2012 NFL Season: Sentiment Analysis of Game Outcomes and Predicting Wins Using Fans’ Tweets

Courtney Peterson, Erin Berg, John Whalen, Keith Carroll
Boston College

Intro

- Predicting accurate game outcomes is desired ability
- Powerful connection between teams & fans
- Explored connection between fans and their teams, and if teams’ fans can predict game outcomes
- Used post-game fan tweets during 2012 season
- Hypothesis: fans’ tweets are more likely to be positive after wins and negative after losses, and post-game tweets can depict game outcome

Goals

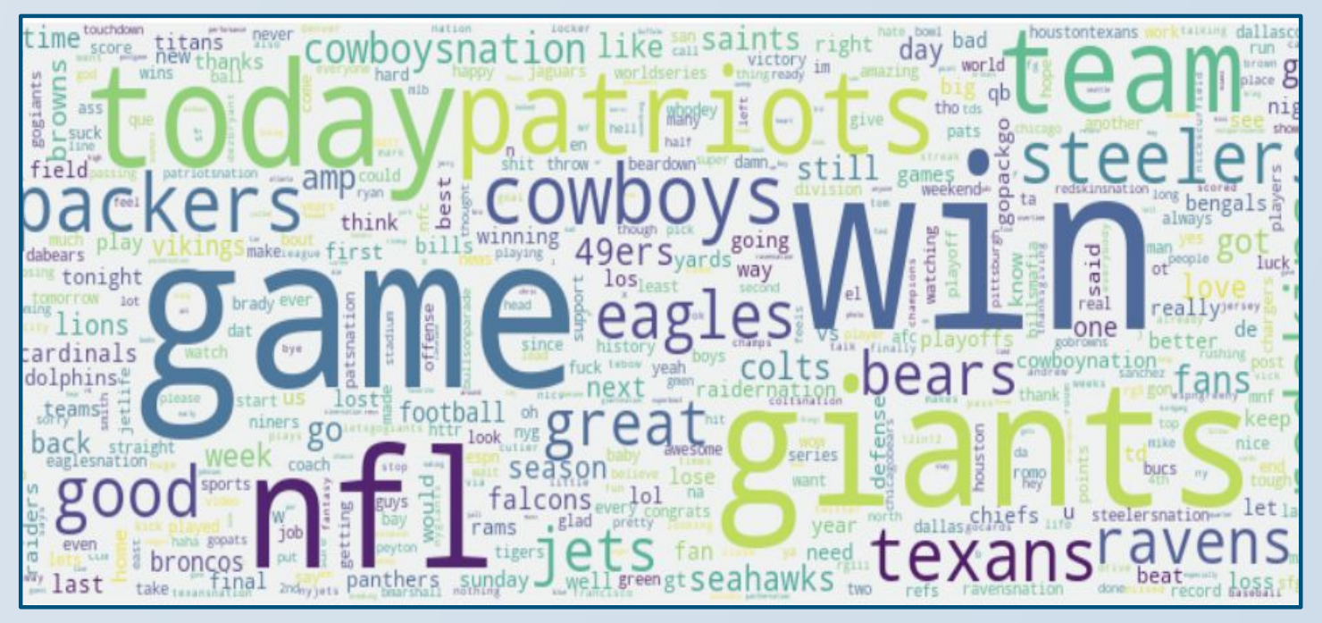
1. **Analysis:** Find connection between fan sentiment and the game outcome
2. **Classification:** Predict who ended up having won based on the tweets after the game

Data

NFL Tweet Dataset

- Tweets were assigned to a team if they contained hashtags corresponding to exactly one team
- Tweets from 256 games and 32 teams
- Tweets made up of 1 hour before the start of the game and 4-28 hours after
- Total of 75,294 tweets (Hydrated with Hydrator)
- Added label of 1 or 0 to instances from given scores

great start to a wonderful season 😏 #CowboysNation	1
Pablo Sandoval just told McDonald to feel his belly #believeinyourself #SFGiants #Giants #PANDAMODE #burleytubaccah	0



