# EDF 6938 Final Summary

*Cheryl Calhoun*

*December 5, 2015*

The purpose of this project was to capture twitter traffic leading up to a football game, then analyze the twitter text using the functions `classify_polarity` and `classify_emotion` to determine if we can develop a predictive model for whether or not the team will win the game, win the game by 7 points, win the game by less than 7 points, or beat the spread.

## Data Collection

The data collection process occurred in three phases. This was due to the retroactive nature of the project. By the time we started data collection, data for games prior to Game 09 were no longer available via the Twitter API. The total number of tweets collected are listed in Table 1.

- Game 01-08:
  - tweets were captured using the copy and paste method from a twitter advanced search on the game specific hashtag.
  - date range = game day -7 through game day -1
- Game 09:
  - tweets were captured using the #GoGators hashtag and then filtered for the game specific hash tag.
  - date range = game day -3 through game day -1
- Game 10-12:
  - tweets were captured using the game specific hashtag.
  - date range = game day -7 through game day -1

Table 1: Total Tweets Collected for Each Game

| Game | TotalTweets |
|---|---:|
| Game01 | 75 |
| Game02 | 105 |
| Game03 | 117 |
| Game04 | 374 |
| Game05 | 151 |
| Game06 | 78 |
| Game07 | 419 |
| Game08 | 1232 |
| Game09 | 312 |
| Game10 | 8677 |
| Game11 | 8714 |
| Game12 | 5592 |

## Schedule Results

The results and spread data, as displayed in Figure 2, were collected using COVERS (http://www.covers.com/sports/ncaaf). This data was entered into an Excel spreadsheet and exported to Schedule.csv.

Table 2: Univeristy of Florida 2015 Football Game Results and Spread Data Table

| Game | DATE | OPPONENT | RESULT | UF | OP | SPREAD | COVER | BEAT | HASHTAG |
|---|---|---|---|---:|---:|---:|---:|---|---|
| Game1 | 2015-09-05 | New Mexico State | W | 61 | 13 | -34.0 | 14.0 | yes | NMSUvsUF |

| Game | DATE | OPPONENT | RESULT | UF | OP | SPREAD | COVER | BEAT | HASHTAG |
|---|---|---|---|---|---|---|---|---|---|
| Game2 | 2015-09-12 | East Carolina | W | 31 | 24 | -20.5 | -13.5 | no | ECUvsUF |
| Game3 | 2015-09-19 | Kentucky | W | 14 | 9 | -3.5 | 1.5 | yes | UFvsUK |
| Game4 | 2015-09-26 | Tennessee | W | 28 | 27 | 1.0 | 2.0 | yes | TENNvsUF |
| Game5 | 2015-10-03 | Ole Miss | W | 38 | 10 | 6.5 | 34.5 | yes | MISSvsUF |
| Game6 | 2015-10-10 | Missouri | W | 21 | 3 | 11.5 | 29.5 | yes | UFvsMIZZ |
| Game7 | 2015-10-17 | LSU | L | 28 | 35 | 6.0 | -1.0 | no | UFvsLSU |
| Game8 | 2015-10-31 | Georgia | W | 27 | 3 | -1.5 | 22.5 | yes | UFvsUGA |
| Game9 | 2015-11-07 | Vanderbilt | W | 9 | 7 | -21.0 | -19.0 | no | VANDYvsUF |
| Game10 | 2015-11-14 | South Carolina | W | 24 | 14 | -7.0 | 3.0 | yes | UFvsSC |
| Game11 | 2015-11-21 | Florida Atlantic | W | 20 | 14 | -28.5 | -22.5 | no | FAUvsUF |
| Game12 | 2015-11-28 | Florida State | L | 2 | 27 | 2.5 | -22.5 | no | FSUvsUF |

### Creating the Full Data Set

The steps used to create the dataframe that was used for analysis are as follows.

