Twitter: temporal events analysis*

Extended Abstract

Giambattista Amati
Fondazione Ugo Bordoni
Rome, Italy
gba@fub.it

Simone Angelini
Fondazione Ugo Bordoni
Rome, Italy
sangelini@fub.it

Giorgio Gambosi
University of Rome "Tor Vergata"
Rome, Italy
giorgio.gambosi@uniroma2.it

Daniele Pasquini†
University of Rome "Tor Vergata"
Rome, Italy
daniele.pasquini@uniroma2.it

Gianluca Rossi
University of Rome "Tor Vergata"
Rome, Italy
gianluca.rossi@uniroma2.it

Paola Vocca
University of Tuscia
Viterbo, Italy
vocca@unitus.it

ABSTRACT

We perform a temporal analysis of the Twitter stream to investigate the evolution of unique events based on the burst of popularity of associated hashtags. We derive a classification of events according to the different patterns corresponding to the peak of the volume of exchanged message and to how these events propagate on any social network with the same characteristics as Twitter. We first provide a precise definition of *unique events* and correlate them to hashtags. With reference to a specific interval of time, the most popular - with respect to number of tweets- hashtags are then detected using the Seasonal Hybrid ESD (S-H-ESD) technique introduced by Twitter. After identifying the unique hashtags among the 1000 most popular, we have identified, through an unsupervised Machine Learning algorithm applied to the historical temporal series of hashtags limited around the maximum peak, the temporal patterns (clusters) of the events. Finally, using the Twitter features, for each cluster, we have studied both the process at the origin of the event and how they evolve over the network.

*This work was partially conducted in the Big Data Laboratory of ISCOM-MISE (Istituto Superiore delle Comunicazioni, Ministero dello Sviluppo Economico). Giorgio Gambosi and Gianluca Rossi were partially supported by the University of Rome "Tor Vergata" under research programme "Mission: Sustainability" project ISIDE (grant no. E81I18000110005)
†Supported by a grant from ISCOM-MISE (Istituto Superiore delle Comunicazioni, Ministero dello Sviluppo Economico).

Goodtechs '18, November 28–30, 2018, Bologna, Italy

ACM ISBN 978-1-4503-6581-9/18/11…$15.00
https://doi.org/10.1145/3284869.3284902

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Twitter, Event analysis, social network temporal evolution,LDA, Anomalies Detection

ACM Reference Format:
Giambattista Amati, Simone Angelini, Giorgio Gambosi, Daniele Pasquini, Gianluca Rossi, and Paola Vocca. 2018. Twitter: temporal events analysis: Extended Abstract. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs '18), November 28–30, 2018, Bologna, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3284869.3284902

1 INTRODUCTION

We study the temporal evolution of *unique events* in Twitter. Due to the simple structure of the data, the basic functionalities offered and the availability, albeit limited, of the datasets, the microblogging platform Twitter is a point of reference for researchers to understand particular phenomena and provide some specific insight [1–4]. The study of events is an area essentially oriented to individuation, classification and analysis of significant events that occur on social media and it stands out from other research areas for its complexity and for the difficult and tricky interpretation of the results. The event analysis is performed according to a suitable adaptation of the model proposed by Yang and Leskovec in [18], where unique events are detected on the basis of the popularity (number of tweets) of the hashtags with reference to a specific interval of time. Extending the definition in [8], by *event* we refer to *a fact that causes, in a certain time period, a substantial increase (compared to a given level) of the frequency of messages (user actions) on Twitter, all having the same hashtag*; while for *unique* we denote all events such that the corresponding hashtag, observing the temporal trend,