Word cloud for all tweets

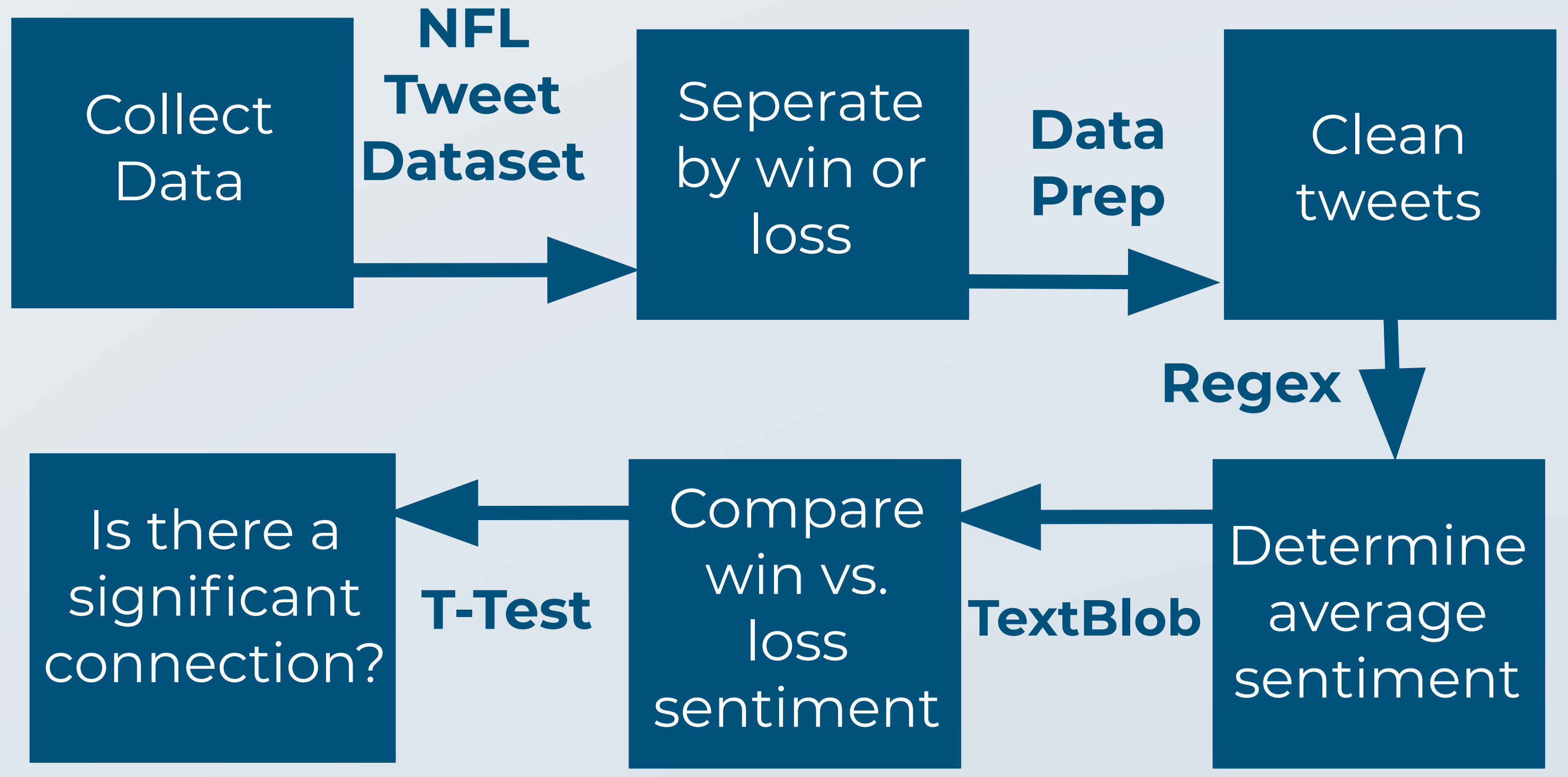
Patriot Hashtags

#patriots #newenglandpatriots #pats #nepats #nepatriots #gopats #letsgopats #gopatsgo #gopatriots #letsgopatriots #gopatriotsgo #patriotnation #patriotsnation #patsnation

