0.463

34	0.19
44	0.124
36	0.1
6	0.073

Table 4.3. Word distributions over the top five topics of an outlier in region #417

Topics	Word					
	P(w t)					
Topic #92	tiff	bell	lightbox	princess	wales	theatre
	0.130	0.095	0.093	0.055	0.046	0.040
Topic #34	Toronto	day	today	time	love	good
	0.027	0.014	0.012	0.012	0.011	0.010
Topic #44	Toronto	tiff	film	festival	party	movie
	0.092	0.072	0.029	0.019	0.018	0.017
Topic #36	grolschza	jauja	laggies	wine	famous	winery
	0.020	0.015	0.013	0.013	0.011	0.009
Topic #6	bae	nyc	pureleaf	maker	series	gotham
	0.025	0.010	0.010	0.008	0.008	0.006

4.1.4 Detection of TIFF events

The detected outliers were further clustered by simultaneously measuring spatial adjacency and semantic similarity. Based on the mechanism of the ST-DBSCAN method introduced in Section 3.2.5, *MinPts* was set as 1 and *eps1* was set as 800m since the geographical area of Toronto is 630.2 km², and $\sqrt{630.2km^2/K} = \sqrt{630.2km^2/1000} \approx 800m$, where *K* is the number of the partitioned regions in k-means clustering.

The threshold (i.e., *eps2*) for cosine similarity between two sets of outlier tweets (i.e., a pair) was estimated by investigating the frequency distribution concerning each interval, as shown in Figure 4.8. As explained in Section 3.2.5, the smaller the cosine similarity is, the more different the two

sets of outlier tweets are. The sharpest decrease occurs when the cosine similarity is larger than 0.7, which reveals that this significant change seems to be used as a distinction. This is what meets the author's requirements to cluster semantically similar outliers rather than all detected outliers within 800m. Therefore, *eps2* for cosine similarity was set as 0.7. It means if $\cos\theta$ is larger than 0.7, the two sets of outlier tweets are identified as semantically similar. In consequence, 673 outlier clusters were generated. Around 13% outliers were regrouped comparing to 774 outliers detected in Section 4.1.2. The number of outliers before and after clustering is summarized in Figure 4.9.

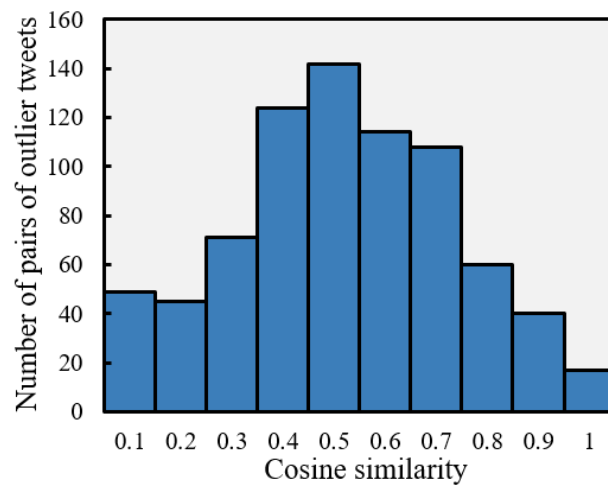


Figure 4.8. Statistic distribution of semantic similarity between pairs of outlier tweets

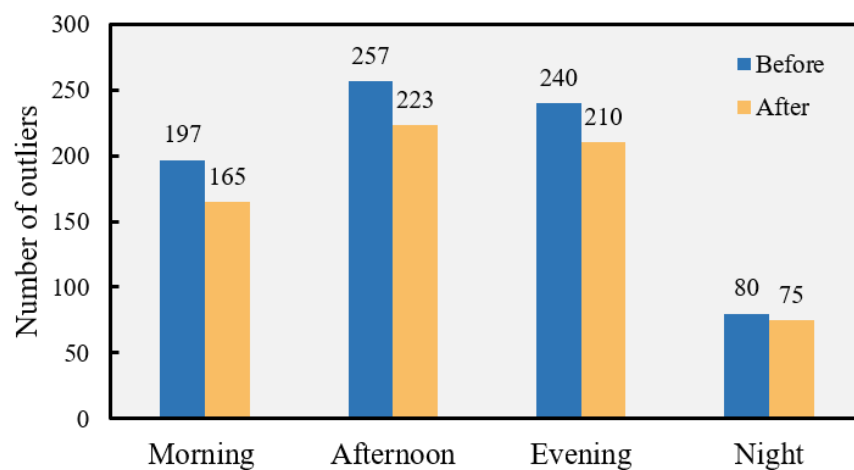


Figure 4.9. Statistics of #outliers before and after clustering by investigating spatial adjacency and semantic similarity

Aiming at investigating whether a detected outlier cluster refers to a TIFF event in the real world, this study evaluated the relevance of each outlier within the cluster to TIFF events based on a regression model in an automatic way. It incorporated the weight distributions over TIFF related topics with the topic distributions over the outlier tweets to estimate the relevance of the outlier to TIFF event. Assuming there are n topics that are related to TIFF event, which contains “tiff” in their word distributions, each topic is assigned with a weight according to the distribution of “tiff” over each topic in a descending order. For example, the weight of the k^{th} topic (W_k) can be computed as:

$$W_k = k * \frac{1}{1 + 2 + \dots + n} = k * \frac{1}{\frac{n * (n + 1)}{2}} = \frac{2 * k}{n * (n + 1)}, k \in (1, n). \quad (24)$$

The significance of the k^{th} topic to TIFF event $T_{k(TIFF)}$ is then computed as:

$$T_{k(TIFF)} = W_k * P_k, k \in (1, n), \quad (25)$$

where P_k refers to the distribution of the k^{th} topic over a set of outlier tweets. Therefore, the relevance of an outlier to TIFF event ($R_{(TIFF)}$) can be measured as:

$$R_{(TIFF)} = T_{1(TIFF)} + T_{2(TIFF)} + \dots + T_{k(TIFF)} + \dots + T_{n(TIFF)}. \quad (26)$$

An outlier cluster includes one or more outliers. The relevance of an outlier cluster to TIFF event $C_{(TIFF)}$ is estimated by taking the average value of the relevance of all outliers that are grouped in the cluster, as shown in Equation (27).

$$C_{(TIFF)} = \frac{R_{1(TIFF)} + R_{2(TIFF)} + \dots + R_{c(TIFF)}}{c}. \quad (27)$$

where c is the number of outliers within a cluster.

Table 4.4. “tiff” distributions over four related topics

Topic ID	P(“tiff” t)
92	0.130
44	0.072
14	0.023
64	0.011

Table 4.5. Word distributions over TIFF related topics

Topics	Word					
	P(w t)					
Topic #14	roy	hall	thomson	tiff	tiffrc	de
	0.065	0.060	0.036	0.023	0.021	0.021
Topic #64	Toronto	day	today	time	love	good
	0.038	0.022	0.021	0.016	0.014	0.011

In this study, among the 100 topics, four topics (i.e., topic #92, topic #44, topic #14, and topic #64) were identified to be most relevant to TIFF event as they all contain “tiff” in their word distributions. Table 4.4 displays “tiff” distributions over the four TIFF related topics, where it was regarded that the higher the distribution of “tiff”, the more relevant to TIFF event. The word distributions over topic #92 and topic #44 have been presented in Table 4.3. Table 4.5 shows the word distributions over the other two topics, i.e., topic #14 and topic #64. As such, according to Equation (24), (25), and (26), the relevance of an outlier to TIFF event can be calculated as follows:

$$R_{TIFF} = 0.4 * P_1 + 0.3 * P_2 + 0.2 * P_3 + 0.1 * P_4, \quad (28)$$

where P_1 , P_2 , P_3 , and P_4 refer to the distributions of topic #92, topic #44, topic #14, and topic #64 over the outlier tweets, respectively. The statistics of the relevance of all the detected outlier clusters to TIFF events is shown with a histogram in Figure 4.10. It indicates that the relevance of most outlier clusters ranges from 0 to 0.1.

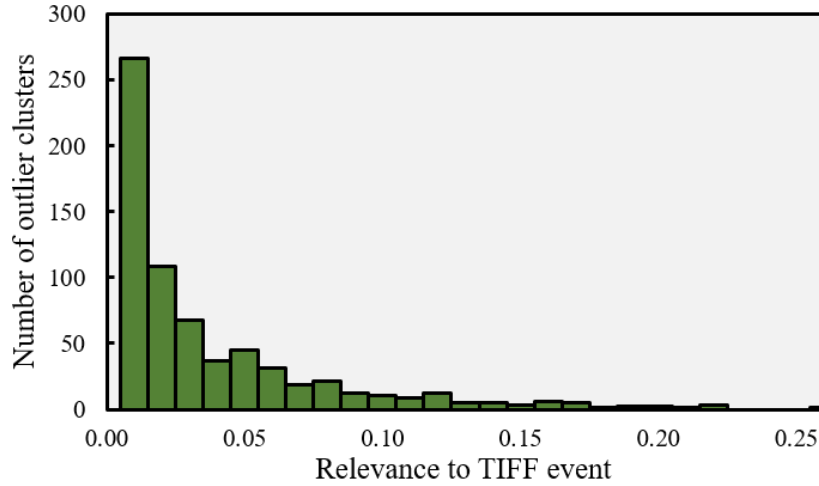


Figure 4.10. Histogram of the relevance of outlier clusters to Toronto International Film Festival (TIFF) event

4.1.5 Evaluation

An outlier cluster refers to a real-world TIFF event if the actual location provided by TIFF schedule locates within the cluster, the actual time overlaps the time interval of the cluster, and the relevance of the cluster to TIFF is more than μ . However, it is difficult to determine a threshold μ to identify the relation of an outlier cluster to a TIFF event based on the relevance computed in Section 4.1.4. As such, sensitivity analysis was conducted to evaluate the performance of TIFF event detection, where four metrics, accuracy, precision, recall, and F₁-score, were used.

Four measures were involved in the performance evaluation including True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP refers to the number of correct detections where detected event and actual event are both positively related to TIFF events. FP refers to the number of incorrect detections where the detected event is positive but the actual event is negative. TN refers to the number of correct detections where detected event and actual event are both negative. FN refers to the number of incorrect detections where the detected event is negative but the actual event is positive. These four measures lay a solid foundation for computing evaluation metrics, namely accuracy, precision, recall, and F₁-score, of which the value all range from 0 to 1. More details about the four evaluation metrics are shown in Table 4.6.

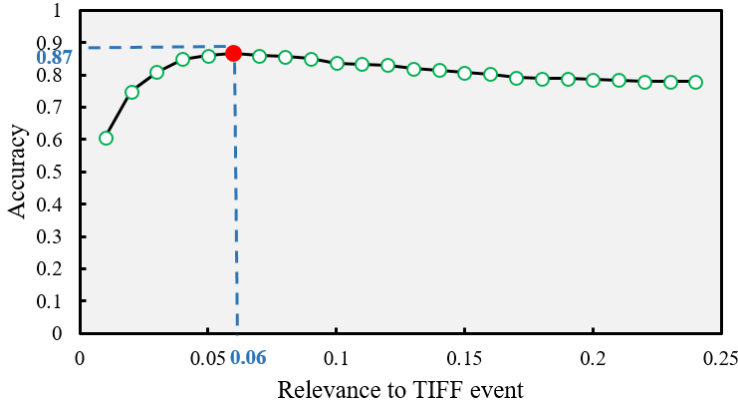
Table 4.6. Details about evaluation metrics

Metric	Definition	Equation	Evaluation for TIFF event detection
Accuracy	The proportion of correct detections.	Accuracy $= \frac{TP + TN}{TP + FN + FP + TN}$	Higher accuracy represents better detection results.
Precision (Specificity)	The percentage of the number of true positive detections to the total number of positive detections.	Precision = $\frac{TP}{TP + FP}$	Value of 1 indicates that every event detected as real-world TIFF event does indeed refer to real-world TIFF event, but without information about the number of real-world TIFF events that are not detected correctly.
Recall (Sensitivity)	The percentage of the true positive detections to the number of all actual positive detections.	Recall = $\frac{TP}{TP + FN}$	Value of 1 reveals that all real-world TIFF events have been detected as TIFF events, but without information about how many other events are also incorrectly detected as referring to real-world TIFF events.
F ₁ -score	The harmonic mean of precision and recall.	F ₁ $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ $= \frac{2TP}{2TP + FP + FN}$	It presents a comprehensive insight into TIFF event detection that balances the meaning of precision and recall.

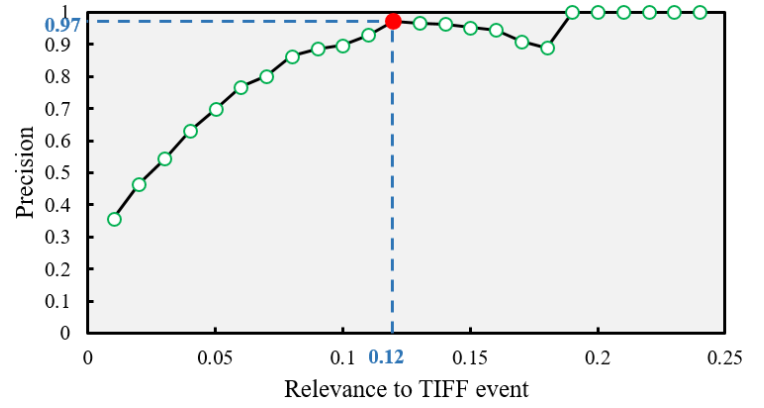
According to the official schedule of TIFF 2014, totally 393 films were shown. These films were grouped into 149 TIFF events by checking their screening time and place (i.e., a certain number of films were integrated as a TIFF event if they were shown in the same venue but different

screening rooms during the same time interval), which were selected as ground truth for evaluation in this study. Specifically, there were 28, 44, 47, 30 TIFF events occurring in the morning, afternoon, evening, and night, respectively. Based on the sensitivity analysis, the result of accuracy, precision, recall, and F_1 -score are illustrated in Figure 4.11 (a), (b), (c), and (d), respectively. The highest accuracy of 87% was reached when the threshold μ was set as 0.06 (i.e., an outlier cluster was identified as a real-world TIFF event if its relevance to TIFF exceeded 0.06), meaning that 87% of the detected outlier clusters were correctly identified through comparing with the ground truth events list. In this case, among the 673 detected outlier clusters, 86 were correctly identified as TIFF events (i.e., TP), 26 were identified as TIFF events while they were not real TIFF events (i.e., FP), 498 were correctly identified as non-TIFF events (i.e., TN), and 63 were identified as non-TIFF events while they were indeed TIFF events (i.e., FN).