present an activity neither continuous nor periodic but only one unambiguous peak. To detect the most popular hashtags we have used the Seasonal Hybrid ESD (S-H-ESD) algorithm, an anomaly detection technique proposed by Twitter [13] , which identifies the peaks on the temporal series considering them as "anomalies" in the temporal evolution. Based on the Generalized ESD test [16], S-H-ESD allows to detect anomalies in a more robust way from the point of statistical view and can be used to detect both global and local anomalies. This is achieved by employing time series decomposition and using robust statistical metrics, viz., median together with ESD. To manage a dataset of 19 million tweets and more than 300 thousand hashtags we used an Hadoop HDFS and Apache Spark on a cluster of 8 servers to perform the following operations: extracting only the significant data and metadata from the whole dataset, building the historical series of hashtags - on a daily basis; and, then, performing all preliminary operations and the dataset cleanup, even on the texts of the individual tweets. After identifying the unique hashtags among the 1000 most popular with the S-H-ESD technique, through an unsupervised Machine Learning algorithm applied to the series historical hashtags - limited around the maximum peak - have been identified the temporal patterns (clusters) of the events. The use of an unsupervised learning technique was preferred to supervised methods so to avoid a priori hypotheses on the profile of the peaks as well as the manual annotation of the dataset to distinguish the different types of events. With a clustering algorithm, we have identified 5 clusters that have been analyzed to derive a connection with specific types of event by extracting the topic from the tweets of each class, with a specific Topic Model used in distributed mode. The analysis of temporal patterns was carried out on unique events on Twitter, where an event is essentially identified from the "burst" of popularity [7, 18] of a hashtag. It must be said that the discovery of precise patterns on the Web, as noted in [18], is not trivial because of the unpredictable behavior of a person, which depends on several factors (among which the interaction between individual users, small groups and corporations). As said above, the clustering was performed with a robust and highly scalable algorithm on the historical series of the hashtags to verify the existence of specific features in each class , also according to the type of event associated (determined by the topics extracted). Finally, using the features of Twitter ([5, 6, 9, 15, 19]) on the individual clusters, we studied the processes at the origin of the popularity profiles on the social network. The result of the analysis clearly shows the duality between the *endogenous event* and the *exogenous event.*

2 DATA SET

For the experiments we use a tweet sample dataset in Italian language obtained by the Tweeter stream filtered using a list of the most used Italian stop words and the Twitter native selection function for languages. The tweets were collected between October 30th 2015 and January 3rd 2016 (66 days). Table 1 shows some preliminary information about our dataset.

Table 1: Dataset general informations

Size	21.33 GB
Number of tweets	19, 000, 000
Number of hashtags	377, 215
Tweets with hashtags	15, 754, 093

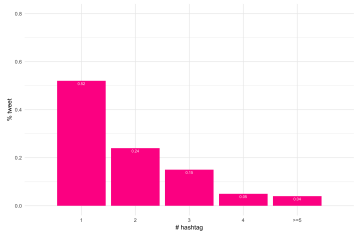

Figure 1: Distribution of tweets on the number of hashtags

3 TECHNOLOGIES

For the upload, extraction and manipulation of the Twitter sampling that was stored on Hadoop HDFS, we used Scala on Apache Spark and Python on a cluster of 8 servers,and Spark RDD and Spark Dataframe as principal data structures. As a result of this elaboration we derived the 1000 most popular hashtags.

4 HASHTAGS ANALYSIS

For our analysis we ignore all tweets without hashtag; from Table 1 it follows that, in this way, we remove less than 20% of tweets from the dataset. Figure 1 shows the percentage of tweets with a given number of hashtags: it is interesting to note that more than 50% of tweets contain only one hashtag.

For every hashtag h, let us consider the number occurrences of h and let us consider for every x between 1 and the total number of tweets, the fraction of hashtags that occur in at least x tweets. Figure 2 shows this relation that is fitted by a log-normal distribution with parameters $(-7.99, 4.26)$ and p-value 0.71. For these reasons in our analysis we consider the 1000 more frequent hashtags.

Let H be the 1000 most popular hashtags. We want to define the *time series* of every h in H in the time interval $\mathcal{T} = [0, \cdots, T]$: for each h in H and t in $\mathcal{T}$ let us define $x_h(t)$ as the number of tweets containing the hashtag h created or retweetted at time t – in short, the *volume* of hashtag h at time t; the sequence $x_h(0), \ldots, x_h(T)$ is the time-series of hashtag h.

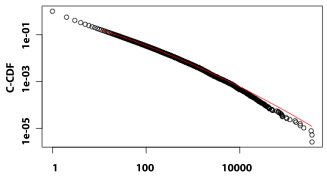

Figure 2: The fraction of the number of hashtag that occur in at least a given number of tweets

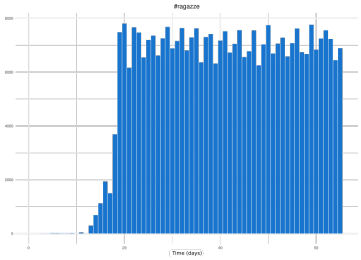

Figure 3: Time series of hashtag *ragazze* (girls)

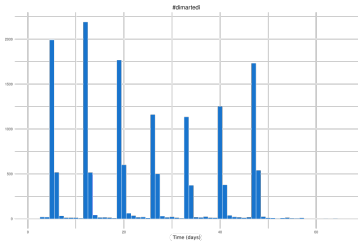

Figure 4: Time series of hashtag *dimartedì* (an Italian weekly TV show)

We are interested in the daily activity around every hashtag, then every element in $\mathcal{T}$ represents a time interval of 24 hours, that is $T = 65$ (see Section 2).

