Situational awareness in social media: lessons learned using information entropy in flood risk management

Sidgley C de Andrade[1], **Camilo Restrepo-Estrada**[2], **Thiago A. G. da Costa**[3]
Jó Ueyama[3], **Alexandre C. B. Delbem**[3], **João Porto de Albuquerque**[4]

[1]*Federal University of Technology - Paraná, Toledo, Brazil*
[2]*Faculty of Economic Sciences, University of Antioquia, Medellín, Colombia*
[3]*Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil*
[4]*Centre for Interdisciplinary Methodologies of the Warwick University, Coventry, UK*

sidgleyandrade@utfpr.edu.br,camilo.restrepo@udea.edu.co,
thiago.gcosta@usp.br,[joueyama,acbd]@icmc.usp.br,J.Porto@warwick.ac.uk

Abstract: Increasing situational awareness using social media data is still a problem for the surveillance of disaster-related events due to the amount of data. In order to address this problem, a number of studies have been conducted on the basis of the Tobler's first law of geography, in which social media messages nearest to events are more relevant than the more distant messages. However, these studies fail to take the explicit content of the messages in terms of quantitative measures into account. A quantitative measure is important to prioritize and rank social media messages using another criterion beyond the geographical distance. This paper conducts a case study in the city of São Paulo, Brazil, for assessing the relationship between the information entropy and the distance to flooded areas of rain- flood-related Twitter messages. The results provide evidence that the entropy measure of the tweets is not governed by the Tobler's law of geography. Nonetheless, our findings do not challenge the Tobler's law assumption, but put forward discussions in terms of the relevance of the social media's content in relation to distance to the affected areas by disasters.

1. Introduction

Significant information about an ongoing extreme event and a disaster can assist in the task of surveillance and thus improve the situational awareness of people involved in disaster management. In this direction, disaster-related social media data has played an important role in gaining situational awareness since it is a valuable source of (near) real-time information provided by eyewitnesses [1, 2].

Many studies have mined useful information from social media data to gain situational awareness in different disaster context. For example, de Albuquerque et al. (2015) carried out a geographic analysis and hand-classification of flood-related tweets. Spinsanti and Ostermann (2013) assessed the importance of forest fires-related tweets. Sakaki, Okazaki and Matsuo (2010) explore earthquake-related tweets using classifier models. In spite the different approaches of these studies, they are driven by the Tobler's first law of geography – "everything is related to everything else, but near things are more related than distant things". That means that they investigated the geographic distance of event-related social media messages to the affected area by the event.

Although recent efforts focus to uncover social media messages that contribute to gain situational awareness, no quantitative measure to extract the importance of the explicit content has been addressed. This is important for prioritizing and ranking the messages using another criteria beyond the geographical distance. For instance, in real situations a decision-maker might choose the messages by the 'informational power' contained in them and not only by the geographical distance criterion. In this direction, the information theory provides the entropy measure that is the amount of information in a process or random variable [6]. Hence, the entropy measure of the social media messages might, theoretically, estimate the 'informational power'. However, to use it in real situations first it is essential to assess the relationship between entropy and distance measures, which raises the following question: "Is there a relationship between entropy and distance measures of disaster-related social media messages?"

This paper assess the relationship between two measures extracted from the rain- flood-related Twitter messages: information entropy and distance to flooded areas. An attempt is made using the Pearson correlation as relationship coefficient of both measures. The remaining of this work is structured as follows: Section 2 presents the case study. Section 3 is devoted to the methodology. Section 4 shows the results obtained. Finally, Section 5 draws some conclusions and makes recommendations for future works.

2. Case study

2.1. *Study area*

The city of São Paulo is the study area because flash floods often affect the city's infrastructure and citizens, leading to fatalities and economic losses of millions of real. In addition, this city has an estimated population of approximately 12 million people [7] and a higher number of Twitter users than other Brazilian cities.

2.2. *Official flood points and reference events*

According to the Center of Emergency Management (CGE)[1], more than 2,000 flood points were reported from February 2015 to January 2017. These flood points are distributed across the city and occur at different time periods with different time durations. In fact, there may be multiple occurrences of flooding during one day and it poses a challenge to spatiotemporal analysis. In order to simplify the case study, we aggregate the flood points using a daily time scale. That means that the reference events are days with occurrence of flood points. After that, we selected a random sample of two days for post-hoc analysis (3 Dezember 2016 and 19 January 2017).

