Early Signals of Trending Rumor Event in Streaming Social Media

Article · July 2017

Shihan Wang
University of Amsterdam

Dirk Helbing
ETH Zurich

Takao Terano
Tokyo Institute of Technology

Early Signals of Trending Rumor Event in Streaming Social Media

Shihan Wang
School of Computing
Tokyo Institute of Technology, Japan
ShihanW@trn.dis.titech.ac.jp

Izabela Moise, Dirk Helbing
Computational Social Science
ETH Zurich, Switzerland
imoise@ethz.ch, dirk.helbing@gess.ethz.ch

Takao Terano
School of Computing
Tokyo Institute of Technology, Japan
terano@dis.titech.ac.jp

***Abstract*—In this study, we propose a mechanism for identifying early signals of trending rumor events (i.e. controversial emerging topics) in streaming social media. The pattern, combining features of both user's attitude and information diffusion, is applied in the sliding windows of social media data streams. By capturing and analyzing frequent patterns within early windows, we found signal patterns appearing at very early stages of trending rumor events (in average, months before their peak time). Our preliminary empirical analysis is applied in two different Twitter datasets. The obtained results indicate the potential of our approach to detect trending rumor event candidates (with high probability of being false) as early as possible in real-time environments.**

I. Introduction

In the past decade, the proliferation of social media services has drastically changed the news dissemination process: along with traditional news sources (television channels and newspapers), an ever-increasing audience turns to social media platforms, such as Twitter, for real-time information sharing. However, the same features (speed and ease of information sharing) that make social media a powerful dissemination tool, have also created a prolific environment for the fast spread of rumors. In psychology, rumors are defined as information or statements that cannot be verified as true or false while spreading among people [1]. Among the many rumors that spread in social media, we focus on those emerging as popular topics and attracting tremendous social attention. We refer to them by *trending rumor events*. In this context, trending rumor event is defined as a group of social media posts indicating one controversial emerging topic (i.e. statement or claim), that is discussed frequently and able to reach a wide audience within a specific time period.

Compared to rumors with limited attention, trending rumor events bring uncertainty into a wider audience and fuel stronger social chaos. Understanding trending rumor events is essential as they fuel mistrust, cause panic and often prompt irrational behaviours. The implications are obvious in emergency and disaster situations, when trending rumor events are categorized as fake information. For example, medical researchers are raising concerns that public health misinformation could impede efforts to limit the spread of a virus. However, the implications go beyond this. Social media platforms and Twitter in particular, provide users with the *service* of fast communication and information sharing. Distorted information spread through these platforms, alters the provided service as it damages its purpose of being beneficial to society, in often cases, achieving the opposite.

In the light of these important implications, we focus on analyzing features of trending rumor events in social media. There are three important keywords that outline our contribution: *early*, *signal* and *streaming*. In particular, we propose an approach to capture significant signals that appear at early stages of trending rumor events, captured in Twitter data streams. By using these signals as potential indicators, we aim at detecting candidates of trending rumor events (with high likelihood of being false) as early as possible in streaming social media. In this context, we overcome the challenge of the streaming environment (limited access to data) by proposing a sliding window mechanism, in which we analyze snapshots of streaming data, captured within a window. In the first step, we rely on previously defined patterns to identify the traces of trending rumors in short term time-series [2]. These patterns combine both users' attitude/sentiment and propagation structure of the rumor diffusion process. In a second step, we build a stream of matched patterns with their appearing windows from individual streams of trending rumor event. Based on the empirical analysis of patterns, we address questions on how early the signal patterns can be captured and once detected, how reliable are they for distinguishing between false and true information. We consider two different Twitter datasets for analysis. Results found that our method is promising for detecting signals of trending rumor events at very early stages, that are overlooked by previous work although they have the potential to distinguish false trending rumor events in real-time streaming data.

II. Related Work

This section introduces the related work on analyzing signals in rumor-related social media (in particular in Twitter and Sina Weibo, less in Facebook [3]) and discusses the previously unaddressed challenges that we tackle in our approach.