We can distinguish at least three types of time-series based on their profile: the time-series with *continuum profile* show a constant level of daily activity (see Figure 3); a *periodic profile* is own of time-series of hashtags associated with events that recur at a fixed time interval such as prime time television shows (see Figure 4); finally the time-series with an *isolated peak* in their profile are produced by hashtags associated to unique events (see Figure 5).

In this paper we will focus on the last class of hashtags because it represents "singular" events more interesting to study. Peaks in time-series are identified by using the Anomaly Detection Seasonal Hybrid ESD (S-H-ESD) algorithm [12, 14]. The algorithm can detect local and global anomalies and is builds upon the Generalized ESD test [17].

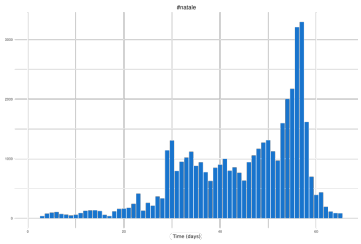

Figure 5: Time series of hashtag *natale* (Christmas)

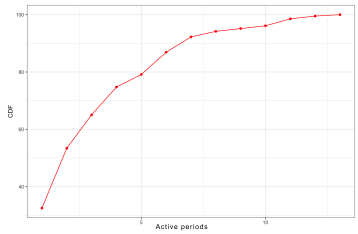

Figure 6: Cumulative distribution function of the number of active periods

Since a hashtag time-series can exhibit different peaks, and since we are only interested in isolated peaks (corresponding to global anomalies), for each hashtag time-series we ignore all the peaks separated from the others for less than one week. Then all peaks but the maximum ones are ignored from all time-series. Finally, we ignore all hashtags (and the respective time-series) without peaks. By applying this method we get 206 hashtags in 1.913.470 tweets.

Active periods

We consider a hashtag to be inactive if it is used in up to 20 tweets within 24 hours. The bound of 20 tweets per day is due to the background noise associated to most popular tweets.

Figure 6 shows the cumulative distribution function of the active periods numbers and Figure 7 shows the cumulative distribution function of the length of active periods. From the first one it follows that more than 90% of hashtags is active for at most 7 days and from the last one follows that the length of the active periods is very short: almost 90% of hashtags have active periods not exceeding 10 days. This suggests that, as one can guess, the hashtags associated with unique events are sporadic, occasional and volatile.

5 CLUSTERING

We will classify the hashtag time series characterized by isolated peaks with the *K-SC* clustering algorithm proposed in [18]. The strenght of this classification is its invariance with respect to the operations of scaling and translation. This means that curves with the same shape but with different

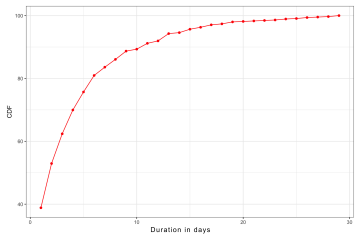

Figure 7: Cumulative distribution function of the length of active periods

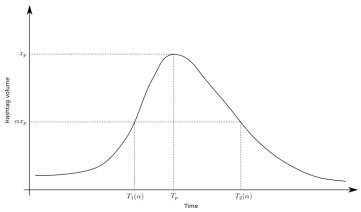

Figure 8: Trend of the time series of a typical hasthag

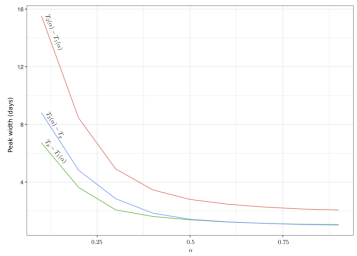

Figure 9: Core width as a function of α

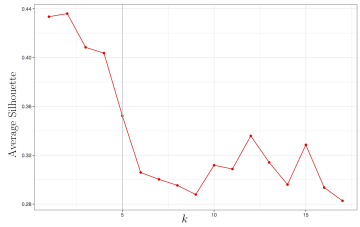

Figure 10: Average Silhouette

size or position can classified in the same way contrary to what happens for the *K-means* classification [10]. All our time series are 65 days long, in order to limit the background noise effects we have chosen to truncate the series and focus on the period around the peak.

