Modern Management based on Big Data I
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200642

Review of the Application of Social Media Data in Disaster Research

Jiting TANG [a,b,c], Saini YANG [a,b,c,1] and Weiping WANG [d,e,f]

[a] *Key Laboratory of Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal University, China*

[b] *State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, China*

[c] *Academy of Disaster Reduction and Emergency Management, Faculty of Geographical Science, Beijing Normal University, China*

[d] *Institute of Transportation Systems Science and Engineering, Beijing Jiaotong University, China*

[e] *State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China*

[f] *Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Ministry of Transport, Beijing Jiaotong University, China*

Abstract. Social media data (SMD) is a new data source in disaster research, which can be used in hazard identification, disaster analysis, risk assessment and emergency rescue. This data-driven disaster research needs to find an appropriate method considering the aspect of data sensitivity. So far, the research in this area is focused on the types of hazard, but rarely considers the relationship between the technical methods and applicable tasks. By emphasizing data and method dependencies, we have attempted to summarize the characteristics of SMD in disaster research, *viz.*, "sociality, rapidity, subjectivity, and un-authenticity", and explore the processing methods in the applications of disaster management. Our work provides ideas and reference to the researchers working in this area from the perspectives of data and research goals.

Keywords. Social media, disaster research, data mining

1. Introduction

Disasters caused by natural hazards are receiving increasing attention globally. They cause enormous casualties and huge economic losses, and adversely affect social stability. Simultaneously, social media popularity for sudden major disasters has also surged. Many individuals employ social media as an effective channel for timely accessible information in emergencies. Government agencies and departments also use social media to disseminate and communicate crucial information in emergencies [1]. Social media data (SMD), as a typical type of big data, is featured with massive data scales, rapid data flow transfers, diverse data types, and low-value density. It can benefit for the whole process of disaster management by quickly obtaining the disaster

[1] Corresponding Author. E-mail: yangsaini@bnu.edu.cn.

information, identifying the scale and location of relief in near real-time, as well as by providing vital information on disaster recovery. This can provide a potential direction for customizing the emergency response, disaster rescue schemes and planning for reduction of the disaster risk in various disasters.

Currently, the SMD-based disaster research is receiving growing attention. However, the existing research in this field is mostly summarized from the perspective of the types of hazard [2][3]. Additionally, the perspective of the applicability and effectiveness of data and methods has not been extensively studied. In this pursuit, we aim to systematize the knowledge about the SMD-based disaster research in recent years, emphasizing the characteristics and processing methods of SMD for disaster research. We hope our work can provide a reference to the researchers working in the area of SMD-based disaster research.

2. Methodology

We employed two different approaches to investigate the application status of SMD in disaster research.

Initially, a Strengths Weaknesses Opportunities Threats (SWOT) analysis [4] was applied to state the general characteristics of SMD in disaster research. SWOT analysis is a powerful method in strategic planning and research, especially in developing a complete understanding of the key factors. It has been widely applied in the field of technology innovations [5], company actions [6], and energy planning [7].

Subsequently, a literature review was conducted. From the perspective of data formats and tasks in the process of disaster management, we studied the research results of SMD-based disaster research in recent years and summarized the processing methods. The data forms included structured numeric data and unstructured data (text and images). Application tasks included disaster identification, disaster spatiotemporal analysis, disaster intensity assessment, public sentiment analysis, rumor identification, communication node analysis, and information display.

3. Characteristics of Social Media Data in Disaster Research

Using the SWOT analysis, we outlined the characteristics of SMD in disasters into sociality, rapidity, subjectivity and un-authenticity (Fig. 1).

Sociality refers to a large number of users about the wide coverage of social media platforms and a huge amount of crowdsourcing data, which is the major advantage of social media-based disaster research and brings many opportunities for big data research and application. Rapidity refers to the swift information production, transmission, and reaction to emergencies in social media, which is an advantage for sudden-onset natural hazards, but sometimes becomes a disadvantage in the research of slow-onset hazards. Subjectivity refers to SMD by means of the subjective factors including user literacy, personal experiences, different mindsets, which is a disadvantage and challenge faced by SMD in disaster research. Un-authenticity refers to the formidability in verifying the authenticity of social media, when the information occurs. Rumor is a major inference in SMD-based disaster research, and also a disadvantage to some extent.

