Visual Dashboard for Real-Time Analysis of Social Media

Huang Linya
Singapore Management University
linya.huang.2020@smu.edu.sg

Lim Jiahui
Singapore Management University
jiahui.lim.2020@smu.edu.sg

Zhang Ying
Singapore Management University
ying.zhang.2020@smu.edu.sg

ABSTRACT
This is the abstract.

It consists of two paragraphs.

1. INTRODUCTION
Detecting abnormal events ,such as disaster or crisis, from microblog social media has become a trend, as social media has played a pervasive role in the way people behave and think. Nowadays, people are also using time-stamped , geo -located data to share live information about what's happening in their surroundings, which enables the public, government and researches to sense abnormal events in community more quickly and take immediate actions.

To better analyzing and visualizing social media texts, several text analytics techniques can be applied, such as word-cloud, topic modeling, network analysis, geospail analysis and so on. In real-world practices, researchers has built various social media text visualization in different domain and lack of an integrated visualization.

In order to build a comprehend visualization dashboard with an interactive user interface, we buit the application based on R shiny - a web application framework to create interactive web applications - and text analytical R packages.

This paper reports our research and development effort to the real-time social media microblog analysis. It consists of XXX sections. Section 1 provides a general introduction of the paper, followed by a motivation and objectives of the paper. Section 3 provides a literature review of related analytical techniques and Section 4 provides the discussion of analytical methods applied in our analysis and development. Then we will discuss the user interface and application design and provides examples of analysis flow. Lastly, the paper concludes with consideration of future improvement work.

2. MOTIVATION AND OBJECTIVES
Streaming online social media can be used to study crime detection problems. Combining geospatial information and suspicious social media post can help prevent and track the potential crimes, which has been widely used in many institutions. Our research is motivated by the lack of integrated and comprehensive real-time social-media dashboard. We aim to apply appropriate text analytics method and visually driven data analysis techniques in R language and R shiny to provide a handy analytic tool to help users understand the social-media posting through various approaches,

1) Exploratory Data Analysis (EDA) and Time-series Analysis by basic statistical and world cloud visualization, which will provide a overview of content being discussed and help to highlight past events that occurred at certain areas;

2) Topic Modeling techniques by Latent Dirichlet Allocation algorithm will be performed to understand topics generated from text data. And topic trend and user engagement of each topic will be provided to understand the trend and public response to the topic;

3) Network Analysis will be performed based on the retweet relationship between users to discover influential authors. We will also analysis the various centrality methods of the network and their distributions to help users explore and identify social relationships, interactions, and communications;

4) Hexagon Binning Map discussed in research paper (Kam, BARSHIKAR, TAN 2012) will be applied to show real-time location-stamped text distribution in the community. By visualizing color with gradient, users can quickly locate the hexagon districts with the posts information and subsequently estimate the risk level in specific areas.

3. LITERATURE REVIEW
Word clouds can provide an overview by distilling text down to those words that appear with highest frequency. In the research paper Word Cloud Explorer (Florian etl 2014), researchers have demonstrated an integrated word cloud dashboard including basic word cloud view, co-occurrence highlighting, term series, information, search panel as well as part-of-speech and named-entities. The extension usage of basic word cloud visualization has been transformed into

No	Dataset	Information
1	Aliba.shp	Aliba geospatial map file with geometry information in linestring format and related location information.
2	csv-1700-1830, csv-1831-2000, csv-2001-2131	Call Center and Blog information during time period from 1700 to 2131 Abila time on January 23,2014, with 5 variables – date, author, message, hotspot location, latitude and longitude

Figure 1: Table on details of dataset

a powerful tool for text analytics. We will selectively apply the general word cloud exploration techniques into real time social media dashboard development. Topic modelling through Latent Dirichlet Allocation has been widely used in the industry for text mining. LDAvis (Chris etl 2014) is one of the most popular interactive web-based topic model visualization proving user interface of selection via number of topics and term and topic relevance. In consideration of the nature of social media streaming text, we will further analysis topic relationship with time series and users.

Network analysis has become increasingly popular to detect the interrelationships between actors of all sorts (Jesse, 2017). With the prevalence of social media, network analysis has expanded from real-life networks to the virtual networks in social media. In France and Christpoher's social network analysis of tweets during the Australian 2010-2011 (2011), they used tweets network to identify active players and their effectiveness in disseminating critical information via UCINet (Borgatti, Everett & Freeman 2002) and Pajek (Batafekj & Mrvar 1998). While the static network visualization failed to provide the users' names and their importance in the graph. To overcome this, we will use visNetwork package in R in our project to create an interactive network and options of various centrality measurements for the users to explore.

