Combining Automatic and Manual Approaches: Towards a Framework for Discovering Themes in Disaster-related Tweets

Leif Romeritch Syliongka[1], Nathaniel Oco[2,1], Alron Jan Lam[1], Cheryll Ruth Soriano[1], Ma. Divina Gracia Roldan[1], Francisco Magno[1], and Charibeth Cheng[1]

[1]De La Salle University
[2]National University

{leif.syliongka,nathan.oco,chari.cheng}@delasalle.ph
{alron_lam,cheryll.soriano,ma.divina.roldan,francisco.magno}@dlsu.edu.ph

ABSTRACT

In this paper, we present a framework that combines automatic and manual approaches to discover themes in disaster-related tweets. As case study, we decided to focus on tweets related to typhoon Haiyan, which caused billions of dollars in damages. We collected tweets from November 2013 to March 2014 and used the local typhoon name "Yolanda" as the filter. Data association was used to expand the tweet set and k-means clustering was then applied. Clusters with high number of instances were subjected to open coding for labeling. The Silhouette indices ranged from 0.27 to 0.50. Analyses reveal that the use of automated Natural Language Processing (NLP) approach has the potential to deal with huge volumes of tweets by clustering frequently occurring words and phrases. This complements the manual approach to surface themes from a more manageable set of tweet pool, allowing for a more nuanced analysis of tweets from a human expert. As application, the themes identified during open coding were used as labels to train a classifier system. Future work could explore on using topic models and focusing on specific content or issues, such as natural calamities and citizen's participation in addressing these.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering, selection process*

H.4 [**Social and Behavioural Sciences**]: *sociology*

Keywords

Discovering themes, clustering, open coding, typhoon Haiyan, tweet analysis.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742125

1. INTRODUCTION

Twitter is one of the most popular networking sites with 600 million active users[1] (n.b., both registered and visiting). Used as a written virtual account of significant events as well as a powerful tool for communication, it is considered as a veritable gold mine of data; most especially in times of disasters. Recent studies [3], [23] have shown that social media activity never stops even during a disaster – affected communities post contents to social medias, which allow individuals to watch videos, see photos and read status updates from ground zero. These types of posts – along with people who are observing the event and are also posting in these mediums – make social media interesting. In this regard, Twitter is useful to organizations concerned with relief efforts for valuable information and for purposes of information dissemination [19]. There is an increasing interest to make sense out of these data, which are often voluminous, unorganized and unstructured. Also, manual analyses are time-consuming and overwhelming. Given these constraints, we sought to develop a framework that combines manual and automatic approaches to extract themes from disaster-related twitter data. In this paper, we therefore seek to address this problem: *how do we apply automatic approaches in doing manual analysis of Twitter Data*?

2. RELATED WORK

Twitter is a type of social networking site that allows individuals to send and read short messages known as tweets. The maximum length of a tweet is 140 characters, making Twitter an ideal medium for self-expressions and announcements [13]. In contradistinction to being merely consumers of media, Twitter also serves as a participatory medium for ordinary users to engage in the creation of content during important news events [7]. It is this reason that a number of studies have focused on analyzing Twitter use and content. In the succeeding texts, we present studies that utilized manual approaches, automatic approaches, and a combination of both approaches.

2.1 Manual Approaches

Manual analysis is often qualitative in nature; emphasizing on in-depth information and knowledge. Previous studies include the

[1] http://twopcharts.com/twitter500million.php

use of open coding on 4,915 tweets to determine the type of information shared by police departments on Twitter [12]; the application of content analysis on a sample of Hurricane Sandy tweets to study the phases of emergency management [32]; studying the potential of Twitter for building a civil society network in the context of disaster relief during typhoon Ketsana [24]; or managing public relations with social media users in the context of organizational relief efforts in Haiti [33]. Given the focus on exploring hidden meanings and discourses behind texts, manual interpretive analysis of Twitter messages may be time consuming and can be applied to a limited amount of data [31]. It is against this scenario that scientists resort to automatic approaches to systematically handle large data and to come up with a limited data pool that can be used for in-depth analysis using qualitative methods.

