VISUALIZATION TECHNIQUES IN REPRESENTING TIME-VARYING SENTIMENT DATA

Daffi Palus
1Department of Computer Science,
Sathyabama University, Chennai, India

Abstract: Microblogging sites like Twitter contain a huge repository of user-generated sentiment about world-events, reviews and products. Sentiment analysis can be very useful if utilized properly. Sentiments cannot be 100% accurate, though it is a useful metric that can be used in relation with other metrics. Existing systems depict the sentiment as a data point over a varying time period. The data point is used to determine the context of the text. This paper proposes the use of a Gauge chart (speedometer) to determine the range of sentiment in web posts. The chart hints whether something is ‘on the whole positive’ or ‘on the whole negative’ without confining itself to a single data point.

IndexTerms - Data Analytics, Data Visualization, Sentiment Analysis, Text Mining, Gauge Charts

I. INTRODUCTION
Sentiment Analysis is the process of extracting the contextual polarity of textual data. It deals with the detection of emotions, opinions, comments, moods and sentiments. It determines whether the given piece of text is contextual positive, contextual negative or contextual neutral. Sentiment analysis can be viewed as a synonym of text-data mining that depends on natural language and computer linguistics. Textual data is broadly categorized as facts and opinions. Factual data do not contain underlying sentiments while opinion words hold the key to showing the contextual sentiment of the text under study.

Visualization aims to depict the type of content, contextual polarity in a piece of written text. This is an effective manner of describing data than a table or some form of numerical representation. In this paper, we make use of the conventional gauge charts to determine the polarity of the text without visualizing at any single numerical point.

II. RELATED WORK
There is a tremendous need for visualizing sentiments from Web sites and Web posts. Marti A. Hearst et al. [1992] proposed the idea of mining the lexical relation of keywords and phrases in textual data called as sentiments and opinions. Tagging based on the subjectivity to identify opinions in a corpus was discussed by Janyce M. Wiebe [2000]. Ahmad Khurshid et al. [2006] proposed sentiment analysis for a multilingual framework. Textual polarity is calculated for any non-English to English translated text using the sentiment of each keyword in its surrounding area to construct a co-location pattern. Alexandra Balahur et al. [2009] argue that open-ended sentiment or opinion mining is more of a difficult activity compared to topic-specific close-ended mining. Sentiment library built on just a group of words is cumbersome and requires large lexicons or sentences to be built using qualified training data and computer linguistics.

M Hao et al. [2011] presented a move from the visualization of yes or no questions and numeric values on scales, to visualization of time-series in sentiment analysis. TwitInfo used a timeline-based visualization to identify heightened activity in data streams and later label them as events and was presented by Adami Marcus et al. [2011]. The limitation was it focused only on tracking an event based on the activity examined and could not be used for actual investigative purposes. The visualization was on a longitudinal plane which made documentation on a large timescale seem tedious. Weiwei Cui et al. [2012], proposed a visual analytics system called TextWheel for streaming news feeds which examines multiple features and keywords contained in the articles, while still keeping intact the dynamic nature of the news content. However, this was not integrated with other data mining methods which caused clutter in the display when too many keywords were used. Christian Rohrdanz et al. [2012] developed a visualization model based on pixel-map calendars and time density plots; to identify and categorize high critical issues that need to be looked into with priority based on the feedback collected which is varying and is time-stamped.

Victor Benjamin et al. [2013], presented sentiment analysis as a combination of content-specific analysis and visualization for comparison of writing styles of different authors. Patrick Hening et al. [2014] used BlogIntelligence to identify and notify trends within weblogs. Weblogs were crawled, analyzed and visualized. Trend identification was accomplished by measuring the frequency of the word or lexicon occurrences.

While visualization frameworks were two dimensional, Maryanne Doyle et al. [2014], used a scatter plot with three axes to represent three dimensional data. TwitInfo used only positive and negative sentiment classifiers with two separate visualizations for sentiment and traffic. Trivis proposed by Doyle, used three sentiment classifiers – positive, negative, neutral and used the same visualization to display sentiment and traffic for multidimensional data. A. Weichselbraun et al. [2014], presented a method for improving the quality and depth of the databases used for mining the opinion hidden and held in user remarks. Ambiguous terms are detected and then its sentiment value is determined independently and the term is then added into the sentiment vocabulary or library to be used for future needs.

III. SENTIMENT MINING
We use any open source web server like Node.js to connect to a corpus. In this study, Twitter is used as the corpora and tweets are used as input to the analysis. Fig.1. shows a typical sentiment analysis system. The relevant feeds are identified by means of the #hash tags.
We use an AFINN-111 wordlist which contains 2477 words and phrases to perform sentiment analysis of random blocks of data. Each word is rated with an integer between minus five (negative) and plus five (positive). Then, the total sentiment value of the comment is calculated. If a word or phrase cannot be found in the AFINN-111 wordlist it is assigned a neutral score ‘0’.