For spatiotemporal microblog visualization, some research has focused on visualization application on disaster events for public response, such as Junghoon et al. (2014), which provided an example of abnormal events detection through microblogs through spatiotemporal visualization, spatial analysis, interactive spatial decision support, temporal pattern and abnormal topic analysis. The heat map is useful in calculating the intensity of data points within the kernel but may have problems in handling large number of point data as discussed in paper DIVAD (Kam etal.2012) which suggested an alternative way by using hexagon binning map. The splitting the existing map by hexagon is an easy approach due to the lack of administrative boundary information.

4. BACKGROUND AND DATASETS

VAST Mini-Challenge 3 poses a social media and text analysis challenge to detect the meaningful event reports, evaluate risk level, public influence and suggest team of first response based on location-stamped posts. It provides the collection of microblogs and emergency calls from the event day and geospatial map of the corresponding areas, details as below:

5. ANALYTICAL METHODS

5.1 Data Preprocessing

After loading the related text files, we use textclean, tm, tidytext packages in R to perform standard text preprocess steps, including lower case transformation, punctuation and stop words removal, retweet and hashtags retrieval by using regular expression techniques.

5.2 Exploratory Data Analysis (EDA)

Temporal pattern distribution of the number of posts from various source such as microblog posts vs. call center messages, tweets vs. retweets given the specific time period. In addition, we will apply basic word cloud and co-occurrence plot to demonstrate straightforward text analysis based on its frequency and co-occurrence rate. With the basic line graph and word cloud, the visualization provides underlying insights and situation awareness overview.

5.3 Topic Modeling Analysis

We apply the LDA function from the topicmodels package to create a LDA model with target number of topics and hyperparameter for topic proportions. This function returns an object containing the full details of the model fitting, including how words are associated with topics and how topics are associated with documents. With this model, we further extract the per-topic-per-word probability "beta" and present the top n most common words based on this beta value for every topic. The second part of topic modeling is to show the trends of the topics over time by assigning each tweet a topic using the document-topic probability gamma. These trends imply the evolving change of topics with time, which useful to detect abnormal distribution peaks in certain time frames. Following by identifying the abnormal topics, we can investigate the tweets of those users most relevant to the topics via user engagement percentages.

5.4 Retweet Network Analysis

By extracting retweets from the original data set, we constructed nodes using the usernames (authors) and edges with directions from the person who retweeted to the person who was being retweeted. With this VisNetwork object, we will be able to plot out the retweets network and draw the histograms of centrality distribution graph. From the distribution graph, we can tell the patterns of the information dissemination. VisNetwork contains different centrality measurements and thus a variety of insights would be derived. Most popular and active users could be identified by in-degree and out-degree centrality. And most influential users can also be detected by closeness and eigenvector centrality.

5.5 Geospatial Analysis

We first plotted location-stamped microblog data on the leaflet map to visually display the social media contents. The location and messages received from call center were also mapped with the Abila shape file (via extraction and st_intersection of cross junctions) and plotted onto the map.

Next, we extracted hexagon binning map of the Aliba area by using MMQGIS grid layer function in QGIS application and exported the Aliba hexagon shape file. Subsequently, we identified the location-stamped posts intersection with the hexagon and counted the number of posts within each hexagon binning areas. By plotting the hexagon layer with color gradient over Aliba map, we can then visualize the

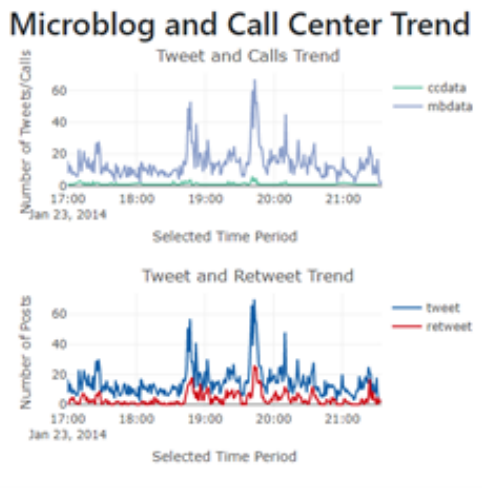

Figure 2: Overall trend of Microblogs and Call-Center messages across time

hexagon binning distribution map and identify "hotspots" area.

