Social Media Text Streaming Visualization

Martin Pustilnik[1,2]

[1] Universidad Nacional General Sarmiento mpustilnik@campus.ungs.edu.ar
[2] Universidad de Buenos Aires
https://www.ungs.edu.ar

Abstract. Social media registers millions of messages per second. This paper aims to develop a visual analysis that allows identifying events of particular interest and the time intervals in which they occur. This analysis uses text mining techniques and sentiment analysis.

Keywords: Social media · text mining · event identification.

1 Introduction

Social networks present insight into particular society problems. For a given instant, and especially when an important event occurs, social media users comment their particular vision over the event. In this context, is not possible to show every message for a given moment during an event, as information flux is big and it is biased.

It is sensible to find some way to display this event considering a general trend. In this paper, we propose a visual metaphor that allows to generalize this problem using data mining techniques.

2 GENERAL IDEA

Whenever social media information is analyzed, it is imperative to determine whether this information will be arriving in real-time or if it is a static collection that should be considered as a whole. In this study, we aim to treat information as a dynamic stream because most event detection scenarios are for emergencies. This visual metaphor might be useful in static scenarios analysis as well, but some caveats should be previously sorted out. This proposal also preserves integrity criteria presented by Tufte [1] and by Ware [2].

As it is not possible to show individual message information, exploratory research should be done first to decide the best way to group messages. Then, once message aggregation is chosen, we proceed to obtain the variables that are of interest in the problem. Besides, it is important to obtain the time window that allows an operator to decide on data. Social media texts are noisy when it comes to spelling and grammar correctness, even more if you consider the use of emoji or slang. Therefore, data should be pre-processed to discard and correct

terms used. Many different techniques could be applied for this, but these are not in the scope of this paper.

The following interactive visualization `shorturl.at/exCH2` was generated as an example.

2.1 Categories

Keyword lists are chosen before grouping them in the categories of interest that are going to be visualized from the data stream. These keywords could be chosen manually or gotten by mining common corpus of social media datasets. The visualization technique is not affected by the words in each category or how each list is obtained. It is important to note that the number of categories should be kept low so as to let a human operator understand them. Then, to localize events related to each category it is needed to perform some text mining and find patterns that imply some correlation between words in each category and data. This mining must convert text information into numbers to be displayed and could be as simple as counting keywords in each text.

Once gathered quantity values for each category, to allow event detection we propose to obtain another variable from the variation in the number of messages for each group between consecutive hours. In this way, a threshold can be defined so this variation is shown altogether with the category value.

2.2 Helping to lower false-positive rate

Written text in social media has an underlying feature associated with the way it is expressed. This feature has a close relation with how the writer felt when the message was published. To quantify this sentiment, we propose to make a positive vs. negative message classification taking account the nuance between them and mapping to the range $[-1; 1]$ considering -1 as totally negative and 1 as totally positive. Then, we take the average sentiment in the time window being analyzed. This sentiment variable helps us to filter false positive values that may occur while looking at categories: if we have some peak in a category but mean sentiment remains unchanged, it is likely to be an unrelated event.

2.3 Message summary

As numeric information is extracted from text, it could potentially hide what it is being said. Therefore, it is fundamental to show which are the most important words used. These words may include contextual data that cannot be automatically retrieved but are useful for operators to understand what is happening. Also, these words should be sorted according to importance.

2.4 Visualization details

To visualize categories information we proceed to put together a composed chart as shown in figure 1: above a multiple line chart, one line per category that shows

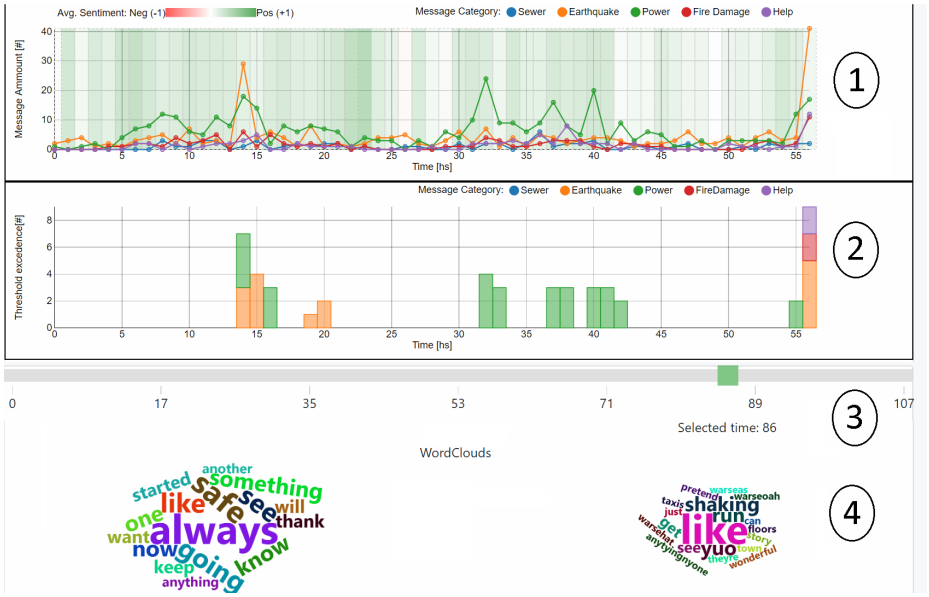

Fig. 1. Overall view. (1) Sentiment Analysis & Categories measurements. (2) Categories difference threshold. (3) Hour slide bar. (4) Top two categories World Cloud's

