The Use of Aspect-Based Sentiment Analysis on Political Classification & Potential Impacts

HON 451

Hunter Berry

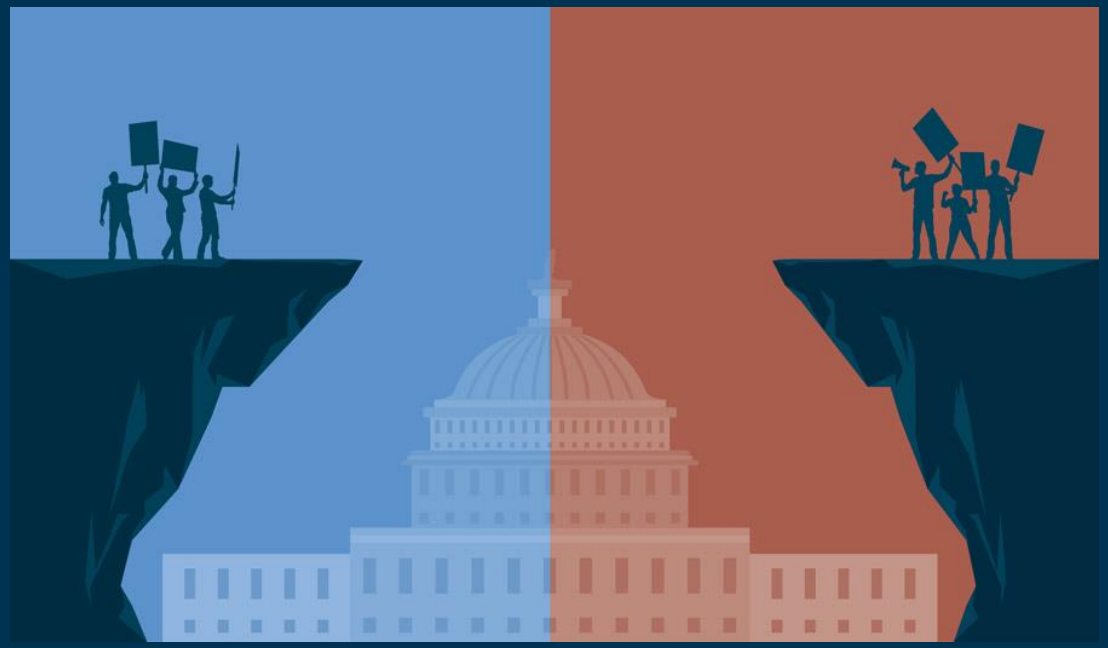
Introduction & Background

“Americans were more ideologically divided than any of the 19 other publics surveyed… These fissures have pervaded nearly every aspect of public and policy response…”