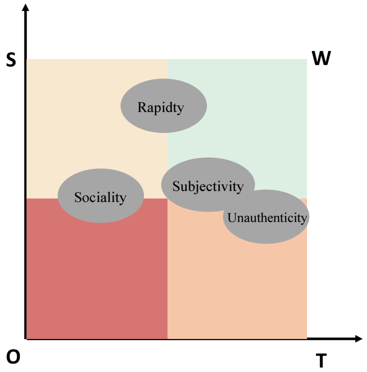


Figure 1. SWOT analysis of SMD in disaster research

3.1. Sociality

Currently, 3.48 billion people worldwide are active on social media [8]. Users on the social media are both consumers and producers [9], and information is generated spontaneously. When a natural hazard occurs, the internet users in the affected area often share their perceptions on the hazard intensity and disaster relief processes. As a result, the researchers can directly collect near real-time disaster information [10]. Furthermore, the sociality of SMD is also reflected in the mutual feedback mechanism between the individuals and the administrators. Individuals provide the disaster-related data promptly by means of crowdsourcing. The relevant government departments and institutes apply SMD to disaster management or research for focusing on the real-time dynamic monitoring, analysis and comprehensive cognition, and attempting disaster relief by disaster insurance and other policies to serve the public. In turn, social media plays a positive role in promoting disaster reduction and supervising disaster management. For example, massive SMD related to Beijing air pollution in 2011 pushed Beijing to become one of the first cities in China with Air Quality Index-PM2.5 monitoring and data release, which accelerated the air pollution governance in Beijing.

3.2. Rapidity

Social media data is updated rapidly spatially and temporally. Since a portion of SMD has location information, there is no need for the secondary spatial positioning processing. Compared to traditional interviews and surveys, this method can get prompt feedback on disaster situations. Even when sudden-onset natural hazards occur, some social media information dissemination can be faster than the broadcasting of a monitoring system. Several institutes in the United States [11], Japan [12], the Netherlands [13], and Australia [14] utilize the promptness of SMD to improve emergency warnings, disasters identification by tracking and monitoring the social media dataflow.

The rapidity of SMD acquirement is conducive to seizing the golden period of post-disaster response after the occurrence of a sudden natural hazard. However, social media is information-centric, which indicates that new hotspots can effortlessly weaken the attention of the original events. Therefore, the unstable social data flow is not beneficial for slow-onset hazards. Rapidity is a double-edged sword in the social media-based disaster research. It is indispensable to consider the limitations and applicability of different hazard types.

3.3. Subjectivity

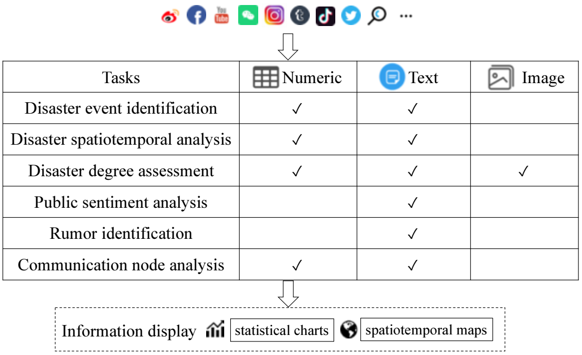
Social media data is a record of netizens, rather than a group of professionals and scientists, so the data may be inaccurate, and the scientific meaning of the data may not be satisfying. The experience and expression of hazards will affect the reliability and validity of disaster-related SMD. Some scholars suggest that people who have experienced disasters have a higher sensitivity to climate change and higher awareness of local natural environment vulnerability [15][16]. There may be unconscious errors when the netizens express about the disasters on social media, and the impact of subjective intention on data quality should not be ignored. To the best of our knowledge, there is no existing work that studies the problem of subjectivity in SMD-based disaster research.