Method

Analysis

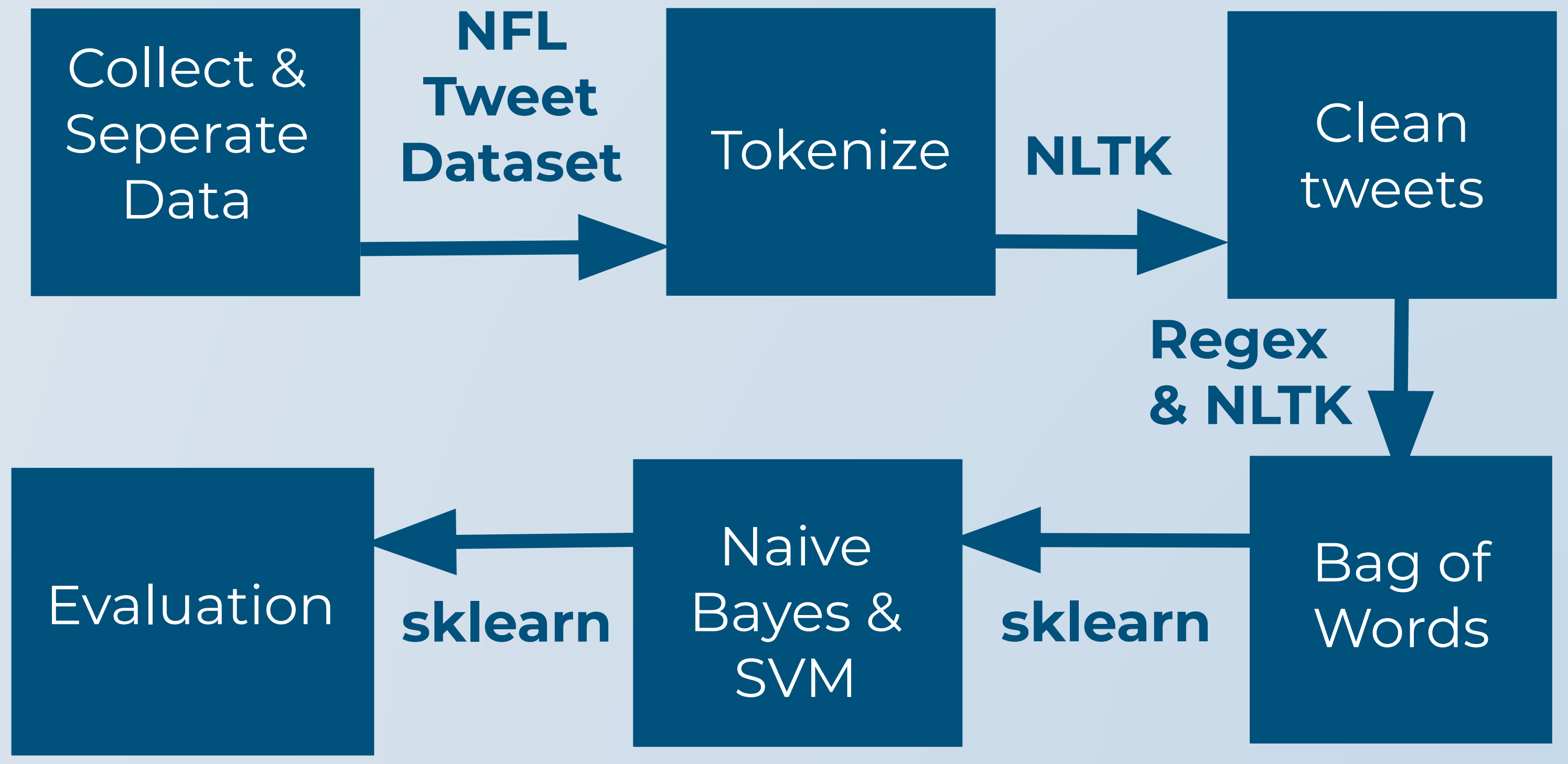
- Pre-Processing Text: Removed punctuation and hyperlinks
- Used TextBlob to determine sentiment of tweets after games won and games lost
- Used T-Test to determine if differences in sentiment after win vs. loss are significant



Collect Data → NFL Tweet Dataset → Seperate by win or loss → Data Prep → Clean tweets → Regex → Determine average sentiment → TextBlob → Compare win vs. loss sentiment → T-Test → Is there a significant connection?