Prior research analyzed rumors in Twitter at the tweet-level and explored the effectiveness of their features in retrieving misinformation [4]. The approach of Castillo et al. [5] proposed the definition of newsworthy topic (a group of tweets) and assessed its credibility automatically. They examined relevant features extracted from four categories (text, author,

0730-3157/17 $31.00 © 2017 IEEE
DOI 10.1109/COMPSAC.2017.115

propagation and topic properties), and particularly emphasized the importance of propagation features. Kwon et al. [6] extended the above-mentioned approach by first understanding the temporal properties in rumor spreading. Afterwards, Zubiaga et al. [7] monitored conversational features across the timeline of Twitter events, while Wu et al. [8] analyzed propagation patterns with temporal behaviors of false rumors in Sina Weibo. Ma et al. [9] focused on a different angle of widely used features, namely the slopes of features between consecutive time intervals. These previous studies analyzed signal features of rumors in online social media, however, their analyses focus on entail historical data by the current time.

In addition, extensive literature explored the time-sensitive signals of rumors, which are essential in real-time detection scenarios. Castillo et al. [10] first captured early features of rumors by considering tweets before the first peak activity in static data. Later, Liu et al. [11] dynamically uncovered the influence of belief features at early stages of false rumors, although the feature vectors were constructed based on an accumulative data collection process, instead of a real-time data stream. Qin et al. [12] proposed an early rumor feature by considering other data sources from traditional media, in addition to social media. The study of Zhao et al. [13] follows a similar direction to ours. They focused on inquiry patterns appearing early in disputed factual claims and examined their significant role in Twitter streams of the Boston bombing events. Extending their definition of rumor cluster, we define the trending rumor event by also integrating the "trending" characteristics (defined in [14]). To overcome the potential limitation of text-based regular expressions, we propose to compute early signals in streaming environment from both context and information propagation perspectives, which represents a novelty in rumor-related social media events.

III. Design and Methodology

This section introduces the design of signals (defined as *patterns*) appearing at early stages in trending rumor events, as well as the methodology that seeks to automatically capture these patterns in streaming social data.

A. Sliding Window Mechanism

The sliding window mechanism is the underlying design principle for both pattern and methodology construction. The sliding window principle is commonly used in mining real-time data streams [15], as it focuses on recent data while decreasing the computational complexity (by limiting the amount of processed data). In our case, instead of processing the entire past history of social media data, we consider only snapshots of recent data streams. Each trending rumor event is decomposed into sliding windows: given a fixed window size, the data stream is split into windows of sub-streams. We also consider the overlapping sliding windows in order to balance the tradeoff between real-time analysis and collecting sufficient data to be able to capture interesting patterns. This allows us to efficiently decrease the interval of detection.

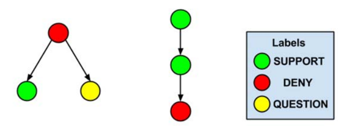

Fig. 1. Example of the patterns

B. Window-based Pattern

Our goal is to obtain a simple pattern definition, for an efficient detection in the data streams. The defined pattern integrates two significant properties: propagation structure and users' attitude toward the trending rumor event.

In a nutshell, Figure 1 represents the patterns as labeled graphs. Within a controversial trending rumor event, various users' attitudes toward the topic are captured by three labels (SUPPORT, DENY and QUESTION). Two main network cascades (star and path) are extracted as structural base for the diffusion process. Based on these features, the patterns in Figure 1 are defined as: $\{SUPPORT \leftarrow DENY \rightarrow QUESTION\}$ and $\{SUPPORT \rightarrow SUPPORT \rightarrow DENY\}$ respectively. The previous work introduced the theoretical base for these patterns as well as their ability to distinguish false rumors from valid news in short-term time series [2].

C. Window-based Methodology

Figure 2 illustrates the proposed methodology for processing raw streams of social media data into detected patterns.

