VAST Challenge 2014 – The Kronos Incident - Mini-Challenge 3

Andrei Rukavina, Mariana Landoni, Paulina Verasay and Maria L. Traverso*
University of Buenos Aires, Argentina

ABSTRACT

Several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance. It is January 23, 2014, and the GAStech employees have been missing for three days. Mini-Challenge 3 poses a streaming analysis challenge. We had access to real-time feeds of microblogs and emergency calls to find the missing employees.

Keywords: VAST 2014 – Mini-Challlenge 3 – Streaming – Tableau – Raw - wordle

1 INTRODUCTION

In order to identify what is happening in the city of Abila and find clues into the disappearance of the GAStech employees we have access to a single data stream.

In our project we analyzed the streaming data using several analytics tools, (e.g. Tableau, Qlikview). In addition to the text analysis we performed text classification using the programming language Python.

2 DATA AND METHODS

The data was from January 23 and it was released in three segments:

- Segment 1 covered the time period from 17:00hs to 18:30hs.
- Segment 2 covered the time period from 18:30hs to 20:00hs
- Segment 3 covered the time period from 20:00hs to shortly after 21:30hs time.

The data stream included two types of records:

- Microblog messages (mbdata). The microblog messages that have been identified by automated filters as being potentially relevant to the ongoing incident. This posts used the @ symbol to designate a username within the body of a message. Hashtags (#) were used to relate the message to specific topics, and "RT" at the start of a message indicated that the current user was re-sending another user's message. Spam and junk messages were also common.
- Call center data (ccdata). Text transcripts of emergency dispatches by the Abila, Kronos local police and fire departments.-

The records have the following attributes:

- Type: mbdata or ccdata.
- Date: date and time when the message was sent.
- Author: user that send the message.
- Message: the message string.
- Latitude: latitude from which the message was sent.
- Longitude: longitude from which the message was sent.
- Location: street at which the incident took place.

We first identified and excluded spam and junk messages by analyzing the number of messages by author. We counted how many messages contained the exact same text for the same user. To quantify text repetitions, we built a spam index by dividing the number of messages with different content sent by a user over the total number of messages sent by that user. A low spam index value indicated many messages with the same text and thus is associated with a high likelihood of spam.

To determine the unfolding events we studied the number of messages sent in different time frames (per second, per minute, in 10 and 15 minutes). If there were peaks of messages in these time frames, we used the hashtags in the messages to identify the different events, and their approximate starting and ending times.

We performed text analysis using a Python lib called TextBlob[1]. First we grouped the microblog data to merge records with their content. After that we executed a sentimental analysis for each message. The results were two different measures defining the sentiments of each microblog text: Polarity and Sentiment.

Finally we analyzed the authors that had a major degree of participation in one of the events using a Tree map. To complement the previous analysis, we created a Word Cloud with the messages of these authors, in order to detect other important participants.

3 VISUAL ANALYSIS

We detected four distinct events by examining the patterns of repeated hashtags: rally of the Protectors of Kronos (POK) in Abila City Park; fire at the Dancing Dolphin apartments; a black van hit and run; and shooting at Gelato Galore ice cream parlor.

The visualization of these events (Built using Tableau), are shown in Figure 1.

To complement the previous visualization we prepared a Stream graph using the open web app RAW, as shown in Figure 2.

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

* email: rukavina.andrei@gmail.com, mariana.landoni@gmail.com, pauliverasay@gmail.com, lautraverso@gmail.com

[1] **TextBlob** is a Simplified Text Analysis library. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more

