# Mixed-Initiative Visual Exploration of Social Media Text and Events - IEEE VAST Challenge 2021 MC3 Award for Strong Human-in-the-Loop Analysis Methodology

Lei Peng Fudan University Yuheng Zhao Fudan University Junting Gao Data Star Observatory

Yijie Hou Fudan University Jiaqi Dong Fudan University Qinghong Wang Data Star Observatory Ziyue Lin Fudan University Siqi Shen Fudan University Siming Chen\* Fudan University

Xingui Lai Fudan University

## ABSTRACT

In this paper we present an interactive visualization system designed for solving IEEE VAST Challenge 2021 Mini-Challenge 3. The system enables interactive classification and mining of social media texts, supports user-defined event analysis, and displays risk of public temporally and spatially. It features mixed-initiative visualization design and human-in-the-loop analysis process.

### **1** INTRODUCTION

IEEE VAST Challenge 2021 Mini-Challenge 3 requires participants to retrospectively detect and evaluate public risks in Abila City during the evening of January 23, 2014. The provided data are microblog records and emergency dispatch records from local police and fire departments. We see this challenge as a social media visual analytics task which places emphasize on risk detection and evaluation [2].

Generally, our visual analytics solution has two integrated interactive steps: microblog classification (step 1) and detailed analysis (step 2). The user may first start from an interactive classification and mining process of microblog records to identify normal records, spam and noises (step 1). After observing crucial information, he or she may then interactively explore user-defined risk events in the spatial temporal scenarios (step 2). During the whole analysis process, the user's knowledge of the situation deepens through the iterative exploration.

#### 2 VISUALIZATION SYSTEM DESIGN

Our visualization interface is mainly built with D3.js [1] and Vue.js. The system interface has a brushable timeline on the left (Fig. 1.A), a filter panel on the right (Fig. 1.C), and a scrollable central view which can switch between the message view (Fig. 1.B1), the entity graph (Fig. 1.B2), and the general analysis view (Fig. 1.B3).

**Timeline:** The timeline is a crucial part of our system. All the other views are linked to the timeline, and the timeline itself can be rendered based on message type, classification result, user-defined keywords, or sentiments.

**Message View:** In the message view, each category of messages is displayed with its corresponding word cloud. We use a combination of machine learning and human interaction to pre-classify relevant messages, chatter and spams. We first encode messages into tensors with the BERT [3] model and then use a lightGBM [4] model (which is trained with 150 manually labeled messages) as a classifier. Since there might be misclassifications, users can directly relabel a certain message by clicking on the relabel button. The two circles below the relabel button on each message card represent sentiment and objectivity of the message.

\*e-mail: simingchen@fudan.edu.cn



Figure 1: The user interface consisting of: (A) Timeline. (B1) Message View. (B2) Entity Graph. (B3) General Analysis View (C) Filter Panel with the event keyword tool, the white list and the black list.

Black List and White List: Users can also do batch operations with the help of the black list and the white list. For example, the user finds that @Clevvah4Evah sent lots of meaningless messages, and @KronosQuoth is a typical chicken soup blogger. Therefore, he or she can add them into the black list to prevent them from appearing in the relevant message set. If the user thinks #AbilaFire must be an important tag directly related to a public risk event, he or she may add it into the white list to avoid potential misclassifications from the model.

**Entity Graph:** From the raw texts, we extract entities such as authors, tags, organizations, locations, and persons to build an entity graph. Two entities are linked together if they both appear in one microblog record. The graph can provide us with event clues and entity relationships.

**Event Keyword Tool:** Having explored with the messages, the word clouds and the entity graph, users may then become curious about several keywords of different events that happen during this evening. Users can now input self-defined events with corresponding keywords into the event keyword tool. The system will render the timeline with the appearance of keywords, generate a bar chart race under in the general analysis view, and highlight every message card that contains a keyword (Fig. 2).

**Map:** We provide a map of Abila City marked with relevant information. The blue rectangles approximately represent the public places that we identify in Abila City. During a certain time period, we mark the locations of microblogs and emergency dispatches



Figure 2: The event keyword tool (A). Users can add self-defined events and corresponding keywords here. The timeline (B) will be rendered by those event keywords, and a bar chart race (C) of keyword appearances will be generated.

on the map. And the hexagons represent the estimated risk in the area. The higher the risk level is, the redder the hexagon will be coloured. To evaluate the risk to the public in a certain area during a certain period, we define two indexes: risk level and population influence rate. Assume that during period T, there are m microblog messages sent and *n* call center actions recorded within area S. Let Neg denote the average negative sentiment level of those m micro blog messages. Then, the risk level in area S during period T is computed as  $Risk(S,T) = m \times e^{Neg} + n \times C$ , where C is a constant. (In practice, we take C = 1.5). Here in practice, we always have  $e^{Neg} < C$ . That is to say, we value more on call center records than microblog messages because they are more directly related to public risk events. The population influence rate is the percentage of the population in Abila who might be affected by a certain risk event. Let N denote the aggregated number of microblog messages and call center records in area S. Let NT denote the number of all the messages and records that are generated during period T. Then, the percentage of the potentially affected population can be defined as  $P(S,T) = \frac{N}{NT}$ . When users hover on a hexagon, a tooltip will display all the useful information about this area including the risk level, population influence rate, negative sentiment level, and original records (Fig. 3).

#### **3** ILLUSTRATIVE CASE

Here is a case showing how a user may track and analyze a certain risk event using the system. After glancing through the word clouds and the entity graph, the user may become interested in the black van. The user may input 'van' as a keyword to the system. Then, he or she can track the appearances of keyword 'van' with the help of the timeline. Along the timeline, the user will first find a car hit, and then a bike hit caused by the black van. Then, the user will see that the black van began to flee in the city. By examing the messages and the map, users will find that the police got involved and stopped traffic so as to catch the van. Therefore, the user may add 'police' or 'cop' as a new keyword. At around 19:40, users will find a police officer got shot by the driver. The driver and the police started a



Figure 3: The map of Abila City marked with microblog and call center records (points), public places (rectangles in blue), and public risks (hexagons) during a certain time period.



Figure 4: Four event details the user can find related to the black van during exploration.

confrontation. Eventually, at 21:24 the driver surrendered and the missing GASTech employees were rescued from the van.

In an exploration loop, the message view and the entity graph provides refined information. The event keyword tool enables userinvolved exploration. The event timeline makes it clear when a certain event happens and how it evolves. And the general analysis view provides users rich and organized details for further analysis. Through the interaction with the whole system, users can keep discovering new interests, and finally get a complete picture of the risk events.

#### REFERENCES

- M. Bostock, V. Ogievetsky, and J. Heer. D<sup>3</sup> data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.
- [2] S. Chen, L. Lin, and X. Yuan. Social media visual analytics. In *Computer Graphics Forum*, vol. 36, pp. 563–587. Wiley Online Library, 2017.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* preprint arXiv:1810.04805, 2018.
- [4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:31486–31554, 2017.