DV Project Report - 2022

CrystalBall: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data

Bharath Kumar Bandaru
Arizona State University
bbandaru@asu.edu

Karthik Chowdhary Nelapati
Arizona State University
Knelapa1@asu.edu

Keerthana Golusula
Arizona State University
kgolusul@asu.edu

Nikhil Sagar Miriyala
Arizona State University
nmiriya1@asu.edu

Sandhya Balaji
Arizona State University
sbalaj17@asu.edu

Tanisha Tembhre
Arizona State University
ttembhre@asu.edu

INTRODUCTION

In this project, our goal was to implement the paper Crystal-Ball: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data. In today's world, we know that social media provides us with a lot of useful data. Social media posts provide information about the volume and scope of conversations about different events, the duration of those conversations, people's feelings about those events, and how those feelings evolve over time. People have started to use social media to plan, coordinate, publicize, and educate others about upcoming events. Examples include concerts and sporting events, social movements/campaigns like Bank Transfer Day and the Occupy Movement, as well as potential event cancellations or changes brought on by things like bad weather. Future events can be predicted and even prepared for by gathering, evaluating, and visually presenting information about them.[1]

This is exactly what we intend to do in our project. Our main objective was to use Twitter data to analyze future events. We used the tweets data to study, analyze and visualize future events based on a given date or location. The target audience for this visual analytics system includes city authorities, members of the emergency services, planners, facility managers, business owners, and anybody else who would be interested in taking part in and contributing to local events. The visualizations can be used by different people to explore different things like when the events would occur, where are the events going to occur, what are the tweets posted about these events, what are the events about, to name a few. This system can empower individuals to explore efficiently and be prepared for upcoming events.

VISUAL DESIGN

Our entire interface is divided into 6 panels. The first panel is called the Events Calendar view. The second panel is called the Map View. The third panel is the Word Cloud View. The fourth panel is called the Social Network View. The fifth panel is known as the Tweets Panel View. The last panel is our extension to the interface called the Events Category View.

All the panels are shown in the figure below.
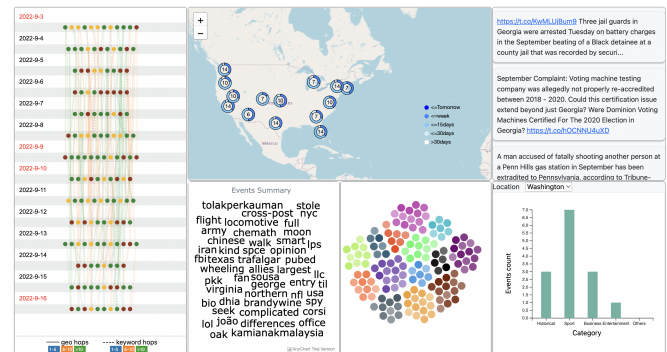
Each of these panels are described in detail as follows:

Events Calendar View

The events calendar view summarizes events happening in the future based on the dates on which they will occur. Each date shows a row of circles, where each circle represents an event. The weekend dates are marked as red. The color of the circle represents the emotion associated with the event, which is deciphered based on the content in the related tweets. The events occurring at the same location are connected by a solid line, while the events sharing the same keywords are connected by a dotted line. Multiple events can be connected by solid or dotted lines. The solid lines and dotted lines are colored based on the number of geo hops and keyword hops respectively as shown in the legend in the panel. Geo hops are the number of lines connecting events having the same location, while keyword hops are the number of lines connecting events having the same keywords. Hovering over the solid lines and events shows the location associated with them. Hovering over dotted lines shows a tooltip which displays the common keywords between the events that the lines connect. Events calendar view is highly coordinated with other views as well. Clicking on the solid lines refreshes the word cloud to show all the words associated with the events that the solid line connects. It also updates the tweets panel view to show the tweets related to that location. Clicking on events updates

the tweet panel to show tweets related to that event. Hovering over an event highlights the associated keywords in red in the word cloud. Hovering over a solid line highlights the donut chart in map view and events in social network view, for the location associated with the solid line. Clicking on a date updates the word cloud to show all the words associated with events occurring on that date.

Map View
The map view summarizes the events based on the location they will occur in. The geographic location of future events can be seen in the map view. The view has a zoom in and zoom out feature which can be used to see event detail and events overview respectively. At the zoomed out level, the donut chart at different locations shows the number of events occurring at that location and the. The color assignment of the segments of the donut chart represents how far into the future the event is going to take place. It ranges from dark blue (tomorrow) to off-white (more than a month away). When the donut chart is clicked, it zooms in to show the details of all the events. Hovering over an event shows its location and date. Map view is highly coordinated with other views. Clicking on the donut chart refreshes the word cloud to show all the keywords associated with events of that location. It also updates the tweet panel to show all tweets related to that location. Hovering over the donut chart highlights all the events associated with that location in the event calendar view and social network view. Clicking on an event in the zoomed in view updates the tweet panel to show the tweets related to that event.