The multi-step workflow of the methodology begins with assuming the trending rumor events have been extracted in overall data stream. While each newly received post is allocated to the relevant event stream, each trending rumor event holds an incremental post stream. The raw post stream of trending rumor event is first decomposed into sub-streams based on the sliding window principle. We further consider for analysis each individual window and we rely on a pattern matching algorithm to detect and identify patterns. During this process, sentiment analysis of the post content is also applied to provide the necessary features for pattern matching. When a pattern is matched, the last step also emits an alarm signal reported at the current time step, carrying a potential indicator for the trending rumor event.

The employed pattern matching algorithm particularly tracks matched patterns in streaming environment [2]. Overall, the algorithm provides a stream of matched patterns and the timestamps at which they are detected. Given a pattern, the algorithm considers each post with diffusion information (i.e. retweet and reply) as an edge, where the direction of information spreading determines the direction of edge. The label of nodes is defined as the users' attitude, which is extracted by relying on sentiment analysis of the post content (see Section V-A).

IV. Twitter data acquisition and preprocessing

Our empirical analysis makes use of two main datasets, described in the following. The first dataset is a clean Twitter

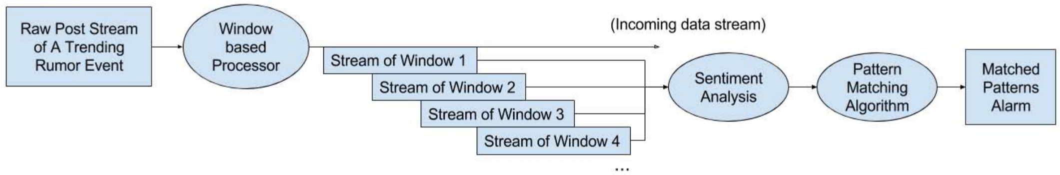

Fig. 2. The overall flow of the methodology

data of 109 trending topics from 2006 to 2009. It was collected and published by the KAIST team [6]. In this data, each trending event contains a full set of tweets, as well as a confirmed credibility label. The validation of false or true labels has been well annotated and evaluated in their work based on both investigation websites and human expertise. In particular, we selected 10 false and 10 true events containing over 500 tweets and a variety of user attitudes. On average, each selected trending rumor event has over 3000 tweets.

Additionally, we acquire and process a real-time streamed Twitter dataset, capturing the conversation on an important and recent topic: the Zika virus. This dataset is referred by the *Zika* dataset, throughout the rest of the paper. As a proxy for identifying trending rumor events, we develop a method that proceeds in two steps: first, we rely on topic modeling and clustering, for identifying the Zika-related topics (i.e. data clusters of tweets) that received a lot of attention in the Twitter conversation; in a second phase, we investigate the credibility of Zika-related topics by consulting several credible websites. This process allows us to collect a list of labeled trending rumor events in the Zika dataset. In this section, we mainly provide an overview of the data acquisition and preprocessing steps for the Zika dataset.

A. Data Acquisition

The Zika dataset we consider in this paper, was collected through the Twitter Streaming API[1], which grants public access to the 1% of public tweets. We define a custom filter using keywords from the official Twitter channels of the Centre for Disease Control and Prevention (CDC) and the World Health Organization (WHO), keywords related to the zika virus: `zika,mosquito,aedes,microcephaly`.

In addition to the text of the tweet, the streaming mechanism also provides relevant information about the tweet, such as posting time and date, geo-tags, author information, retweet counts, and network indicators (number of followers and friends). The dataset contains around 60 million tweets, covering the time period from April 13th to August 31 2016.

B. Data Preprocessing

In the following, we describe the multi-step pipeline that transforms raw tweets into trending rumor events (with a group of tweets and its false or true label) that can be provided as input to complex analysis.

1) Extract Topics in Data: In a first step, we acquire the most discussed topics in the Zika dataset by utilizing topic modeling and clustering techniques. The phase starts with extracting message field from each tweet and transforms the text into a "bag of words" representation. We implemented this first phase on top of Hadoop[2], by filtering out all non-latin characters and removing @mentions, hashtags, URLs, as well as stopwords (most common words used in the english language).

