# Social Media Big Data Approach to Emotional Intelligence

**Olumide Adebayo[1] Jeongkyu Lee[2] (Advisor)**
**[1]Department of Technology Management, University of Bridgeport, CT.**
**[2]Department of Computer Science, University of Bridgeport, CT**

## Brief

This paper aims to research the possibility of detecting *regional sentiment* by analyzing Social Media Big Data (SMBD). It can be said that the vast majority of people with some form of internet access participate in one Social Media platform or the other. The importance of the insight derivable from SMBD cannot be overemphasized. Many theories have been postulated that show that human emotion is extremely powerful and drives our thought process. Noting that our communication is driven by our thoughts makes social media data a huge wealth of information from which an average regional sentiment analysis can be deduced. This sentimental knowledge becomes a valuable tool in policing (security) and marketing (commerce).

## Introduction

A few decades ago, the Internet was a ground breaking phenomenon that enables people across the globe to interact with just the click of a button. Of recent Social Media has taken that ability to communicate and share to a whole new level. It is no secret how Social Media touches every aspect of our daily life. The vast majority of people that have access to the internet have at least one active social media account be it on Facebook, Twitter, G+, Snap Chat, Pinterest, YouTube or LinkedIn, just to mention a few. The volume of data generated by social media platforms makes it a great source of information of different sorts for both commerce and security.

## Methodology

### Data Collection

In conjunction with the Spring XD platform, 2 java-based applications were developed for data collection purposes. The first application was deployed as a streaming application that runs within Spring XD for fetching tweets from Twitter continuously over a 2 month period. The second is a Hadoop MapReduce (MRv2) application that was used to process the raw tweets and perform sentiment analysis on the extracted texts.

MapReduce is a framework that allows for a parallelized distribution of the processing work of large data sets across many compute nodes. It is made of 3 steps namely **Map**, **Shuffle** and **Reduce**. Any computational task that can be reduced to a map() and reduce() step will greatly benefit from the MapReduce framework.

A similar Apache Spark application was developed to compare performance with the MRv2. Spark leverages computer memory for processing large data sets and as such is usually faster MRv2. The output of both applications where queried using Apache Hive. Hive is an infrastructure that runs on top of Hadoop which enables SQL-like access to data stored in HDFS.

All activities were carried out on an Amazon Cloud EC2 instance

### Data Cleaning

Data cleaning was two folds. In the first run, all tweets that had no meaningful geo information were discarded. Those that cannot be identified as originating from the United States were also dropped. . Location attributes were also normalized as shown in the sample.

| Original | Transformed |
|---|---|
| **Not Self-centered, Texas** | TX |
| **Lilburn, ga** | GA |
| **Between some thighs, La** | LA |
| **New Orleans , La!** | LA |

Table 1 Sample State Transformation

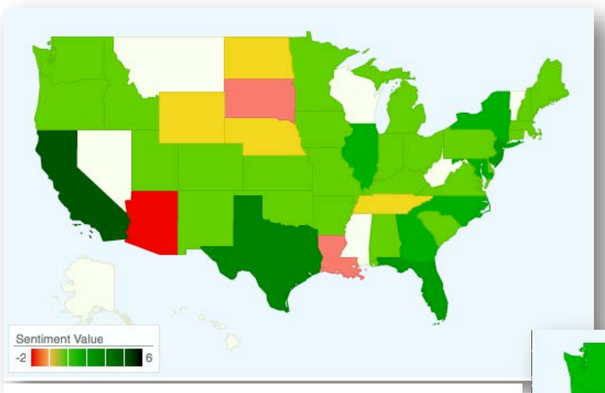
Finally, to simplify visualization, a coding mechanism was applied to sentiment values as shown in the table below. Raw sentiment values calculated from sentiment analysis ranged from -3 to 58. Negative values represent negative sentiments while positive values represent positive sentiments. The farther away from 0 the sentiment value is, the stronger the sentiment. A sentiment value of 0 either means no sentiments were detected or the negative cancelled out positive.

| Sentiment Code | Actual Values | Definition |
|---|---|---|
| -3 | -3 | Very Negative |
| -2 | -2 | Negative |
| -1 | -1 | Somewhat Negative |
| 0 | 0 | Neutral |
| 1 | 1 - 10 | Somewhat Positive |
| 2 | 11 – 20 | Positive |
| 3 | 21 - 30 | More Positive |
| 4 | 31 - 40 | Very Positive |
| 5 | 41 - 60 | Extremely Positive |

Table 2 Sentiment Level Codification

## Results

The output of the MapReduce (MRv2) and Spark App was a triple of (date,state,sentiment] values. Feeding these triples into Google's GeoMap API provided the visualizations shown below. The sample visualizations represent average daily regional sentiment for October 3rd, October 23, November 8th and November 11th 2016. The shades of red represent negativity while the shades of green represent positivity. The darker the shade, the stronger the sentiment. Yellow indicates neutrality.