-Michael Dimock, President of the Pew Research Center

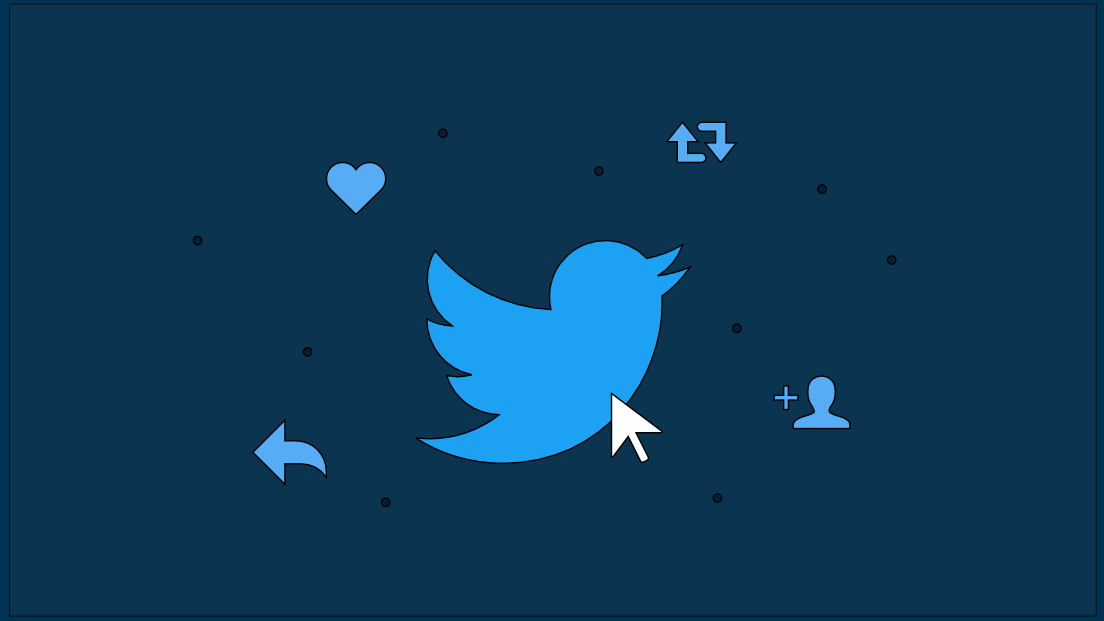
Background

- In the United States, political party is one of the most important "classifications" that belong to a person.
  - Decisions based around party, including where we shop and who we interact with.
- With a growing partisan divide, we've seen many issues in the United States grow more and more violent and discriminatory, using party as the dividing line.
  - BLM / Defund the Police
  - Election / Capital Riots
  - COVID-19 / Coronavirus



Background Cont.

- Twitter has become one of the “hotspots” for political users, using the platform as:
  - a source a news,
  - a source of debate/argumentation, and
  - a source of propaganda/advertising.



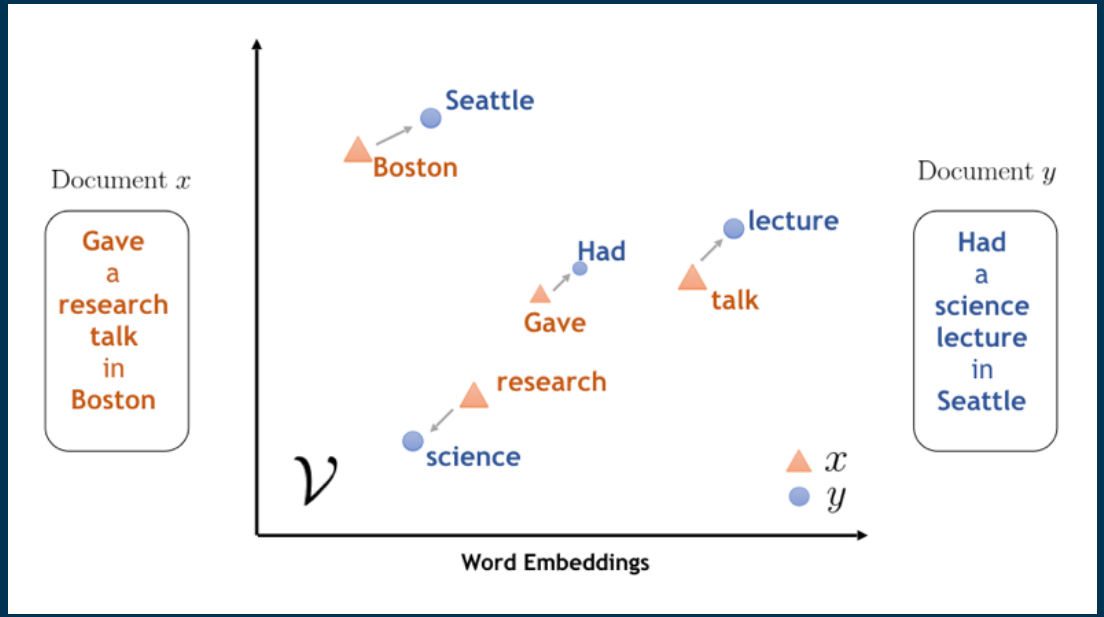
Goals / Overview

- I seek to determine if Twitter can be used to identify someone's political party.
  - If so, what does this mean for the world, and are the results good or bad?
  - What can we do to prevent misuse of this technology?
- Presentation Outline:
  - Data Definitions & Literature Review
  - Methodology & Implementation of Models
  - Results
  - Future Work & Complications
  - Implications
  - Final Thoughts