Then, we use a powerful technique widely used in data mining and information retrieval, topic modeling, to extract the distribution of hidden or *latent* topics from input tweet texts. The Mahout[3] implementation of Latent Dirichlet Allocation (LDA) [16] is applied in our work. The topics we extracted from the dataset, mostly reflect real-world events that created an important echo in the Twitter-sphere. Figure 3 illustrates an example of two topics, in the form of size-coded word-clouds: a larger font for a specific word indicates a higher probability of the word to appear in the topic definition.

In the next step, we cluster the tweets into captured trending events based on their similarity. In a nutshell, each tweet is represented as a vector in the overall topic space (i.e. a distribution of topics). We perform the clustering algorithm on the overall tweet-topic matrix [17]. The resulted clusters identify groups of tweets that are defined by similar topics. Finally, we sort the topics in descending order, based on the number of their tweets. By doing so, we are able to construct a proxy to discovering emerging topics in our dataset.

2) Collect Credible Statements: Based on the topics discovered at the previous step, we further enhance our dataset by collecting Zika-related statements corresponding to the trending topics.

In the collection process, we rely on websites of rumor tracking, popular news and official health organization channels (e.g. snopes, emergent.info, Times, BBC, WHO and CDC). By cross-matching the timestamps of the statements with the time range of our dataset, we compile a list of true and false Zika statements, with a prior confirmed veracity. Within the obtained latent topics which correspond to the verified statements, we further select those with the largest number of tweets and various user attitudes. Thus, we are able to extract several groups of Zika-related tweets which are labeled as true or false. Table I summarizes a list of 6 selected trending rumor

[1]Source: https://dev.twitter.com/rest/public
and https://dev.twitter.com/streaming/overview

[2]Source: http://www.hadoop.org

[3]Source: http://mahout.apache.org/

(a) (b)

Fig. 3. An illustration of topics discovered by running LDA on the Zika dataset. Topic (a) captures events from sports world, that triggered a lot of echo in the news and social media: by mid July, shortly before the start of the Olympics, the Canadian tennis player Milos Raonic announced his official withdrawal from participation to the Rio 2016 Olympic Games. Several other athletes (Tomas Berdych and Rory Mcllroy) expressed similar concerns amid fears and risks associated with the virus. Topic (b) highlights the first travel advisory issued by public health officials at the beginning of June. the Zika virus has prompted public health officials to warn pregnant women to avoid traveling to a part of the continental US, in response to a growing outbreak of the mosquito-borne disease in South Florida. 10 new cases have been reported in Miami, with strong evidence that mosquito control efforts were not working as well as officials hoped.

events, with their associated labels, which we further analyze in the experimental phase.

V. Experiment and Results

In the experimental phase, we seek to address two main questions: how early can we detect these rumor signals; once detected, how useful are they for distinguishing false trending rumor events. We investigate these two questions first in the clean dataset, then in the Zika dataset, which provides an experience closer to the real-world online data.

A. Data Processing and Setup

As explained in Section III, the data stream of each trending rumor event is transformed into sliding windows of sub-streams. Starting with collecting tweet frequency in each window over a trending event, we examine two important parameters: the window size and the length of interval between two windows. The observation shows that the latter value is less significant than the window time. In addition, the trend of tweet frequency in sliding windows remains relatively constant when the window length is unchanged. Based on these initial remarks, we set values for the window size and interval length to 5 and 1 hour respectively.

In order to build the features defined in the patterns (see Section III), we rely on the official Twitter APIs for obtaining the *retweet*, *mention* and *reply* metadata. Sentiment analysis is implemented based on the Natural Language Toolkit (`nltk`) [18], able to learn positive (SUPPORT) or negative (DENY) opinion from the text of each tweet. We also identify QUESTION tweets based on simple lexical patterns [13], [19]. A NEUTRAL label is also attached to tweets expressing factual information, without clear attitude indicators. Overall, the results of sentiment label show a consistent distribution with previous work [2], [20] and reliable for further experiments.