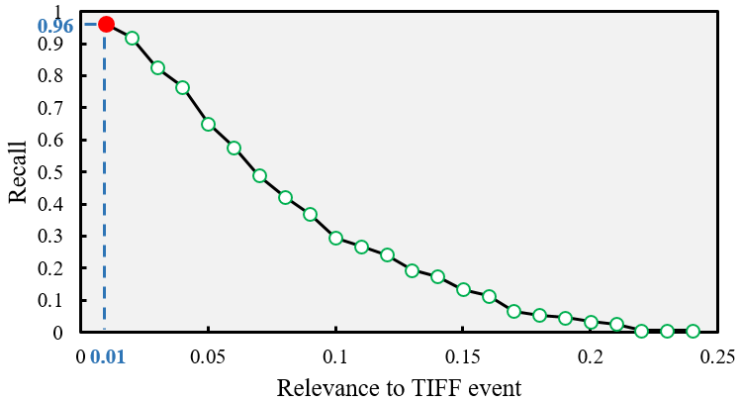
The precision keeps increasing on condition that μ is not more than 0.12, and achieves the highest precision of 97%, which means that the higher μ , the more TP outlier clusters. It decreases as μ is between 0.13 and 0.18. This may be because there appeared some outlier clusters that were detected as TIFF events in this experiment, but they were not in the real world, namely FP increases. For example, many people were interested in TIFF and posted a number of TIFF related tweets to express their excitement in the evening at home, while they were normally inactive in Twitter during this period. Thus, the burst in the number of tweets and the number of users, and significant semantic relation to TIFF made it possible to be detected as TIFF events, but its location did not overlap with any TIFF venues. Such outlier cluster was labelled as FP. Recall usually changes inversely with precision, meaning recall increase/decreases as precision decreases/increases. As indicated in Figure 4.11 (c), recall decreases with the increase of μ , since a higher μ will result in more FN. F_1 -score, calculated as a combination of precision and recall, has the highest value of 0.69 when μ equals to 0.04.



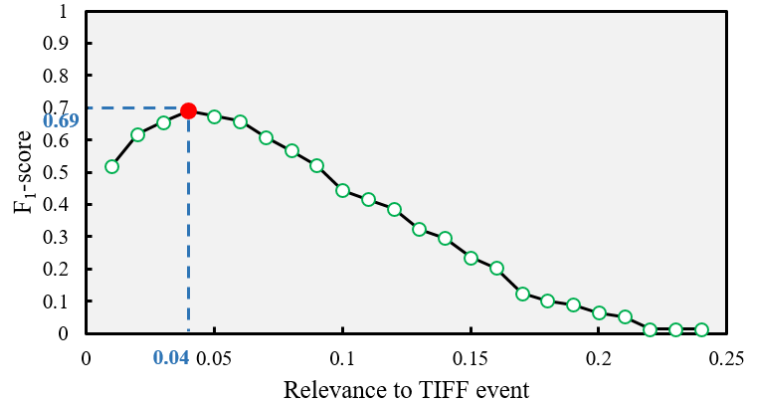
(a) Accuracy



(b) Precision



(c) Recall



(d) F₁-score

Figure 4.11. Evaluation of social event detection results measured by accuracy, precision, recall, and F₁-score

Additionally, a comparison experiment were further conducted with R. Lee and Sumiya (2010)'s work, where the boxplot method was adopted to detect spatiotemporal outliers for identifying social events but without semantic analysis. An example of outlier detection result in region #417 is shown in Figure 4.12, where a social event was detected if the number of tweets and the number of users were both identified as outliers within the same time interval. 23 pairs of outliers (i.e., 5 in the morning, 9 in the afternoon, 5 in the evening, and 4 in the night) were detected using the boxplot method while 40 outliers were detected by the Mahalanobis Distance method (Figure 4.6). As a result, 242 outliers were detected using the boxplot method, including 42, 71, 77, and 52 in

the morning, afternoon, evening, and night. In this case, all of these detected outliers were regarded as candidate TIFF events. By comparing with the official schedule, 50 among 149 real TIFF events were successfully detected. Therefore, the recall of the detection result is $50/149 = 34\%$. In contrast, the author's approach with additional content analysis reaches the recall of 58% when it comes to the highest accuracy of 87%. It is obvious that the proposed spatial-temporal-semantic approach has a better capability of accurately extracting more real-world social events.

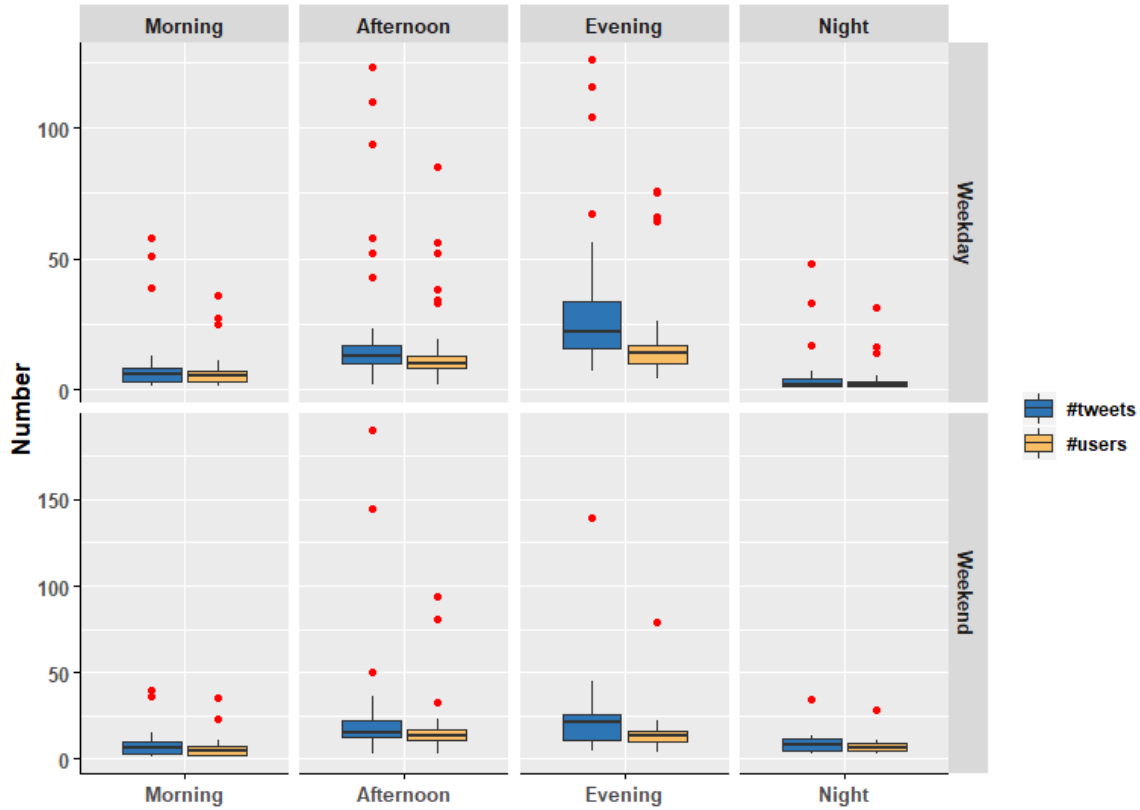


Figure 4.12. Detection of outliers (red point) in region #417 on weekday and weekend (boxplot)

4.2 Traffic event detection based on the classification methods

4.2.1 Study area and dataset

A case study was conducted using the Twitter data collected from April 1, 2014 to March 31, 2015 through the Twitter Streaming API by defining a bounding box enclosing Toronto, Ontario,

Canada. A total number of 17,170,543 tweets geotagged with coordinates or Twitter places were obtained and stored in MongoDB, where the spatial distribution is visualized in Figure 4.13. There appears to be a higher concentration of tweets being posted in the downtown Toronto area.



Figure 4.13. The geography of Toronto, Ontario, Canada

Initially, a full text search was made in MongoDB to return tweets with the occurrence of traffic related keywords. A total of 59 keywords used in this query were generated by counting their frequency in the reviewed studies (Xu et al., 2018). If a keyword appeared twice or more, it would be included in the 59-keyword list. A sample of top ten keywords with their frequency are summarized in Table 4.7. Based on these single keywords, 468,446 tweets were obtained, which were further processed by the NLTK ¹⁰ Python package. As a result, a tweet was simplified to a list of stemmed words for the convenience of mining association rules after tokenization, lowercase normalization, stop words removal, and word stemming.

Table 4.7. Top ten traffic related keywords ranked by their frequency

Keywords	Frequency	Keywords	Frequency
Accident	10	Street	7

¹⁰ <https://www.nltk.org/>

Traffic	9	Congestion	6
Crash	9	Delay	6
Road	8	Incident	5
Blocked	7	Closed	5

4.2.2 Association rules mined by Apriori algorithm

Among all the initial queried tweets, the tweets posted in November 2014 were manually labeled as positive or negative. The 907 positive tweets were selected as training data to mine the association rules using the Apriori algorithm. A python package named apyori was used for this process, where the thresholds for the minimum support (*min_support*) and the minimum confidence (*min_confidence*) are required to be predetermined.

Instead of empirically setting the two thresholds, a sensitivity analysis was deployed to examine how the number of mined association rules changes with different values of *min_support* and *min_confidence*. As demonstrated in Figure 4.14, there exists a significant difference between the blue solid line indicating *min_support* equals to 0.01 and other lines. The minimum support reflects the lowest frequency of a word in all tweets. Any word with a support less than the minimum support will not be considered as a potential word to generate frequent wordsets. As such, with a limited amount of training data, the *min_support* was set as 0.01 in order to engage as many words as possible to mine the association rules. With regard to the blue solid line, the number of association rules decreases sharply as *min_confidence* is over 0.2, compared to the trend between 0.1 and 0.2, which changes gently after 0.6. The author took a look at both of the mined association rules with *min_confidence* of 0.2 and 0.6, and found that the majority of meaningful association rules were excluded when *min_confidence* equalled to 0.6. In this study, in order to extract distinct and meaningful association rules as many as possible, *min_confidence* was defined to be 0.2.

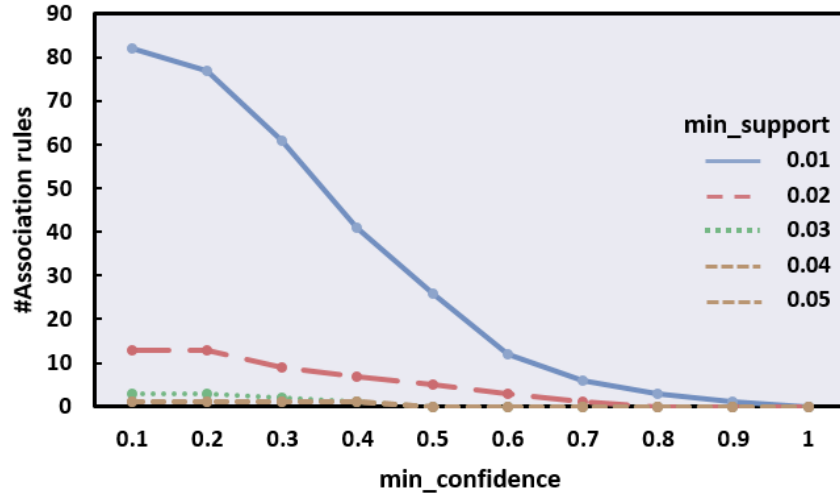


Figure 4.14. The number of association rules changing with different support and confidence values

Given *min_support* and *min_confidence* to be 0.01 and 0.2, frequent two-wordset, three-wordset, four-wordset, and five-wordset were further mined for association rule representation. The result of 5-wordset was empty. With respect to each mined three-wordset and four-wordset, at least one selected combination of two-word terms could be found in the pool of two-words. In other words, the tweets queried by the three-wordset and four-wordset were all covered by the tweets queried by the two-wordset. Therefore, association rules were mined using two-wordset. As a result, a total of 77 association rules were generated, where some examples are provided in Table 4.8. The basic form of the words rather than the stemmed words are presented to make it easier to understand. For example, the association rule was built between *word*₁ “close” and *word*₂ “highway”, indicating that these two words were frequently co-used in a tweet reporting a traffic event.

Table 4.8. A list of mined association rules

<i>word</i> ₁	<i>word</i> ₂	<i>word</i> ₁	<i>word</i> ₂
close	highway	serious	injury
collision	condition	strike	vehicle

involve	crash	injure	crash
crash	hwy	snow	drive
safe	drive	northbound	close

Considering that the above association rules mined from one-month labelled positive tweets may not cover all cases, the author further extended them by referring to word abbreviation (e.g., ‘st’ and ‘street’, and ‘rd’ and ‘road’), synonyms (e.g., ‘injure’ and ‘hurt’, and ‘collision’ and ‘hit’,), hyponyms (e.g., “weather’ and ‘snow’, ‘rain’, and ‘wind’) and common sense knowledge (e.g., ‘westbound’ and ‘eastbound, ‘northbound’, and ‘southbound’). Both synonyms and hyponyms were generated using the WordNet tool (Miller, 1998). In this manner, the number of association rules was extended to be 504. Furthermore, a text search engine of Whoosh¹¹ in Python language was used to extract positive tweets as well as filter negative tweets from the initial queried result. A total number of 8,714 tweets were finally obtained.

4.2.3 Event classification

As explained in the Section 1.3, traffic anomalies referring to traffic events mainly result from traffic incidents and severe traffic conditions. The association rules-filtered tweets were classified into three categories (Dabiri & Heaslip, 2019).