Table 1. presents the algorithm used in the proposed system. PosisentiVal stores the positive sentiment value, NegsentiVal stores the negative sentiment value and sentiVal indicates a neutral score. The TotsentiVal is calculated and is plotted on the gauge chart.

Table 1. Pseudo code for Sentiment Mining

```
/* Sentiment Mining */
FOR each segmented word in the Tweet:
  IF the segmented word is in the Sentiment Analyzing Dictionary/AFINN-111 wordlist:
    IF the segmented word is in the Positive Sentiment section:
      PosisentiVal = value (segmented word)
    ELSEIF the segmented word is in the Negative Sentiment section:
      NegsentiVal = value (segmented word)
    ELSE:
      sentiVal = 0
  ENDIF
ENDFOR
TotsentiVal = sum (PosisentiVal, NegsentiVal, sentiVal)/ number of words in Tweet

/* Sentiment Visualization – Gauge Chart */
Create Gauge Chart with min and max ranges
FOR every Topic T being tweeted
  Plot the value (TotsentiVal) on the gauge chart
ENDFOR
```

IV. VISUALIZATION

Data, whether qualitative and quantitative can be understood better when shown as visualizations than numbers. Visualization has become a common trend today in newspaper, journals, and more importantly in the business world where data have to be consumed at tremendous pace in order to arrive at decisions and for strategic planning. This is where the need for using visualization techniques arises. The best data visualization methods are those that reveal something new and interesting about the underlying patterns and relationships embedded within the raw data.

The gauge chart is used in business intelligence for a quite some time. A gauge chart (typical speedometer used in cars) was first used to determine the approximate volume of liquid inside a tank. They focus to determine the range of a given value.

Basic user interface components consist of a dial which is circular or arc-shaped, pointer and direction of movement as given in Fig.2. The face of the dial is colored with the colors ranging from red on one side of the dial to yellow and subsequently fading to green on the other side of the dial. The pointer can move in both directions, right to left and left to right.

The position of the pointer on the dial indicates the contextual polarity of the textual feed. If the pointer is on the green colored part of the dial, then the feed is context-positive. If the pointer is on the yellow colored part of the dial, then the feed is context-neutral. If the pointer is...
on the red colored part of the dial, then the feed is context-negative. The movement of the pointer on the dial indicates whether the feed is moving towards the positive zone or towards the negative zone.

![Sentiment Gauge Chart](image)

**Fig.2. Sample Gauge chart reading**

**V. RESULTS AND DISCUSSION**

The performance aspects of the proposed system are compared against several related systems, RadViz, SentiView and StoryViz. The performance comparison is presented in Table.2.

<table>
<thead>
<tr>
<th>System</th>
<th>Time Varying</th>
<th>Multi-Dimension</th>
<th>Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>RadViz</td>
<td>No</td>
<td>Explicit</td>
<td>No</td>
</tr>
<tr>
<td>SentiView</td>
<td>Yes</td>
<td>Explicit</td>
<td>Yes</td>
</tr>
<tr>
<td>StoryViz</td>
<td>No</td>
<td>Implicit</td>
<td>No</td>
</tr>
<tr>
<td>Proposed</td>
<td>No</td>
<td>Single Dimension</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table.2. Performance Comparison

Fig.3, Fig.4 and Fig.5 presents the relationship between the three factors – time, tweet count and sentiment. The hash tag #Modi was used for the study. The tweet count and sentiment were recorded in periodic intervals of time. Tweet count increases proportionally with time. Sentiment values vary according to the time and number of tweets analyzed by the system.

![Tweet Count vs Time](image)

**Fig.3. Comparison of Tweet Count vs Time**

![Sentiment vs Time](image)

**Fig.4. Comparison of Sentiment vs Time**

![Sentiment vs Tweets](image)

**Fig.5. Comparison of Sentiment vs Tweet Count**

Unlike the existing systems, the proposed system does not use any single data point to represent the sentiment at any time ‘t’. Instead, it represents the contextual polarity of the written text in the form of a gauge chart.
VI. CONCLUSION
This paper has proposed the use of the conventional gauge charts to describe the contextual polarity of the textual data in Twitter feeds. It does not confine itself to any single numerical point; however it does not consider the time varying nature of feeds. Gauge charts that show the varying degree of sentiment against varying time-periods can be considered for future studies.

VII. ACKNOWLEDGMENT
I thank Dr. Andrews (Sathyabama University) for his constructive criticisms that helped improve the clarity of the paper. Data for this paper was sourced from Twitter feeds.

REFERENCES
[14] Doyle, Maryanne and Smeaton, Alan F. and Bermingham, Adam, “TriVis: visualising multivariate data from sentiment analysis”, 8th Annual Irish