A Preliminary Meta-analysis of Social Media Use during Disaster

Yi-Min Chang Chien* , Alexis Comber and Steve Carver

School of Geography, University of Leeds, Leeds, LS2 9JT, UK
*Email: gyymcc@leeds.ac.uk

Abstract

The past few years have witnessed the increasing public use of social media at an unprecedented rate as well as the explosive growth of diverse services and platforms. Analysis of social media has been used in the disaster and emergency management domains. To further extend the empirical knowledge and practices about social media use pertaining to disaster, this paper combines text mining and a statistical model applied to abstracts downloaded from Scopus. The aim is to quantify and evaluate changes in the temporal and thematic trends in the scientific analysis of social media in the context of disasters and emergencies. As an initial step, this research considers social media in general and examines different social media platforms. It then examines the links between text mined topics link across time in an attempt to describe the evolutions of ideas in social media analyses associated with disasters and emergencies. These methods, explored here, will be used to undertake a deeper and wider meta-analysis of the literature in this area.

Keywords: Social media, Disaster, Preliminary meta-analysis, Text mining, Latent Dirichlet Allocation.

1. Introduction

The past few years have witnessed the increasing use of social media with millions of users around the world generating volume of data at an unprecedented rate as well as explosive growth of social media services and platforms. (Croitoru et al., 2015) According to their different affordances, they could be categorised into microblogs (e.g., Twitter, Tumblr, and Weibo), social networking services (e.g., Facebook, Google+, and LinkedIn), and multimedia content sharing services (e.g., Flickr and Youtube). This information somehow provides additional content (e.g. location) and context (e.g. topics and sentiment accord) in a sense of linking the cyber and physical spaces. Therefore, it offers opportunities to study human dynamics and activities. That is, analysis of social media can be used to observe socio-cultural expressions in space to understand people's actions, reactions, and interactions in spatiotemporal coverage.

Analysis of social media has also raised concerns in the disaster emergency domain due to the characteristics of timely dissemination to meet the requirement of rapid allocation in disaster rescue and relief. Additionally, this information somehow provides additional content (i.e., space) and context (i.e., topics and sentiment accord) which can be used further to promote current models of disaster response and recovery by the employment of these individual-level perspectives in geosocial analysis. (Croitoru et al., 2014) For example, geosocial analysis of tweets could be used to

localise the impact area of the Virginia earthquake in 2011 (Crooks et al., 2013), of the wildfire of Colorado in 2012 (Panteras et al., 2015), and of the UK floods in 2014 (Saravanou et al., 2015).

Nowadays, a plethora of research seeks to incorporate social media into disaster management to supplement traditional geospatial datasets. There is a need of meriting additional research to understand empirical knowledge and practices about social media use pertaining to disasters and what remains to be investigated. This paper starts to develop a preliminary meta-analysis of social media use in relation to disaster through text mining of publications from 2000 to 2016 unveiling the hidden topics and concepts described in research papers. The aim is to quantify and evaluate changes in the temporal and thematic trends in the scientific analysis of social media in the context of disasters and emergencies.

2. Method

2.1. Data Extraction

In order to conduct a preliminary meta-analysis of disaster-related research on data analysis of social media, A Scopus search was performed through 2000 to 2016 with certain keywords. The limiting search criteria sought article titles or abstracts or keywords that contains the terms “disaster” or “hazard” or “emergency” coupled with the term of “social AND media.” The query resulted in 4057 articles.

2.2. Data Cleaning

To extract or infer the topics contained in the downloaded articles, a *Latent Dirichlet Allocation* (LDA) was used. (Blei et al., 2003) The first step was data cleaning through tokenization, removal of stopwords, and stemming. Tokenization segments a document into its atomic elements. Certain English stopwords, such as conjunctions and pronouns, numbers, punctuation, whitespaces, and any words less than three characters long were removed from the token list. Then, the words were stemmed to reduce topically similar words according to their etymological root. The cleaned and stemmed abstracts were then categorised into 16 documents based on the year of publication in a “bag-of-words.” Another corpuses of each year were created without the process of removing stopwords and stemming for the term frequency analysis.