Classification

- Created X & y sets by extracting the tweet and associated outcome
- Word Tokenization
- Additional Pre-Processing Text: Removed stop words and downcased
- Feature Engineering: Bag of Words
- Model: Naive Bayes & SVM
  - 30% of dataset for testing & 70% for training
- Model Evaluation: Accuracy and F1-Score



Collect & Seperate Data → NFL Tweet Dataset → Tokenize → NLTK → Clean tweets → Regex & NLTK → Bag of Words → sklearn → Naive Bayes & SVM → sklearn → Evaluation

Results

Analysis

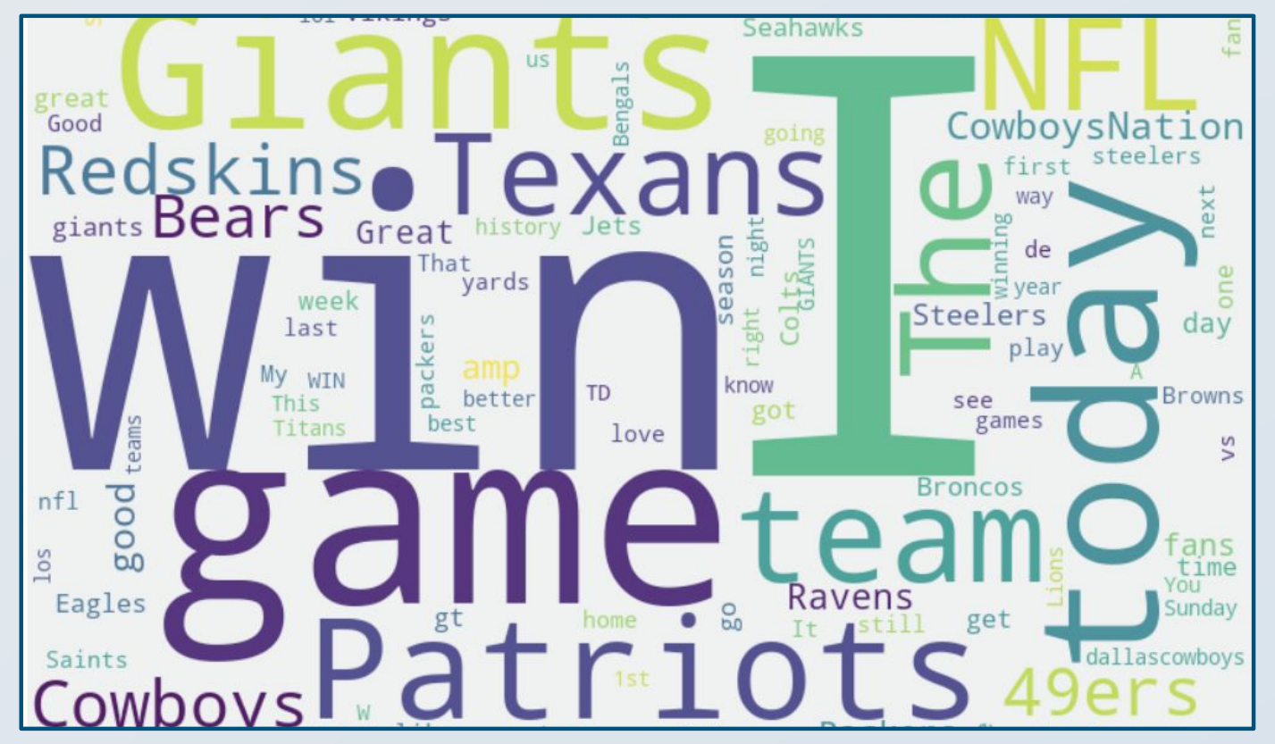
- Average Polarity for Wins: 0.157 & Losses: 0.052
- P-Value: 0.0, T-value : 49.281
  - P-value indicates we can conclude won games lead to more positive tweets then those after lost games