1. **Non-traffic events:** Tweets that do not report real-world traffic events.
2. **Traffic incidents:** Tweets reporting non-recurrent crashes, disabled vehicles, road work, special events, traffic signal problems, and bad weather conditions.
3. **Severe traffic conditions:** Tweets reporting traffic congestion, rush hours, traffic delay due to high traffic volume and jammed traffic.

The 8,714 association-rule queried tweets were all manually labeled by going through tweets content and identifying a tweet to be positive or negative to a real-world traffic event, i.e., positive tweet or negative tweet. Those tweets with fuzzy information and uncertainty, neither explicitly describing traffic anomalies nor extremely irrelevant to traffic, were discarded during the labelling

¹¹ <https://whoosh.readthedocs.io/en/latest/index.html>

process. Thus, the labelled tweets could be used as ground truth data to investigate the effectiveness of the classification model proposed by NB method, SVM method, and LR method. The “train_test_split” function provided by the Python package scikit-learn was first used to randomly split all the association-rule queried tweets into training dataset and test dataset based on the provided default partition ratio¹², where 75% of tweets (6,535 tweets) were used as the training dataset and the remaining 25% tweets (2,179 tweets) were used as the test dataset.

Accordingly, the grid-search technique engaging a ten-fold cross validation approach was applied to find the optimal parameter values to build three classification models from the training dataset. To be specific, the training dataset was randomly partitioned into ten equal sized folds, where nine folds were used to train the model and the remaining one fold was used for validation in each pass. This cross-validation process repeated ten times until each fold of data was used once as validation data. Subsequently, the trained models were applied to predict the event type a tweet in the test dataset referred to.

In the classification model training process, the grid search involving the cross validation method were from an open-source Python library called scikit-learn¹³, which serves as an efficient tool in for data mining and data analysis. Thus, the event type of test dataset can be predicted using the trained classification models.

Based on the same four metrics (i.e., accuracy, precision, recall, and F_1 -score) introduced in Section 4.1.5, the predicted results of each event type were quantitatively evaluated. With regard to each type of traffic events, the details about the basic measures (i.e., TP, TN, FP, and FN) and derived metrics are summarized in Table 4.9. The accuracy, precision, recall, and F_1 -score are all between 0 and 1. A higher accuracy indicates better detection results. The precision of 1 means that all the events predicted as event type E_t indeed refer to E_t , but lacks information about the number of events actually in E_t that are incorrectly detected. Recall is also called sensitivity, of which a higher value reveals that the classifier is more sensitive to positive detections. A recall of 1 reveals that the events in E_t are all correctly detected, but lacks information about the number of

¹² https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

¹³ <https://scikit-learn.org/stable/>

events that do not belong to E_t , but are detected as E_t . The value of precision usually increases with the decrease in recall, which represents an inverse correlation. Normally, they are not used separately. A combined measure of F₁-score provides a comprehensive view of the detection performance by balancing the meaning of precision and recall.

Table 4.9. Explanation to evaluation metrics

Basic measures	<p>TP: The number of correct detections where predicted event and actual event both refer to the same event type E_t.</p> <p>TN: The number of correct detections where predicted event and actual event both do not refer to event type E_t.</p> <p>FP: The number of incorrect detections where the events that are predicted as E_t do not actually refer to E_t.</p> <p>FN: The number of incorrect detections where the events that are not predicted as E_t are actually E_t.</p>								
Metrics	<table> <tr> <td data-bbox="358 1031 885 1144">Accuracy is the percentage of correct detections.</td><td data-bbox="885 1031 1430 1144"> $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$ </td></tr> <tr> <td data-bbox="358 1144 885 1312">Precision is the number of true positive detections divided by the number of all positive detections.</td><td data-bbox="885 1144 1430 1312"> $\text{Precision} = \frac{TP}{TP + FP}$ </td></tr> <tr> <td data-bbox="358 1312 885 1501">Recall is the number of true positive detections divided by the number of all actual positive events.</td><td data-bbox="885 1312 1430 1501"> $\text{Recall} = \frac{TP}{TP + FN}$ </td></tr> <tr> <td data-bbox="358 1501 885 1680">F₁-score is the harmonic mean of precision and recall.</td><td data-bbox="885 1501 1430 1680"> $F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ $= \frac{2TP}{2TP + FP + FN}$ </td></tr> </table>	Accuracy is the percentage of correct detections.	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	Precision is the number of true positive detections divided by the number of all positive detections.	$\text{Precision} = \frac{TP}{TP + FP}$	Recall is the number of true positive detections divided by the number of all actual positive events.	$\text{Recall} = \frac{TP}{TP + FN}$	F ₁ -score is the harmonic mean of precision and recall.	$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ $= \frac{2TP}{2TP + FP + FN}$
Accuracy is the percentage of correct detections.	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$								
Precision is the number of true positive detections divided by the number of all positive detections.	$\text{Precision} = \frac{TP}{TP + FP}$								
Recall is the number of true positive detections divided by the number of all actual positive events.	$\text{Recall} = \frac{TP}{TP + FN}$								
F ₁ -score is the harmonic mean of precision and recall.	$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ $= \frac{2TP}{2TP + FP + FN}$								

Through comparing the predicted event type with the prior labelled event type, the performance of event classification is shown in Table 4.10, regarding the above four evaluation metrics that were generated by averaging of the three event types. It can be clearly seen that SVM method

outperforms NB method and LR method as it has the highest accuracy and F₁-score. In other words, the SVM method is capable of accurately identifying the type of traffic events that a tweet is referring to. The classification result from the SVM method was used for the following validation of traffic events with vehicle traffic data (Section 4.2.4).

Table 4.10. Traffic event classification results of three different methods

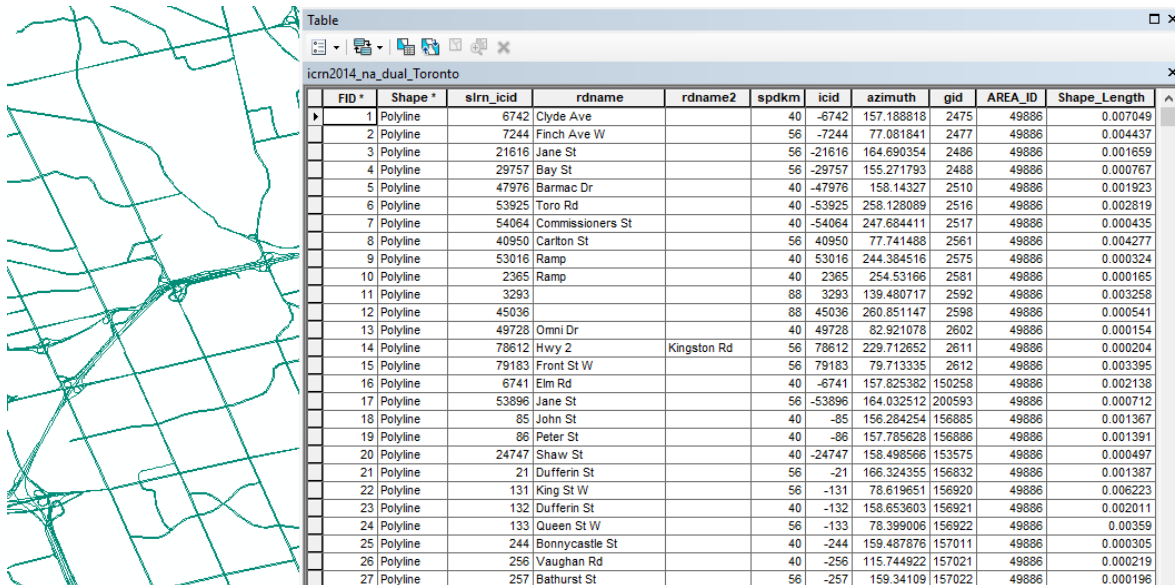
Method	Accuracy	Recall	Precision	F ₁ -score
NB	0.72	0.35	0.55	0.31
SVM	0.78	0.63	0.71	0.64
LR	0.76	0.57	0.78	0.59

4.2.4 Validation with vehicle travel speed data

The above detected traffic events were validated with the vehicle travel speed data provided by MTO. The MTO traffic data covering the Toronto area shows that the hourly average travel speed at each road link were summarized from passenger vehicles and trucks in the year of 2014 and 2015. Each record in the traffic speed data consists of the road link identifier (i.e., *icid*), date string (i.e., *datestr* arranged by year-month-day), day of week (i.e., *dow*, from Monday to Sunday), the hour period in a day (i.e., *period_1hr*), the number of vehicles (i.e., *sample*), total hourly travel speed (i.e., *tot_gpsspd*), and average hourly travel speed computed by dividing total hourly travel speed by the number of vehicles (i.e., *ats*). A capture of sample records in the data file of csv format is shown in Figure 4.15 (a).

icid	datestr	dow	period_1hr	sample	tot_gpsspd	ats
-79944	20140613	5	1300	5	167.371776	33.4743552
-78564	20140510	6	1300	1	61.155072	61.155072
-77275	20141009	4	1500	6	131.966208	21.994368
-79386	20140821	4	1700	1	24.14016	24.14016
-75824	20140904	4	1300	1	45.061632	45.061632
-74637	20140619	4	1100	2	88.51392	44.25696
-71149	20140610	2	800	3	123.919488	41.306496
-64864	20141002	4	900	2	93.341952	46.670976
-59324	20140507	3	2300	4	231.745536	57.936384
-53902	20140805	2	1000	3	102.998016	34.332672
-48146	20141124	1	900	5	262.323072	52.4646144

(a) MTO travel speed data in csv format



FID *	Shape *	slrn_icid	rdname	rdname2	spdkm	icid	azimuth	gid	AREA_ID	Shape_Length
1	Polyline	6742	Clyde Ave		40	-6742	157.188818	2475	49886	0.007049
2	Polyline	7244	Finch Ave W		56	-7244	77.081841	2477	49886	0.004437
3	Polyline	21616	Jane St		56	-21616	164.690354	2486	49886	0.001659
4	Polyline	29757	Bay St		56	-29757	155.271793	2488	49886	0.000767
5	Polyline	47976	Barnum Dr		40	-47976	158.14327	2510	49886	0.001923
6	Polyline	53925	Toro Rd		40	-53925	258.128089	2516	49886	0.002819
7	Polyline	54064	Commissioners St		40	-54064	247.684411	2517	49886	0.000435
8	Polyline	40950	Carlton St		56	40950	77.741488	2561	49886	0.004277
9	Polyline	53016	Ramp		40	53016	244.384516	2575	49886	0.000324
10	Polyline	2365	Ramp		40	2365	254.53166	2581	49886	0.000165
11	Polyline	3293			88	3293	139.480717	2592	49886	0.003258
12	Polyline	45036			88	45036	260.851147	2598	49886	0.000541
13	Polyline	49728	Omni Dr		40	49728	82.921078	2602	49886	0.000154
14	Polyline	78612	Hwy 2	Kingston Rd	56	78612	229.712652	2611	49886	0.000204
15	Polyline	79183	Front St W		56	79183	79.713335	2612	49886	0.003395
16	Polyline	6741	Elm Rd		40	-6741	157.825382	150258	49886	0.002138
17	Polyline	53896	Jane St		56	-53896	164.032512	200593	49886	0.000712
18	Polyline	85	John St		40	-85	156.284254	156885	49886	0.001367
19	Polyline	86	Peter St		40	-86	157.785628	156886	49886	0.001391
20	Polyline	24747	Shaw St		40	-24747	158.498566	153575	49886	0.000497
21	Polyline	21	Dufferin St		56	-21	166.324355	156832	49886	0.001387
22	Polyline	131	King St W		56	-131	78.619651	156920	49886	0.006223
23	Polyline	132	Dufferin St		40	-132	158.653603	156921	49886	0.002011
24	Polyline	133	Queen St W		56	-133	78.399006	156922	49886	0.00359
25	Polyline	244	Bonnycastle St		40	-244	159.487876	157011	49886	0.000305
26	Polyline	256	Vaughan Rd		40	-256	115.744922	157021	49886	0.000219
27	Polyline	257	Bathurst St		56	-257	159.34109	157022	49886	0.000196

(b) MTO travel speed data in shapefile format

Figure 4.15. MTO travel speed data

Another shapefile format data is also provided in Figure 4.15 (b), which contains the road link identifier, the geospatial information, and geographic features (i.e., road length and road name) of each road link. The common field of road link identifier makes it possible to relate these two types of data so that each road link includes hourly travel speed value as well as geospatial information,

acting as reference information for the geo-codable traffic events to be located to the nearest road link.

As mentioned before, tweets collected through Twitter Streaming API with a bounding box can be tagged with either a Twitter place or a pair of coordinates. In this study, the location of traffic events was inferred by referring to their tagged GPS coordinates. A total of 4,750 tweets among the 8,714 tweets were removed since they were tagged with a certain Twitter place and the attribute of “coordinates” was null. In order to keep the locational reliability of detected traffic events, the tweets classified as non-traffic events (933 tweets) and the tweets posted by organizational accounts (1,405 tweets) were further filtered from the GPS tagged tweets. Thus, a total of 1,626 tweets indicating traffic events (1,120 traffic incidents and 506 severe traffic conditions) were left for validation with the travel speed data.

In terms of the geo-codable traffic events, the average hourly travel speed of their nearest road links were all extracted. Suppose a traffic event TE occurred at road link l during hour h on date d , the travel speed extracted at the same road link and time was defined as actual travel speed a_{TE} . Similarly, the travel speed at this road link and time without the occurrence of traffic events was defined as typical travel speed t_{TE} . A set of t_{TE} of the same road link, hour, and day of week composed the typical travel speed collection T_{TE} , which together with the actual speed a_{TE} was then standardized in Equation (29-30) as basis to investigate whether there existed a significant difference between them. $Avg(T_{TE})$ and $Std(T_{TE})$ were computed by the math function of mean and standard deviation, respectively.

$$a'_{TE} = \frac{a_{TE} - Avg(T_{TE})}{Std(T_{TE})} \quad (29)$$

$$T'_{TE} = \frac{T_{TE} - Avg(T_{TE})}{Std(T_{TE})} \quad (30)$$

The distribution of standardized actual travel speed and distribution of typical travel speed are both presented in Figure 4.16, which illustrates that the two distributions are obviously different.