2.3. *Social media data*

We developed a crawler-twitter tool to retrieve public tweets through Twitter Stream API. Moreover, we defined two bounding-box filters covering the city of São Paulo, one north (-46.95,-23.62,-46.28,-23.33) and one south (-46.95,-23.91,-46.28,-23.62). A total of 11,848,923 million tweets were retrieved from 7 November 2016 to 28 February 2017 (UTC time), where only 891,367 were geotagged (7,52%). After retrieving the tweets, we filtered the geotagged ones based on a set of keywords (Table 1) – using a substring-searching approach. The filtered tweets hereafter are referred as "on-topic tweets".

Table 1. Keywords in Brazilian-Portuguese with their English meaning in parentheses. The keywords were chosen based on preliminary analysis of the tweets. Keywords with grammar mistakes were take into account as long as the frequency was equal or greater than 10 (e.g. "chuvendo").

alagamento (flood), alagado (flooded), alagada (flooded), alagando (it's flooding), alagou (flooded), alagar (to flood), chove (rain), chova (rain), chovia (had been rained), chuva (rain), chuvarada (rain), chuvosa (rain), chuvoso (rainy), chuvona (heavy rain), chuvinha (drizzle), chuvisco (drizzle), chuvendo (it's raining), dilúvio (heavy rain), garoa (drizzle), inundação (flood), inundada (flooded), inundado (flooded), inundar (to flood), inundam (flood), inundou (flooded), temporal (storm), temporais (storms)

3. Methodology

The methodology comprises three steps: (i) to determine the flooded areas, (ii) to calculate the entropy of the on-topic tweets and the distances of them to the flooded areas, and (iii) to perform the correlation of both measures.

3.1. *Determining the flooded areas*

The Density-based spatial clustering of applications with noise (DBSCAN)[2] algorithm was used to determine the flooded areas on 3 Dezember 2016 and 19 January 2017. Thus, the terms 'flooded area' and 'cluster' are used interchangeably in this study. To derive the flooded areas from the flood points it was necessary to configure at least two DBSCAN parameters: size and number of points of the epsilon neighborhood. The size of the epsilon was defined being the ratio between the maximum autocorrelation distance of the flood points and the earth's equatorial radius, whilst the number of points was fixed at 1. The maximum autocorrelation was estimated at 216.43 meters using the semi-variogram function, which describes the spatial dependence of the frequency of the flood points as function of the distance between ones. For each month we extracted the 'range' parameter of the semivariogram funcion and later we calculated the average. Figure 1 illustrate the flooded areas obtained by the DBSCAN algorithm.

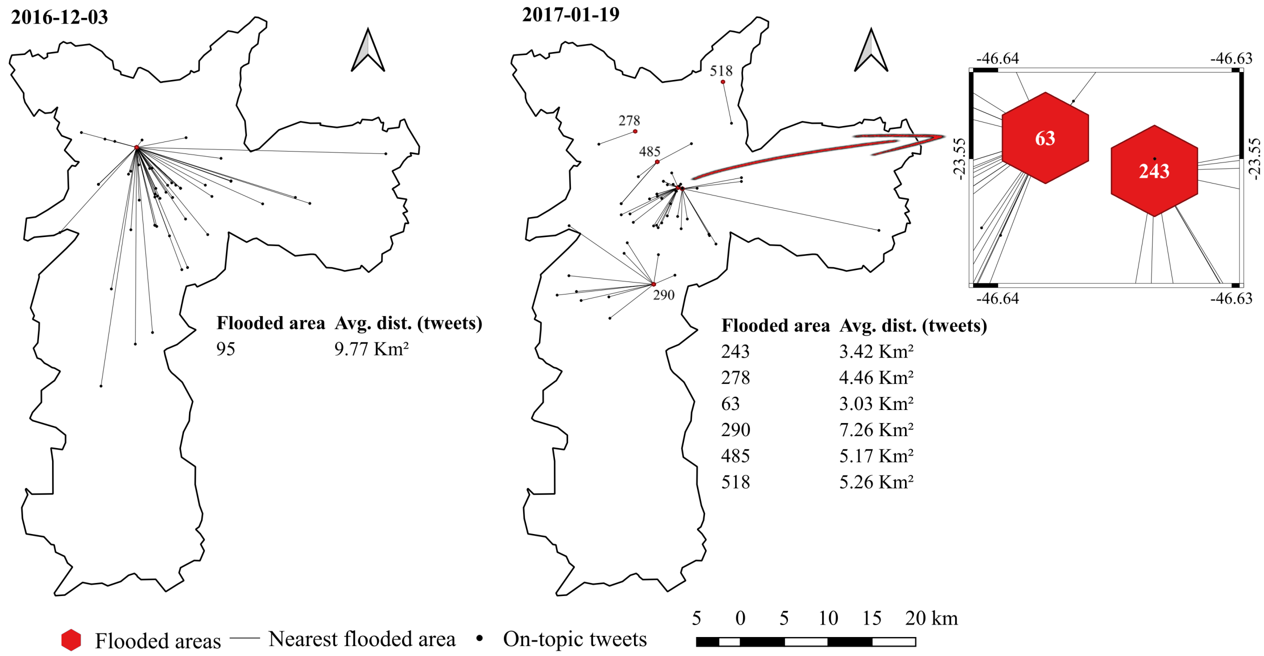
3.2. *Calculating the entropy and geographical distance*