Word Cloud View
The Word Cloud view is used to summarize keywords related to all the future events. In the very beginning, when the interface is loaded, the word cloud view displays all the major keywords related to all events. Size of the keywords indicate the frequency of the keywords. This view mainly helps the users to read what the events are about without having to actually read the tweets itself. This view is highly interactive and coordinated with the other views too. Following are the interactions of the word cloud view: When the user clicks on a date in the Events Calendar view, the word cloud updates the keywords to summarize all the events that are going to occur on the selected date. When the user clicks on an event in the Events Calendar view or in the Social Network view, the wordcloud highlights all the keywords that are related to the selected event. When the user clicks on a given location in the Map view, the word cloud updates to show all the keywords related to the events that are going to occur in the selected location.

Social Network View
The Social Network view is used to display the events information based on the tweet count. In this view, each node represents an event and the nodes that are color-coded based on their location of occurrence. When one hovers over the node, we display a tooltip which shows the number of tweets that have been made for that given event. This view therefore, is highly useful to the user to determine the popularity of a given event based on the tweet count. This view is also interactive: when the user clicks on a given node/event in this view, the words pertaining to that selected node are highlighted in the word cloud view.

Tweets Panel View
This view is useful for getting detailed information about the events. It shows all the tweets related to the events. It updates according to the selection done in other views. Clicking on the solid lines in the event calendar view or donut chart in the map view updates the tweet panel view to show the aggregated tweets related to the location for the selection. Similarly, clicking on the events in the event calendar view and map view shows all the tweets related to that individual event in the tweets panel. These can be used to make sense of the events and may contain external links which can give more information about the event.

Extension - Events Category View
This view summarizes the events according to their categories. The events are divided into multiple categories such as historical, sports, business, entertainment and others. The panel provides a dropdown which is a list of locations to choose from. Clicking on the location in the dropdown updates the bar chart to show the number of events occurring for each category for the selected location. This view can enable the user to find the location which has more events of their interest and then use other views to get more information about those events.

DATA DESCRIPTION
- Since the visualization entirely focuses on identifying and visualizing future events based on tweets from the users, we have utilized 'tweetPy', which is a python library provided by Twitter that can be utilized for extracting tweets from twitter based on the query of our choice. In order to use the tweetPy API, we need to create a twitter developer account and generate API keys for the project being worked on.

- The query used for extracting the tweets using the API, query=<Location September>, extracts the tweets which consist of both the location and September as part of the tweet. Additionally, to ignore the duplicate tweets, a filter provided by Twitter, '-filter:retweets', was used.

- For the purpose of our project, we have considered 14 different states within the United States, to make the visualization look cleaner.

- Once the tweets were extracted based on each location, 'dateFinder' library within Python was utilized to filter and identify the dates within the tweets.

- For each location, the tweets corresponding to a particular date, identified by 'dateFinder', were grouped to convert the tweets into events.

- In order to visualize the word cloud with respect to the most frequent words in the tweets, and to predict the sentiment and category of an event, data clean up processes were performed.

- Each of the tweets was cleaned up, using data pre-processing techniques, such as stop word removal, lemmatization, tokenization and stemming.

- Post tokenization, keywords for each of the events were identified before stemming was performed using 'Yake' python library.

- To identify common keywords based on the event keywords, the Union-Find algorithm was applied at the event level.

- A pre-trained Bert model from Kaggle was utilized to perform sentiment analysis prediction.

- Similarly, a pre-trained model from Kaggle, which applied Logistic Regression technique, was utilized to perform category prediction.

- Finally, all the required information, such as the event location, date, keywords, common keywords, sentiment and category, was organized into multiple tables (Events, Keywords and Common Keywords) in order to work with the visualization with ease.

CASE STUDIES

In this section, we will report different kinds of findings users can achieve from the visualizations in our project.

