Predicting an Election from Tweets

Michaël Juillard, Mikhail Vorobiev, Chiara Ercolani

A Network Tour of Data Science

18th January 2017

Project Description

- Idea : Twitter = opinions
- Goal : Predict election results with Tweets
- Dataset : Tweets about the US Senate Election 2016
- Workflow :
  - Gather tweets about candidates
  - Perform Sentiment analysis
  - Train a machine learning algorithm for prediction

Web Scraping

- Data mining from Twitter
- Time frame : two months before the elections
- Twitter API limitation
- Data stored in .csv files

Data Analysis I
Sentiment Analysis

- Sentiment analysis tool : Pattern
- Natural language processing on English language
- Returns polarity and subjectivity values

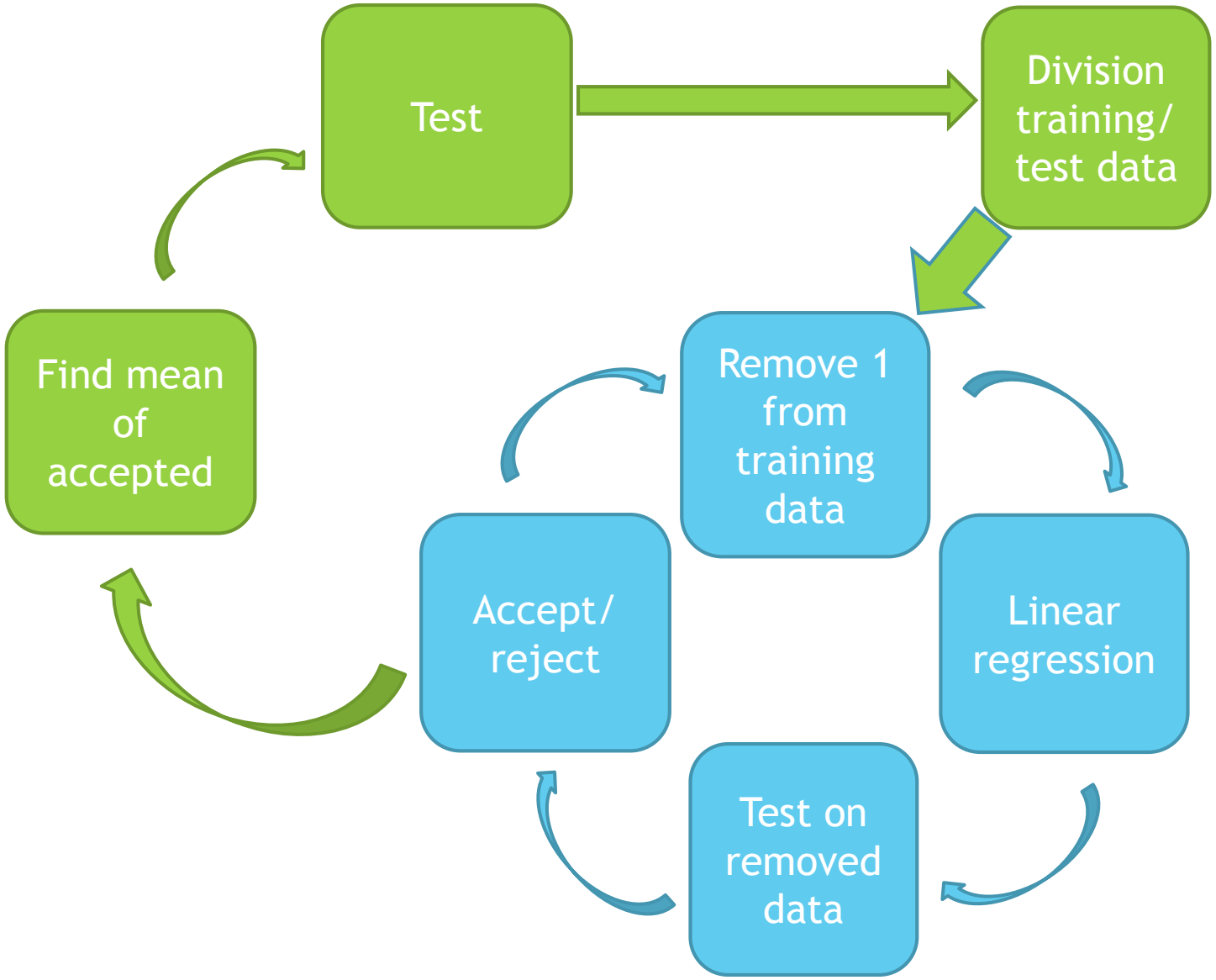
Data Analysis II
Mean and Unique User Identification

- Average computed on the polarity values during the same day
- Average was weighted with number of likes and retweets
- Identification of the number of unique authors of tweets about a candidate

Data pre-processing

- Adding extra-features
- 65 features for 38 datasets => need for reduction
- Fitting of mean values with AR model
- Test multiple orders to find best prediction

Machine Learning



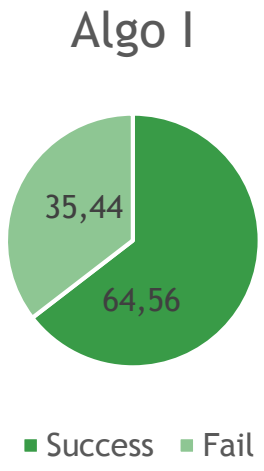
Algorithm I and II

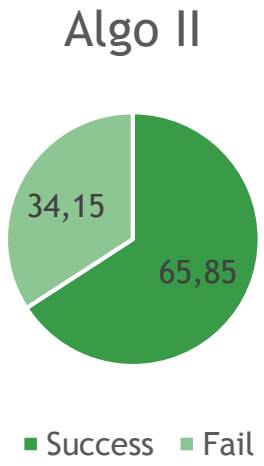
- Candidates considered individually (38 data sets)
- 30 for training, 8 for testing
- Acceptance if both prediction and true data >/< 50%

- Candidates considered by pairs won-lost (19 pairs)
- 15 pairs for training, 4 for testing
- Acceptance if prediction of win/lose correct

Result analysis

- Maximum accuracy :
  - I : 64.5% => 22 features
  - II : 65.8% => 24 features
- Individuality vs Pair
- Number of features
- Autoregression advantage
- Number of candidates





Issues

- "*I hate how @Trump denies the work of @Obama*" VS "*I hate how @Obama denies the work of @Trump*"
- Slang and the sentiment analysis
- Popularity of candidates
- Twitter population : a good sample ?
- Twitter population ≠ Voting population
- Message ≠ Opinion

Conclusions

- Poor correlation
- Difficulty of using twitter data
- Natural language processing
- Uncertainty of election