Once the flooded areas were determined, the haversine distance[3] was calculated between the centroid of flooded areas and the coordinates of the on-topic tweets. Due to the existence of two or more flooded areas in a day, the distance of the on-topic tweets was considered only to the nearest flooded area (see Equation 1).

[1] http://cgesp.org/

[2] https://en.wikipedia.org/wiki/DBSCAN

[3] https://en.wikipedia.org/wiki/Haversine_formula



Flooded area	Avg. dist. (tweets)
95	9.77 Km²

Flooded area	Avg. dist. (tweets)
243	3.42 Km²
278	4.46 Km²
63	3.03 Km²
290	7.26 Km²
485	5.17 Km²
518	5.26 Km²

Fig. 1. Reference events and the distribution of the on-topic tweets in relation to flooded areas. The red hexagons correspond to flood points clusterized using the DBSCAN algorithm (i.e., the flooded areas with radius of 216.43 m). The black dots and lines illustrate the on-topic tweets and their distance to the flooded areas. The tables show the average distance of the on-topic tweets to the flooded areas.

$$d(x_i^j, e_k^j) = min\left(2 \cdot r \cdot arcsin\sqrt{\sin^2\left(\frac{e_{k,lat}^j - x_{i,lat}^j}{2}\right) + \cos(x_{i,lat}^j)\cdot\cos(e_{k,lat}^j)\cdot\sin^2\left(\frac{e_{k,lon}^j - x_{i,lon}^j}{2}\right)}\right) \qquad (1)$$

where x_i^j corresponds to the on-topic tweet i on day j, e_k^j the centroid of the flooded area k on day j, and the $d(\cdot)$ is the shortest haversine distance between x_i^j and e_k^j on day j. The lat and lon indexes correspond to latitude and longitude coordinates, respectively.

After that, the 'informational power' of the on-topic tweets was calculated using the information entropy (Equation 2). We only consider the keywords in Table 1 to calculate the words frequencies of the on-topic tweets.

$$H(x_i) = \sum_{\forall w_{ij} \in x_i} f_j \, log_2\left(\frac{1}{f_j}\right) \qquad (2)$$

where f_j is the relative frequency of the keywords w_j contained within the on-topic tweets x.

Example. Let $x_1 = \{$"chova chuva..."$\}$ and $x_2 = \{$"chuva caindo do céu!"$\}$ be two on-topic tweets. The relative frequency of the keywords "chove" and "chuva" is $f_{chova} = \frac{1}{3}$ and $f_{chuva} = \frac{2}{3}$, respectivelly. As a result, the entropy of the on-topic tweets is $x_1 = \frac{1}{3} \cdot \log_2\left(\frac{1}{\frac{1}{3}}\right) + \frac{2}{3} \cdot \log_2\left(\frac{1}{\frac{2}{3}}\right)$ and $x_2 = \frac{2}{3} \cdot \log_2\left(\frac{1}{\frac{2}{3}}\right)$.

3.3. Analysing the correlation

Finally, Pearson's correlation coefficient was used to assess the relationship between the entropy and the distance to the flooded areas of the on-topic tweets. We chose to use a simple linear correlation method due to the limited number of on-topics tweets per flooded area. Nonetheless, for a more detailed investigation, it should be considered other relationships such as robust and non-linear correlations.

