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

Sai Likhith Yadav Seera
Computer Science
School of Computing and Augmented Intelligence
Arizona State University
Tempe, United States of America
sseera@asu.edu

***Abstract*—Social media is one of the most popular platforms for people to share and communicate their ideas and opinions on different topics from the past, present, and future. One of the common topics is discussing and informing others about future events. Using this concept, we created our project to predict future events using Twitter data.**

***Index Terms*—Twitter data, future events, analysis, prediction, and views**

I. Introduction

In our project, we implement Crystalball, an interactiv visual analysis system that identifies and ranks future event from Twitter data. We used seven parameters, such a timestamps, location coordinates, hashtags, links shared by user in their tweet, and information in the tweets, to identif and characterize future events using tweet informatio We have implemented five visualizations to describe th characteristics of future events derived from tweet data. W have included a case study analyzing future events derive from Twitter data.

II. Description of Solution

For our project, we have used a system pipeline [1] that describes the Crystal ball implementation as shown in figure 1.

The Twitter Streaming API is used to collect and analyze data online. This API pulls thousands of streaming tweets from around the world every day. These tweets are then stored in MongoDB and utilized as required. Then, a stream of tweets is sent for entity extraction, which identifies the provided tweet's location, date, and geocodes. Algorithms such as Stanford SUTIME, TweetNLP, and OpenStreetMap Nominatim are used to extract and map the date and time stamp, location, and geocode of a streaming tweet. The entity extraction section identifies the location and time reference of the tweet stream. These discovered times and dates are then assigned a calendar date. If a tweet contains references to future times, dates, or locations, it is saved in the database for further processing. If not, the tweets are not considered for further analysis. Further, these extracted tweets will be utilized for future event identification. These tweets about future events account for a small portion of all processed tweets. As a result, we extract small future event signals from large amounts of noisy data.

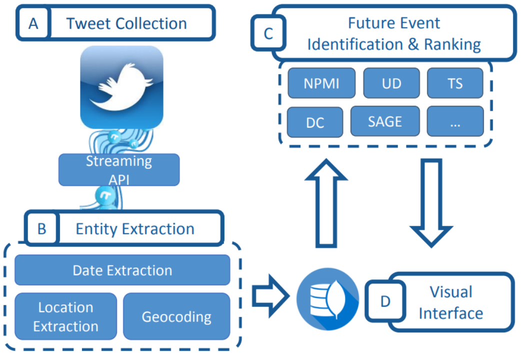


Fig. 1. System pipeline of Crystal ball.

Normalized Pointwise Mutual Information (NPMI) and Pointwise Mutual Information (PMI) are the attributes used for the extraction of these signals, which provide information about future events. In this case, NPMI and PMI are calculated to determine the correlation between future event attributes such as location and time from the tweet data. The higher the NPMI score, the higher the correlation between the location-time pair. These correlated tweets are then classified using six metric measures. The Link Ratio, Hashtag Ratio, User Credibility, and User Diversity all indicate the informativeness of an event tweet. The other two measures that show event-tweet cohesiveness are Degree Centrality and Tweet Similarity.

Tweets that contain links and hashtags are more likely to be informative, which includes blogs, pages, or articles related to future events. Link Ratio is calculated as the ratio of tweets containing links to tweets that are related to a possible future event. The hashtag ratio is computed as the ratio of tweets containing hashtags to tweets related to a possible future event. Informative tweets are more likely to be posted by credible users. The Twitter Follower-Friend (TFF) ratio has been considered to measure the user's credibility. TFF is the ratio of followers to friends. User diversity is related to the informativeness of the sources of tweets. If one user tweets regarding one potential future event, then it would likely be a bot that is programmed to send periodic tweets. We assume that if the tweets are associated and connected, it is more likely that they refer to the same future event. To measure the degree of centrality between these tweets, we consider the retweets and mentions regarding one possible future event and then calculate DC by using Freeman's General Formula for Centralization. There are scenarios when the content of all tweets related to the same future event is similar, but they are not connected. In this case, degree centrality fails. To address this case, Twitter similarity has been introduced. Here, the similarity of a set of tweets linking to one future event is considered and calculated using Levenshtein distance.