Furthermore, a quantitative analysis was conducted based on the Kolmogorov-Smirnov (K-S) test with the null hypothesis that “*the actual speed holds the same distribution as the typical speed*”. The P -value resulted from K-S test was 5.2267E-15, which indicates the probability of finding the extreme observations when the null hypothesis is true. Thus, such small P -value value suggests rejecting the null hypothesis. In other words, the significant difference between the actual speed and typical speed implies that the geo-codable traffic events have high possibility of being correctly identified.

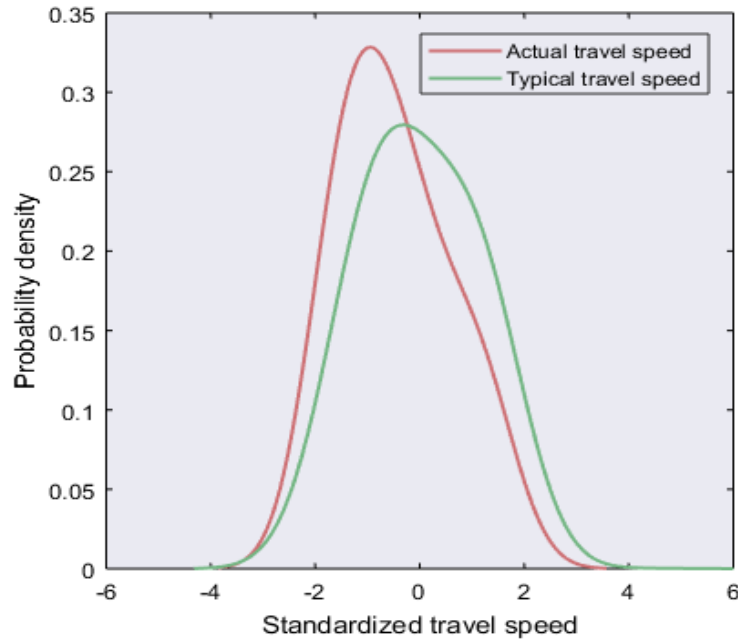


Figure 4.16. The distribution of standardized actual travel speed and typical travel speed based on classification method

In order to measure how many detected traffic events refer to real-world traffic events, a left tailed Z-test was conducted in terms of each traffic event. The null hypothesis H_0 and alternative hypothesis H_1 were made by

$$\begin{aligned} H_0: & \text{actual travel speed} \geq \text{the mean of typical travel speed} \\ H_1: & \text{actual travel speed} < \text{the mean of typical travel speed} \end{aligned} \quad (31)$$

where H_0 is the null hypothesis. The Z score of a traffic event (Z_t) can be estimated by its standardized actual speed a'_{TE} . According to the concept of left-tailed Z-test, the P -value of a traffic event is determined by

$$P_{TE} = Probability(Z \leq Z_t) = Probability\left(Z \leq a'_{TE}\right). \quad (32)$$

Given a certain significance level SL ,

$$\begin{aligned} P_{TE} &\geq SL, \text{ fail to reject the null hypothesis} \\ P_{TE} &< SL, \text{ reject the null hypothesis.} \end{aligned} \quad (33)$$

In this study, rejecting the null hypothesis H_0 indicates that a detected traffic event actually refers to a real-world traffic event. The percentage of correctly detected traffic events (detection rate) varies with the determination of the threshold for the Significance Level SL . Instead of giving the significance level a specific threshold, different thresholds for SL were assigned to explore how the detection rate was influenced. As displayed in Figure 4.17, when the threshold is set as 0.5, there is a possibility of around 70% to reject the null hypothesis, which means that approximately 70% of the detected traffic events are true traffic events in reality. At least 27% of the detected traffic events are true traffic events as the threshold is determined to be a small value of 0.1.

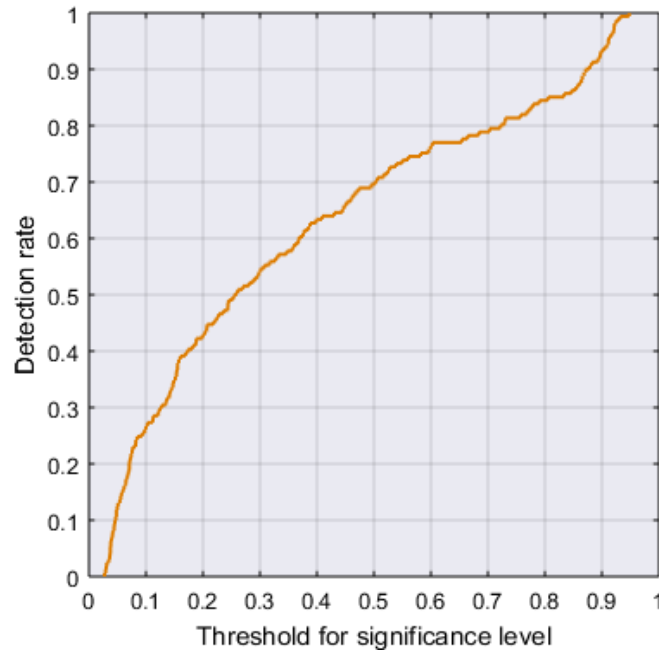
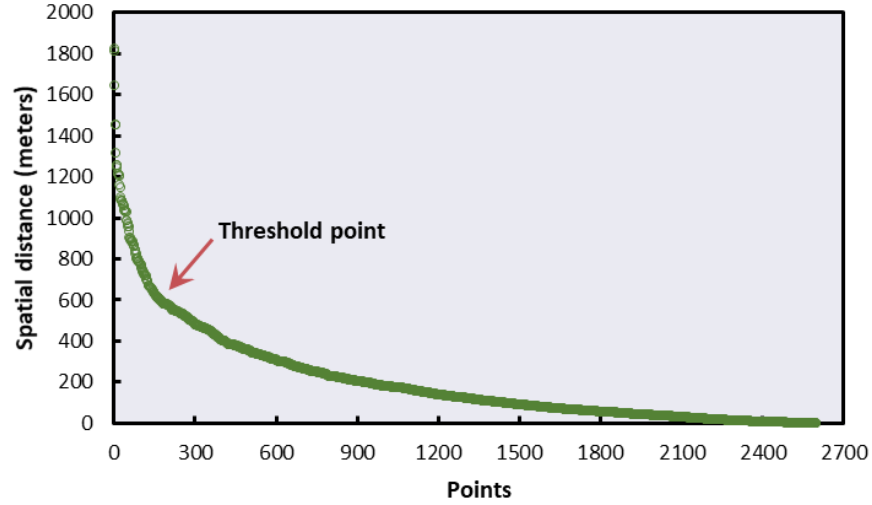


Figure 4.17. The detection rate changing with different thresholds for significance level based on the classification method

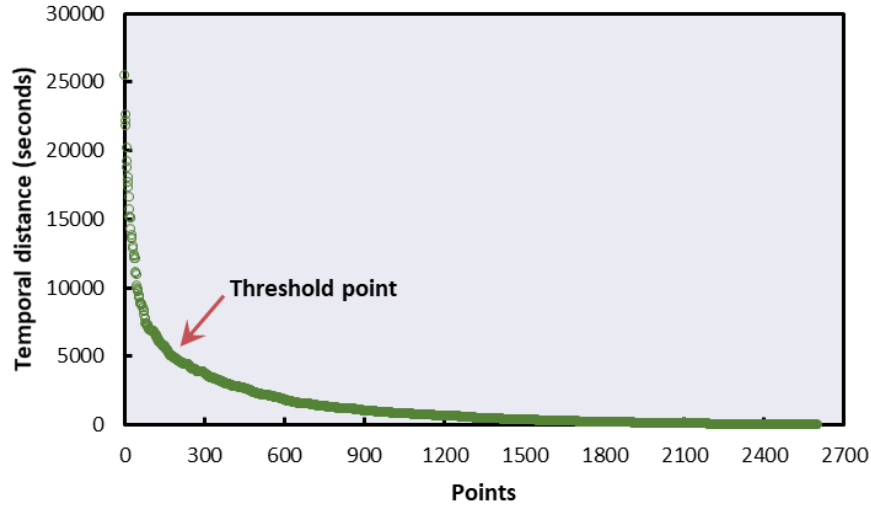
4.3 Traffic event detection based on the clustering method

4.3.1 Parameter estimation

By removing the tweets without GPS coordinates and the tweets posted by organizational accounts from the association rule-filtered tweets, a total of 2,559 tweets were left to generate spatiotemporal clusters for traffic event identification.



(a) Sorted 2-dist graph based on spatial distance



(b) Sorted 2-dist graph based on temporal distance

Figure 4.18. Sorted 2-dist graph

In this study, *MinPts* was set to be 2 due to the limited Twitter data (up to 1% of all tweets) and data sparsity. The sorted 2-dist graph in terms of spatial distance and temporal distance are plotted in Figure 4.18 (a) and (b), respectively. By selecting the threshold points separating abrupt changes from flat changes (i.e., the occurrence of first valley), *eps1* and *eps2* were approximately determined as 600 meters and 5,000 seconds. As a result, a total of 361 spatiotemporal clusters were generated. In other words, 361 traffic events were detected by the clustering-based method.

4.3.2 Spatiotemporal distribution of clusters indicating traffic events

An overall spatiotemporal distribution of the detected traffic events is presented in three-dimensional view in Figure 4.19. Traffic events were visualized using spatiotemporal clusters (i.e., cylinders), where the height was determined by the start time and the end time, and the area enclosed the data points composing the cluster. It can be generally seen that the traffic events concentrated near the center of the city (downtown Toronto area) and most of them occurred during the daytime. Furthermore, more details about the spatial and temporal distributions of the detected traffic events are illustrated in Figure 4.20 and Figure 4.21, respectively.

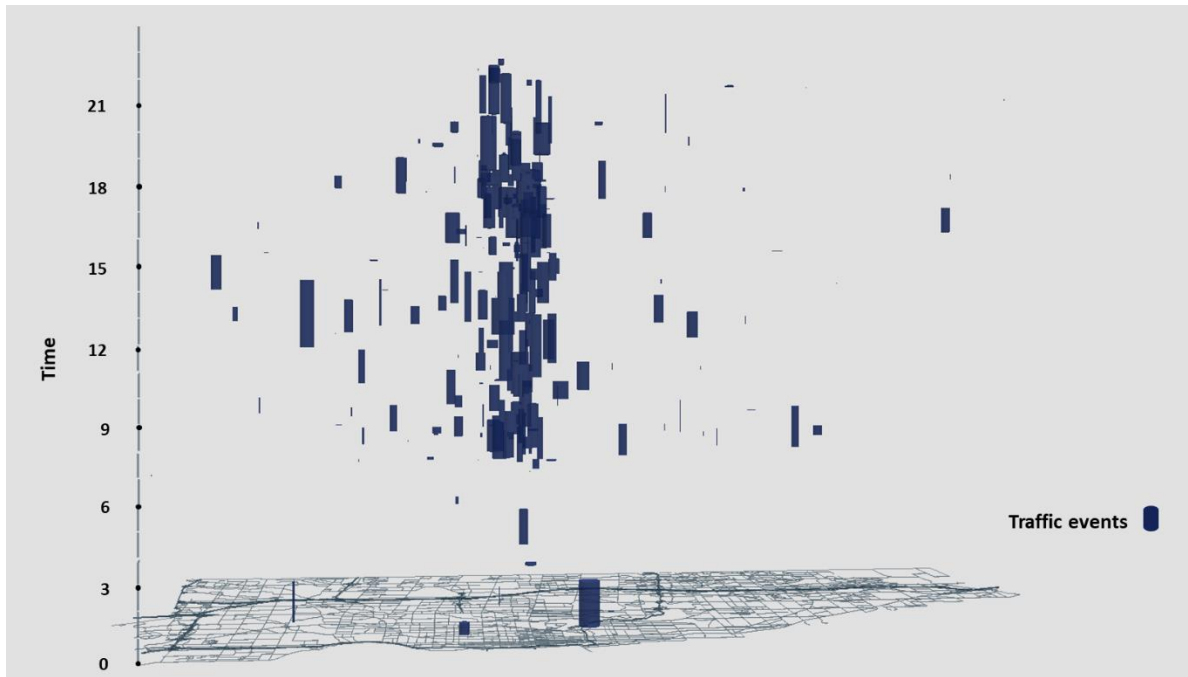


Figure 4.19. Spatiotemporal distribution of detected traffic events

As shown in Figure 4.20, the downtown area is a hot spot for the occurrence of traffic events. This detection result makes sense since there is usually a higher traffic flow at locations with limited space, and there is also a higher possibility of capturing traffic relevant posts from the denser number of active users across space. Apart from that, the second most traffic events were mainly reported around highways. Despite the sparse tweets posted on highways in comparison with

downtown area, the generated clusters indicate that it is very likely that traffic delays or traffic jams last long, drawing people's attention to complain.