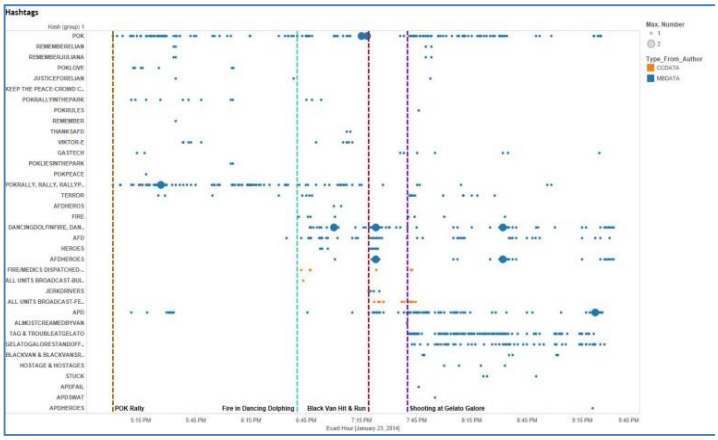

***Figure 1**: Timeline of events*

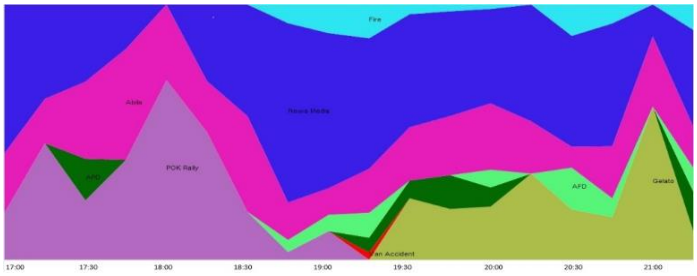

***Figure 2:** Stream graph*

Adding the sentiment analysis to the time series of the messages we generated what we called The Mood Analysis Graph. The subjectivity (grey line) and polarity feelings (green and red bars) are displayed using a dual axis against time as shown in Figure 3.

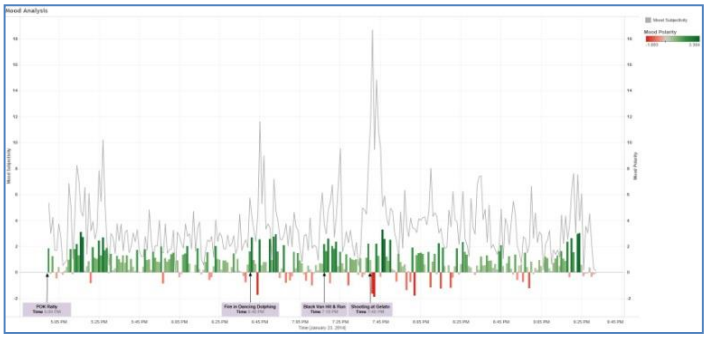

***Figure 3:** Mood Analysis*

The grey peaks show high subjectivity in the messages, the highest peak took place when the shooting at Gelato Galore started. The other peaks occurred during the speeches at the rally, when the fire started at the Dancing Dolphin apartments and during the hit and run incident.

Over all, the messages show positive feelings. The strongest negative feelings appeared when the fire and shooting started, consistently with the fear expressed in the texts.

Based on these analysis, we considered the shooting at Gelato Galore, to be the most likely event that provided additional clues to the investigation of the GAStech disappearances.

We analyzed the authors that had a major degree of participation in this event using a Tree map. We created a Word Cloud using Wordle™, with the messages of these authors, in order to identify the words that had more frequency.

As shown in Figure 4, we can see that the words which have a bigger size give an idea of the place (TAG-Trouble At Gelato) and the event's participants (Van, Police, APD-Abila Police Department, Cop, Guy, Terrorist, Hostages), the middle size words give an idea of the situations (Shooting, Fire, Terror, Standoff).

***Figure 4:** Word Cloud*

Lastly, applying Shneiderman's mantra for visual data analysis, (overview first, zoom and filter, then details-on-demand) we read the messages of the important participants in order to figure out what was happening at Gelato Galore parlor.

4 CONCLUSION

The presented approach is a combination of visual and interactive analysis. The visual analysis has proven to be useful to determine different events by their time frame and the mood of what was going on, however to find out about the missing employees it was necessary to read some of the messages.

The visual exploration has been enhanced by the Timeline of events, the Stream graph and the Sentiment graph to detect events. The Tree map and the Word Cloud helped us to identify the important participants in the "Shooting at Gelato Galore" event.

Since we considered the shooting at Gelato Galore to be the event that provided additional clues to the investigation, we focused our analysis in these messages. We found out that inside the black van involved in the event, there were two kidnapped women, possible GAStech employees. With this information we searched into the other events and found out that someone called "Rachel" had been missing for several days.

After the analysis of the background documents from Challenge 1, we concluded that one of the hostages is Rachel Pantanal, Executive Assistant of GAStech CIO and that Isia Vann is one of the kidnappers since we found out that he send emails to Rachel that we considered were harassment towards her.

REFERENCES

[1] Vast 2014 Mini Challenge 3, The Kronos Incident: http://vacommunity.org/VAST+Challenge+2014

[2] Ben Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336-343, Washington. IEEE Computer Society Press, 1996.

[3] Python™: https://www.python.org

[4] TextBlob for Python: https://textblob.readthedocs.org/en/dev/index.htmld

[5] Wordle™: http://www.wordle.net

[6] RAW: http://raw.densitydesign.org

[7] QlikView software: http://www.qlik.com

[8] Tableu software: http://www.tableausoftware.com