RankSVM is a feature ranking method that describes the relationship between explanatory variables and the response variable, which determines a future event's overall quality or rank. Lower-quality events primarily consist of advertisement messages or bot-generated tweets.

III. Results

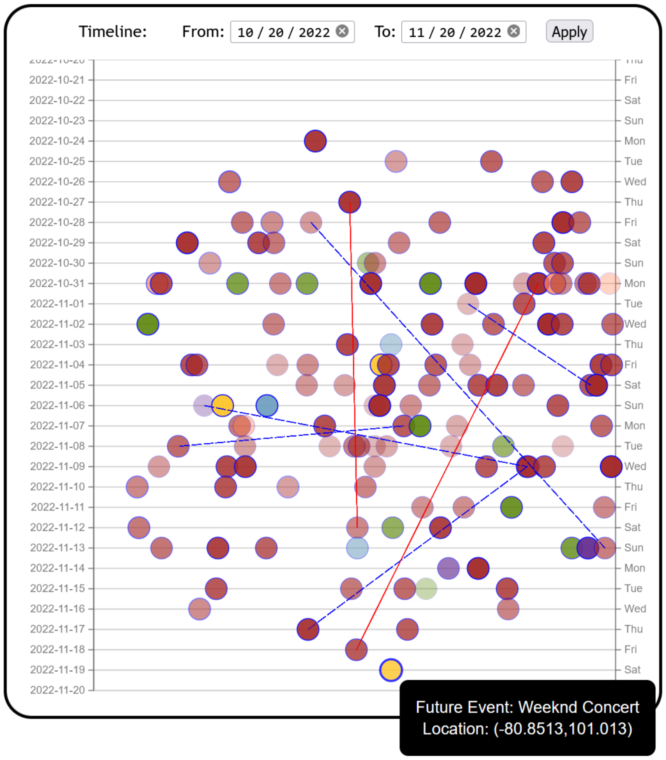
We have implemented the visualization system using HTML, JavaScript, and the D3 library (version 7), with Python used for pre-processing. We developed a system that included the following visualizations to forecast future events:

- Event Calendar View
- Map View
- Word Cloud View
- Event List view
- HeatMap View

The HeatMap view was an add-on to our work.

A. Event Calendar View

The Event Calendar View uses Twitter data to depict upcoming events by date and their relationships which is shown in figure 2 [5]. The view is divided into several rows, with the date scale displayed on the left and the corresponding day on the right. The horizontal axes represent the timestamp in milliseconds. Several events that occur within each row that corresponds to date and time are displayed in the shape of circles. The association between future events is shown through the use of location and keywords. A solid red line connecting the two points represents future events that will occur at the same location. A blue dotted line connecting two points indicates upcoming events that share at least eight keywords. Hovering over these lines reveals future events and shared keywords within these events. When you click on the blue dotted line, a list of hashtags connecting the two events appears in the word cloud. The colors in these event circles represent six emotions (anger in red, disgust in purple, fear in orange, joy in green, sadness in blue, and surprise in yellow). The circle's opacity, which ranges from 0 to 1, represents the uncertainty of emotion. Hovering over the circular marks would provide us with event information. When you click on an event, the Event List View (using the Rank SVM model) for that event appears in the form of a circular radar chart with three bar charts associated with it. We have added an extension to this view which is a date range that allows the user to scale the date axis according to the dates given.



Timeline: From: 10 / 20 / 2022 To: 11 / 20 / 2022

Future Event: Weeknd Concert
Location: (-80.8513,101.013)

Fig. 2. Event-Calendar View.

B. Map View

The map view depicts the location of the forthcoming event. A donut chart represents each event location, accounting for the number of events held at that location. The donut chart is separated into five categories for more information on the event: tomorrow, less than or equal to a week, less than or equal to two weeks, less than or equal to 30 days, and greater than 30 days. A blue color interpolation scale is used to color these five types. When the event location is clicked, a spiral chart appears, displaying the order in which events occur at that specific location. The word cloud view of keywords associated with that location appears when you click the arcs of the donut chart as shown in figure 3 [3] [6].