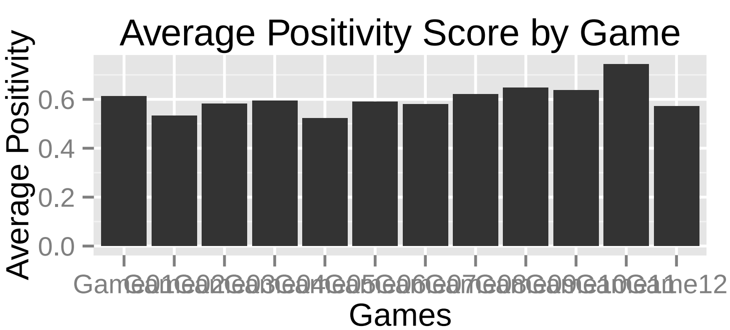
1. Create dataframe `Sentiment` from original `GameTweets` dataframe.
2. Add in the results from the `classify_polarity` analysis.
3. Add in the results from the `classify_emotion` analysis.
4. Join game results data from `Sched` frame (Table2 above).

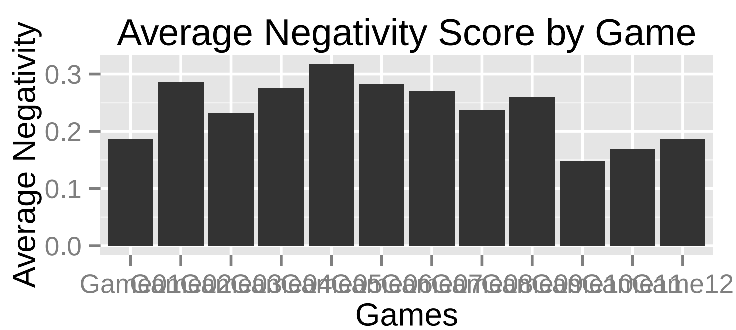
The resulting dataframe contains the following fields:

```
##  [1] "Game"       "screenName" "date"       "text"       "POS"
##  [6] "NEG"        "POS/NEG"    "SBEST_FIT"  "ANGER"      "DISGUST"
## [11] "FEAR"       "JOY"        "SADNESS"    "SURPRISE"   "EBEST_FIT"
## [16] "GAMEDATE"   "OPPONENT"   "H.A"        "RESULT"     "UFSCORE"
## [21] "OSCORE"     "MARGIN"     "SPREAD"     "COVER"      "BEAT"
## [26] "HASHTAG"
```

### Polarity or Subjectivitiy Analysis

The initial analysis of polarity or subjectivity, using the `BEST_FIT` results do show some correlation to game results as is shown in the Figure 1: Average Positivity Score by Game and Figure 2: Average Negativity Score by Game.




### Beat the Spread Model

The data was further analyzed using the `glm` (generalized liner model) function to create a predictor model for each of the scenarios below. These include:

- WIN - Did the Gators win?

- WINPlus - Did the Gators win by greater than 7 points?
- WINLess - Did they loose by more than 7 points?
- BEAT - Did they beat the spread is included in out game schedule above?

Displayed below are the results from the Beat the Spread Model as displayed using the `exp(coef())` function and `confint()` functions. All of the coefficients are statistically significant except for fear. Even thought the results are statistically significant, I don't feel this is a valuable model. One issue is that this model shows that whether or not the `BEST_FIT` subjectivity is positive or negative, it adds positively to the model. Intuitively, it would seem that a negative `BEST_FIT` should add negatively to the model.

```
##       (Intercept)  SBEST_FITneutral SBEST_FITpositive  EBEST_FITdisgust
##        17.0111666         0.3959834         0.5290705         0.2731225
##     EBEST_FITfear      EBEST_FITjoy  EBEST_FITsadness EBEST_FITsurprise
##         0.6439475         0.1635192         0.3146303         0.3795191
```

```
##                       2.5 %     97.5 %
## (Intercept)        2.486517  3.2056609
## SBEST_FITneutral  -1.135635 -0.7199774
## SBEST_FITpositive -0.836105 -0.4406506
## EBEST_FITdisgust  -2.040539 -0.4975689
## EBEST_FITfear     -1.176291  0.3919628
## EBEST_FITjoy      -2.154602 -1.4937308
## EBEST_FITsadness  -1.573401 -0.7537842
## EBEST_FITsurprise -1.416271 -0.5279942
```

### University of Florida


Figure 1:

**Results:** A lot more work in cleaning up the text file is needed before this wordcloud accurately represents the text content in the tweets database. I've included it here because I think it creates a fun introduction to the project and it helped me to work on developing my text analysis skills.