Oct 3rd 2016

Fig. 1 Sentiment Distribution on Oct 3rd, 2016

Oct 23rd 2016

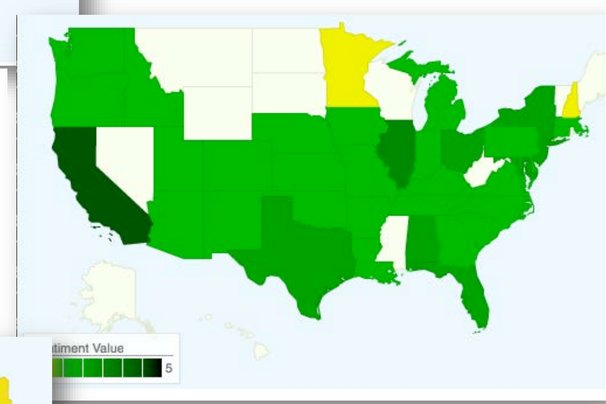



Fig. 2 Sentiment Distribution on Oct 23rd, 2016

Nov 8th 2016

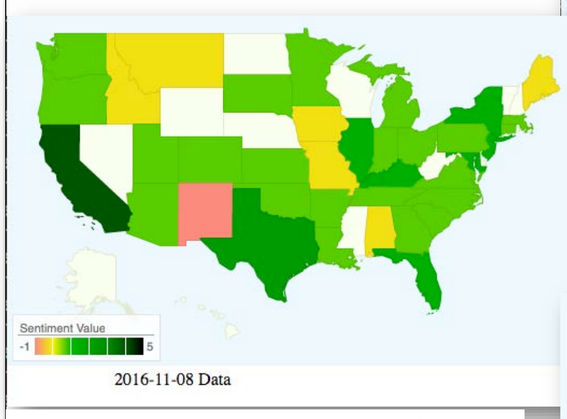



Fig. 3 Sentiment Distribution on Nov 8th, 2016

Nov 11th 2016

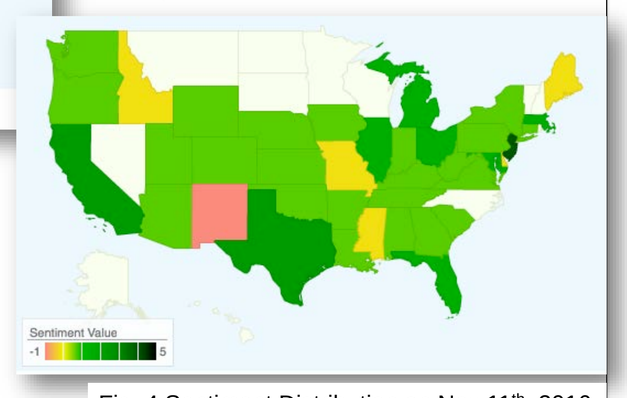



Fig. 4 Sentiment Distribution on Nov 11th, 2016

## Limitations

There are number of limitations to this study that are worth highlighting for the sake of completeness. Due to time constraint, the research used data gathered over short period only. This period included the 2016 general elections in the United States. It will be great to see what the outcome will be when data is gathered over, say, a two year period.
Facebook is the #1 social media site [9] but the research opted to use Twitter, the #3 social media site, for reasons stated earlier. Using data from the top 5 social media sites might enhance or refute the final outcome.

## Conclusion

The need for regional sentiment is of great importance because it is a tool for literarily taking the pulse of a state, nation or continent. It becomes a potentially security issue when a region trends negative even though all other situations (economy, political, etc.) are normal. This can help to raise a red-flag for security officials. The same can be said of commerce. A regional sentiment analysis will point out quickly to retailers, manufacturers, vendors where to apply more incentives as well.

The ability to execute this analysis in real-time and take action in real-time as well, creates tremendous opportunity with instant feedback.

**References**

[1] D.-H. Shin, "Demystifying big data: Anatomy of big data developmental process," *Telecommunications Policy*.
[2] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management,* vol. 35, pp. 137-144, 4// 2015.
[3] M. Mayeh, R. Scheepers, and M. Valos, "Understanding the role of social media monitoring in generating external intelligence," in *ACIS 2012: Location, location, location: Proceedings of the 23rd Australasian Conference on Information Systems 2012*, 2012, pp. 1-10.
[4] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," Business horizons, vol. 54, pp. 241-251, 2011.
[5] (2016). Top 15 Most Popular Social Networking Sites. Available: http://www.ebizmba.com/articles/social-networking-websites
[6] A. K. Jose, N. Bhatia, and S. Krishna, "Twitter sentiment analysis," *Major Project Report, NIT Calicut,* 2010.
[7] A. Sarlan, C. Nadam, and S. Basri, "Twitter sentiment analysis," in *Information Technology and Multimedia (ICIMU), 2014 International Conference on*, 2014, pp. 212-216.
[8] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford,* vol. 1, p. 12, 2009.
[9] A. Z. Khan, M. Atique, and V. Thakare, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE),* p. 89, 2015.
[10] A. Go, L. Huang, and R. Bhayani, "Twitter sentiment analysis," *Entropy,* vol. 17, 2009.
[11] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Psychology,* vol. 66, 2015.