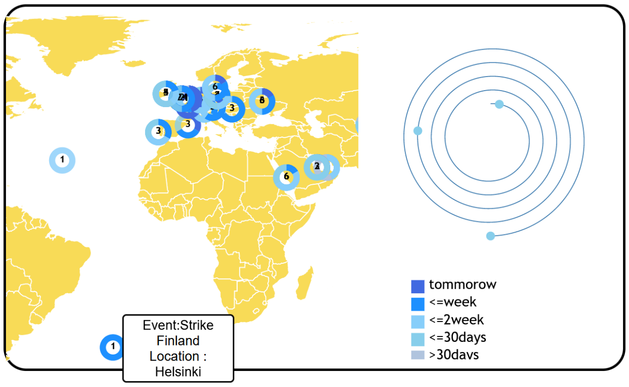


Fig. 3. Map View.

C. Word Cloud View

The word cloud view shows the keywords across all the events. The size of the words indicates how frequently such terms appear in tweets. On clicking the donut chart in the map view, we can see the keywords used for events in that location as shown in figure 4. When we click on the blue-dotted line connecting two events in the event calendar view, the word cloud displays the common terms and their relative frequency. The layout for the word cloud was taken from the word cloud library [2].

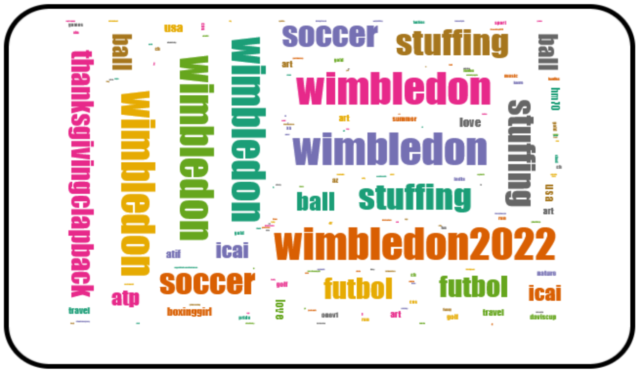


Fig. 4. Word Cloud View.

D. Event List view

Two timeline bar charts, one emotion analysis bar chart and a circular radar chart comprise the event list view. When an event is selected from the event calendar view, the circular radar chart and three bar charts are displayed as shown in figure 5 [4]. The first timeline bar chart shows the number of tweets for the past 30 days from the date on which the event occurs. The second timeline bar chart shows the number of tweets during the event's 24-hour period. The third bar chart depicts the likelihood of various emotions like joy, surprise, sadness, anger, disgust, and fear for the given future event. The circular radar chart represents measures associated with a given future event. These measures include Link Ratio (LR), Hashtag Ratio (HR), Normalized Pointwise Mutual Information (NPMI), User credibility (UC), User Diversity (UD), Degree Centrality (DC), and Tweet Similarity (TS). NPMI is the feature used for event identification. LR, HR, UC, and UD are measures that provide event tweet informativeness. The event tweet's cohesiveness is produced by DC and TS values.

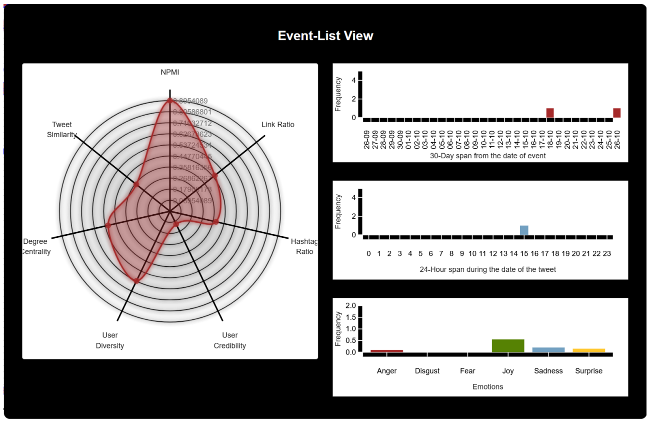


Fig. 5. Event-List View.