2.2 Automatic Approaches

Automatic approaches utilize mathematical representations to analyze data; expressing results in terms of numbers and their equivalent meaning. An example is sentiment analysis. Being one of the "hottest research areas in computer science" [6], a number of papers have been written on the topic. Examples include the analysis of election-related tweets using a sentiment lexicon to predict election outcome [22], [35]; and agenda setting via Twitter during Presidential elections [36], or in framing political issues [25]. Other studies that utilized automatic approaches include the classification of disaster-related tweets into resource coordination, urgent rescue needed, urgent rescue resolution, damage reporting, missing people, and media storm coverage using Naive Bayes and Support Vector Machines or SVM [19]; the use of Walktrap algorithm to detect community structures [9]; detection of flood-related tweets using SVM [2]; and the use of Naïve Bayes and SVM classification combined with clustering for studying emergency situations [38]. One study [14] also surveyed a number of automatic approaches and found key weaknesses; while certain approaches can handle voluminous data, such as the case in sentiment analysis studies, they are dependent on the training data in which they were trained.

2.3 Combined Approaches

A number of earlier works have combined manual and automatic approaches to address each approach's weakness. One instance is the application of software-aided semantic content analysis [20] to study dimensions of information quality. In this related study, CATPAC – a software that utilizes neural networks – was used to identify frequently occurring phrases and keywords. Another study [16] combined manual coding and automated sentiment analysis on 150,000 microblog postings to study word of mouth branding while others utilized tools as aid in doing manual analysis and presenting results [28], [37]. In these research works, automatic approaches have been utilized as guides for manual analysis. There is also an existing work that utilized crowd sourcing to annotate data [15].

3. CASE STUDY

To address the research problem, we developed a framework that combines manual and automatic approaches to discover themes. It involves (1) filtering the data using certain keywords; (2) clustering the filtered tweets; and (3) labeling the clusters using open coding with language models as guide.

Table 1. Number of tweets

Month	Number of Tweets	Yolanda Tweets
NOV 2013	2,250,703	1357
DEC 2013	2,686,864	1932
JAN 2014	4,726,773	419
FEB 2014	3,663,172	354
MAR 2014	2,349,156	200
Total	15,676,668	4,262

Table 2. Confidence values

Keyword	NOV	DEC	JAN	FEB	MAR
victims	1.91	4.00	2.96	3.22	2.34
typhoon	1.67	2.84	1.29	1.76	0.95

As a case study, we focused on typhoon Haiyan, a category 5 typhoon with a maximum of 145 mph 10-minute sustained winds. It caused billions of dollars in damages and more than 6,000 fatalities[2] in the Philippines alone. Known locally as Yolanda, it made landfall in the Philippines on November 07, 2013. Millions of tweets were sent before, during, and after the typhoon. The succeeding texts detail the filtering process.

3.1 Collection

Twitter data, from November 2013 to March 2014, that have geolocations (latitude, longitude) originating from Metro Manila were collected using a program based on Twitter4J[3], a Java library for the Twitter API. From these tweets, we selected those containing the term "Yolanda". The number of tweets is shown in Table 1. Considering several holidays, tweets for November and December (see second column) are surprisingly low.

3.2 Expanding the Tweet Set

We decided to expand the tweet set by mining related keywords using data association, a process of learning relations between variables in a dataset [1]. A relation or an association rule, defined in (1), works on the notion of an antecedent *X* resulting in a consequent *Y* with particular support and confidence.

$$\{X\} \rightarrow \{Y\} \quad (1)$$

For this study, we applied the data association technique used in a related work [34]. The data set is first modeled in terms of 6-grams and association rule mining is applied on the 6-grams: where *X* is "yolanda" and *Y* is another keyword. Among the different keywords, two were selected because of their high confidence values, shown in Table 2. Confidence, shown in (2), is defined as the support for occurrences of *X* and *Y* over the support for occurrence *X*. A higher confidence value means *Y* is usually seen with *X*.