2.3. Text mining and statistical analysis of abstracts

A statistic provides information on the most popular social networking sites as of February 2017, ranked by number of active accounts listed in

Table *1*. (KALLAS, 2017) To understand which type of social media has been mainly used regarding disaster in the scientific analyses a frequency matrix was constructed describing the occurrence of 15 popular social media platforms in each of the 16 documents representing the corpus of abstracts for each year (2000 to 2016).

rank	social media site	active users (in millions)	rank	social media site	active users (in millions)
1	Facebook	1,860	9	Tumblr	115
2	Youtube	1,000	10	Flickr	112
3	Instagram	600	11	Google+	111
4	Twitter	313	12	LinkedIn	106
5	Reddit	234	13	Vk	90
6	Vine	200	14	ClassMates	57
7	Pinterest	150	15	Meetup	30
8	Ask.fm	160			

Table 1: social media sites ranked by number of active users. (KALLAS, 2017)

Combined with a term's Inverse Document Frequency (IDF), which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents, the Term Frequency-Inverse Document Frequency (TF-IDF) was calculated to measure how important a word is to a document in a corpus. (Comber et al., 2014)

A Latent Dirichlet Allocation analysis was run on the corpus with the R *topicmodel* package. A topic model was derived based on the assumption of 5 topics. (Ponweiser, 2012) Then, the composite relationships between topics and years are calculated and visualized by integrating the relationships of consistency, co-occurrence and linking semantics to topics over years.

3. Initial Result

Figure 1 shows that much of the previous research conducted, particularly on Twitter, followed by Facebook. It is also noteworthy that the use of Instagram which was launched in 2010 have gradually grown in popularity in this domain whilst Flickr, the classic photo sharing program that has been around since 2004 relatively remains stable.

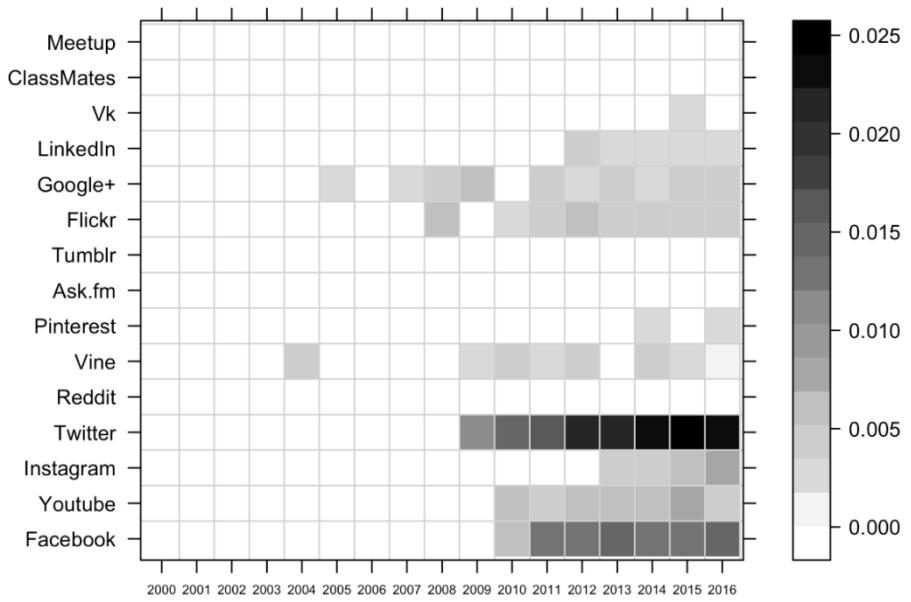

Figure 1: The change of TF-IDF value for each social media platform from 2000 to 2016.

Figure 2 shows that the resulting output of 5 topics containing associated terms from the LDA model which provide insight into the semantic concepts of them. This then suggests 3 distinct topic groups which could be inferred by corresponding terms: Topics 4 (use, disast, inform, emerg), Topics 3 (use, social, studi, inform) and Topics 1, 2, and 5 (health, risk, use, social, risk, disast). Note, that Topic 1 included the term of "health" which is worth investigating whether it implies the application during post-disasters.

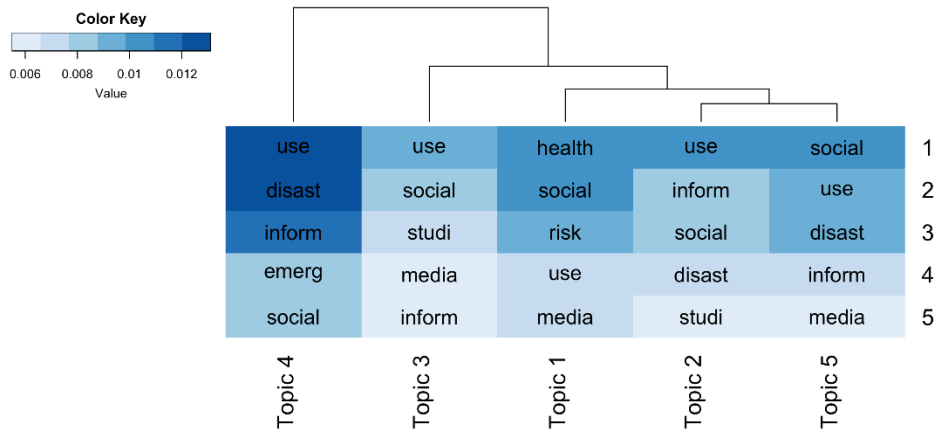

Figure 2: The frequency of occurrence for each social media platform.