E. HeatMap View

The Heat Map View is an extension of our work that compares the frequency of tweets for each of the eight categories for all six emotions using Twitter data as shown in figure 6 [7]. The color hue in the heat map cells represents the number of tweets for a certain group on an emotion. The more intense the color, the more tweets fall within that pair (category and emotion), and vice versa. We have included a tooltip that displays the number of events present for the selected emotion-category pair.

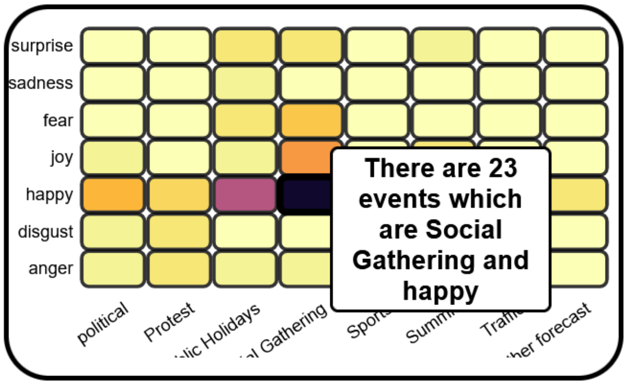


Fig. 6. Event-Calendar View.

Figure 7 depicts the visual system with all views integrated.

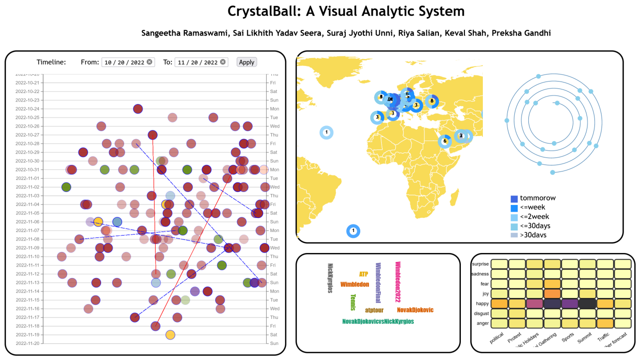


Fig. 7. Crystalball visual system.

IV. Contributions and List of Team Members

The contributions that I have made to this project are as follows:

- Implemented Event-Calendar View
- Implemented the extension part- The Heatmap view
- Integrated Event-Calendar view with the Word cloud and Event-List views.
- Implemented Event-List view.
- Preprocessed the derived attributes for the Rank-SVM model.
- Documenting our project's report and proposal documents.

The following members of my team helped me through my project and made it successful.

- Keval Shah
- Preksha Gandhi
- Sangeetha Ramaswami
- Suraj Unni
- Riya Salian

V. Learnings

I have learned the following lessons during the course of my project:

- How to create an event view in a Google calendar that resembles that of the Event-calendar view.
- How to discern various graphs and glyphs based on the data provided.
- Interactions among various charts using various event listeners.
- Implementing different charts within the tooltip of a chart using the same data.

References

[1] Isaac Cho,Ryan Wesslen,Svitlana Volkova, William Ribarsky, Wenwen Dou. CrystalBall: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data. In Proceedings Of IEEE Conference on Visual Analytics Science and Technology(VAST),2017. https://www.researchgate.net/profile/Isaac-Cho/publication/324598122

[2] Word Cloud Generator, https://www.jasondavies.com/wordcloud/

[3] Spiral for John Hunter, http://bl.ocks.org/syntagmatic/3543186

[4] Radar Chart Redesign, http://bl.ocks.org/nbremer/21746a9668ffdf6d8242

[5] For the building blocks of the Event-Calendar view, https://blog.logrocket.com/using-d3-to-create-a-calendar-app/

[6] For the implementation of Map view, https://d3-graph-gallery.com/backgroundmap.html

[7] For the implementation of the Heatmap view, https://d3-graph-gallery.com/graph/heatmap_style.html