Definitions & Literature Review

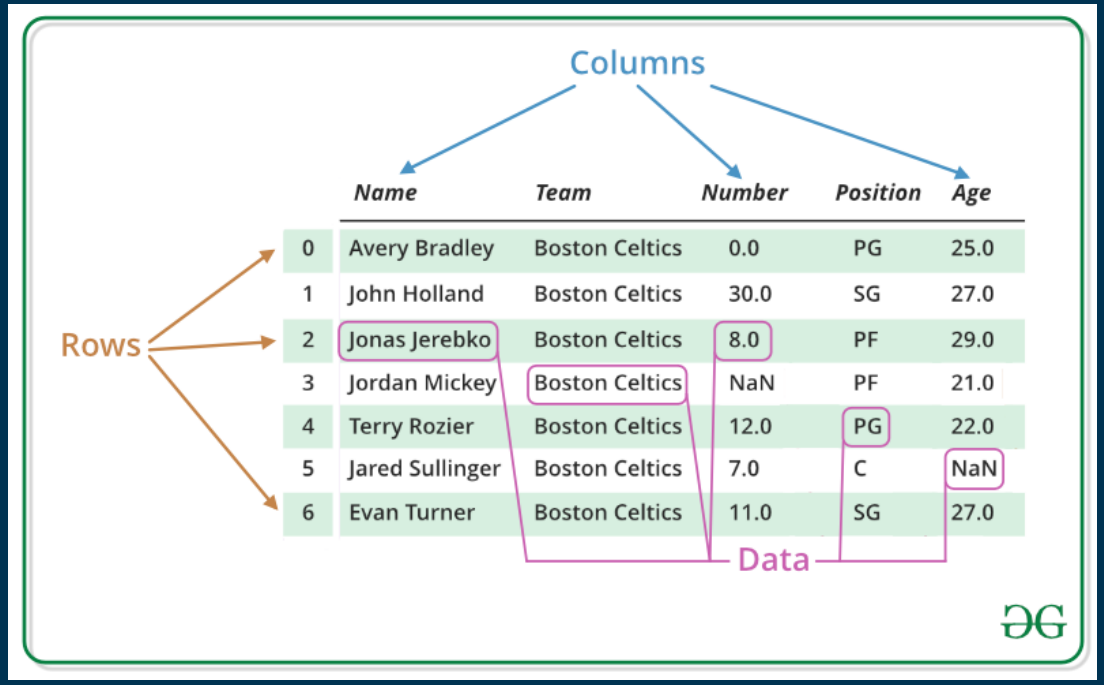
Misc. Terms and Definitions

- Features
  - The data that allows us to make a prediction. Think the “X” or “independent variable.” We have lots of them in our data – things like sentiment towards “national security” or the number of times a user used the word “horrendous.”
- DataFrame
  - Similar to a CSV/Excel spreadsheet. Data is separated into rows and columns, each of which can be filtered and queried.
- Vectorization
  - Mapping words and phrases to vectors of real numbers. Allows for comparison of different words and phrases.
- Bag-of-Words
  - Similar to a frequency table, a bag-of-words is a dataset of words in their simplest form. We count the words used overall and the words used in a specific instance to analyze frequency.



Terms & Definitions Cont.