[2] Data from the National Disaster Risk Reduction and Management Council: http://www.ndrrmc.gov.ph/

[3] http://twitter4j.org/en/index.html

$$confidence = \frac{\sup(X \cup Y)}{\sup(X)} \quad (2)$$

Table 3. Number of tweets for the expanded set

Month	Number of Tweets
NOV 2013	3,488
DEC 2013	2,782
JAN 2014	593
FEB 2014	483
MAR 2014	261
Total	7,607

Table 4. Number of tweets per cluster

Month	Cluster	Number of tweets
NOV 2013	C1	3282
	C2	160
	C3	46
DEC 2013	C1	304
	C2	2471
	C3	7
JAN 2014	C1	19
	C2	86
	C3	488
FEB 2014	C1	55
	C2	393
	C3	35
MAR 2014	C1	212
	C2	12
	C3	37

Only these were selected as expanding the tweet set with too many keywords may introduce too much noise (i.e., tweets not related to typhoon Haiyan). Using "Yolanda", the two related keywords, and the international name "Haiyan" as filters, the tweet set was expanded. Table 3 shows the number of tweets per month. The expanded tweet set contains a total of 7,607 tweets. It can be noted that the numbers start to decline after two months. This reflects the intensity of tweets at the time when the disaster struck up to a month after the disaster when information dissemination and coordination was at its peak in providing relief services to victims. It may also indicate a decline in attention on Typhoon Yolanda among twitter users starting the first quarter of 2014. Further analysis of the content of tweets is necessary to determine the nature of tweets as time progresses.

After filtering, the tweets are pre-processed. This entails the (1) removal of insignificant white spaces; (2) proper tokenization of tweets; (3) conversion of all tweets to lower case; (4) transforming each tweet into values; and (5) preparation of tweets for clustering.

3.3 Pre-Processing

To remove insignificant white spaces and to tokenize the tweets, regular expressions and the following Moses scripts [18] were used:

- tokenizer.perl – inserts spaces between words and punctuations
- clear-corpus-n.perl – removes white spaces

All tweets were then converted to lower case and the tweets were transformed into values using Term Frequency-Inverse Document Frequency or TF-IDF [30], a weighting scheme that takes into account the spread of a term throughout the entire set of documents. Terms are then pruned and a vector space model [21] is produced. In this process, the following are removed in the order specified:

- Function words;
- Terms with low document frequency values (DF);
- Terms with high IDF values;
- URLs (terms beginning in "http"); and
- Other special characters (e.g., terms with less than two characters and are non-alphanumeric).

These were deemed as adding noise, irrelevant or too frequently occurring, and do not provide any discriminating information.

3.4 k-means Clustering

Once the data has been transformed into a vector space, *k*-means clustering [10] – a process of partitioning a tweet set into *k* clusters – is applied on each month. The tweet set was separated per month for purposes of time series analysis. To minimize the bias of the model to the seed documents, simulated annealing [5] was used. The presence of the bias could deter the results of the clustering due to local maxima. For this research, ten simulations were performed per tweet set per value of *k*. R[4], a computer environment, was used to automate these processes. The number of tweets per cluster per month is shown in Table 4. As *k*-means clustering groups related tweets together, only those clusters with the highest number of tweets per month were selected for manual analysis as these contain the majority of the discussion. The selected clusters were underlined in the table.

3.5 Labeling using Open Coding

To label the selected clusters, open coding was used; language models and a list of action words were utilized as a guide. The following texts show how we came up with the labels.

3.5.1 Language Modeling

Language modeling is the process of creating smaller representations of a document and is composed of n-grams and the number of times they appeared in the document. The standard formulation [27] for an n-gram is as follows: given a sentence *S* composed of different words, $S=\{s_1...s_o\}$, word sequences $T=\{t_1...t_n\}$ is an n-gram of *S* if there exists a strictly incrementing

[4] http://www.r-project.org/

sequence $p_1...p_n$ of indices of S such that for all $q=1...n$, $S_{p_q} = T_q$. 1-grams up to 6-grams of each selected cluster were generated using SRILM[5]. In addition, English and Tagalog action words were identified in the unigrams using a tagger dictionary [26].

3.5.2 Open Coding

Open coding involves breaking down the data into distinct ideas, events or objects and assigning codes [8]. Once the codes of the data have been formulated, they are again analyzed and grouped together forming categories. The categories are used in formulating labels or themes. With language modeling and part-of-speech tagging as guides, the n-grams and list of action words already act as codes, simplifying the process. Also, the rank and frequency counts of each n-gram provide a certain level of importance; giving emphasis on unique keywords that emerge frequently.