We define the *core* of a time series as the time interval around the peak where the volumes are at least a defined fraction of the maximum volume encountered in the peak. The time interval outside the core will be truncated from the time series. Let x_p be the maximum of the time series, that is the volume at the peak; T_p be the time when the peak occurs and α be the value between 0 and 1 such that αv_p is the minimum volume defining the core. The core of the time series is the interval between $T_1(\alpha)$ and $T_2(\alpha)$ where $T_1(\alpha) < T_p$, $T_2(\alpha) > T_p$ and for all t between $T_1(\alpha)$ and $T_2(\alpha)$, the volume at t is at least αx_p (see Figure 8).

Now we will see how to choose the appropriate value of α. Given α, the core width of a time series is $T_2(\alpha) - T_1(\alpha)$; the left core width is $T_p - T_1(\alpha)$ and the right core width is $T_2(\alpha) - T_p$. Figure 9 shows the average values of the three core widths. With values of α smaller that 0.5 the right core width is bigger than the left core width, this implies that the slope on the left of the core is greater than the slope on the right. We have set up the core size to 21 days: this number is obtained by observing that in the average with two weeks we guarantee a minimum peak volume of at least 12% of the maximum value (see Figure 9); we add a further week to cover any anomalous values.

In the *K-Means* algorithm the number of clusters must be specified as parameter; this is also true for all the other its variants including th *K-SC* algorithm. In order to choose the most appropriate number of clusters k we ran the *K-SC* algorithm for various values of k and measured the quality of the clustering with the Average Silhouette metric [10].

Figure 10 shows the Average Silhouette as a function of the number of clusters k; the best clusters are obtained in correspondence of higher Average Silhouette values. In our case the best scores correspond to small values of k. We chose $k = 5$ as a compromise between clustering quality and appropriate number of clusters. The number of iterations is 100.

Figure 11 shows the 5 clusters returned by the algorithm. For every cluster it is indicated the percentage of hashtags that it contains, the volume of the maximum and the volumes of the left and right side of the peak (in percentage). Cluster 4 is the bigger (46.12% of the hashtags) and is characterized by the concentration of almost all its volume in a single day. The temporal pattern of Cluster 1 is almost similar to Cluster 4 although the former has a more pronounced tail, highlighting a more lasting relative interest from community members. Clusters 3 and 4 have an opposite trend: the former reveals a large relative volume after the peak, while in the latter it is revealed before the peak. Finally, cluster 5 is rather symmetrical. It is also interesting to note that almost 70% fall into two clusters (1 and 4) whose volumes are all concentrated around the peak.

6 TOPIC EXTRACTION

In the next step we examine the topics that characterize the five clusters that we have identified. We define a single document d_h for every hashtag h by aggregating all the tweets containing hashtag h. The collection of documents related

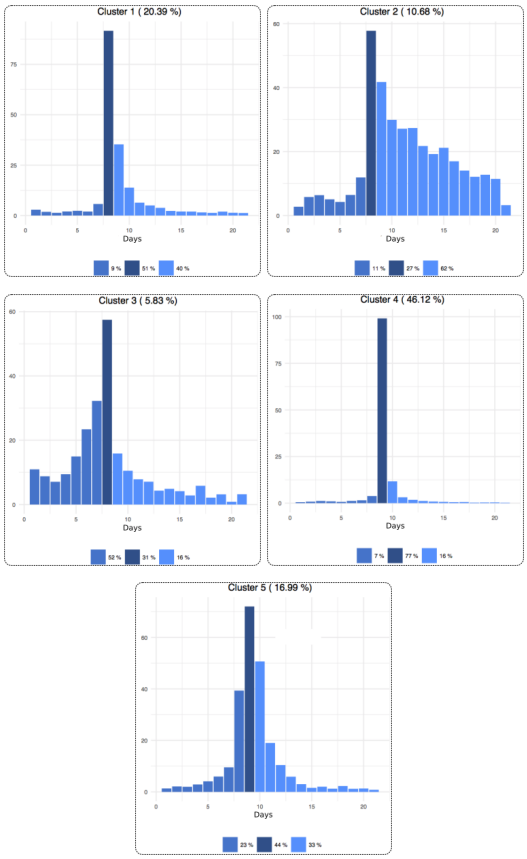

Figure 11: The 5 clusters found by the *K-SC* algorithm

with cluster i (with $i = 1, \ldots, 5$) contains all the documents d_h such that hashtag h is in cluster i; let us call this collection D_i. The five collections $D_1, \ldots, D_5$ are used as input of the Latent Dirichlet Allocation (LDA) algorithm to extract the topic of the five clusters The number of topics k is a parameter of the algorithm, its value is chosen by running 200 iteration of LDA for different numbers of topic to find the one with the maximum log-likelihood. For the α and β parameters we chose the values $50/k+1$ and 1.1, respectively. For each cluster we got a list of topics identified with the 5 most likely words. For lack of space we show only the list related to cluster 1.