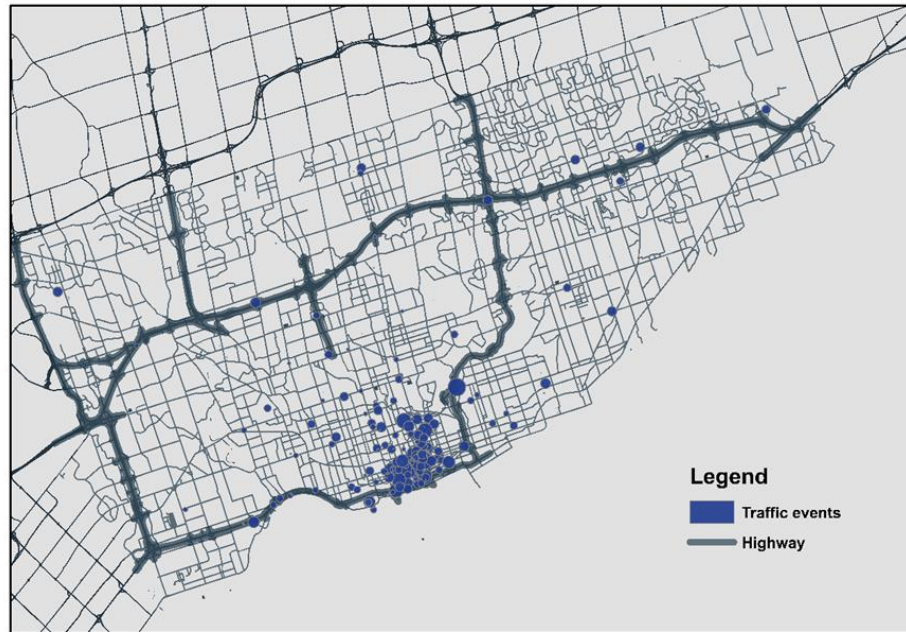


Figure 4.20. Spatial distribution of traffic events

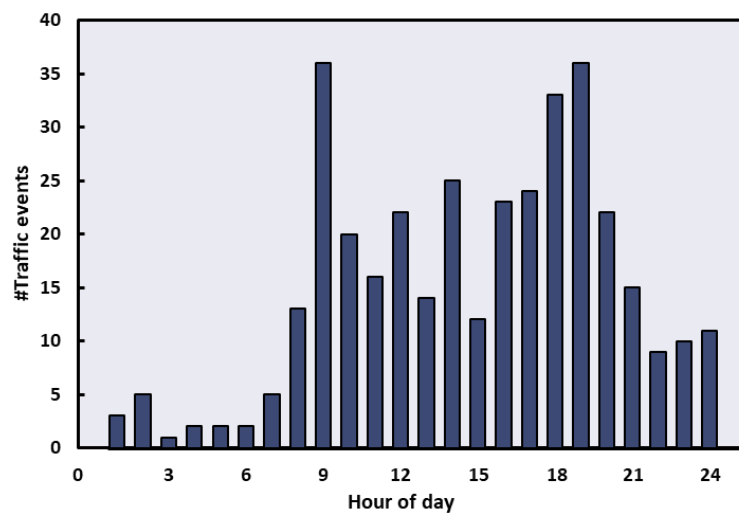


Figure 4.21. Temporal distribution of traffic events

To measure the temporal pattern of the detected traffic events, the number of spatiotemporal clusters reporting traffic events were counted in each hour of the day. The statistic result is shown in Figure 4.21, where two peaks are obviously found during 8am and 8pm. These two peaks almost align with the acknowledged morning and afternoon peak hours when the worst traffic conditions usually occur.

4.3.3 Content of clusters

The content of clusters were inferred by K representative keywords using TextRank model, which was realized using the Python library of pytextrank¹⁴. K was set to be ten in this study. Five clusters were randomly selected to illustrate how their content was summarized by the top ten single keywords and/or multi-word keywords, of which the details are presented in Table 4.11.

Table 4.11. Content of some clusters summarized by TextRank model

Clusters	Raw tweets	Keywords	Scores
#55	<ul style="list-style-type: none"> Following Toronto Police to the scene if the #deer jamming up traffic on the Gardiner Expressway 	the gardiner	0.081
		expressway	0.059
	<ul style="list-style-type: none"> Can someone please shoot this thing so we can get on with our day already? #toronto #traffic @cp24 @TorontoPolice 	toronto traffic	0.053
		toronto police	0.046
	<ul style="list-style-type: none"> For well over an hour a baby fawn has completely closed all westbound lanes of the gardiner expressway. #DeerWatch 	westbound lanes	0.040
		gardiner	0.039
	<ul style="list-style-type: none"> It appears the Toronto Zoo official has just arrived. Both lanes of the Gardiner are now closed. #DeerWatch 	expressway	0.033
		deer	0.030
		toronto zoo	0.027
		official	0.027
		traffic	0.027
		jamming	0.027
		gardiner	0.027

¹⁴ <https://pypi.org/project/pytextrank/>

- Traffic is now moving again along the Gardiner.
Bambi has been taken to the Toronto Zoo where she'll be assessed.

#197	• King Street Road closure being set up from Peter	tiff14 traffic	0.047
	all the way to University for #TIFF14 traffic chaos	chaos ensues	0.044
	ensues.	closing peter &	
	• Oh my Fuck. #Toronto has 2nd worst traffic in	john	0.035
	n.america, w/spadina @ the heart of it. Closing	john st	
	Peter & John for #TIFF should be illegal!!	peter	0.034
	• King St closed for fun, films and stars	king street road	0.027
	@TIFF_NET Fest director Cameron Bailey	closure	
#216	@cameron_tiff swarmed by paparazzi	st closed	0.027
	• King street is so closed right now haha @ TIFF	toronto	0.027
	Bell Lightbox	tiff	
	• King St. between Blue Jays Way and John St.	blue jays way	0.024
	closed to pedestrian. Smart idea. #TIFF14	cameron bailey	
	• I am late for an interview about Toronto's transit &	oakwood	0.096
	traffic problems, because of a massive traffic jam	allen & eg	
	downtown. Moved 2 blocks in 25 mins.	new lane closures	0.081
#218	• Traffic moving extremely slowly EB on Eglinton	eglinton west	0.061
	West near Oakwood. New lane closures?	construction	0.048
	• Love how @CrosstownTO construction closes	metromorning	0.048
	new lanes without warning and suddenly I'm 15	eg	0.044
	minutes later than normal.	lanes	0.041
	• @cityoftoronto Why move light @ Allen & Eg?	new lanes	0.041
	Traffic was finally moving well now it's disastrous	closures	0.040
	again! @Metrolinx #toronto @metromorning	king and yonge	0.192

	<ul style="list-style-type: none"> • Crash!! Streetcar takes out bus in front of me. Down to 7 lives left! • Streetcar hits bus at King and Yonge and shuts down King Street @CP24 @TTCnotices • @CityNews delay at King and Yonge. TTC hit bus • 30 mins in traffic from CBC to city hall. #Toronto #stupifying 	king street	0.096
		king	0.096
		city hall	0.057
		ttcnotices	0.048
		citynews	0.048
		streetcar hits bus	0.040
		street	0.039
		cp24	0.037
		cbc	0.044
#263	<ul style="list-style-type: none"> • @TTCnotices: 504 King route holding westbound on King at Portland due to autos in collision blocking the rail. #TTC • Issues with streetcars heading West on King St #TTC #Toronto • the TTC has to be a huge practical joke, right? why else would there be so much inefficiency DURING RUSH HOUR. • @TTChelps will you guys ever not suck? consistently terrible service during RUSH HOUR. is this a massive troll? 	toronto	0.120
		king	0.073
		streetcars	0.062
		ttc issues	0.059
		rush hour	0.048
		portland	0.045
		westbound	0.045
		huge practical	0.036
		joke	0.032
		massive troll	0.032
		ttc	0.032

It is clear that the content of the clusters can be represented based on the extractive summarization. For instance, from the keywords extracted in cluster #197, it can be inferred that there was a road closure between Peter Street and John Street, and King Street was very likely to be affected due to TIFF (i.e., Toronto International Film Festival) event. Similarly, in cluster #218, the streetcar hitting the bus engaged Yonge Street, King Street, and roads around city hall in trouble.

4.3.4 Validation with vehicle travel speed data

A total of 361 spatiotemporal clusters indicating traffic events were detected by adopting the clustering-based method. As described in Section 3.4.2, the location and time of a detected event were estimated by the average longitude and latitude and the average posted time of all engaged tweets. Base on the estimated location, the detected traffic event was located to the nearest road link. Accordingly, the typical travel speed and actual travel speed of this road link were separately obtained by referring to its estimated time including date, hour, and day of week, which were further standardized for the hypothesis test.

Similar to Section 4.2.4, the author first examined whether there existed a significant difference between the standardized typical travel speed and the standardized actual travel speed through K-S test. The computed P value equalled to $5.5238\text{E-}46$ provided reliable guidance to reject the null hypothesis. In other words, the distributions of standardized typical travel speed and actual travel speed travel were supposed to follow significantly different patterns, which were intuitively presented in Figure 4.22.

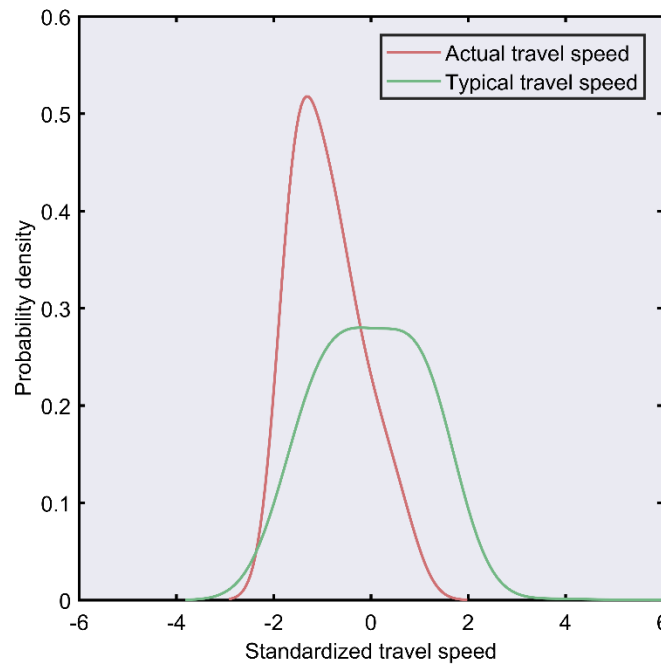


Figure 4.22. The distribution of standardized actual travel speed and typical travel speed based on clustering method

The left tailed Z-test was further conducted to explore how the detection rate changed with different thresholds for the significance level SL with regard to the spatiotemporal clustering method. Meanwhile, the detection result of the classification-based method was also added in the same figure to make a comparison. As shown in Figure 4.23, the overall performance of the clustering-based method is better than the classification-based method. Given a certain threshold for significance level SL , it reaches a higher detection rate by adopting the clustering-based method. Based on the clustering method, as many as 86% of the detected traffic events are correct with investigation of the spatiotemporal characteristics of traffic relevant posts when the threshold for significance level is determined as 0.5. In comparison, by applying the classification-based method to classify the traffic relevant posts into the predetermined categories, 70% of the identified traffic events are likely to be true traffic events at the same significance level. It reveals that the clustering-based method is more capable of accurately extracting traffic events from geosocial media data.

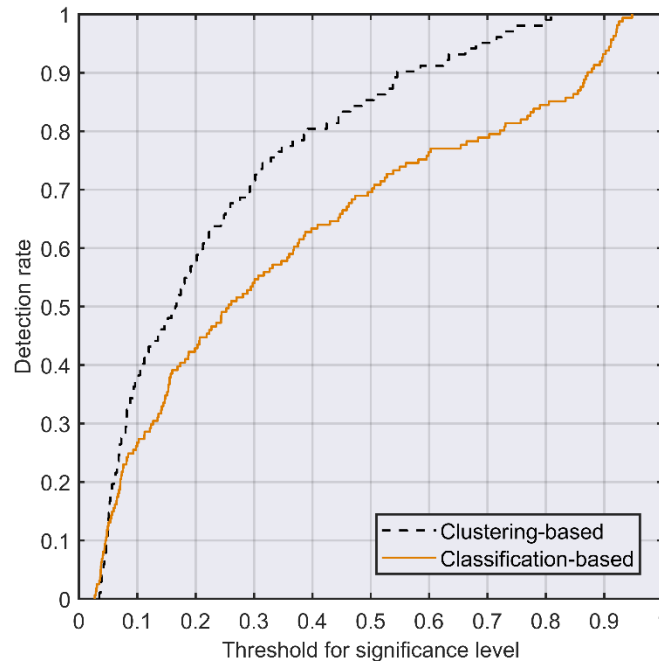


Figure 4.23. Comparison between classification-based method and clustering-based method

4.4 Discussion

4.4.1 Social event detection

The proposed methodology not only can detect TIFF events with a highest accuracy of 87% based on only 1% of Twitter data, but also illustrates its capacity of being generally applied to extract social events with spatiotemporal burst through a distinct example. The 2014 Cabbagetown Festival of the Arts was a fall festival annually held in Toronto's historic Cabbagetown district, which had programs such as buskers performing, sports zone, Cabbagetown arts and crafts sale, and live entertainments. It was held between Carlton Street and Parliament Street (south of Wellesley Street to Gerrard Street) on September 6 and 7, 2014¹⁵. In the morning of September 6, two regions, region #132 and region #165, were detected as outliers, which geographically overlapped with Parliament Street and Gerrard Street. The geographical distance between these two regions was 578.4 meters, and their semantic similarity was 0.89. Comparing to the thresholds that were defined in Section 4.4, region #132 and region #165 were grouped to one outlier cluster. The topic distributions (top five topics are selected) over the two outliers are shown in Table 4.12. It reveals that topic # 71 in two outliers are both with the significantly highest distribution. It can be simply inferred that topic #71 can reflect the content of the outlier cluster. Table 4.13 presents the detailed word distributions over the 10 topics discussed in region #132 and region #165 (some details shown in Table 4.3 are not repeated here). Topic #71 obviously refers to the Cabbagetown Festival. It again proves the capability of the proposed method in detecting social events from spatial, temporal, and semantic perspectives.

Table 4.12. Distributions of top five topics over the detected outliers

Region #132		Region #165	
Topic ID	P(t)	Topic ID	P(t)
71	0.496	71	0.466
34	0.259	34	0.189

¹⁵ <https://cabbagetownto.com/events/cabbagetown-festival-of-the-arts-2014-09-06/>

44	0.074	51	0.102
85	0.033	24	0.098
49	0.003	44	0.049

Table 4.13. Word distributions over topic #71

Topics	Word					
	P(w t)					
Topic	cabbagetown	festival	cabbagetownfestival	cabbage	naturopathy	osteopathy
#71	0.066	0.056	0.052	0.020	0.020	0.020
Topic	plane	loserkaraoke	throwback	marquis	turner	boring
#51	0.024	0.017	0.017	0.014	0.014	0.012
Topic	riotfest	downsview	park	riot	Toronto	riotfestto
#24	0.100	0.071	0.069	0.048	0.048	0.046
Topic	airport	Toronto	yyz	international	wind	kpa
#85	0.062	0.062	0.046	0.046	0.042	0.041
Topic	class	school	lol	campus	classes	prof
#49	0.071	0.028	0.013	0.013	0.011	0.011

Unlike the existing studies that mostly lacked automatic analysis of event content and/or integration of redundant events, this study reported a new approach with which event content was generated using a machine-learning based method. Further, correlation analysis was performed between abnormal neighborhoods for social event detection. Although the proposed method can successfully detect social events with relatively high accuracy, there are still some limitations that require further research efforts.