4. DISCUSSION

Among the different phases of emergency management [32], analyses of the different clusters reveal only recovery (see Table 5 for the different labels). In the Philippines, recovery pertains to the short term and long term actions to rehabilitate the affected community and reconstruct infrastructure to restore services. Sample n-grams are shown in Table 6. The different n-grams for NOV 2013 C1 pertain to identifying people affected by the typhoon. These mostly indicated concern over the disaster victims and the need for help to be extended to them. Mobilizing assistance for the victims was carried over to Cluster 2 DEC 2013 C2. In said cluster, however, concrete assistance emerge in the form of fund-raising and holding events such as concerts and benefit shows for the typhoon victims. With the advent of a new year, the damage brought by the typhoon had been fully assessed. This shows the focus of the n-grams on death tolls in JAN 2014 C3. It took some time as damage was widespread. To show gratitude to the international response provided, a thank you campaign was launched in February; explaining the predominance of expressions of gratitude ("thank you") in FEB 2014 C2 n-grams. Finally, in March 2014, we noted an attention shift away from discussion about the victims or about the state of disaster affected communities. A majority of the tweets focused on the "celebrification" of disaster assistance, as the posts highlighted not the act of humanitarian assistance, but on the celebrities involved and the spectacle that the fund-raising activities raised. "Celebrification" is a term coined to refer to the active use of celebrity factor on behavior, and on the perception by an individual of a person or an idea, as an effect of the 'parasocial' interaction with the personalities of popular entertainment media [4], [11]. This was apparent in the attention given to the amount of donations, as well as the focus on popular personalities providing assistance to the affected communities in the month of March 2014.

[5] Stanford Research Institute Language Modeling toolkit: http://www.speech.sri.com/projects/srilm/

Table 5. Labels per cluster

Month and Cluster	Label
NOV 2013 C1	Victim Identification; Mobilizing Assistance for Disaster Victims
DEC 2013 C2	Raising Funds for Typhoon Victims
JAN 2014 C3	Accounting the Damage
FEB 2014 C2	Expressing Appreciation
MAR 2014 C1	Celebrification and Disaster Assistance

Table 6. Sample n-grams and their rank and frequency count

Month	n-gram	Rank	Count
NOV 2013 C1	victims	12	1,065
	help	27	325
	victims of	3	445
	victims of typhoon	5	175
	help the victims of	21	33
DEC 2013 C2	charity	33	217
	benefit	50	159
	concert	54	147
	for the victims	4	130
	# 1heart1beat1voiceconcert - a benefit	7	87
JAN 2014 C3	death	48	26
	toll	49	26
	yolanda victims	3	68
	now at 6,190	3	21
	death toll in phl	6	12
FEB 2014 C2	thank	30	35
	thank you	9	34
	# phthankyou	15	27
	<s> thank you to	9	3
	a big thank you to everyone	5	4
MAR 2014 C1	survivors	27	18
	after	37	11
	bruno mars	4	6
	bruno mars donates	20	4
	$ 100k for children	28	3

5. EVALUATION AND LIMITATION

To evaluate the different clusters, we used a measure for cohesion (i.e., which tweets lie within a cluster and which are marginal) – the Silhouette index or SIL [17]. Defined in (3), where $a(i)$ is the average similarity of tweet i with other tweets in the cluster, and $b(i)$ is the minimum average distance of tweet i in other clusters, SIL is a value between -1 and 1 with 1 being the best value. The number of tweets and Silhouette indices per cluster per month is shown in Table 7. The Silhouette indices ranged from 0.27 to

0.50. The results indicate that there are varieties in discussions per cluster. It is important to note that although automatic processes allow the processing of voluminous data, they do not necessarily constitute complete data [29]. We used filtering processes that facilitate a dataset representing only a portion of the sample universe (i.e., Metro Manila; using the keywords *Yolanda*, *Haiyan*, *victims*, and *typhoon*; disregarding tweets that do not identify location). The tweet pool also does not recognize tweets repeatedly sent by robots. Finally, the clusters are identified on the basis of frequency. In the context of disasters, however, there are certain words that naturally emerge as frequent (i.e., typhoon, disaster, victim). Thus, higher values for *n* (e.g., 7-gram and 8-gram) could also be meaningful for analysis because they provide more contexts. There is also a need to prune out function words to facilitate the emergence of more meaningful data for further manual analysis. Nonetheless, automated methods for data analysis are becoming increasingly valuable due to the possibility of generating insights from voluminous datasets that would appear to be onerous using other methods. Thus, the potential of automated methods has been demonstrated in terms of managing datasets and generating first-level thematic analyses from data obtained from a broad time frame.