(1) speech, Mattarella (*Italian President*), president, Italy, politics;
(2) Grillo (*blogger and politician*), Italy, dictator, blog, Renzi (*ex Prime Minister*);
(3) dignity, unity, Pope, money, History;
(4) Renzi, Italian, Italy, press conference, greeting
(5) champions league, Juventus (*soccer team*), Roma (*soccer team*), Italian, soccer market;
(6) paralyzed, Wednesday, competition, tomorrow, hired;
(7) elections, upcoming, party, wins, electoral;
(8) Pompei, money, excavations, Renzi, position;
(9) Turkey, Isis, Erdogan, city, Turkish;
(10) Paris, terrorism, dead, safety, war;
(11) good night, readers, night, book, evening;
(12) Renzi, bridge, water, Messina, channel;
(13) Renzi, Mannoia (*singer*), concert, Fiorella (*Mannoia's first name*), excluding;
(14) match, young people, San Remo, names, festival;
(15) Christmas, Sant'Egidio (*charity organization*), lunch, mercy;
(16) Left, Renzi, Fassina (*politician*), born, politics;
(17) Cinema, star, wars, force, risveglio;
(18) banquets, brave, excavations, square, Renzi;
(19) Apple, IPhone, tax authorities, agreement, Samsung;
(20) San Remo, young people, solidarity, Scialpi (*singer*), discrimination;
(21) Livorno, balance, company, deficit, garbage;
(22) graduation, Poletti (*ex Minister of Labor*), young people, schedule, Italy;
(23) stamp, license plate, proposal, law, bicycle;
(24) expo, Milan, success, Italy, end.

The topics mostly concern unexpected events often related to politics. There are also topics regarding sports (Champions League) and entertainment (Sanremo festival). Also cluster 2 is related to unexpected events like the crack of Etruria bank. Cluster 3 is completely different; t is mainly related to scheduled events like Christmas, New Year, an event dedicated to books, Telethon and the Vatileaks process. Cluster 4 collects daily events: Black Friday, opening of the Holy Door in Vatican and sport events. Finally, cluster 5 is characterized by mixed events with reduced impact in public opinion.

7 CONCLUSIONS

Differently from other works, in this paper we adopt a highly scalable algorithm for clustering events according the historical temporal series. Each cluster characterize unique events on Twitter. On our data sets we have identified five different clusters. The events considered were previously selected using a technique for Anomaly Detection: the Seasonal Hybrid ESD (S-H-ESD) technique, first proposed by Twitter and also tested on Netflix, has been preferred with respect to the classic ESD since it turns out to be more statistically robust. Successively, we used LDA for performing an analysis of the extracted topics, by aggregating the tweets for each hashtag, it was possible to identify the types of events associated to a specific temporal pattern, by semantic annotation. Since it was out of the scope of our research, this step has been accomplished manually. In particular, since in cluster 3 we have a high message volume before the peak, it can be associated with *programmed event*. It is worth noting that the

same temporal pattern is typically associated to endogenous events. Instead, cluster 4 is associated to a temporal pattern referring to a *single day event*, which receive the attention of users only in the day in which they occur, and, considering the high proportion retweet, stem from endogenous processes. Clusters 1 and 2, on the other hand, are associated to unexpected events that impact differently on users. While in the first case we can interpret the phenomenon as an endogenous propagation, in the second - observing the proportion of tweets with url - it can be assumed that the events are guided from external factors, injected into the network through the mass media. Finally, cluster 5 shows a *symmetric* behavior, since it corresponds to *mixed* events, in which both endogenous processes and exogenous processes contribute to the propagation information. Clearly the performed analysis can be further refined. With more massive datasets, using K-SC clustering algorithm it is potentially possible to identify new clusters and therefore new temporal patterns. Additionally, it might be interesting to study the events propagation analyzing the Follow Graph of each individual user [1]. As further line of investigation, it would be interesting to use Natural Language Processing and Named Entity Recognition algorithms in combination with LDA to improve the results of the topic model.