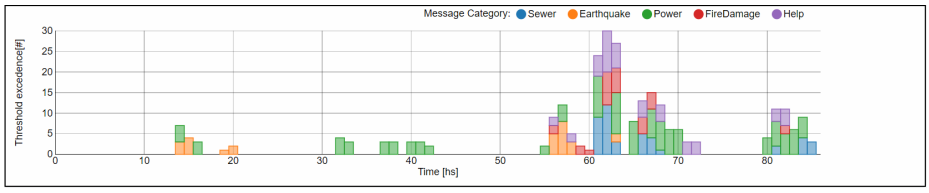

Fig. 2. Categories difference

how many messages have been detected in the time window and below a stacked bar plot that shows categories that have surpassed the threshold for hourly change (see figure 2. All categories can be toggled to be shown or not, enabling to analyze specific categories.

To visualize the sentiment information we proceed to map positive values to green values and negative values into red ones. The nuance is mapped to an alpha channel value and it is shown in the background of the message count chart to express the idea that is underlying data (see figure 3).

Finally, to visualize the summary we proceed to show a classic word cloud as it is easily understood and well known for a wide audience. The clouds can be associated with other data like GIS and category.

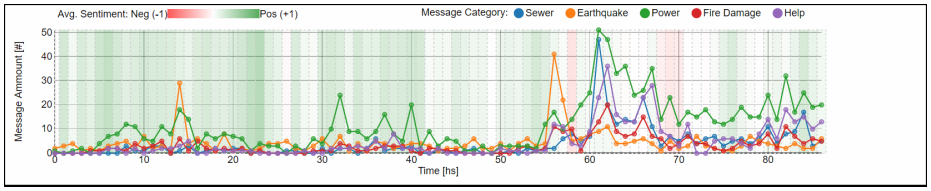

Fig. 3. Categories quantity

3 Example

To show a working example of this technique we choose data from VAST Mini Challenge 3 [3]. This challenge required to determine the type of problems that are occurring across the fictional city of "St. Himark" and to advise the City on how to prioritize the distribution of resources in natural disaster events. The data source in VAST Mini Challenge 3 consists of text messages from a made-up local social network ("Y*INT") of the "St.Himark" community.

3.1 Dataset

The dataset consists of 41,868 plain text messages in 108 consecutive hours ranging from April 6, 2020, to April 12, 2020. Each message also has information about the location, timestamp, and account. It is known that there is a major earthquake sometime within the recorded messages. There is an average of 387 messages/hour and for the sake of simplicity in this example, discretization is hourly.

3.2 Tasks and Questions

Lists of 20 manually pre-selected keywords were separated in five groups of categories for analysis: Sewer Repair Needed, Earthquake felt, Power Outage, Fire Damage, Help. Thresholds were selected to show categories only if more than five messages were detected. With the visualization in 1 conditions across the city were characterized for recommendation on how resources should be allocated.

We can successfully identify the time of the earthquake using techniques described in sections 2.2 and 2.1 and detect locations affected using data provided from message summary from section 2.3. We can successfully take the pulse of the community and detect how has the earthquake affected life in the city and what is the community experiencing using data from 2.2 and 2.3.

The scroll bar suggested in section 2.4 is used to simulate a data flow or a static compilation according to the selected value.

4 Discussion

4.1 Detecting mood changes

With the aforementioned approach, we can represent mood changes at any time and we could measure either massive events or more subtle ones given the correct keywords are chosen. Adding sentiment analysis and mapping the results to easy reading values are valuable resources for decision making. Further investigation should be done to evaluate whether it is valuable to add a different kind of sentiment detection.

4.2 Thresholds

Section 2.1 shows only significant events. This result is achieved by selecting arbitrary database dependent thresholds. These thresholds may be chosen from previous knowledge of the data or an automatic technique could be developed to detect outliers that should be shown.

4.3 Static collection or dynamic stream

In this visualization, we emulate arriving information with a sliding bar that changes time. This approach is meant to show how this technique responds to variable data but generates a considerable distortion when there is not enough data input. An alternative to "first hour's distortion problem" is to use variable size time windows that, if necessary, could show a wide range or a shorter range but acknowledging information loss. Pre-trained algorithms and general usage corpus may be used to sort out this problem.

Analyzing data as a static collection may have better Precision and Recall detecting events in forensics analysis but is not helpful in real life scenarios.

5 Conclusion

This technique shows successful event detection and therefore provides effective help in real-time decision making for resource allocation.

This visualization was tested only for VAST Mini Challenge 3, but it can be easily adapted to other event types like traffic incidents and networking outages. A possible future work could be testing this technique over other social media steaming, like Twitter® or Instagram®.

References

1. R. Tufte: The Visual Display of Quantitative Information. In: GraphicsPress, Cheshire, Conn., 2nd ed. ed., 2001.
2. C. Ware: Information Visualization: Perception for Design. MorganKaufmann Publishers Inc., San Francisco, 2nded., 2004. https://doi.org/10.1016/B978-155860819-1/50001-7
3. IEEEVIS. Vast-challenge 2019, mini-challenge 3: Voice from the people,2019., https://vast-challenge.github.io/2019/MC3.html.