$$SIL(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases} \quad (3)$$

Table 7. Evaluation Results

Month	Cluster	Number of tweets	SIL
NOV 2013	C1	3282	0.35
	C2	160	0.37
	C3	46	0.36
DEC 2013	C1	304	0.02
	C2	2471	0.27
	C3	7	1.00
JAN 2014	C1	19	0.22
	C2	86	-0.04
	C3	488	0.48
FEB 2014	C1	55	0.06
	C2	393	0.46
	C3	35	0.35
MAR 2014	C1	212	0.50
	C2	12	-0.04
	C3	37	-0.11

6. APPLICATION

Using the themes found in the typhoon-related tweets, a web application[6] that uses Support Vector Machine (SVM) was developed. Such a web application may be useful if there is interest in being able to quickly look for tweets that fall under certain themes. For example, some agencies may be interested in finding tweets about victims and the assistance they need, in which the tweets under the theme "Victim Identification; Mobilizing Assistance for Disaster Victims" will be relevant. The web application then looks for tweets in real-time and categorizes them. As such, the concerned agencies only need to monitor the application for updates. Of course, automatic classification will not be perfect, and tweets may be classified erroneously. Despite this, automatic classifiers still provide a faster and easier way of looking for tweets as opposed to manual methods. One possible extension is a Philippine map with victims' needs plotted out on that map.

7. ACKNOWLEDGMENTS

This project was supported in part by the University Research Coordination Office of De La Salle University [06 IR U 1TAY14-3TAY14].

8. REFERENCES

[1] Agrawal, R., T. Imielinski, and A. Swami. 1993. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.

[2] Beduya, L. J. and K. J. Espinosa. 2014. Flood-Related Disaster Tweet Classification Using Support Vector Machines. Proceedings of the 10th National Natural Language Processing Research Symposium.

[3] Caragea, C., N. McNeese, A. Jaiswal, G. Traylor, H.-W. Kim, P. Mitra, D. Wu, A. Tapia, L. Giles, B. Jansen, and J. Yen. 2011. Classifying Text Messages for Haiti Earthquake. 8th International Conference on Information Systems for Crisis Response and Management.

[4] Centeno, D. 2010. Celebrification in Philippine Politics: Exploring the Relationship Between Celebrity Endorsers' Parasociability and the Public Voting Behavior. Social Science Diliman,6(1): 66-85.

[5] De Vicente, J., J. Lanchares, and R. Hermida. 2003. Placement by Thermodynamic Simulated Annealing. Physics Letters A, 317(5–6): 415-423.

[6] Feldman, R. 2013. Techniques and Applications for Sentiment Analysis. Communications of the ACM, 56(4): 82-89.

[7] Freeman, M. 2011. Fire, Wind and Water: Social Networks in Natural Disasters. Journal of Cases on Information Technology, 13(2): 69–79.

[8] Glaser, B. 1978. The Grounded Theory Perspective II: Description's Remodeling of Grounded Theory Methodology. The Sociology Press, CA.

[6] Testing the classifier with 10-fold cross validation yielded 92.74% accuracy, a kappa statistic of 0.81, and an average F-Measure of 0.923.

[9] Gonzales, H. A. and K. J. Espinosa. 2014. Community Structure Detection and Analysis in Disaster Related Tweets. Proceedings of the 10th National Natural Language Processing Research Symposium.

[10] Hartigan, J. A. and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C, 28(1): 100-108.

[11] Hartley, J. 2008. Television Truths: Forms of Knowledge in Popular Culture. Wiley-Blackwell, Malden, MA.

[12] Heverin, T. and L. Zach. 2010. Twitter for City Police Department Information Sharing. Proceedings of the 73rd American Society for Information Science and Technology Conference.

[13] Honeycutt, C. and S. Herring. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. Proceedings of the 42nd Hawaii International Conference on System Sciences.

[14] Imran, M., C. Castillo, F. Diaz, and S. Vieweg. 2012. Processing Social Media Messages in Mass Emergency: A Survey. arXiv.org. Available: http://arxiv.org/abs/1407.7071

[15] Imran, M., C. Castillo, J. Lucas, P. Meier, and S. Vieweg. AIDR: Artificial intelligence for disaster response. 2014. In Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web.