4. Results

Figure 2 shows a weak correlation between the information entropy and the distance to the flooded areas of the on-topic tweets. Moreover, the Pearson's r found are not statistically significant (p-values ≥ 0.05), which means that the hypothesis "there is a relationship between the entropy and the distance measures of the on-topic tweets" is rejected when the linear relationship is considered. Other evidence suggest that the magnitude of the correlation is inversely proportional to the number of on-topic tweets. For example, the slope line of the flooded areas 290 and 485 are more inclined than the flooded areas 95, 243 and 63. Although Figure 2 only considers tweets up to 8 km from the flooded areas, a similar kind of behavior can be observed at greater distances.

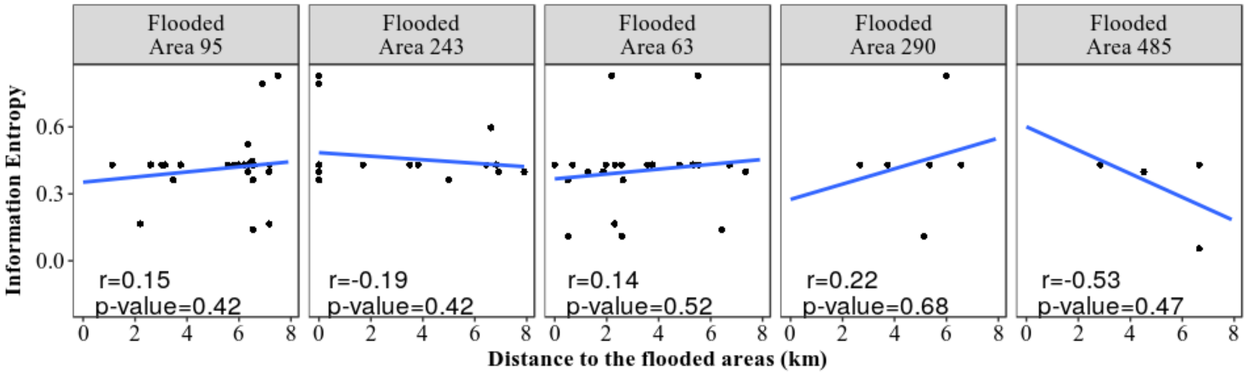


Fig. 2. Pearson's correlation coefficient between information entropy and distance to the flooded areas of the on-topic tweets. The black dots correspond to on-topic tweets.

5. Discussions and Conclusions

This work carried out a case study to assess the relationship between the information entropy of rain- flood-related tweets and the distance of ones to the flooded areas. The results are still incipient, however, provide evidences that the content of the tweets are not governed by the Tobler's law of geography in scenarios with high resolution and multiple occurrences of events. Although we have not intention of challenge the Tobler's law assumption, further in-depth investigations are necessary and future works should incorporate more scenarios and alternative methods to assess the relevance of social media content. Furthermore, other type of contents such as images and videos should be taken into account.

6. Acknowledgments

SCA would like to thank the FAPESP (grant no. #2017/15413-0) and FAPPR/SETI (grant no. Edital 18/2015) agencies for its financial support. TAGC is grateful for the financial support from the CAPES.

7. References

[1] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: a survey", ACM Comput. Surv. 47 (4) (2015) 1–38.

[2] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. 2010. "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In Proc. of the SIGCHI Conference on Human Factors in Comput. Systems (CHI'10). ACM, New York, NY, USA, 1079-1088.

[3] Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015) A Geographic Approach for Combining Social Media and Authoritative Data towards Identifying Useful Information for Disaster Management. International Journal of Geographical Information Science, 29, 667-689. doi: 10.1080/13658816.2014.996567.

[4] L. Spinsanti, F. Ostermann, "Automated geographic context analysis for volunteered information", Applied Geography, (43), 2013, 36–44.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. "Earthquake shakes Twitter users: real-time event detection by social sensors". In Proc. of the 19th Int. Conf. on World Wide Web ('10). ACM, New York, NY, USA, 851–860.

[6] Gray, R. M. (1990). "Entropy and Information Theory". https://doi.org/10.1007/978-1-4757-3982-4

[7] IBGE "Censo Demográfico 2010". Brazilian Institute of Geography and Statistics, Rio de Janeiro, 2010.