# Interim Project Report

Prerit Auti and Antara Bhowmick

9 March 2018

## 1 Objective

The main objective of this project is to implement the visualization presented in **PEARL: An Interactive Visual Analytic Tool for Understanding Personal Emotion Style Derived from Social Media** [9]. As mentioned in the proposal, we will only implement some of the views - we won't be implementing the split view or the emotion outlook, volatility and resilience filters because although they're referred to in the paper, there aren't many details given and the authors have not implemented either of them in their online demo [1].

For the milestone, we intended to get the datasets, complete the data processing subsystem and implement the following visualizations as given in Fig 1:

- (a) Overview of the timeline
- (e) Sliding window over timeline
- (c) Scatterplot of emotional words from the tweet
- (d) Raw tweet data

### 1.1 Approach

In the paper, the server side of the PEARL system, that is, the data processing subsystem was implemented in Java. However, we opted Python for all data processing and pre-processing. The authors of the paper developed the visualization using D3.js[2] and we also chose to do the same.

Initially we aimed to finish processing the data and then move on to developing the visualization. However the data processing subsystem ended up being very involved; since we already knew the format of the final data, we decided to work on data processing and visualization simultaneously by using dummy data.

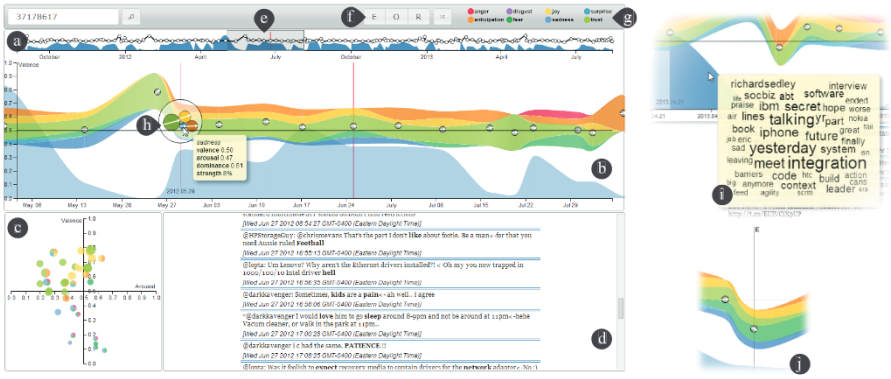

Figure 1: PEARL user interface [9]

# 2 Implementation

## 2.1 Datasets

The main data source in this paper is Twitter - the system collects all tweets given a Twitter ID. The paper does not elaborate further on how this data is collected. We decided to implement this step using the official Twitter API via Tweepy, which is a Python wrapper. Unfortunately, the API places certain restrictions on the amount of data that can be downloaded given an ID, so we weren't able to download a full history like in the paper. Nonetheless, we downloaded a little over 3200 tweets each for 4 users using this method.
Furthermore, the paper uses lexicons for processing the data. The NRC lexicon[6] used for the Plutchik model [8] is provided on request for non-commercial purposes, and the WordNet lexicon [5] is publicly available. However, the official ANEW dataset [3] that the paper uses required for the VAD model [4] is not easily procured. Instead of waiting for the official dataset, we chose to use a combination of ANEW data compiled by two different sources.

## 2.2 Processing

The paper breaks down data processing into two major steps: emotion analysis and mood analysis. In emotion analysis, each tweet is broken down into 'emotional words' that have emotion categories from Plutchik's model [8] as well as VAD scores [4]. If either of the values are missing, synonyms of the words are obtained from WordNet and the corresponding labels and averaged VAD scores are used. In case there aren't any synonyms that appear in the lexicon, the original word is removed from the dataset.
This was relatively easy to implement. The paper goes into sufficient detail about the processing as well as the format of the data expected. Since the size of the dataset was quite large, and referenced other large datasets it was a fairly

time-consuming process.

In mood analysis, the tweets are grouped together to form tweet-segments based on certain criteria (emotional proximity, semantic proximity and temporal proximity). The paper states that it uses a constrained co-clustering approach, similar to temporal topic segmentation [7] to segment the tweets. Additionally, it also stated the output format as before.
Surprisingly, this was quite difficult to implement. The output format defined in emotion analysis was not a convenient input to this algorithm. Moreover, the algorithm wasn't well-defined and we had to make assumptions for parameters (such as thresholds for VAD values). At present, the clusters obtained aren't satisfactory (unevenly distributed tweets, most too short for significant analysis) and the parameters need to be tweaked further. We may also look into different clustering algorithms.

### 2.3 Visualization

As stated before, the visualizations are implemented using D3.js. Since we weren't able to get proper tweet segments, we used dummy data to visualize the results.

The time-line view was replicated using an area chart. In the final visualization the data points will be the starting date for each cluster. To mimic the effect, we used data-points that were non-uniformly distributed across the time-line.

For the scatter-plot of emotional tweets, each emotional word from a tweet segment is represented on graph where the x axis denotes Arousal values and y axis denotes Valence. The size of the point represents the Dominance value. Additionally, each point is an equally distributed pie chart with every slice representing the emotion categories that the word belongs to. The slices are colored according to the colors given in Plutchik's emotion wheel. Currently, we have implemented the scatter-plot with dummy data.

## References


[1] PEARL online demo. Accessed: 2018-03-09.

[2] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011.

[3] Margaret M. Bradley, Peter J. Lang, Margaret M. Bradley, and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings, 1999.

[4] Albert Mehrabian. *Basic dimensions for a general psychological theory : implications for personality, social, environmental, and developmental studies.* Cambridge : Oelgeschlager, Gunn Hain, 1980. Includes index.

[5] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41, 1992.

[6] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[7] Shimei Pan, Michelle X Zhou, Yangqiu Song, Weihong Qian, Fei Wang, and Shixia Liu. Optimizing temporal topic segmentation for intelligent text visualization. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 339–350. ACM, 2013.

[8] Robert Plutchik. The nature of emotions. 89:344, 01 2001.

[9] J. Zhao, L. Gou, F. Wang, and M. Zhou. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 203–212, Oct 2014.