And really #Dbacks Goldschmidt is mad enough as it is since #Giants Tim Lincecum is not pitching in this series.
Polarity: -.14
Lost Game

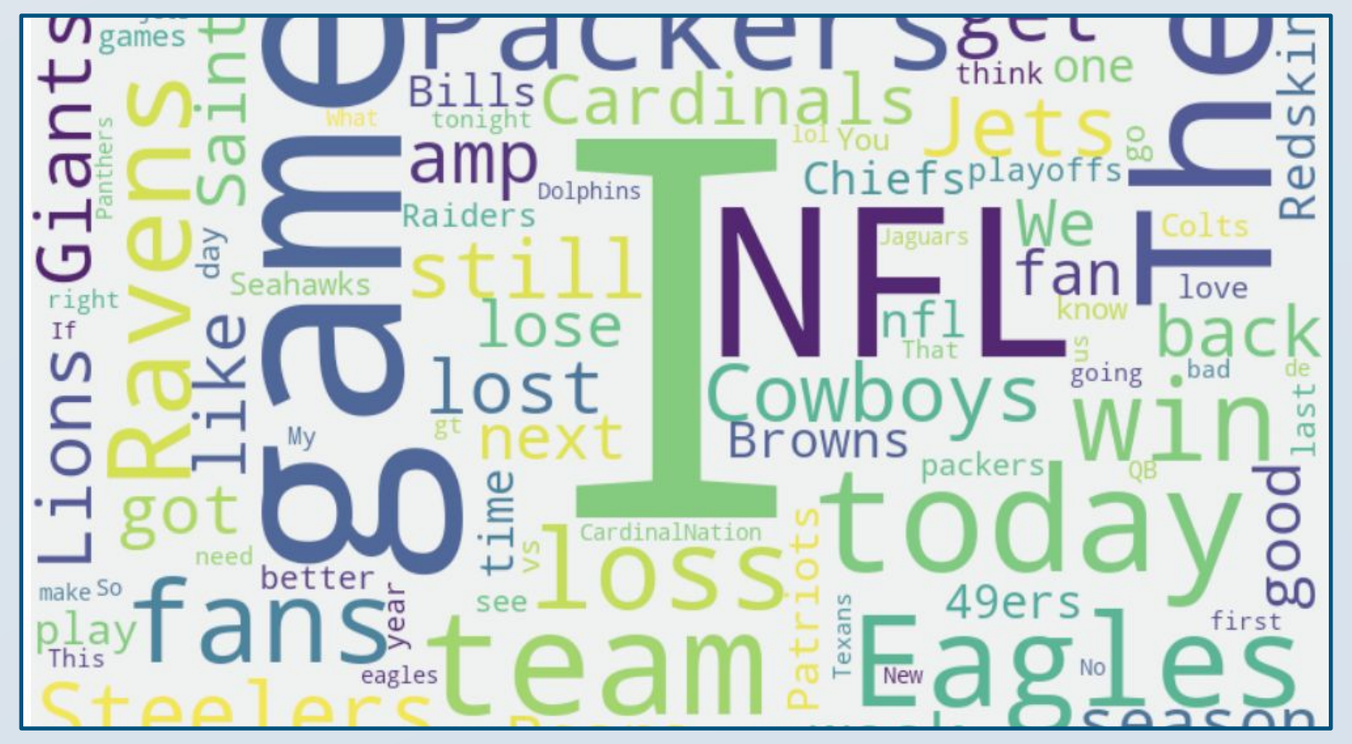
#Cowboysnation @dallascowboys
First win of the season baby!!!!!
Finally after so many games we finally beat the Giants!!!!
Polarity: .3
Won Game