The proposed method in this study only considered and captured active users' behaviors from Twitter text data for social event detection. On one hand, the text-only event summarization method may neglect certain information that only exist in other types of multimedia data such as images and videos. A certain correlation is usually embedded among texts, videos, and images, which can be fully exploited for better event summarization (Bian, Yang, Zhang, & Chua, 2015).

On the other hand, the event related information witnessed by those who do not have an account or seldom post tweets might be missed. As a result, such a lack of data coverage may not provide a comprehensive detection. Potential solution toward such limitation can be addressed by exploring the capability of other types of geosocial media data (e.g., Flickr, Instagram, Foursquare, etc.), or retrieving event related information from these multiple social media platforms. This thus enhances the ability of the proposed method for social event detection, since information absent from one site may be filled by the others. For example, one person posts the event through Twitter, while another person may post the same to Facebook due to their preference.

The research area was partitioned into K (i.e., 1,000) sub-regions using k-means clustering method, where the number K was estimated by referring to the attributes of street blocks. Other spatial clustering method, e.g., DBSCAN, and different parameters can be applied to examine if any significant changes occur in detection results. In terms of the spatiotemporal outlier detection, Mondays, Tuesdays, Wednesdays, Thursdays and Fridays were classified as weekdays, where Mondays and Fridays are not usually considered as normal weekdays since they often present different patterns from Tuesdays, Wednesdays and Thursdays in transportation area. The patterns representing social activities can be further examined regarding each day of week to check if the transportation differentiate has any influence on social event detection results. Moreover, a day was divided equally into four time intervals, which were Morning (6am-12pm), Afternoon (12pm-6pm), Evening (6pm-12am), and Night (12am-6am), in this study. However, an event that happens during 5pm-8pm may be split into two events in the detection process. To overcome such drawback, the data-driven method acts as an alternative. With regard to a general study, the time windows can be estimated from the research data through a temporal (or spatiotemporal) clustering method (e.g., DBSCAN). In this manner, the temporal windows is scalable in accordance with the clusters, well aligning with the time span of actual events.

Data noise is a general concern in social media-based studies. Arguably, it is reported that about 40% tweets are just “pointless babbles” (Pear analytics, 2009). In this study, noise tweets that were posted by robots were removed, which counted for 18% of all crawled tweets. Further, a large number of abbreviations, mixed languages, wrong grammars, and non-ruled sentence structures are frequently found in tweets due to their limited length and unlimited writing styles. The

language complexity of tweets, including uncreditable information (Castillo, Mendoza, & Poblete, 2011), polluted content (K. Lee, Eoff, & Caverlee, 2011), and meaningless messages (Hurlock & Wilson, 2011), may negatively affect the performance of topic modeling. Schofield, Magnusson, Thompson, and Mimno (2017) and Schofield and Mimno (2016) found that text pre-processing tools including document duplication, removing stop words (e.g., determiners, conjunctions, and prepositions), and stemming, seemed to have no measurable effect or have a negative impact on the performance of topic modeling. Instead, it was suggested as an alternative option to implement these pre-processing tools after running the topic modeling, namely conducting post-processing methodology to decide on a most suitable tool for their application.

With the consideration of spatial adjacency and semantic similarity of the detected outliers, around 100 outliers were regrouped to outlier clusters. It proves the aforementioned assumption about the correlation between outliers, i.e., spatially adjacent and semantically similar outliers within the same time interval are likely to indicate the same social event. However, as discussed in Section 4.1.5, some outlier clusters may be detected positively related to TIFF events, but actually have a negative relation to TIFF since they are far away from TIFF venues. In this case, it results in a higher number of false positives, and thus a lower precision. Taking POIs (Point of Interests) information as additional features for the identification of social events seems to be a potential approach to deal with this problem, as social events gathering many people tend to occur near or around POIs with high popularity.

Comparing with the experiment that applied R. Lee and Sumiya (2010)'s method in the same case study, the proposed spatial-temporal-semantic approach was able to correctly detect 24% more real TIFF events, which illustrates the effectiveness of additional analysis of event content and spatial-semantic correlation. With regard to the detection results of the existing work, they vary with distinct methods and case studies. R. Lee and Sumiya (2010) selected 15 daily events occurring at city level as ground truth, among which 6 events were detected. A performance of recall was $9/15 = 60\%$. Gao, Cao, He, et al. (2013), which improved R. Lee and Sumiya (2010)'s work by representing event content with highly-frequent keywords, could detect 13 events among the prepared 20 events from Weibo data, namely a recall of $13/20 = 65\%$ was reached. In this thesis, the recall was 58% when it reached the highest accuracy of 87%. It seems that the author's

approach has a slight lower performance of recall than the existing methods applied in different case studies, which may be mainly due to the spatiotemporal granularity of ground truth events. Based on the similar spatiotemporal partition, a higher percentage of the listed events is supposedly to be obtained if they refer to events lasting the whole day and covering several partitioned regions, since they have a higher error tolerance than the ones defined in one partitioned region during one specific period. On the other hand, by comparing with the fine-grained ground truth events in this study, it demonstrates that the author's approach has the ability to precisely identify the occurring time and place for social events.

4.4.2 Traffic event detection

This work provided an association rule-based method to deal with the problem that the number of negative samples is far more than the number of positive samples in the query process. Moreover, a hybrid approach was generated to detect traffic events based on space, time and semantics comparing to the existing studies that mostly placed emphasis on the semantic analysis. The promising detection results illustrate the capability of the proposed approach for traffic event identification. More detailed observations and insights are discussed as follows.

The geosocial media data used for the traffic event detection is also Twitter data. The imbalance problems discussed in Section 4.4.1 exist in this study as well. For example, the other types of event information that is embedded in images or videos, and/or posted on the other social media platforms is missed. As such, the traffic related information of multiple social media platforms can be integrated to capture real-world traffic events.

With regard to the classification-based method, SVM method held the best performance among the three classification methods. This may be due to the SVM method being able to handle data with high dimensional feature space, few independent features (most of the features are somewhat relevant to each other or the target prediction variable), and the sparse document vector containing lots of zero values while learning text classifiers (Joachims, 1998).

Based on the clustering method, the generated spatiotemporal clusters indicating traffic events mainly occurred during morning and afternoon peak hours and located in downtown area and highways. On one hand, it indicates that traffic events have a higher possibility to happen around these locations and periods than other arterials and hours. On the other hand, it is likely that the traffic anomalies within these time and space windows involve more people and draw their concerns, which contributes to more relevant posts for traffic event identification.

In this study, the location of detected events was directly inferred using their tagged GPS coordinates. There may exist differences between the location of the tweets reporting the traffic events and the exact traffic event location if people post after they leave the event location. The location information in tweet texts can be extracted as a reference. This tweet texts based approach normally builds a place name dictionary and creates entity annotations to generate candidate place names first, and then converts them to a pair of longitude and latitude based on string matching method.

By conducting a comparison between the classification-based method and the clustering-based method, the latter outperformed the former regarding the percentage of correct detections at the same significance level. Based on the classification method, a total of 1,626 traffic events were detected, while 361 spatiotemporal clusters were grouped to represent traffic events based on the clustering method. To be specific, the classification-based method generates a higher percentage of false positive detections, namely some tweets are identified as traffic events but they do not refer to real-world traffic events. The clustering-based method generates a higher percentage of true positive detections and false negative detections. In other words, some tweets that refer to actual traffic events are discarded while clustering. It means some insignificant traffic events that do not draw several people's attention and involve few posts may be missed by the clustering-based method. Through the validation with vehicle travel speed data, the higher detection rate regarding the same significance level reveals that the clustering-based method can more accurately estimate the location and time of detected events than the classification-based method.

Chapter 5 Conclusions and future work

This chapter briefly summarizes the work described in the previous chapters, concerning detecting small-scale events using geosocial media data from spatial, temporal, and semantic perspectives, where social events and traffic events are selected as two representatives. The experiment results are generally discussed to draw conclusions, followed by the description of future work, improving the research outcomes and generally applying the proposed approaches to other types of small-scale events.

5.1 Conclusions

Timely sensing and detecting small-scale events guides relevant departments and persons to handle the abnormalities. Specifically, detecting traffic events efficiently enables drivers and traffic authorities to come up with responsive plans to manage traffic flow as well as road safety. Detecting spatiotemporal social events provides valuable information for people to make right plans to get involved in (e.g., showing great interest in a sport game) or avoid (e.g., considering a protest may lead to traffic congestion) one or more events. Social media platforms pave a free-cost way for users to publish what is going on with timestamp and geolocation information if the internet is available. They act as abundant source to capture small-scale events relevant information, such as traffic anomalies and social events occurring in the real world, due to the fact these social media contents are shared to the public in a real-time manner.

In this thesis, a spatial-temporal-semantic method has been first proposed to detect social events using geosocial media data. Spatiotemporal outliers were extracted if the pattern (i.e., the number of tweets and the number of users increase) was found to be abnormal within a geographical region during a time period. Topic modeling method was used to analyze the content of the outliers. Considering the characteristics of social events, spatially adjacent and semantically similar outliers were clustered to represent social events in the real world.

The proposed method was tested on detecting TIFF events from Twitter data, which was held during 4-14 September 2014 in Toronto, Canada. The experiment shows that the event content can

be effectively inferred by investigating topic distributions over outlier tweets and word distributions over topics. About 13% decrease in the number of outlier clusters positively supports the assumption about the correlation between outliers that geographically close and semantically coherent outliers may indicate the same social event. An accuracy of 87% was reached by comparing with the ground truth data from official TIFF schedule. Moreover, the proposed approach outperforms the existing methods by successfully extracting more real-world TIFF events from Twitter data, as well as illustrates the ability of precisely identifying the occurring time and place for fine-grained events.

In addition, the capability of the proposed method being generally applied for detecting social events with spatiotemporal burstiness was illustrated. This study sheds light on the emergency management in urban planning. Having a good knowledge of what is happening in a city helps relevant authorities make suitable strategies/policies to deal with the anomalies caused by the gathering, especially for the unplanned events, which accelerates the development of smart cities.

Furthermore, an integrated approach was proposed to detect traffic events from Twitter data. Tweets were collected through Twitter Streaming API using a geo-bounding box. A list of single traffic related keywords were first summarized to implement a query from the raw tweets. Each queried tweet was converted into a set of stemmed word tokens while removing stop words based on NLP tools. Considering the imbalance between positive tweets and negative tweets existed in the collection of initial queried tweets, association rules embedded in the positive tweets were further mined to retain positive tweets as well as discarding negative tweets. A classification-based approach and a clustering-based approach were subsequently proposed to extract traffic events from the association rule-filtered tweets. With regard to the classification-based approach, the association rule-filtered tweets were mapped into a high dimensional vector feature space shaped by the tf-idf extracted features. Three text classification models trained by NB method, SVM method, and LR method were separately used to classify the vectorized tweets into non-traffic events, traffic incidents, and severe traffic conditions. The clustering-based approach integrating space, time, and semantics explored the spatiotemporal characteristics of association rule-filtered tweets in order to group candidate tweets into clusters for traffic event identification. Accordingly,

the content of detected events was automatically summarized by a set of representative terms using TextRank method.

The proposed approaches were applied in Toronto, Canada using one-year Twitter data from April 1, 2014 to March 31, 2015. The result shows that most of the detected traffic events were located in the downtown area and along highways during the morning and afternoon peak hours. Through comparing with the hourly average travel speed data, the classification-based method was able to map around 70% of the detected traffic events to the real world correctly using the limited sample of Twitter data (less than 1% of entire data) at the significance level of 50%. While as many as 86% of the detected events referred to real-world traffic events as the spatiotemporal pattern of traffic relevant posts were integrated for consideration. This outperforming result well proves the assumption that the concentration of tweets that potentially represent traffic anomalies in spatial and temporal dimensions is likely to indicate a traffic event, and demonstrates the clustering-based method helps estimate event location and time.

In summary, upon extracting traffic events from geosocial media data with the proposed approaches, effective measures will be taken to deal with the affected traffic. The air pollution caused by the severe condition of urban traffic can be reduced to improve the quality of health for citizens. In addition, the abundant semantic of geosocial media messages provides efficient ways to explore the reasons behind the traffic events, which can be used as favorable feedback for road networks expansion planning, speed limits setting, and road signs replacement. As such, all of these changes provide a data-driven approach that sheds light on the future development of smart transportation and smart cities.

5.2 Future work

In the future, the general workflow for small-scale event detection will be applied to other types of small-scale events, such as security events. In addition, the capability of other kinds and types of geosocial media data will be explored for small-scale event detection, as well as integrating multiple social media data to better sense what is going on in the city. In addition, sentiment information in short text messages and images can be recognized using sentiment analysis

technology, such as valid Twitter sentiment analysis API or machine learning classification methods, to aid identify different types of events since peoples' emotions (e.g., positive, neutral, or negative) tend to be affected by different topics. As humans are critical components of small-scale events, social network analysis can be added as an additional feature to evaluate their impact on event occurrence and evolution by examining their connections on social media platforms. In this manner, human interactions in network space as well as geospatial space are able to be captured for small-scale event identification. The approaches proposed in this research lay a methodological foundation for detecting small-scale events from streaming geosocial media data in a real-time or near real-time manner. These approaches will be further investigated and adjusted to achieve good performance in real-time or near real-time detection.

More specifically, in the case of social event detection, spatiotemporal clustering method will be applied to automatically generate spatiotemporal clusters for outlier detection. In addition, POIs features with popularity will be considered for social event identification to improve the detection performance. With regard to traffic event detection, the location information buried in geosocial media messages can be used as supplemented references to infer the event location. Deep learning approaches will be tested for event classification given more data is collected.

