ABSTRACT

We present City Sentinel, an in-house built visual analytic software capable of handling a large collection of textual documents by combining diverse text mining and visualization tools. We applied this tool for the Vast Challenge 2011, Mini Challenge 1 over millions of tweet messages. We demonstrate how City Sentinel aided the analyst in retrieving the hidden information from the tweet messages to analyze and locate a hypothetical epidemic outbreak.

Index Terms: H.1.2 [Models and Principles]: User/Machine Systems—Human information processing; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 DATA MINING OVER TWEET SETS

In this paper we describe the City Sentinel Epidemic Visualizer, an in-house-built tool developed for the Challenge. The tool deploys text classification and clustering based on the tf.idf representation and cosine similarity (see e.g. [1]) of the tweets to help the analyst in defining tweet sets that can then be visualized along the geographic and temporal dimensions as movies.

2 OPERATIONS OVER TWEET SETS

The central concept of our application is a Tweet Set that can be defined by both tweet text and metadata to capture important tweets posted at the time and in the proximity of certain event. Initial tweet sets are primarily defined by clustering (Section 3) and using filters for keywords, square areas and time intervals. Active learning (Section 4) was crucial in cleansing and defining the final sets representing the key events in the data. Finally we could also manipulate the tweet sets by the following operations: closure for sets representing the key events in the data. Finally we could also manipulate the tweet sets by the following operations: closure for sets representing the key events in the data.

3 TWEET TRENDS BY FILTERED CLUSTERING

To find the cause of the epidemic outbreak, we required an overview of the events happened in Vastopolis based on trends appearing in the microblog data. In order to identify trends, we selected relatively small time intervals of a few hours or days and clustered all the tweets at that time by k-means (see e.g. [3]). The output was automatically filtered if a cluster contained keywords with significantly higher frequency than in the whole data set, the respective cluster was presented to the user along with its keywords. The filtered clusters represented both epidemic symptoms as well as important a priori unknown events such as a technology convention and truck accidents.

4 CLASSIFICATION WITH ACTIVE LEARNING

Active learning (for a survey see e.g. [2]) was used to cleanse the tweet sets and define coherent collections of events for visualization. For example, by clustering one can easily identify “flu” as a key term. However a large fraction of the messages containing “flu” are otherwise irrelevant for the epidemic outbreak and we would like to remove noise such as the “Fried Chicken Flu” episode.

Our active learning procedure starts out with a tweet set to be cleansed. From this we randomly selected a small subset of approximately 30 tweets and asked the user to manually annotate it (see Figure 1). Next, in iterations, new models were built and new lists were presented as long as the user became satisfied with the result. We used the k-NN classifier (see e.g. [3]).

5 VISUAL INTEGRATION

City Sentinel helps the analyst by providing a simple user interface to define tweet sets. Using of City Sentinel is quite straightforward after understanding its main concept, the manipulation and visualization of tweet sets. To start the investigation, the analyst can ask for a clue by using clustering, or directly specify keywords. Then, manipulating the message sets can be managed on a simple GUI. Finally, visualization by time and location helps the user to understand the information obtained by the process.

The application has two main panels: the Studio and the Map. The Studio lets the user specify a screenplay of tweet sets, each represented by a selected color. New tweet sets can be obtained from existing ones by applying the set operations (filter, union, etc.). The user can then visualize the tweets in the Map panel.

We developed various ways to visualize the text of a selected tweet set. The messages can be displayed in a list, their distribution in time can be viewed in a bar chart, and their most important words can be represented by a word cloud.

Spatio-temporal visualization of a tweet set is implemented over the Map panel as a movie, it can be played showing the tweets as colored dots at their exact location, and a dynamic word cloud showing the most important words that were tweeted at the given time (see Figure 2). While playing a movie, the weather data including wind direction is also shown at the top of the map.

6 ROLE OF CITY SENTINEL IN FINDING THE ANSWER FOR THE QUESTIONS OF THE CHALLENGE

To identify the source of the epidemic outbreak, first we clustered tweets in one-day intervals. Several clusters represented events including a truck accident at a bridge, or people with certain symptoms such as “vomit”, “diarrhea”, “flu”, “pneumonia”, “flem”, “blood”, “shortness of breath”.

We found several suspicious events which might be related to an epidemic outbreak: for example, an explosion in Smogtown, a technology convention and baseball matches. These events can easily be visualized in the Map panel, with a dynamic word cloud providing feedback on the quality of the current tweet sets.

Next we filtered for the previously found keywords of symptoms. Some of the sets were still noisy, for example filtering for “blood”
resulted tweet set containing tweets about “True Blood”. This problem can be detected by observing the word cloud: one can realize irrelevant frequent words “true” and “Sookie”, both relating to the TV Series. Other challenging problem was to separate a tweet of a sick person from messages about a friend’s illness, a key distinction for the usability of the location of the tweeter for tracking the epidemic spread. We applied active learning to generate clean tweet sets of key events and symptoms. It took about ten minutes for an analyst to train a classifier.

By screenplays of the cleansed tweet sets, we managed to compete our hypotheses about the possible causes of the outbreak. The most probable screenplays show two main epidemic spread right after a truck accident. We could see that abdomen-related symptoms mainly appear in the south-west part of Vastopolis, east from the accident, corresponding to the current wind direction, while lungs-related symptoms show up at first only in Downtown and Eastside, down by the river flow (See Figure 2). Hence wind and river may be the two carriers of the epidemic that outbroke at the time of the accident.

7 CONCLUSION AND FUTURE WORK

We have developed an interactive, flexible, easy-to-use interface to spatio-temporal text mining algorithms aided with visualization. With this tool, an analyst can find useful pieces of information within a large dataset with relative ease.

City Sentinel is a general tool in the sense that it handles records with text message, geographical coordinate, and a time label. Domain specific knowledge used when solving the Mini Challenge 1 came solely from the user interactions.

REFERENCES