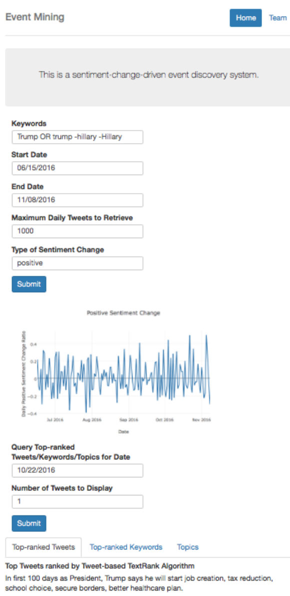
Analytics and Data Science Institute

datascience.kennesaw.edu

# Data Science Research Series

## A Sentiment-Change-Driven Event Discovery System




## BACKGROUND

In situations from marketing campaigns to presidential elections, people's decisions are driven by their sentiments. Therefore, it is beneficial for strategy makers to know 3Ws of people's sentiment changes, namely, what happens, when it happens, and what its effect is. For that, we present a system that can automatically discover events that have significantly driven people's sentiment changes towards a target in a timely manner.

## APPROACH

The system architecture can be found in Figure 1. There are four components in this system, including Tweets Sampling, Sentiment Sensor, Sentiment Filter, and Event Discovery.

- Tweets Sampling: Establish the target based on the application scenario and collect Tweets related to the target in a certain time period.
- Sentiment Sensor: Measure people's daily sentiment changes towards the target.

- Sentiment Filter: Select Tweets to analyze further based on people's sentiment change direction. If there is daily positive ratio increase, then we will only study Tweets labeled with positive. Otherwise, we will only study Tweets labeled with negative.
- Event Discovery: Discover events at people's sentiment change time points with TextRank and Topic Modeling algorithms. A problem-dependent module, External Source, can be added, to eliminate potential bias from Tweets.

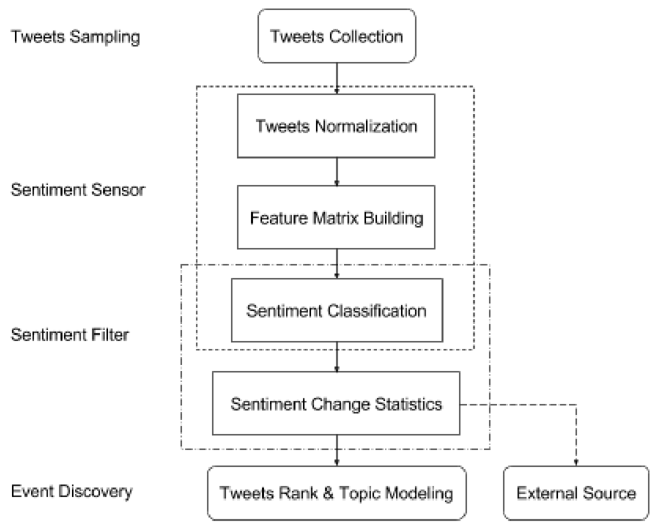



Figure 1. System Architecture

The system was applied to study people's sentiment changes during the 2017 U.S. Presidential Election by aiming to answer the following questions.

1. When people talk about a candidate, are their words positive, negative, or neutral?
2. How have people's sentiments changed towards a candidate over time?
3. What has driven those significant sentiment changes?

## RESULTS

### Tweets Collection

- Targets: Trump and Clinton
- Time: Jun. 16, 2015 – Nov. 8, 2016
- Daily Tweets: 1000 for each candidate
- Total Tweets: 1,020,672

### Tweets Normalization

Each Tweet is normalized following the flow in Figure 2.

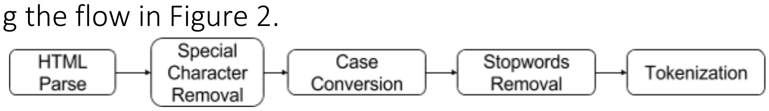



Figure 2. Tweet Normalization Flow

### Feature Matrix Building

In order to apply machine learning algorithms to Tweets, each Tweet is vectorized into numeric values using the tf-idf document-term technique.

## Sentiment Classification

Each Tweet is classified to be positive, negative, and neutral, using the sentiment analysis API, called Sentiment140, which was developed by Stanford University. Then daily positive and negative ratio is calculated. Suppose the number of positive, negative, and total tweets mentioning the target daily to be $n\_pos, n\_neg, n\_tot$, then positive ratio is $n\_pos \;/\; n\_tot$ and negative ratio is $n\_neg \;/\; n\_tot$ .

Figure 3 shows the daily positive ratio for Trump and Clinton since Candidacy and Primary respectively. As shown, the blue line is above the orange line overall. To eliminate the potential influence from other candidates, we will only focus on the analysis since Primary in later stage.