By considering all possible combinations of the four labels and two shapes, we construct a total number of 112 candidate patterns. In the case of "star" shape, the two children nodes are symmetric with respect to the propagation structure. Thus, we consider the patterns $\{SUPPORT \leftarrow DENY \rightarrow QUESTION\}$ and $\{QUESTION \leftarrow DENY \rightarrow SUPPORT\}$ to be equivalent.

B. Window-based Experiment 1

To address the first question, we start with the clean dataset and we proceed as follows: for each trending rumor event, we consider all possible patterns and detect matched patterns independently in the stream of each window.

We further count the frequency of matched patterns and collect these values across the succession of sliding windows. Fig 4 shows the comparison between tweet frequency and matched pattern frequency in sliding windows of ten false trending rumor events.

This Figure indicates that in all of the ten false trending rumor events, we were able to identify the matched patterns before the peak window (with the largest amount of tweets). Here, one unit of X-axis corresponds to one window. In some cases (i.e. LadyGG and ObamaAT), the first matched patterns appear in very early windows of the trending events, even if the tweet frequency is still low. The same trend are also found in the true trending rumor events. This observation indicates that our pattern is capable to capture signals of trending rumor events at very early stage.

C. Window-based Experiment 2

In a second set of experiments, we further explore the early signal patterns in trending rumor events. We use a frequent pattern mining technique to extract patterns, that not only appear early in the trending rumor events, but can also distinguish false from true trending rumor events.

Our approach considers each window as a list of matched patterns. By relying on the `pymining` python library [21], we explore the frequent patterns that appear in time windows earlier than the peak window. Following the study of peak time in social media topics [10], we denote the *peak window* as the period of time with the highest tweet frequency for the considered event (this is also referred to as the fully flourish of the trending rumor event). We explore two different time periods: 1) the first window in which we ever detect matched patterns; 2) sliding windows within 20 hours before the peak window, as they are the closest to the fully flourish.

Our analysis shows that the first window with matched patterns appears much earlier than 20 hours before the peak. In average, patterns are first identified about 5 months before the peak in trending rumor events, and at least 38 hours. Also, trained trending rumor events generally hold more patterns in the earlier windows (the first detected ones) than within 20 hours near the peak window. This observation indicates that sliding windows at early stages also contain recognizable properties, which are captured by our designed patterns.

Moreover, there is a similar trend of frequent patterns across windows in two different periods: for false trending

TABLE I
A LIST OF SELECTED TRENDING RUMOR EVENTS IN THE ZIKA DATASET

Trending Rumor Event	**Veracity**	**Tweet Amount**
Brazil's Rio Olympic will cause globe Zika spread.	False	41892
Miami outbreaks Zika, CDC issued travel warning.	True	7125
Congress rejected the Zika funding Obama proposed.	True	23412
First female-to-male Zika sexual transmission reported.	True	6331
Milos Raonic out of Olympic because of the fear for Zika.	True	3729
Zika vaccine go ahead for human trail.	True	2164