3.4. Un-authenticity

Social media data has a wide coverage and a fast propagation without verification [17], which indicates that the Internet may amplify or attenuate the risk and social impact of disasters [18]. The propagation speed of rumor is six times that of fact, and the probability of fake news being forwarded is 70% higher than that of real news [19]. Disaster-related rumors are often more likely to attract the attention, which may not only cause negative public emotions but also result in a large-scale panic and economic losses. This conscious data error can severely interfere with disaster analysis and emergency management.

4. Processing Methods of Social Media Data in Disaster Research

The use of SMD for disaster research is one of the applications of data mining technology. It must be noted that the performance of different processing methods depends on the quality and availability of its underlying data, since it is a well-known weakness of any data-driven approach. The data formats and general tasks used during the research are depicted in Fig. 2.



Tasks	Numeric	Text	Image
Disaster event identification	✓	✓	
Disaster spatiotemporal analysis	✓	✓	
Disaster degree assessment	✓	✓	✓
Public sentiment analysis		✓	
Rumor identification		✓	
Communication node analysis	✓	✓	

Figure 2. Application tasks of SMD-based disaster research

4.1. Disaster Event Identification

By monitoring the abnormal fluctuation of social media related to disasters in real-time, we can identify a disaster event and broadcast disaster information promptly. Lee et al. monitored the number of updated posts, active users, and the regional mobile population in a certain space. A large change in this dataflow with the historical data suggests an emergency in an area [20]. However, this method does not filter the noise in social media. Sakaki et al. introduced a particle filter and Kalman filter in the signal processing algorithm to preprocess noise text data [12]. Robinson et al. pre-defined earthquake-related keywords and then monitored the fluctuation, and focused on the continuous alarms based on the time series to reduce the false alarm rate [21]. Abhik et al. calculated the text similarity based on the TF-IDF algorithm for clustering and then identifying the disaster sub-events [22]. Yin et al. optimized the method of online incremental clustering and offline merging to detect multiple hazard types [23].

4.2. Disaster Spatiotemporal Analysis

The spatial pattern and temporal process of disasters can be analyzed in a data-driven manner.

Temporal analysis normally extracts the social media information published in the first instance, and then analyzes the situation or trend of key indicators in a continuous period for a disaster information theme (such as a hot spot of social attention, hazard intensity and affected areas). The methods for time series analysis include the trend fitting [24], ARIMA model [25], and the gray forecast model [26]. However, the information release time may not necessarily be the same as the disaster occurrence time, and the bias between the two is often ignored in the current research.

Spatial analysis of the geographical location is generally obtained in an event to analyze the coverage of a hazard or disaster trend. There are three common steps: providing the longitude and latitude information directly; converting the coordinates according to the IP address of the equipment; extracting the words involving the place names, administrative area codes, postal codes or Points of Information (POI) in the text, and then analyzing them into coordinates by matching with the toponym database [23]. But the toponym ambiguity during extracting the words from text has not been solved completely, and there is still a bias in the fine-grained positioning. Besides, due to the regional differences in user distribution, the spatiotemporal distribution of the disaster events is possibly biased towards the population gathering areas [27]. Some studies introduce the user activity weighting [28] or disaster-related ratio weighting [29] to alleviate the spatial heterogeneity of user distribution.

4.3. Disaster Intensity Assessment

Social media data related to disasters generally describes the impact of the disaster. The challenging part of this task is latent semantic analysis. There is few social media-based term thesaurus or corpus with enough data in the field of natural hazard and emergency management, and the direct use of existing data sets in other fields may reduce the accuracy of disaster intensity classification. Some researchers share tweet labeled sample sets of several disasters such as CrisisNLP [30] or CrisisLex [31].

The near real-time disaster assessment is mainly carried to identify the affected area and estimate the loss of the exposure. The static affected area can be identified by spatial clustering algorithms [27]; the dynamic tracking affected area can be identified by spatial logical growth model [32]. The loss of disaster bearing body can be estimated by combining with emotional analysis [33] or extracting specific location and affected degree description of damaged infrastructure [34].