The plots are governed by time slider to show how the "hotspots" and messages changes across the evening. A more specific analysis can be done by selecting either the microblogs, call center or both.

6. USER INTERFACE AND APPLICATION DESIGN

7. RESULTS ANALYSIS AND DISCUSSION

The Shiny Dashboard was designed to incorporate interactivity and usability with different parameters for better analysis and understanding of the problem. In this section, we will focus on general analysis of events that occurred throughout the evening.

7.1 EDA and Topic Modeling

7.1.1 EDA

From initial exploratory analysis, we observe from the spikes that the frequency of microblogs and calls received are in synced. The number of retweets and tweets also corresponds to each other, most likely by people affected during that period of time.

The top trending hashtags observed are "#POKRally" and "#KronosStar", while "#DancingDolphinFire" and "#standoff" interestingly appears towards the end of the timeline. A deeper analysis of tagged user shows that top users such as "**???**" and "**???**", which denotes the official accounts of Abila media outlets that are frequently tagged or retweeted by public.

7.1.2 Topic Modeling

Topic Modeling using microblog data and with 10 topics of interest across the entire evening shows that the main events occurred were:

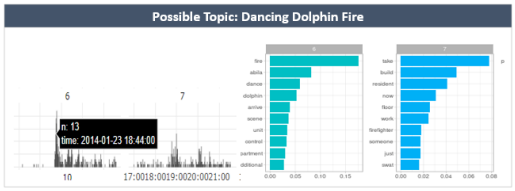

Figure 3: Topic Modeling analysis with Time

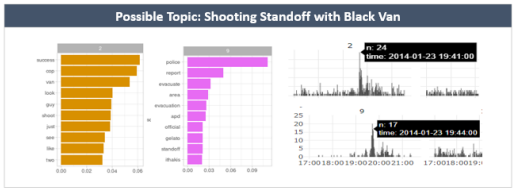

Figure 4: Topic Modeling analysis with Time

- Fire at Dancing Dolphin building (from 6:40pm to about 8pm)
- Black Van incident and Shooting Standoff at gelato galore (from 7:40pm onwards)
- POK Rally event (from 5pm to 6pm, but tweets about it last the entire evening)

A brief approximation of the starting time for these events are denoted with spikes in time-series analysis.

7.2 Network Analysis

7.3 Geospatial Analysis

Given the event estimated starting time, we will filter the time accordingly so as to examine the flow of event as well as response time for call center. The location mapping allows detailed evaluation on whether the tweet message and call message are within the same vicinity and then compared with the timing for each.

7.3.1 Event 1: POK Rally (around 5-6pm)

From the map, we noted on the estimated location of where POK Rally occurs. By observation, there are almost no calls received about the Rally within the vicinity and time range. This shows that the rally was most likely peaceful and not violent or chaotic.

7.3.2 Event 2: Fire at Dancing Dolphin building (around 6:40pm-8pm)

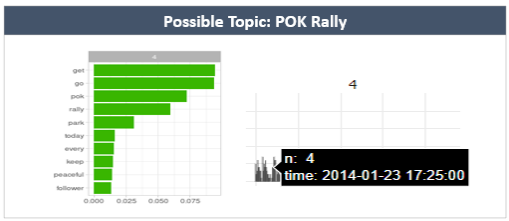

Figure 5: Topic Modeling analysis with Time

Figure 6: POK Rally Messages

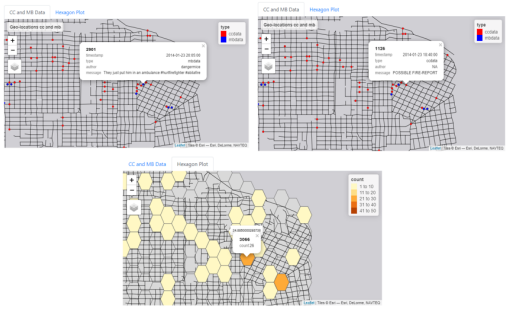

Figure 7: Microblog and Call messages for Fire Incident

We noted that the affected area has sudden increase in both microblogs and call messages received denoted by the hexagonal plot. Judging from the call message within the affected area, the first call received is at 6:40pm, and is in-synced with the estimated first microblog message related. Hence, the response time can be considered prompt, denoting the severity of the event.

8. CONCLUSION AND FUTURE WORK

9. REFERENCES