REFERENCES

[1] Giambattista Amati, Simone Angelini, Marco Bianchi, Gianmarco Fusco, Giorgio Gambosi, Giancarlo Gaudino, Giuseppe Marcone, Gianluca Rossi, and Paola Vocca. 2015. Moving Beyond the Twitter Follow Graph. In *KDIR 2015 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 1, Lisbon, Portugal, November 12-14, 2015*, Vol. 1. IEEE, 612–619. https://doi.org/10.5220/0005616906120619

[2] Giambattista Amati, Simone Angelini, Francesca Capri, Giorgio Gambosi, Gianluca Rossi, and Paola Vocca. 2016. Modelling the temporal evolution of the retweet graph. *IADIS International Journal on Computer Science & Information Systems* 11, 2 (2016).

[3] Giambattista Amati, Simone Angelini, Francesca Capri, Giorgio Gambosi, Gianluca Rossi, and Paola Vocca. 2016. Twitter Temporal Evolution Analysis: Comparing Event and Topic Driven Retweet Graphs. In *BIGDACI 2016 - Proceedings of the International Conference on Big Data Analytics, Data Mining and Computational Intelligence, Volume 1, Funchal, Madeira, Portugal, July 2 – 4, 2016.*

[4] Giambattista Amati, Simone Angelini, Francesca Capri, Giorgio Gambosi, Gianluca Rossi, and Paola Vocca. 2017. On the Retweet Decay of the Evolutionary Retweet Graph. In *Smart Objects and Technologies for Social Good: Second International Conference, GOODTECHS 2016, Venice, Italy, November 30 – December 1, 2016, Proceedings*. Springer International Publishing, Cham, 243–253. https://doi.org/10.1007/978-3-319-61949-1_26

[5] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, and Matei Zaharia. 2015. Spark SQL: Relational Data Processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*. ACM, New York, NY, USA, 1383–1394. https://doi.org/10.1145/2723372.2742797

[6] Sitaram Asur, Bernardo A. Huberman, Gábor Szabó, and Chunyan Wang. 2011. Trends in Social Media : Persistence and Decay. *CoRR* abs/1102.1402 (2011).

[7] Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105, 41 (2008), 15649–15653. https://doi.org/10.1073/pnas.0803685105 arXiv:http://www.pnas.org/content/105/41/15649.full.pdf

[8] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. 2012. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 93–102. https://doi.org/10.1109/VAST.2012.6400485

[9] Wanqiu Guan, Haoyu Gao, Mingmin Yang, Yuan Li, Haixin Ma, Weining Qian, Zhigang Cao, and Xiaoguang Yang. 2013. Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events. 395 (04 2013).

[10] L. Kaufman and P.J. Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. Number Book 59 in Wiley Series in Probability and Statistics. Wiley-Interscience. https://doi.org/10.1002/9780470316801

[11] A. Kejariwal. 2015. Introducing practical and robust anomaly detection in a time series. Retrieved Tuesday, 6 January 2015 from https://www.google.com/search?q=Introducing+practical+and+robust+anomaly+detection+in+a+time+series&ie=utf-8&oe=utf-8&client=firefox-b-ab

[12] A. Kejariwal. 2015. Introducing practical and robust anomaly detection in a time series. https://blog.twitter.com/engineering/en_us/a/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.html

[13] Stephen Kelly and Khurshid Ahmad. 2015. Propagating Disaster Warnings on Social and Digital Media. In *Intelligent Data Engineering and Automated Learning – IDEAL 2015*, Konrad Jackowski, Robert Burduk, Krzysztof Walkowiak, Michal Wozniak, and Hujun Yin (Eds.). Springer International Publishing, Cham, 475–484.

[14] S. Kelly and K. Ahmad. 2015. Propagating Disaster Warnings on Social and Digital Media. In *Intelligent Data Engineering and Automated Learning – IDEAL 2015*, Konrad Jackowski, Robert Burduk, Krzysztof Walkowiak, Michal Wozniak, and Hujun Yin (Eds.). Springer International Publishing, Cham, 475–484.

[15] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. 2012. Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 251–260. https://doi.org/10.1145/2187836.2187871

[16] Bernard Rosner. 1983. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* 25, 2 (May 1983), 165–172.

[17] B. Rosner. 1983. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* 25, 2 (1983), 165–172. http://www.jstor.org/stable/1268549

[18] J. Yang and J. Leskovec. 2011. Patterns of Temporal Variation in Online Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 177–186. https://doi.org/10.1145/1935826.1935863

[19] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *Advances in Information Retrieval*, Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 338–349.