Classification

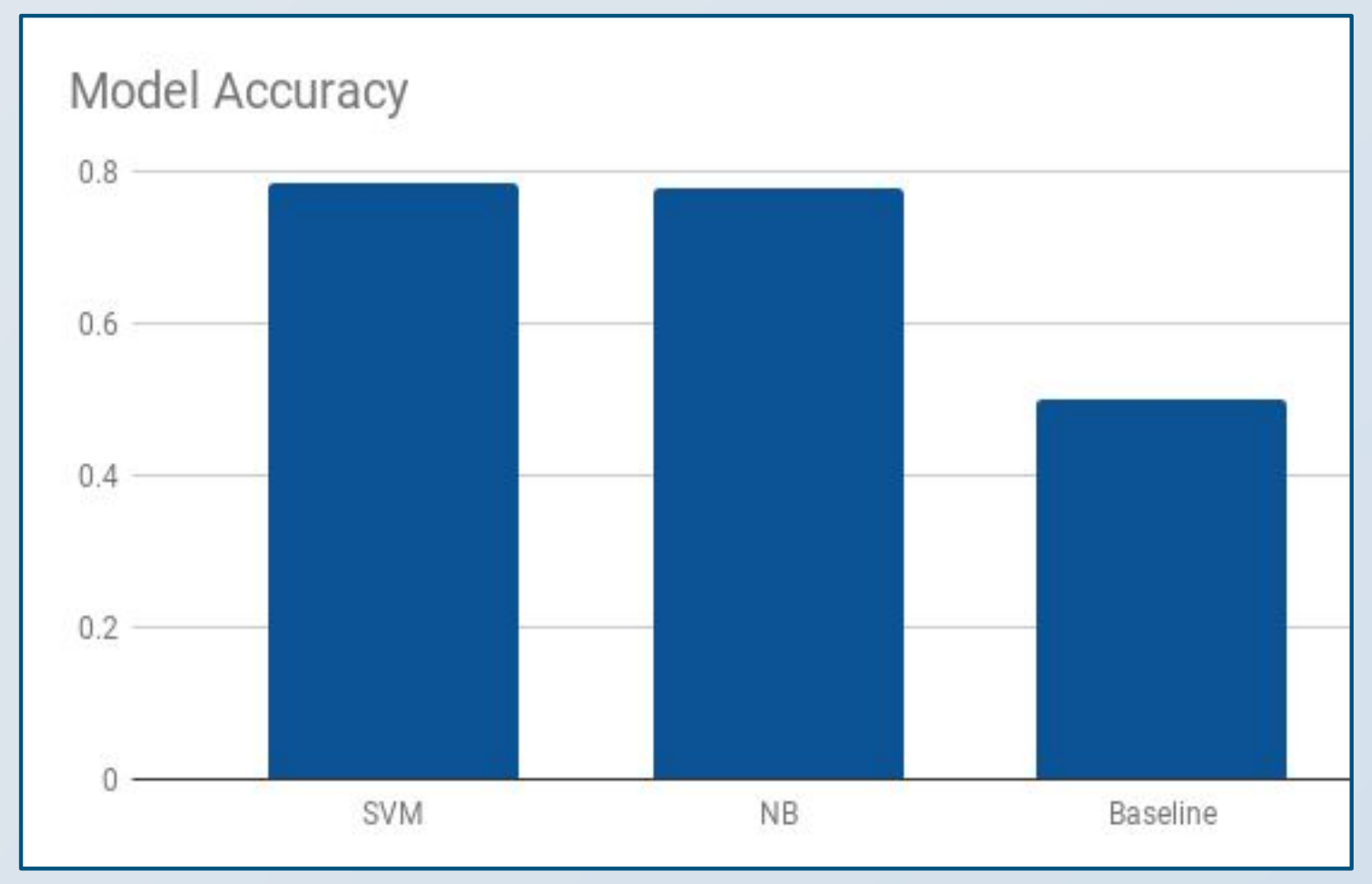
- Accuracy of 0.7769 for NB & 0.7861 for SVM
- Baseline: 50%
  - Equal # of tweets from lost and won games