Figure 3 shows that the links between topics and years indicating a time evolution of these topics associated with publications in this domain. The width of the edges denotes the strength of the link quantified by the posterior probability in the LDA model. These show only very weak temporal patterns.

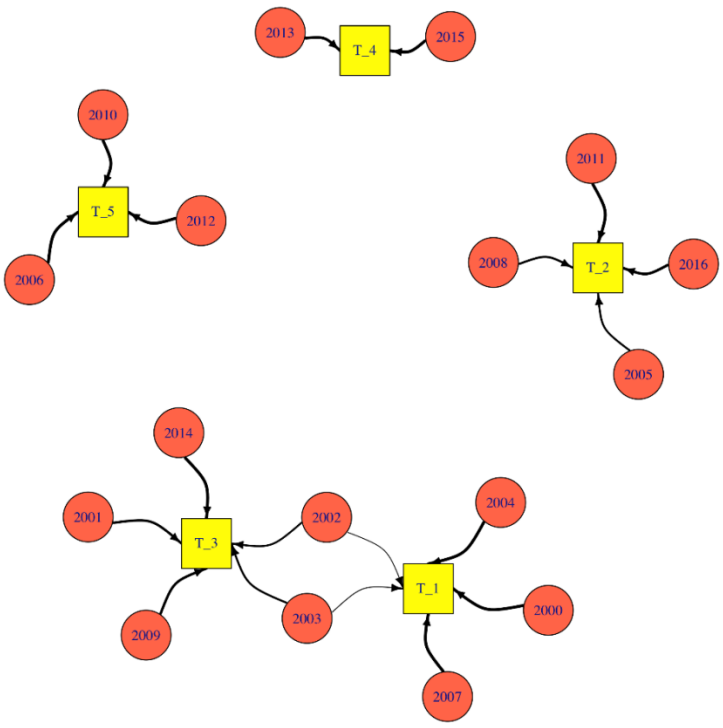

Figure 3: The links between topics and years.

4. Discussion

The findings of this paper indicate only coarse temporal trends in the topics associated with social media as reported in the scientific literature in the analysis of social media in the context of disaster regarding temporal dimension. There are a number of areas of further work which will be undertaken by extending the methods and dataset and outlined below:

- Analysis of the spatial trends in scientific analysis of social media for disaster management: For example, does location of contributors corresponds to disasters, especially occurring in the "digital divide" regions?
- Analysis of the thematic trends in scientific analysis of social media: the well-known problems with social media analyses not the least of which are cognitive and semantic variations within and between groups.
- Evaluations on the specificity of analysis using data from specific platforms: For example, do analyses of Flickr and Foursquare POIs have any specificity?
- The geographic scope of analyses of social media: For instance, how do the localised and global disasters (e.g. El Nino) influence the nature of the social media analyses?

The overall aim of this further work is to develop a series of highly sensitive search strategies and meta-analyses using topic modelling and text mining, linked to gazetteers and formal ontologies in order to obtain deeper insights into the participation of regional studies and data. In parallel, the semantics of a domain ontology should be adopted to link topics for yielding better results.

5. Acknowledgements

The lead author is in his first year of studying for a PhD and the authors would like to thank Taiwan's Ministry of National Defence for funding this research.

6. References

BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research,* 3**,** 993-1022.

COMBER, A., SCHADE, S., SEE, L., MOONEY, P. & FOODY, G. 2014. Semantic analysis of citizen sensing, crowdsourcing and VGI.

CROITORU, A., CROOKS, R., RADZIKOWSKI, J., STEFANIDIS, A., VATSAVAI, R. & WAYANT, N. 2014. Geoinformatics and Social Media. *Big Data.* CRC Press.

CROITORU, A., WAYANT, N., CROOKS, A., RADZIKOWSKI, J. & STEFANIDIS, A. 2015. Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems,* 53**,** 47-64.

CROOKS, A., CROITORU, A., STEFANIDIS, A. & RADZIKOWSKI, J. 2013. #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS,* 17**,** 124-147.

KALLAS, P. 2017. *Top 15 Most Popular Social Networking Sites* [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/.

PANTERAS, G., WISE, S., LU, X., CROITORU, A., CROOKS, A. & STEFANIDIS, A. 2015. Triangulating Social Multimedia Content for Event Localization using Flickr and Twitter. *Transactions in GIS,* 19**,** 694-715.

PONWEISER, M. 2012. Latent Dirichlet allocation in R.

SARAVANOU, A., VALKANAS, G., GUNOPULOS, D. & ANDRIENKO, G. Twitter floods when it rains: A case study of the uk floods in early 2014. 24th International Conference on World Wide Web, WWW 2015, 2015. Association for Computing Machinery, Inc, 1233-1238.