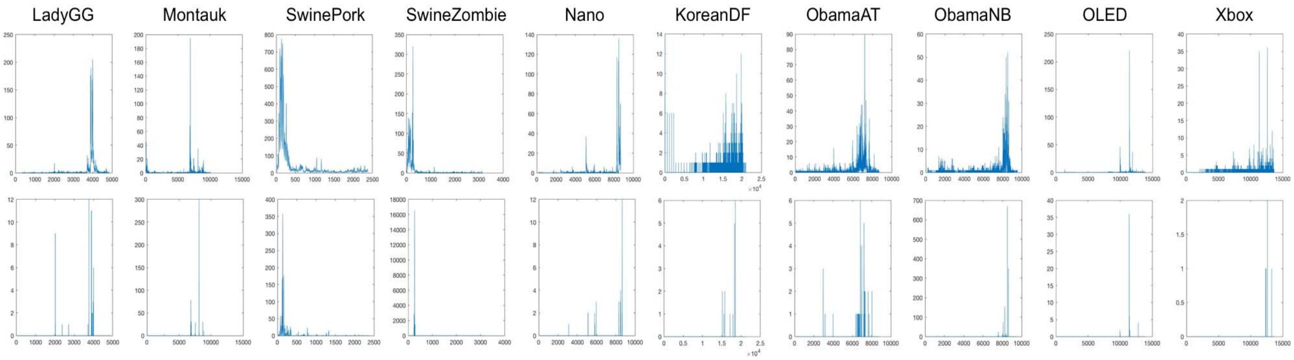

Fig. 4. Tweet frequency (up) and matched pattern frequency (down) in sliding windows of 10 false trending rumor events

TABLE II
NUMBER OF EVENTS WITH QUESTION OR DENY PATTERNS

	With	**Without**
False Trending Rumor Events	8	2
True Trending Rumor Events	4	6

rumor events, frequently detected patterns have QUESTION or DENY labels, while only few valid events have QUESTION or DENY related patterns. Such similarity indicates the potential use of these patterns to evaluate false trending rumor events in the very early sliding windows. In particular, the previously ignored DENY label plays an important role in the frequent patterns. By considering either QUESTION or DENY related patterns, our work can contribute to a better recall of trending rumor event early detection.

The number of different trending rumor events with and without QUESTION or DENY related patterns are summarized in Table II. As shown, 8 out of 10 false trending rumor events hold QUESTION or DENY related patterns, which contributes to a good recall of 80%, with reasonable 66.7% precision and 70% accuracy.

D. Window-based Experiment 3

In a third phase, we conduct the same experiments on the Zika dataset. Figure 5 depicts the tweet and pattern frequencies for one false trending rumor event (left) and five true trending rumor events (right five). By comparing two frequencies, the Figure shows that the first pattern is detected earlier than peak time, for all events: in average, the first patterns are detected around 24 days (i.e. 378 hours) before the peak window.

Another essential result emphasises the importance of the DENY label. In a nutshell, the approach captured patterns related to DENY or QUESTION, for only one true trending rumor event. In the case of the false trending event, DENY related patterns are identified in both the first matched window and 20 hours before the peak. The most frequent patterns within 20 hours before peak are $\{$*DENY* $\leftarrow$ *NEUTRAL* $\rightarrow$ *NEUTRAL*$\}$ (with frequency 9) and $\{$*QUESTION* $\leftarrow$ *QUESTION* $\rightarrow$ *QUESTION*$\}$ (with frequency 7). This result is consistent with previous findings and further indicates the importance of DENY patterns, previously overlooked by related research, as it contributes to a better recall of the early detection of rumors.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a method for detecting early signals of trending rumor events, in streaming social media. By analyzing snapshots of data stream contained within sliding windows, our method captures early patterns, integrating both users sentiment and network basic structures through which trending rumor events spread. The experiment results of two Twitter datasets are performed to validate the stability of our approach. In a nutshell, the first matched patterns can be identified much earlier than the peak window of trending rumor events in both datasets (24 days and 38 hours averagely earlier than the peak window respectively). Also, DENY-related patterns show an important role in distinguishing false trending rumor events in early stage, which was previously overlooked in related work.

To put our contribution into perspective, we provide a brief overview of how we plan to integrate this work into the