Image data in social media by satellite remote sensing, unmanned aerial vehicle (UAV) or ground shooting by individuals are widely used to extract the disaster information. The image of satellite remote sensing or UAV can be used to study the spatial distribution of population, houses, crops, transportation and other disaster situations. Data processing methods are deep learning algorithms and image processing algorithms including the Digital Terrain Model (DTM), Digital Elevation Model (DEM) and the Digital Surface Model (DSM) model [35]. The image data uploaded by netizens is heterogeneous to social media, which can be used for the disaster degree assessment, such as the storm intensity, flood depth and inundation area. However, there is no existing disaster image dataset, and the quality of social media image is inconsistent. Hence, the image data from ground shooting in social media has not been widely used for quantitative assessment of disasters, but can provide a rough qualitative estimation by visual inspection. Besides, due to the lack of samples, the transfer learning based on small samples [36] also brings a new direction for the application of SMD in disaster research.

4.4. Sentiment Analysis

Public sentiment reflected in disaster-related posts can be used to identify the disaster situations, relief attitudes, and potential risks. Sentiment analysis interprets the meaning or polarity of larger text units (sentences, paragraphs, articles) through the semantic composition of smaller elements. The process takes a language as a digital signal, using some word vector algorithm to represent words, and then use the traditional machine learning algorithms including Bayesian, support vector machine (SVM), random forest or deep learning neural network algorithms including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional Encoder Representation from Transformers (BERT) to train the vector features in text for emotional polarity classification or regression[37].

Besides, emoticons are frequently used in social media, such as "candle" to express solidarity with the disaster victims, and "bawling" to express the uneasiness. Emoticons can reflect public sentiment changes in the process of disaster, but existing studies often filter emoticons directly.

4.5. Rumor Identification

Rumor identification is a technical difficulty of social media monitoring. The common method is to classify SMD simply by trigger words and initial release time. However, social media texts are diversified, matching templates cannot cover all linguistic phenomena and ensure the consistency of rules.

Researchers say that they can rely on computer technology to find rumors through big data statistics [38] or combined with other data sources [39], and estimate the data credibility [40]. Social media itself provides a mechanism to suppress the rumor spreading, and users can question the information from low credibility sources [41]. Zhao et al. used the planned behavior theory and the normative activation theory in the field of psychology to identify the rumors in social media [42]. Wang et al. studied social media users' rumor awareness and response behavior in the disasters by content analysis method in the field of communication and decision tree model [43]. It is very formidable to supervise the social media content and put an end to rumor. At present, it is still a major bottleneck in SMD-based disaster research.

4.6. Communication Node Analysis

To explore the communication mechanism of disaster information in social media, network diagram is a useful tool. We can identify opinion leaders by forwarding and comment data [38] by refining different groups by clustering the whole network or core network, identifying the information propagation path in different disaster management stages by the time evolution of network diagram, and analyzing public behavior by node changes of spatial mobility [44].

4.7. Disaster Information Display

The disaster information forms based on SMD mainly includes the statistical charts and the spatiotemporal maps. The statistical charts focus on the event trend [24], keywords [25], hot topics [13], website statistics, social network map [38], and communication path. Spatiotemporal maps focus on the location of disaster events, the spatial distribution of public opinions [45], disaster-affected area, and rescue points. The method employs the extraction of the spatial attributes from SMD, convert into latitudes and longitudes, and then use GIS spatial interpolation or programming for visual display.

5. Conclusion

Social media data plays a supporting role and has immense potential in disaster analysis and emergency management, as a result of which it is attracting increasing attention. The social media-based disaster research is a data-driven approach, so the consideration of its underlying data type and quality is essential. Considering the strong dependence between data and methods, this review comprehensively summarizes the characteristics and application tasks of social media data in disaster research. Social media data has the advantages of sociality and rapidity for onset hazards but also has the disadvantage of subjective errors and rumor distortion in an application. Current major social media-based research tasks include disaster identification, disaster spatiotemporal analysis, disaster intensity assessment, public sentiment analysis, rumor identification, communication node analysis, and information display. However, there are still many unsolved technical difficulties. We hope that our study stimulates more researchers to actively participate in this field.