- *Case study for exploring memorial services in remembrance of 9/11 attacks*

  The data we collected predicted events for the month of September. Since this month happens to have the 9/11 anniversary, we saw that there were quite some tweets related to the attacks. Events such as commemoration events, ceremonies, tributes and memorial services were organized in order to remember the ones who lost their lives in the attacks. The emotion related to these events are rightly captured as negative in the Events Calendar view as the tweets express sadness, shock, fear or anger. A user can easily decipher the nature of the events by looking at the color in Events Calendar View and click on a specific event to get more details or information pertaining to it. If a user wants to find out memorial services that they can attend in their location of interest, they can easily do so by exploring the Map View.
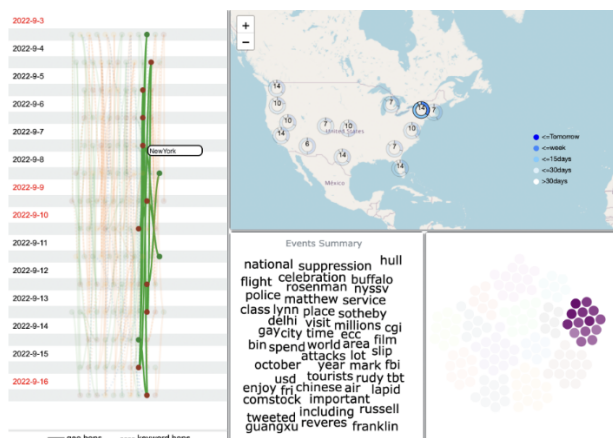
- *Case study for enabling a user to explore events of interest in a new place*

  In this case study, we use CrystalBall to explore events from the perspective of a user visiting a new place, for a short duration. The system can be provide a lot of useful information if one has limited time and needs an overview of events nearby. It further allows them to filter out the events they are interested in attending. As can be seen in the figure below, a user can choose the map view as a starting point for their exploration. For a given place, say Washington, clicking on the donut chart will provide the user with all the information about different events occurring there. This not only gives an overview of events but zooms into another radial chart which contains details of individual events. Clicking on these events allows user to see summary, tweets and popularity of the event by seeing its interactions in other views. This makes it very convenient for the user to plan their itinerary in a limited duration.
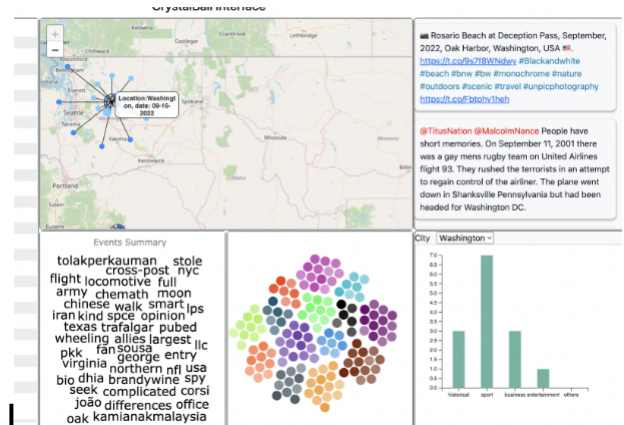
DISCUSSIONS

Lessons learnt:

- We got to work on various new visualizations apart from the ones in the assignments. We got to learn how to apply interactions better between these new kind of visualizations. It was a challenging yet an enriching experience to work in a team, understand each others' codes and make the visual interactions seamless.

- We got a chance to work with machine learning models and to learn the intricacies of data transformation to derive and visualize insightful information.

- We got to appreciate the working of CrystalBall system and understand its utility in detail while we implemented it and added our own extensions.

- Through this project, we were able to design and develop a highly interactive visualization system on our own. We learnt how powerful visualizations can be to convey information in an effective manner to the user. This project also helped us apply the concepts we learnt in class and enhanced our understanding of those concepts.

Potential improvements:

- The system can also use blogs and articles to predict events in conjunction with tweets. Blogs and articles from reputable sources can be much more reliable in predicting future events.

- Twitter can be home to some very hostile and in-appropriate material . We can show moderated tweets in the tweet panel which is devoid of any insensitive or offensive content. Of course, this kind of moderation can be situational but when we consider professional events like tech-meetups and tech-conventions, it is better to only use moderated tweets to be shown the tweet panel.

- The above argument can also be applied to the word cloud and can prevent any appearance of in-appropriate words in the word cloud.

ACKNOWLEDGEMENT

REFERENCES

[1] I. Cho, R. Wesslen, S. Volkova, W. Ribarsky and W. Dou, "CrystalBall: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data," 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), 2017, pp. 25-35, doi: 10.1109/VAST.2017.8585658.
[2] https://docs.anychart.com/BasicCharts/TagCloud
[3] https://leafletjs.com/examples/quick-start/
[4] https://www.latlong.net/category/states-236-14.html
[5] https://www.kaggle.com/code/lbronchal/sentiment-analysis-with-svm
[6] https://www.kaggle.com/code/khoulaalkharusi/semi-supervised-document-classification
[7] https://d3-graph-gallery.com/graph/shape.html
[8] http://using-d3js.com/0508links.html
[9] https://bl.ocks.org/bumbeishvili/63841a5dcf018b90