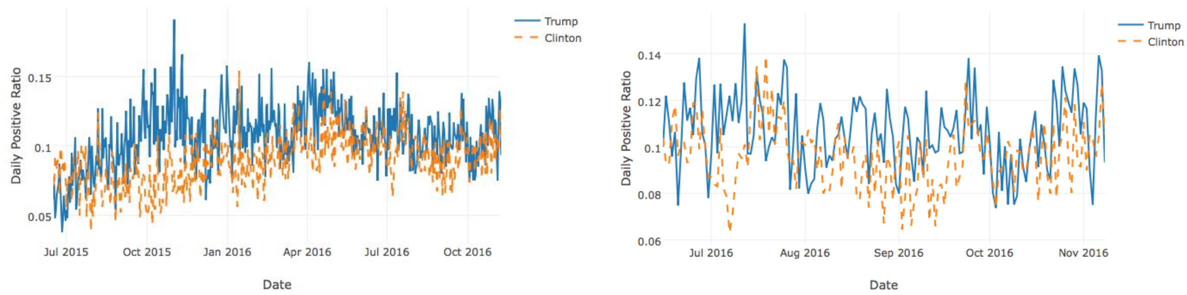



Figure 3. Daily Positive Ratio since Candidacy (left) and since Primary (right)

## Sentiment Change Statistics

Daily sentiment ratio change is calculated. Suppose the positive sentiment ratios on the current day and previous day to be $p\_1$ and $p\_2$, then positive sentiment ratio change is $(p_1 - p_2)/p\_2$. Figure 4 shows the daily positive ratio change since Primary. Table 1 shows top daily positive changes for Trump and Clinton.

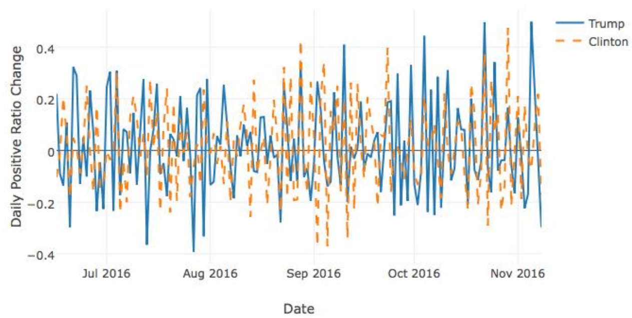



Figure 4. Daily Positive Ratio Change Since Primary

| Candidate | Date | Positive Ratio Change |
|---|---|---|
| Trump | Oct. 22, 2016 | 50% |
| Trump | Jul. 27, 2016 | -39% |
| Clinton | Oct. 29, 2016 | 47% |
| Clinton | Sept. 02, 2016 | -37% |

Table 1. Top Daily Positive Ratio Change

### Tweets Rank & Topic Modeling

To automatically discover events at people's sentiment change time points from a large number of Tweets, three algorithms are used, as listed in Table 2.

| Algorithm | Name | Output |
|---|---|---|
| TS1 | Tweet-graph TextRank | Top ranked tweets |
| TS2 | Word-graph TextRank | Top ranked keywords |
| TT1 | Nonnegative Matrix Factorization (NMF) | Multiple topics with each represented by keywords |

Table 2. Algorithms used in Event Discovery

For days with top positive ratio changes, the following events have been discovered, as shown in Table 3.

| Candidate | Date | Event |
|---|---|---|
| Trump | Oct. 22, 2016 | Top 1 Tweet: "In first 100 days as President, Trump says he will start job creation, tax reduction, school choice, secure borders, better healthcare plan."<br>Effect: Positive Increase |
| Trump | Jul. 27, 2016 | External Source: "Trump calls on Russia to find Clinton's missing emails."<br>Effect: Positive Decrease |
| Clinton | Oct. 29, 2016 | External Source: "FBI reviews emails related to Clinton's case."<br>Effect: Positive Increase |
| Clinton | Sept. 02, 2016 | Top 1 Tweet: "BREAKING FBI NEWS: Hillary Clinton Lost Laptop With Classified Data."<br>Effect: Positive Decrease |

Table 3. Events Driving People's Sentiment Changes

## CONCLUSIONS

By using the sentiment classifier as sensor and filter, we can successfully detect events when they happen and measure their importance based on people's sentiment changes. Moreover, Tweet-based TextRank algorithm along with NMF and the word-based TextRank can be combined to automatically provide overviews of events.

## CITATION FOR FULL ARTICLE

Lili Zhang, Ying Xie and Guoliang Liu. "A Sentiment-change-driven event discovery system." In Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017, pp. 1035-1041. ACM. DOI: [10.1145/3106426.3109038](10.1145/3106426.3109038)