Acknowledgments

This work was financially supported by National Key Research and Development Program of China (No. 2018YFC1508903), National Key Research and Development Program of China (No. 2016YFA0602403), Creative Research Groups of National

Natural Science Foundation of China, (No.41621061), Fundamental Research Funds for the Central Universities (No. 2019RC043), China Postdoctoral Science Foundation (No. 2019M660435).

References

[1] Obar J A, Wildman S. Social media definition and the governance challenge: An introduction to the special issue[J]. Telecommunications Policy,2015,39(9).

[2] Aerts J C J H, Botzen W J, Clarke K C, et al. Integrating human behaviour dynamics into flood disaster risk assessment[J]. Nature Climate Change, 2018, 8(3):193-199.

[3] Alexander D E. Social Media in Disaster Risk Reduction and Crisis Management[J]. Science & Engineering Ethics, 2014, 20:717–733.

[4] Helms M M, Nixon J C. Exploring SWOT analysis – where are we now?[J]. Journal of Strategy and Management, 2010, 3(3): 215-251.

[5] Hajizadeh Y. Machine learning in oil and gas; a SWOT analysis approach[J]. Journal of Petroleum Science and Engineering, 2019, 176: 661-663.

[6] Bui Trung Thuc. A Study of Improving the Competition for a Construction Company Using Five-Force Model and SWOT Analysis[J]. Quaternary International, 2014, 348:247-265.

[7] Ervural B C, Zaim S, Demirel O F, et al. An ANP and fuzzy TOPSIS-based SWOT analysis for Turkey's energy planning[J]. Renewable & Sustainable Energy Reviews, 2018: 1538-1550.

[8] We are social, Hootsuite. Digital Report 2019[J]. Recuperado de https://wearesocial. com/global-digital-report-2019, 2019.

[9] Ritzer G, Jurgenson N. Production, Consumption, Prosumption: The nature of capitalism in the age of the digital 'prosumer'[J]. Journal of Consumer Culture, 2010,10(1):13-36.

[10] Rowe M, Angeletou S, Alani H. Predicting discussions on the social semantic web. [C]// Extended Semantic Web Conference on the Semanic Web: Research & Applications. Springer-Verlag, 2011,6644:405-420.

[11] Earle P, Guy M, Buckmaster R, et al. OMG Earthquake! Can Twitter Improve Earthquake Response?[J]. Seismological Research Letters, 2010, 81(2):246-251.

[12] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, 2010, S:851–860.

[13] Abel F, Hauff C, Houben G J, et al. Twitcident: fighting fire with information from social web streams[C]// International Conference on World Wide Web. ACM, 2012.

[14] Robinson B, Power R, Cameron M. An Evidence Based Earthquake Detector using Twitter. Proceedings of the Workshop on Language Processing and Crisis Information, 2013:1-9.

[15] Gruebner O, Lowe S R, Tracy M, et al. Mapping concentrations of posttraumatic stress and depression trajectories following Hurricane Ike[J]. Scientific Reports, 2016,6(1):32242.

[16] Spence A, Poortinga W, Butler C, et al. Perceptions of climate change and willingness to save energy related to flood experience[J]. Nature Climate Change, 2011,1(1):46-49.

[17] Zubiaga A, Hoi G W S, Liakata M, et al. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. [J]. PloS one,2016,11(3).

[18] Deng Y, Wang M.Characteristics of Public Sentiment on Risk in the Era of New Media — a Case Study of the Social Ripple Effect of Foggy Weather[J]. China Soft Science, 2014(08):61-69.

[19] Vosoughi S, Mostafa N M, Roy D. Rumor Gauge: Predicting the Veracity of Rumors on Twitter[J]. ACM Transactions on Knowledge Discovery from Data, 2017, 11(4):1-36.

[20] Lee R, Wakamiya S, Sumiya K. Discovery of unusual regional social activities using geo-tagged microblogs[J]. World Wide Web, 2011, 14(4):321-349.

[21] Robinson, Bella, Power, et al. An Evidence Based Earthquake Detector using Twitter. Proceedings of the Workshop on Language Processing and Crisis Information, 2013:1-9.