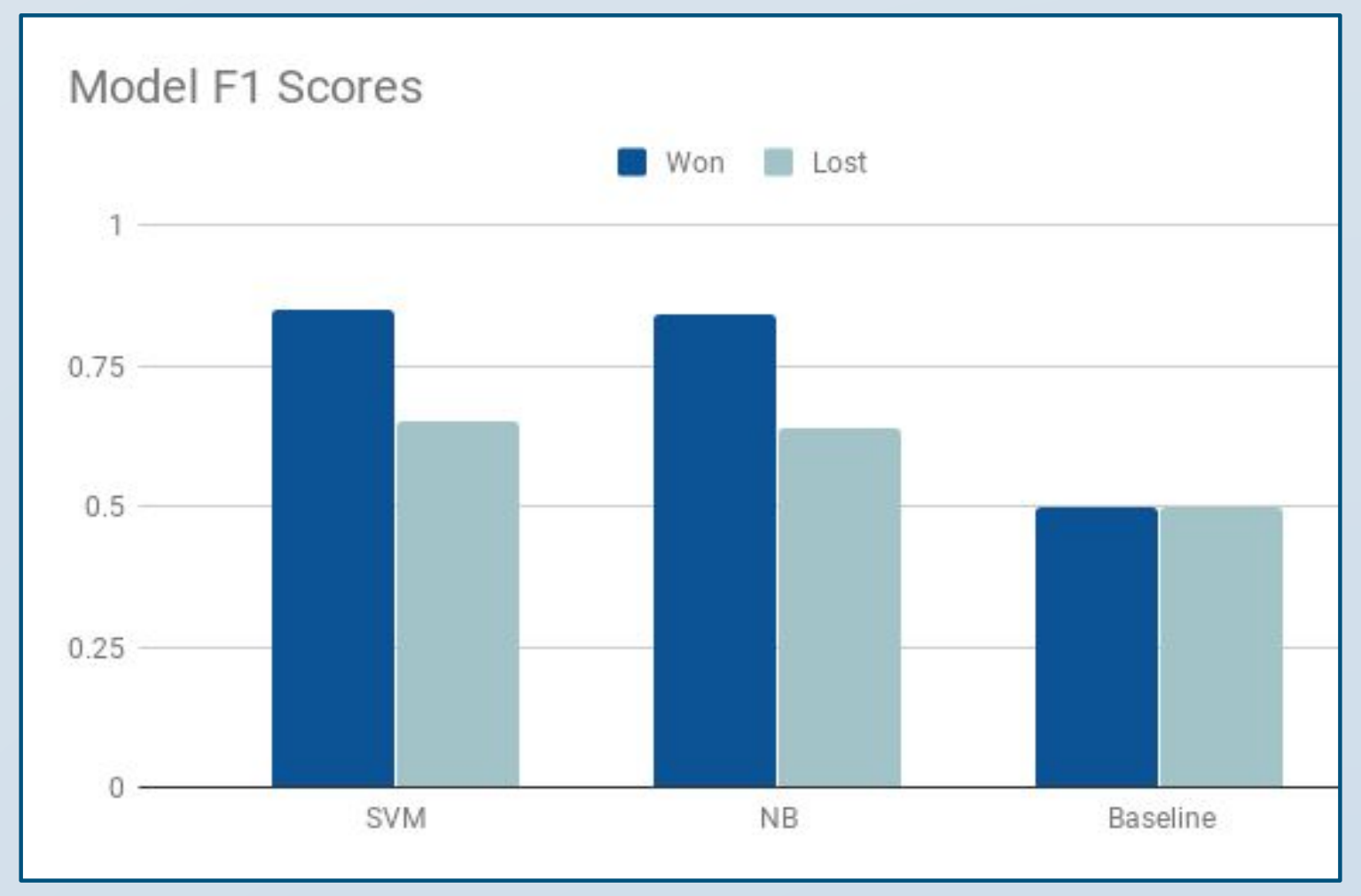


Word cloud for tweets after won game



Word cloud for tweets after lost game





Conclusion

- Best Model: SVM by .0092
- We can in fact predict game outcomes
- Limitations
  - Determining team has its complications
    - Ex: Tweeter of “Sucks to be a #Giants fan” is wrongly categorized as a Giants fan
  - Assuming Sentiment Analysis tool is correct
- Future Steps
  - Filtering by location as well as hashtags
  - Cross-Validation
  - Results from before and during game

Libraries, Applications, & Datasets Implemented: NFL Tweet Dataset DocNow/hydrator TextBlob pandas NLTK scikit-learn

GitHub Repository: https://github.com/courtneypeterson56/NFL-SentimentAnalyzer