[16] Jansen, B., M. Zhang, K. Sobel, and A. Chowdury. 2009. Twitter Power: Tweets as Electronic Word of Mouth. Journal of the American Society for Information Science and Technology archive, 60(11): 2169-2188.

[17] Kaufman, L. and P. Rousseeuw. 1990. Finding Groups in Data – An Introduction to Cluster Analysis. Probability and Mathematical Statistics. John Wiley and Sons, Inc., NY.

[18] Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Demonstration Session.

[19] Lam, A.J., I. Paner, J. M. Macatangay, D.D. Delos Santos. 2014. Classifying Typhoon Related Tweets. Proceedings of the 10th National Natural Language Processing Research Symposium.

[20] Li, J. and H.R. Rao. 2010. Twitter as a Rapid Response News Service: An Exploration in the Context of the 2008 China Earthquake. The Electronic Journal on Information Systems in Developing Countries, 42(4): 1-22.

[21] Manning, C., P. Raghavan, and H. Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.

[22] Metaxas, P., E. Mustafaraj, and D. Gayo-Avello. How (Not) To Predict Elections. Proceedings of the 3rd IEEE International Conference on Social Computing and the 3rd IEEE International Conference on Privacy, Security, Risk and Trust.

[23] Meier, P. 2012. How the UN Used Social Media in Response to Typhoon Pablo (Updated). iRevolutions: From innovation to Revolutions. Available: http://irevolution.net/2012/12/08/digital-response-typhoon-pablo/

[24] Morales, X.Y.Z. 2010. Networks to the Rescue Tweeting Aid and Relief During Ondoy. M.A. Thesis, Georgetown University.

[25] Neuman, W.R., L. Guggenheim, S. Mo Jang, and S. Young Bae. 2014. The Dynamics of Public Attention: Agenda Setting Meets Big Data. Journal of Communication, 64(2): 193-214.

[26] Oco, N. and R. E. Roxas. 2012. Pattern Matching Refinements to Dictionary-Based Code-Switching Point Detection. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation.

[27] Oco, N., L. R. Syliongka, J. Ilao, and R. E. Roxas. 2013. Dice's Coefficient on Trigram Profiles as Metric for Language Similarity. In Proceedings of the 16th Oriental COCOSDA.

[28] Pablo, Z.C., N. Oco, M. D. G. Roldan, C. Cheng, and R. E. Roxas. Toward an Enriched Understanding of Factors Influencing Filipino Behavior during Elections through the Analysis of Twitter Data. Philippine Political Science Journal, 35(2): 203-224.

[29] Parks, M. 2014. Big Data in Communication Research: Its Contents, and Discontents. Journal of Communication, 64(2): 355-360.

[30] Rajaraman, A. and J. D. Ullman. 2011. Mining of Massive Datasets. Cambridge University Press, Cambridge, UK.

[31] Richards, L. 2005. Handling Qualitative Data: A Practical Guide. Sage, London.

[32] Skinner, J. 2013. Natural Disasters and Twitter: Thinking from both sides of the Tweet. First Monday, 18(9).

[33] Smith, B. 2010. Socially Distributing Public Relations: Twitter, Haiti, and Interactivity in Social Media. Public Relations Review, 36(4): 329–335.

[34] Syliongka, L.R. and N. Oco. 2014. Using Language Modeling and Data Association to Perform Named Entity Recognition. Proceedings of the 10th National Natural Language Processing Research Symposium.

[35] Tumasjan, A., T. Sprenger, P. Sandner, I. Welpe. 2011. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. Social Science Computer Review, 29(4): 402-418.

[36] Vargo, C., L. Guo, M. McCombs, and D. Shaw. 2014. Network Issue Agendas on Twitter During the 2012 Presidential Elections. Journal of Communication, 64(2): 296-316.

[37] Verbeke, M., B. Berendt, L. d'Haenens, and M. Opgenhaffen. 2014. When Two Disciplines Meet, Data Mining for Communication Science. Proceedings of the 2014 Annual Meeting of International Communication Association.

[38] Yin, J., A. Lampert, M. Cameron, B. Robinson, and R. Power. 2012. Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems, 27(6): 52-59.