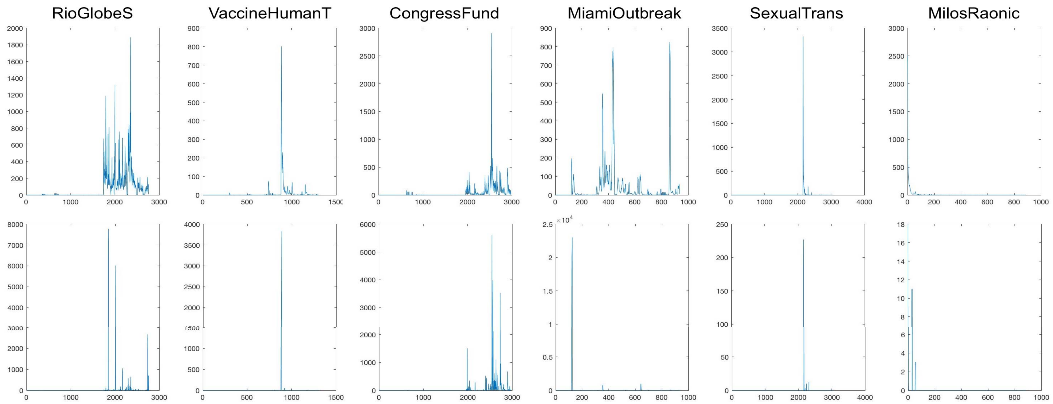

Fig. 5. Tweet frequency (up) and matched pattern frequency (down) in sliding windows of Zika events (the name corresponds with statements in Table I)

overall framework of detecting trending rumor event in real-time (which is part of on-going work). We model behaviour of the framework as a state-transition process, with three possible states: normal (the initial state of the framework), alarm and detected. The state transition is triggered by two decision mechanisms, that process streams of data and raise certain flags based on the observed features and patterns. The first decision mechanism is implemented by the early signal alarm: if this initial step detects the presence of early signals in trending rumor events, it triggers a state transition from normal to alarm state. While the framework is in alarm state, a second decision mechanism further investigates the flags risen by the early detection. Each candidate might transit into two states: the normal one, if there was not sufficient data or patterns for a reliable decision making; or in the detected state, if the framework is able to validate (with high accuracy) a trending rumor event as false. Using this early detection framework of trending rumor events, we expect to take immediate actions to restrict and stop its spread by limiting the audience it can reach or emphasising accurate information.

ACKNOWLEDGMENT

The research is partially supported by Japan Society for the Promotion of Science KAKENHI (Grant Number 25240048) and by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem' http://www.sobigdata.eu.

REFERENCES

[1] G. W. Allport and L. Postman, "The psychology of rumor." 1947.

[2] S. Wang and T. Terano, "Detecting rumor patterns in streaming social media," in *Intl Conf on Big Data*. IEEE, 2015, pp. 2709–2715.

[3] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng, "Rumor cascades." in *ICWSM*, 2014.

[4] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1589–1599.

[5] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Procs of the 20th Intl Conf on World wide web*. ACM, 2011, pp. 675–684.

[6] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Data Mining (ICDM), 2013 IEEE 13th Intl Conf on*. IEEE, 2013, pp. 1103–1108.

[7] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie, "Towards detecting rumours in social media," *arXiv preprint arXiv:1504.04712*, 2015.

[8] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *31st Intl Conf on Data Engineering*. IEEE, 2015.

[9] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Procs of the 24th ACM Intl Conf on Information and Knowledge Management*. ACM, 2015.

[10] E. M. M. S. H. S. Gayo-Avello, Panagiotis Takis Metaxas, D. Peter Gloor, C. Castillo, M. Mendoza, and B. Poblete, "Predicting information credibility in time-sensitive social media," *Internet Research*, vol. 23, no. 5, pp. 560–588, 2013.

[11] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Procs of the 24th ACM Intl Conf on Information and Knowledge Management*. ACM, 2015, pp. 1867–1870.

[12] Y. Qin, D. Wurzer, V. Lavrenko, and C. Tang, "Spotting rumors via novelty detection," *arXiv preprint arXiv:1611.06322*, 2016.

[13] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Procs of the 24th Intl. Conf. on World Wide Web*, 2015, pp. 1395–1405.

[14] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 1155–1158.

[15] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *Procs of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2002, pp. 1–16.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[17] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[18] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[19] B. Li, X. Si, M. R. Lyu, I. King, and E. Y. Chang, "Question identification on twitter," in *Procs of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 2477–2480.

[20] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" in *Procs of the 1st workshop on social media analytics*. ACM, 2010.

[21] C. Borgelt, "Simple algorithms for frequent item set mining," in *Advances in machine learning II*. Springer, 2010, pp. 351–369.