Appendices

A complete list of topic distributions over an outlier in region #417 are summarized as below. Those topics with a probability of zero are not listed.

Appendix I. Topic distributions over an outlier in region #417

Topic ID	P(t)	Topic ID	P(t)
92	0.463	53	0.01
34	0.19	89	0.006
44	0.124	14	0.003
36	0.1	29	0.003
6	0.073	48	0.003
94	0.016	50	0.003

Appendix II. Word distributions over topics in Appendix I

Topics	Word P(w t)				
Topic #92	tiff	bell	lightbox	princess	wales
	0.130	0.095	0.093	0.055	0.046
	theatre	film	day	topfive	thisisyourfilmfestival
	0.040	0.021	0.011	0.01	0.009
Topic #34	Toronto	day	today	time	love
	0.027	0.014	0.012	0.012	0.011
	good	don	people	back	great
	0.010	0.01	0.008	0.008	0.008
Topic #44	Toronto	tiff	film	festival	party
	0.092	0.072	0.029	0.019	0.018
	movie	street	night	premiere	red

	0.017	0.015	0.015	0.013	0.012
Topic #36	grolschza	jauja	laggies	wine	famous
	0.020	0.015	0.013	0.013	0.011
	winery	discover	stands	setonking	lisandro
	0.009	0.009	0.009	0.009	0.009
Topic #6	bae	nyc	pureleaf	maker	series
	0.025	0.010	0.010	0.008	0.008
	gotham	na	caaarrl	lord	teaser
	0.006	0.006	0.006	0.006	0.006
Topic #94	west	queen	apple	coming	bills
	0.019	0.016	0.015	0.014	0.011
	buffalo	Tuesday	race	soknacki	neighbourhood
	0.009	0.009	0.009	0.009	0.008
Topic #53	bgfest	cube	pr	youlculuk	miyerine
	0.022	0.015	0.015	0.011	0.011
	nya	triflic	karijobe	beverleyhotelto	mbartyzel
	0.011	0.011	0.009	0.009	0.009
Topic #89	ryan	thevoices	ryanreynolds	editor	reynolds
	0.034	0.022	0.022	0.022	0.02
	voices	horror	kendrick	theeditormovie	brofromanother
	0.016	0.012	0.012	0.01	0.01
Topic #14	roy	hall	thomson	tiff	tiffrc
	0.065	0.06	0.036	0.023	0.021
	de	le	bluecarpet	onrc	thompson
	0.021	0.019	0.018	0.016	0.014
Topic #29	al	thehumbling	pacino	drake	bellwoods
	0.022	0.018	0.017	0.017	0.009
	trinity	person	ptxofficial	squirrel	bieber
	0.009	0.009	0.007	0.007	0.007
Topic #48	rosewater	jon	stewart	wild	foxcatcher

	0.04	0.036	0.036	0.02	0.018
	tatum	channing	reese	cheryl	gael
	0.018	0.016	0.014	0.013	0.011
Topic #50	bttoronto	seneca	ceic	ryerson	relationship
	0.022	0.018	0.013	0.012	0.01
	building	basically	jeffler	bt	newham
	0.01	0.01	0.008	0.008	0.008

References

- Abdelhaq, H., Gertz, M., & Armiti, A. (2017). Efficient online extraction of keywords for localized events in twitter. *GeoInformatica*, 21(2), 365–388.
<https://doi.org/10.1007/s10707-016-0258-x>
- Abdelhaq, H., Gertz, M., & Sengstock, C. (2013). Spatio-temporal characteristics of bursty words in Twitter streams. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 194–203). ACM.
<https://doi.org/10.1145/2525314.2525354>
- Abdelhaq, H., Sengstock, C., & Gertz, M. (2013). EvenTweet: Online localized event detection from Twitter. *Proceedings of the VLDB Endowment*, 6(12), 1326–1329.
<https://doi.org/10.14778/2536274.2536307>
- Acuna, E., & Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical Paper, Department of Mathematics, University of Puerto Rico at Mayaguez*, (May), 1–25. Retrieved from <http://paperout.pdf>
- Agha, N., & Taks, M. (2015). A theoretical comparison of the economic impact of large and small events. *International Journal of Sport Finance*, 10(3), 199–216.
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, 1215*, 487–499. Retrieved from papers2://publication/uuid/E20EA258-DABA-4B88-A7D4-C4DED7E4C0ED
- Alfarrarjeh, A., Agrawal, S., Kim, S. H., & Shahabi, C. (2017). Geo-spatial multimedia sentiment analysis in disasters. In *2017 International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 193–202). <https://doi.org/10.1109/DSAA.2017.77>
- Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and Its Applications*, 7(3), 176–204.
- Analytics, P. (2009). *Twitter study–August 2009*. Retrieved from <https://pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf>
- Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing*

- in Science and Engineering*, 15(3), 72–82. <https://doi.org/10.1109/MCSE.2013.70>
- Aziz, M. V. G., Prihatmanto, A. S., Henriyan, D., & Wljaya, R. (2015). Design and implementation of natural language processing with syntax and semantic analysis for extract traffic conditions from social media data. In *2015 5th IEEE International Conference on System Engineering and Technology (ICSET)* (pp. 43–48). Shah Alam, Malaysia: IEEE.
- Banweer, K., Graham, A., Ripberger, J., Cesare, N., Nsoesie, E., & Grant, C. (2018). Multi-stage collaborative filtering for tweet geolocation. In *LocalRec@ SIGSPATIAL* (pp. 4–1). <https://doi.org/10.1145/3282825.3282831>
- Batty, M., Desyllas, J., & Duxbury, E. (2003). The discrete dynamics of small-scale events: Agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, 17(7), 673–697. <https://doi.org/10.1080/1365881031000135474>
- Bian, J., Yang, Y., Zhang, H., & Chua, T. S. (2015). Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2), 216–228. <https://doi.org/10.1109/TMM.2014.2384912>
- Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, 60(1), 208–221. <https://doi.org/10.1016/j.datak.2006.01.013>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
- Boettcher, A., & Lee, D. (2012). EventRadar: A real-time local event detection scheme using Twitter stream. In *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on* (pp. 358–367). IEEE. <https://doi.org/10.1109/GreenCom.2012.59>
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hyper textual web search engine. *Computer Networks and ISDN Systems*. <https://doi.org/10.1109/ICCEE.2009.59>
- Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liang, L., & Ratti, C. (2010). The geography of taste: analyzing cell-phone mobility and social events. In *International conference on pervasive computing* (pp. 22–37). Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-12654-3>

- Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675–684). Hyderabad, India: ACM.
- Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (pp. 143–152). IEEE.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27. Retrieved from <http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/papers/libsvm.pdf>
- Chaurasia, N., & Tiwari, A. (2014). On the use of Brokerage approach to discover influencing nodes in terrorist networks. In *Social Networking* (pp. 271–295). Springer, Cham.
- Chen, P.-T., Chen, F., & Qian, Z. (2014). Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields. *2014 IEEE International Conference on Data Mining*, 80–89. <https://doi.org/10.1109/ICDM.2014.139>
- Chen, S., Wang, W., & van Zuylen, H. (2010). A comparison of outlier detection algorithms for ITS data. *Expert Systems with Applications*, 37(2), 1169–1178. <https://doi.org/10.1016/j.eswa.2009.06.008>
- Cheng, T., & Wicks, T. (2014). Event detection using twitter: A spatio-temporal approach. *PLoS ONE*, 9(6), 1–10. <https://doi.org/10.1371/journal.pone.0097807>
- Conway, M., & McInerney, L. (2008). Jihadi video and auto-radicalisation: Evidence from an exploratory YouTube study. In *European Conference on Intelligence and Security Informatics* (pp. 108–118). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-89900-6_13
- Cordeiro, M. (2012). Twitter event detection: combining wavelet analysis and topic inference summarization. In *Proceedings of Doctoral Symposium on Informatics Engineering* (pp. 11–16).
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 29(20), 273–297. <https://doi.org/10.1111/j.1747-0285.2009.00840.x>
- Costas, H. C., Vilas, A. F., Vicente, M. M., & Díaz Redondo, R. P. (2018). Discovering geo-dependent stories by combining density-based clustering and thread-based aggregation

- techniques. *Expert Systems with Applications*, 95, 32–42.
<https://doi.org/10.1016/j.eswa.2017.11.019>
- Cui, J., Fu, R., Dong, C., & Zhang, Z. (2014). Extraction of traffic information from social media interactions: Methods and experiments. *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 1549–1554. <https://doi.org/10.1109/ITSC.2014.6957913>
- D’Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-time detection of traffic from twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(4), 2269–2283. Retrieved from www.ijiset.com
- Dabiri, S., & Heaslip, K. (2019). Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, 118, 425–439.
<https://doi.org/10.1016/j.eswa.2018.10.017>
- Daly, E. M., Lécué, F., & Bicer, V. (2013). Westland row why so slow? Fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 203–212). Santa Monica, California, USA: ACM.
- De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1), 1–18.
[https://doi.org/10.1016/S0169-7439\(99\)00047-7](https://doi.org/10.1016/S0169-7439(99)00047-7)
- Domínguez, D. R., Díaz Redondo, R. P., Vilas, A. F., & Khalifa, M. Ben. (2017). Sensing the city with Instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, 78, 319–333. <https://doi.org/http://dx.doi.org/10.1016/j.eswa.2017.02.018>
- Endarnoto, S. K., Pradipta, S., Nugroho, A. S., & Purnama, J. (2011). Traffic condition information extraction & visualization from social media twitter for android mobile application. In *2011 International Conference on Electrical Engineering and Informatics (ICEEI)* (pp. 1–4). Bandung, Indonesia: IEEE. <https://doi.org/10.1109/ICEEI.2011.6021743>
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226–231.
- Fan, R. E., Chang, K. W., Hsieh, C. J., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, 1871–1874. Retrieved from <http://www.csie.ntu.edu.tw/>

- Feng, W., Zhang, C., Zhang, W., Han, J., Wang, J., Aggarwal, C., & Huang, J. (2015). STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on* (pp. 1561–1572). IEEE. <https://doi.org/10.1109/ICDE.2015.7113425>
- Fu, K., Lu, C.-T., Nune, R., & Tao, J. X. (2015). Steds: Social media based transportation event detection with text summarization. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1952–1957). Las Palmas, Spain: IEEE. <https://doi.org/10.1109/ITSC.2015.316>
- Fujisaka, T., Lee, R., & Sumiya, K. (2010). Discovery of user behavior patterns from geo-tagged microblogs. In *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication* (p. 36). <https://doi.org/10.1145/2108616.2108660>
- Gao, X., Cao, J., He, Q., & Li, J. (2013). A novel method for geographical social event detection in social media. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service* (pp. 305–308). <https://doi.org/10.1145/2499788.2499819>
- Gao, X., Cao, J., Jin, Z., Li, X., & Li, J. (2013). GeSoDeck a geo-social event detection and tracking system. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 471–472).
- Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using Oriented Violent Flows. *Image and Vision Computing*, 48(2016), 37–41. <https://doi.org/10.1016/j.imavis.2016.01.006>
- Getz, D., & Page, S. (2016). *Event studies: Theory, research and policy for planned events*. Routledge.
- Giridhar, P., Abdelzaher, T., George, J., & Kaplan, L. (2015). On quality of event localization from social network feeds. *2015 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2015*, 75–80. <https://doi.org/10.1109/PERCOMW.2015.7133997>
- Gratton, C., & Taylor, P. (2000). *Economics of sport and recreation*. E & FN Spon Ltd.
- Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 24(24). Retrieved from <https://cloudfront.escholarship.org/dist/prd/content/qt44x9v7m7/qt44x9v7m7.pdf>

- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1), 5228–5235.
<https://doi.org/10.1073/pnas.0307752101>
- Gu, Y., Qian, Z. (Sean), & Chen, F. (2016a). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67, 321–342. <https://doi.org/10.1016/j.trc.2016.02.011>
- Gu, Y., Qian, Z. (Sean), & Chen, F. (2016b). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67, 321–342. <https://doi.org/10.1016/j.trc.2016.02.011>
- Gutiérrez, C., Figuerias, P., Oliveira, P., Costa, R., & Jardim-Goncalves, R. (2015). Twitter mining for traffic events detection. In *2015 Science and Information Conference* (pp. 371–378). London, UK: IEEE. <https://doi.org/10.1109/SAI.2015.7237170>
- Han, Z., Li, S., Cui, C., Han, D., & Song, H. (2019). Geosocial media as a proxy for security: a review. *IEEE Access*, 7(October), 154224–154238.
<https://doi.org/10.1109/ACCESS.2019.2949115>
- Hawkins, D. M. (1980). *Identification of Outliers* (Vol. 11). London: Chapman and Hall.
<https://doi.org/10.1007/978-94-015-3994-4>
- Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80–88). acm.
- Hossny, A. H., & Mitchell, L. (2018). Event detection in twitter: A keyword volume approach. *IEEE International Conference on Data Mining Workshops, ICDMW*, 1200–1208.
<https://doi.org/10.1109/ICDMW.2018.00172>
- Hu, M., Wongsuphasawat, K., & Stasko, J. (2017). Visualizing social media content with sententree. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), 621–630.
<https://doi.org/10.1109/TVCG.2016.2598590>
- Huang, W., Fan, H., & Zipf, A. (2017). Towards detecting the crowd involved in social events. *ISPRS International Journal of Geo-Information*, 6(10), 305.
<https://doi.org/10.3390/ijgi6100305>
- Hurlock, J., & Wilson, M. L. (2011). Searching Twitter: Separating the tweet from the chaff. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Searching* (pp. 161–168).

- Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2), 237–253. <https://doi.org/10.1007/s10844-017-0458-3>
- Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1398, 137–142. <https://doi.org/10.1007/s13928716>
- Kamran, S., & Haas, O. (2007). A multilevel traffic incidents detection approach : identifying traffic patterns and vehicle behaviours using GPS data. In *Intelligent Vehicles Symposium, 2007 IEEE* (pp. 912–917). Istanbul, Turkey: IEEE. <https://doi.org/10.1109/IVS.2007.4290233>
- Kaneko, T., & Yanai, K. (2016). Event photo mining from Twitter using keyword bursts and image clustering. *Neurocomputing*, 172, 143–158. <https://doi.org/10.1016/j.neucom.2015.02.081>
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Transactions on Pattern Analysis & Machine Intelligence*, 7, 881–892.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68. <https://doi.org/10.1016/j.bushor.2009.09.003>
- Kelley, M. J. (2013). The emergent urban imaginaries of geosocial media. *GeoJournal*, 78(1), 181–203. <https://doi.org/10.1007/s10708-011-9439-1>
- Khalifa, M. Ben, Díaz Redondo, R. P., Vilas, A. F., & Rodríguez, S. S. (2017). Identifying urban crowds using geo-located Social media data: a Twitter experiment in New York City. *Journal of Intelligent Information Systems*, 48(2), 287–308. <https://doi.org/10.1007/s10844-016-0411-x>
- Klausen, J. (2015). Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq. *Studies in Conflict and Terrorism*, 38(1), 1–22. <https://doi.org/10.1080/1057610X.2014.974948>
- Kong, L., Liu, Z., & Huang, Y. (2014). SPOT: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, 7(13), 1681–1684.

<https://doi.org/10.14778/2733004.2733060>

Krumm, J., & Horvitz, E. (2015). Eyewitness : Identifying local events via space-time signals in Twitter feeds. In *Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 20).

<https://doi.org/10.1145/2820783.2820801>

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., & Shoor, I. (2017).

Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies*, 77, 275–291. <https://doi.org/10.1016/j.trc.2017.02.003>

Kulldorff, M., Athas, W. F., Feuer, E. J., Miller, B. A., & Key, C. R. (1998). Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health*, 88(9), 1377–1380. <https://doi.org/10.2105/AJPH.88.9.1377>

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3), 0216–0224. <https://doi.org/10.1371/journal.pmed.0020059>

Kumar, A., Jiang, M., & Fang, Y. (2014). Where not to go? Detecting road hazards using Twitter. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 1223–1226). Gold Coast, Queensland, Australia: ACM.

Kurniawan, D. A., Wibirama, S., & Setiawan, N. A. (2016). Real-time traffic classification with Twitter data mining. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 165–169). Yogyakarta, Indonesia: IEEE. <https://doi.org/10.1109/ICITEED.2016.7863251>

Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven months with the Devils : A long-term study of content polluters on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 185–192).

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* (pp. 1–10). San Jose, California: ACM.

Lee, R., Wakamiya, S., & Sumiya, K. (2011). Discovery of unusual regional social activities

- using geo-tagged microblogs. *World Wide Web*, 14(4), 321–349.
<https://doi.org/10.1007/s11280-011-0120-x>
- Li, Q., Shah, S., Thomas, M., Anderson, K., & Liu, X. (2016). How much data do you need ? Twitter decahose data analysis. In *The 9th International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction*. Washington, DC, USA.
- Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012). TEDAS: A twitter-based event detection and analysis system. In *2012 IEEE 28th International Conference on Data Engineering (ICDE)* (pp. 1273–1276). Washington, DC, USA: IEEE.
<https://doi.org/10.1109/ICDE.2012.125>
- Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ... Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119–133.
<https://doi.org/10.1016/j.isprsjprs.2015.10.012>
- Liang, Y., Caverlee, J., & Cao, C. (2015). A noise-filtering approach for spatio-temporal event detection in social media. In *European Conference on Information Retrieval* (pp. 233–244). Springer, Cham. https://doi.org/10.1007/978-3-319-16354-3_25
- Lin, C., He, Y., Everson, R., & Rüger, S. (2012). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), 1134–1145. <https://doi.org/10.1109/TKDE.2011.48>
- Liu, M., Fu, K., Lu, C.-T., Chen, G., & Wang, H. (2014). A search and summary application for traffic events detection based on Twitter data. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 549–552). Dallas, Texas: ACM. <https://doi.org/10.1145/2666310.2666366>
- Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.
- Mahmud, J., Nichols, J., & Drews, C. (2014). Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, 5(3), 1–21.
<https://doi.org/10.1145/2528548>
- Mai, E., & Hranac, R. (2013). Twitter interactions as a data source for transportation incidents. In *Transportation Research Board 92nd Annual Meeting (No. 13-1636)*. Washington, DC, USA. Retrieved from <http://docs.trb.org/prp/13-1636.pdf>

- Mair, D. (2016). #Westgate: A case study: How al-Shabaab used twitter during an ongoing attack. In *Violent Extremism Online* (pp. 81–102). Routledge.
<https://doi.org/10.1080/1057610X.2016.1157404>
- Manning, C., Prabhakar, R., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100–103. <https://doi.org/10.1109/LPT.2009.2020494>
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Stroudsburg, PA, USA: Association for Computational Linguistics.
<https://doi.org/10.3115/v1/P14-5010>
- Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). TwitInfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 227–236). ACM.
<https://doi.org/10.1145/1978942.1978975>
- Matheson, V. A. (2006). Is smaller better? A comment on “comparative economic impact analyses” by Michael Mondello and Patrick Rishe. *Economic Development Quarterly*, 20(2), 192–195.
- Mccallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Retrieved from citeulike-article-id:1062263
- Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. Retrieved from <http://www.aclweb.org/anthology/W04-3252>
- Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter’s Streaming API with Twitter’s firehose. In *ICWSM* (pp. 400–408). https://doi.org/10.1007/978-3-319-05579-4_10
- Nguyen, H., Liu, W., Rivera, P., & Chen, F. (2016). TrafficWatch: Real-time traffic incident detection and monitoring using social media. In *PAKDD 2016: Advances in Knowledge Discovery and Data Mining* (pp. 540–551). Springer, Cham. <https://doi.org/10.1007/978-3-319-31753-3>
- Ozdikis, O., Oğuztüzün, H., & Karagoz, P. (2017). A survey on location estimation techniques

- for events detected in Twitter. *Knowledge and Information Systems*, 52(2), 291–339.
<https://doi.org/10.1007/s10115-016-1007-z>
- Ozdikis, O., Ramampiaro, H., & Nørnvåg, K. (2018). Spatial statistics of term co-occurrences for location prediction of tweets. In *European Conference on Information Retrieval* (pp. 494–506). Springer, Cham. <https://doi.org/10.1007/978-3-319-76941-7>
- Park, E. S., Turner, S., & Spiegelman, C. H. (2007). Empirical Approaches to Outlier Detection in Intelligent Transportation Systems Data. *Transportation Research Record: Journal of the Transportation Research Board*, 1840(1), 21–30. <https://doi.org/10.3141/1840-03>
- Park, H., & Haghani, A. (2016). Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C*, 70, 69–85.
<https://doi.org/10.1016/j.trc.2015.03.018>
- Pozdnoukhov, A., & Kaiser, C. (2011). Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 1–8). ACM. <https://doi.org/10.1145/2063212.2063223>
- Ranneries, S. B., Kalør, M. E., Nielsen, S. A., Dalgaard, L. N., Christensen, L. D., & Kanhabua, N. (2016). Wisdom of the local crowd: Detecting local events using social media data. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 352–354). ACM.
<https://doi.org/10.1145/2908131.2908197>
- Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 75, 197–211.
<https://doi.org/10.1016/j.trc.2016.12.008>
- Ribeiro Jr., S. S., Davis Jr., C. A., Oliveira, D. R. R., Meira Jr., W., Gonçalves, T. S., & Pappa, G. L. (2012). Traffic Observatory: A system to detect and locate traffic events and conditions using Twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 5–11). Redondo Beach, California: ACM. <https://doi.org/10.1145/2442796.2442800>
- Ribeiro, S., & Pappa, G. L. (2018). Strategies for combining Twitter users geo-location methods. *GeoInformatica*, 22(3), 563–587. <https://doi.org/10.1007/s10707-017-0296-z>
- Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate Data. *Journal of the American Statistical Association*, 91(435), 1047–1061.

- Sakaki, T., Matsuo, Y., Yanagihara, T., Chandrasiri, N. P., & Nawa, K. (2012). Real-time event extraction for driving information from social sensors. In *2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)* (pp. 221–226). Bangkok, Thailand: IEEE. <https://doi.org/10.1109/CYBER.2012.6392557>
- Salem, A., Reid, E., & Chen, H. (2008). Multimedia content coding and analysis: Unraveling the content of jihadi extremist groups' videos. *Studies in Conflict and Terrorism*, 31(7), 605–626. <https://doi.org/10.1080/10576100802144072>
- Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding text pre-processing for Latent Dirichlet Allocation. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics* (pp. 432–436). Retrieved from <http://www.cs.cornell.edu/~xanda/winlp2017.pdf>
- Schofield, A., & Mimno, D. (2016). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics*, 4, 287–300. Retrieved from <https://transacl.org/ojs/index.php/tacl/article/view/868>
- Schulz, A., & Ristoski, P. (2013). The car that hit the burning house: Understanding small scale incident related information in microblogs. *AAAI Technical Report / WS, 13-04*, 11–14. Retrieved from http://ub-madoc.bib.uni-mannheim.de/35450/1/The_Car_That_Hit_The_Burning_House_Understanding_Small_Scale_Incident_Related_Information_in_Microblogs.pdf
- Schulz, A., Ristoski, P., & Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. In *Extended Semantic Web Conference* (pp. 22–33). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41242-4_3
- Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., & Adini, B. (2014). Twitter in the cross fire - The use of social media in the Westgate mall terror attack in Kenya. *PLoS ONE*, 9(8), e104136. <https://doi.org/10.1371/journal.pone.0104136>
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7), 424–440. Retrieved from <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4912206>
- Sureka, A., Kumaraguru, P., Goyal, A., & Chhabra, S. (2010). Mining YouTube to discover extremist videos, users and hidden communities. In *Asia Information Retrieval Symposium* (pp. 13–24). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17187-1_2

- Tejaswin, P., Kumar, R., & Gupta, S. (2015). Tweeting Traffic : Analyzing Twitter for generating real-time city traffic insights and predictions. In *Proceedings of the 2nd IKDD Conference on Data Sciences* (p. 9). Bangalore, India: ACM.
<https://doi.org/10.1145/2778865.2778874>
- Thom, D., Bosch, H., Koch, S., Worner, M., & Ertl, T. (2012). Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In *Visualization Symposium (PacificVis), 2012 IEEE Pacific* (pp. 41–48). IEEE.
<https://doi.org/10.1109/PacificVis.2012.6183572>
- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
- Walther, M., & Kaisser, M. (2013). Geo-spatial event detection in the Twitter stream. In *European conference on information retrieval* (pp. 356–367). Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-36973-5_30
- Wang, S., He, L., Stenneth, L., Yu, P. S., & Li, Z. (2015). Citywide traffic congestion estimation with social media. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 34). Seattle, Washington: ACM.
- Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from Twitter posts. In *5th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP)* (pp. 231–238). College Park, MD, USA: Springer International Publishing.
- Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., & Chaovalit, P. (2011). Social-based traffic information extraction and classification. In *2011 11th International Conference on ITS Telecommunications (ITST)* (pp. 107–112). St. Petersburg, Russia: IEEE. <https://doi.org/10.1109/ITST.2011.6060036>
- Warren, R., Smith, R. E., & Cybenko, A. K. (2011). Use of mahalanobis distance for detecting outliers and outlier clusters in markedly non-normal data: A vehicular traffic example. *SRA INTERNATIONAL INC DAYTON OH*.
- Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 2541–2544). Glasgow, Scotland, UK: ACM.
<https://doi.org/10.1145/2063576.2064014>

- Xia, C., Hu, J., Zhu, Y., & Naaman, M. (2015). What is new in our city? A frame work for event extraction using social media posts. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 16–32). Springer, Cham. <https://doi.org/10.1007/978-3-319-18038-0>
- Xie, K., Xia, C., Grinberg, N., Schwartz, R., & Naaman, M. (2013). Robust detection of hyper-local events from geotagged social media data. In *Proceedings of the Thirteenth International Workshop on Multimedia Data Mining* (p. 2). <https://doi.org/10.1145/2501217.2501219>
- Xu, S., Li, S., & Wen, R. (2018, June). Sensing and detecting traffic events using geosocial media data: A review. *Computers, Environment and Urban Systems*. <https://doi.org/10.1016/j.compenvurbsys.2018.06.006>
- Yazici, M. A., Mudigonda, S., & Kamga, C. (2017). Incident detection through Twitter organization vs. personal accounts (No. 17-03884).
- Zandbergen, P. A., & Barbeau, S. J. (2011). Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. *Journal of Navigation*, 64(3), 381–399. <https://doi.org/10.1017/S0373463311000051>
- Zhang, C., Liu, L., Lei, D., Zhuang, H., Hanraay, T., & Han, J. (2017). TrioVecEvent : Embedding-based online local event detection in geo-tagged tweet streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 595–604). ACM. <https://doi.org/10.1145/3097983.3098027>
- Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., ... Han, J. (2016). GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 513–522. <https://doi.org/10.1145/2911451.2911519>
- Zhang, S., Cheng, Y., & Ke, D. (2017). Event-Radar: Real-time local event detection system for geo-tagged tweet streams. Retrieved from <http://arxiv.org/abs/1708.05878>
- Zhang, S., Tang, J., Wang, H., & Wang, Y. (2015). Enhancing traffic incident detection by using spatial point pattern analysis on social media. *Transportation Research Record: Journal of the Transportation Research Board*, 2528, 69–77. <https://doi.org/10.3141/2528-08>
- Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., & He, X. (2016). A new method for violence detection in surveillance scenes. *Multimedia Tools and Applications*, 75(12), 7327–7349. <https://doi.org/10.1007/s11042-015-2648-8>

- Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys and Tutorials*, 12(2), 159–170.
<https://doi.org/10.1109/SURV.2010.021510.00088>
- Zhang, Z., Ni, M., He, Q., Gao, J., Gou, J., & Li, X. (2016). An exploratory study on the correlation between Twitter concentration and traffic surge. *Transportation Research Record*, 35, 36. Retrieved from the possibility of detect traffic from Twitter
- Zhao, L., Chen, F., Lu, C.-T., & Ramakrishnan, N. (2015). Spatiotemporal event forecasting in social media. <https://doi.org/10.1137/1.9781611974010.108>