[22] Abhik D, Toshniwal D. Sub-event detection during natural hazards using features of social media data[C]// International Conference on World Wide Web, 2013:783-788.

[23] Yin J, Lampert A, Cameron M, et al. Using Social Media to Enhance Emergency Situation Awareness[J]. IEEE Intelligent Systems, 2012, 27(6):52-59.

[24] Kryvasheyeu Y, Chen H, Obradovich N, et al. Rapid assessment of disaster damage using social media activity[J]. Science Advances, 2016, 2(3): 500779.

[25] Khare A, He Q, Batta R. Predicting gasoline shortage during disasters using social media[J]. OR Spectrum, 2019:1-34.

[26] Bai H, Yu G. A Weibo-based approach to disaster informatics: incidents monitor in post-disaster situation via Weibo text negative sentiment analysis[J]. Natural Hazards, 2016, 83(2):1177-1196.

[27] Bakillah M, Li R Y, Liang S H L. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan[J]. International Journal of Geographical Information Science, 2015,29(2):258-279.

[28] Liang C, Lin G, Zhang M, et al. Assessing the Effectiveness of Social Media Data in Mapping the Distribution of Typhoon Disasters[J]. Journal of Geo-Information Science,2018, 20(6): 807-816.

[29] Guan X, Chen C. Using social media data to understand and assess disasters[J]. Natural Hazards, 2014, 74(2):837-850.

[30] Imran M, Mitra P, Castillo C. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages[J]// Language Resources and Evaluation Conference (LREC). 2016:1638-1643.

[31] Olteanu A, Castillo C, Diaz F, et al. Crisislex: A lexicon for collecting and filtering microblogged communications in crises[C]//Eighth International AAAI Conference on Weblogs and Social Media. 2014.

[32] Wang Y, Ruan S, Wang T, et al. Rapid estimation of an earthquake impact area using a spatial logistic growth model based on social media data[J]. International Journal of Digital Earth, 2018:12(11):1265-1284.

[33] Bo T, Li X, Chen S, et al. Research of seismic intensity rapid assessment based on social media data[J]. Earthquake Engineering and Engineering Vibration, 2018,38(5):206-215.

[34] Xu J, Chu J, Nie G, et al. Earthquake disaster information extraction based on location microblog[J]. Journal of Natural Disasters, 2015,24(5):12-18.

[35] Doshi J. Residual Inception Skip Network for Binary Segmentation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018,1: 216-219.

[36] Seo J, Lee S, Kim B, et al. Revisiting Classical Bagging with Modern Transfer Learning for On-the-fly Disaster Damage Detector[J]. arXiv preprint arXiv: 1910.01911, 2019.

[37] de Diego I M, Fernández-Isabel A, Ortega F, et al. A visual framework for dynamic emotional web analysis[J]. Knowledge-Based Systems, 2018, 145: 264-273.

[38] Takahashi T, Igata N. Rumor detection on twitter[C]// Joint International Conference on Soft Computing & Intelligent Systems. IEEE, 2012:452-457.

[39] Lewandowsky S, Ecker U K H, Seifert C M, et al. Misinformation and Its Correction: Continued Influence and Successful Debiasing[J]. Psychological Science in the Public Interest, 2012,13(3):106-131.

[40] Carlos C, Marcelo M, Barbara P. Predicting information credibility in time-sensitive social media[J]. Internet Research, 2013,23(5):560-588.

[41] Mendoza M, Poblete B, Castillo C. Twitter Under Crisis: Can we trust what we RT?[J]. Proceedings of the First Workshop on Social Media Analytics, 2011:71-79.

[42] Zhao L, Yin J, Song Y. An exploration of rumor combating behavior on social media in the context of social crises[J]. Computers in Human Behavior, 2016, 58:25-36.

[43] Wang B, Zhuang J. Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters[J]. Natural Hazards, 2018, 93(3):1145-1162.

[44] Lu X, Brelsford C. Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami[J]. Scientific reports, 2014, 4(1): 6773.

[45] Sakai T, Tamura K. Real-time analysis application for identifying bursty local areas related to emergency topics[J]. SpringerPlus, 2015, 4(1):162.