- Supervised Learning
  - Most common type of method for classification problems. Map input to output labels or map input to a continuous output.
- Unsupervised Learning
  - Unsupervised learning is a type of machine learning where the machine is not given any specific labels and is told to classify and section the data as it sees fit.
- Labels
  - Labels are the things we want to classify. In the case of this study, we have two labels – Republican (R) and Democrat (D).



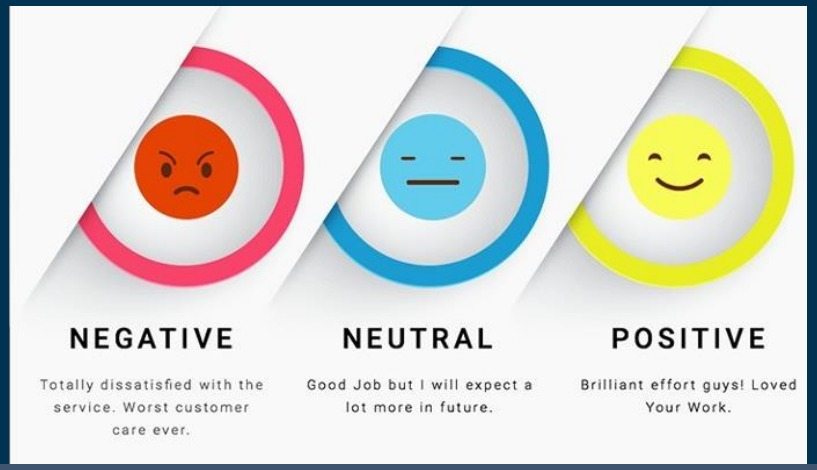
Naïve Bayes Classifier

- Naïve Bayes is a probabilistic method of classification.
- $P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$
  - A: Label
  - B: Feature
- Probability of a certain classification being given when provided a set of features.
- Extremely Successful in Recent Studies
  - 2016 Presidential Election – Short Term Events, ~75% Accuracy (Ding 2017)
  - Indonesian Election – 76% Accuracy (Hasan et. al 2018).

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

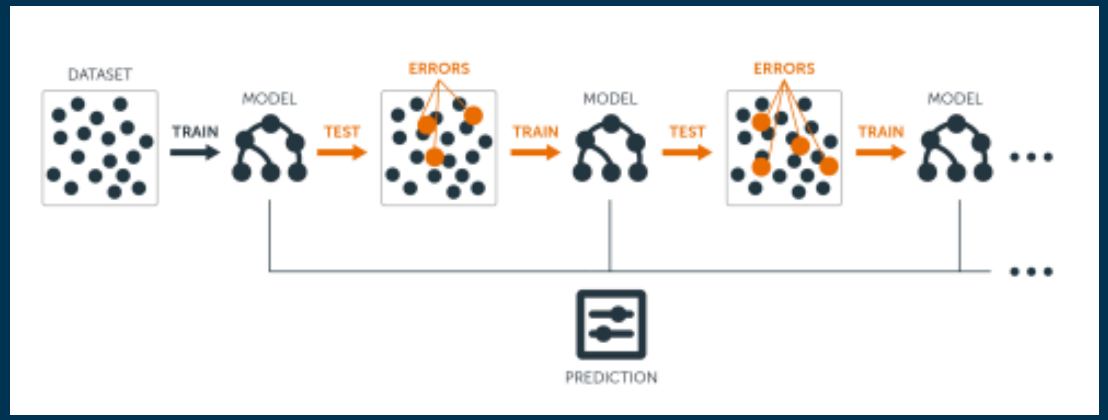
Sentiment Analysis

- Uses the terms and phrases in each sentence or set of words to determine the “sentiment” or opinion.
- Can be a little simplistic, hence why aspect-based sentiment analysis (ABSA) is often used.
  - Sentiment towards a specific topic or idea.
- Promising Results
  - 2011 Irish Elections – 60% SA Accuracy
  - Zainuddin Twitter Study – 70% ABSA Accuracy



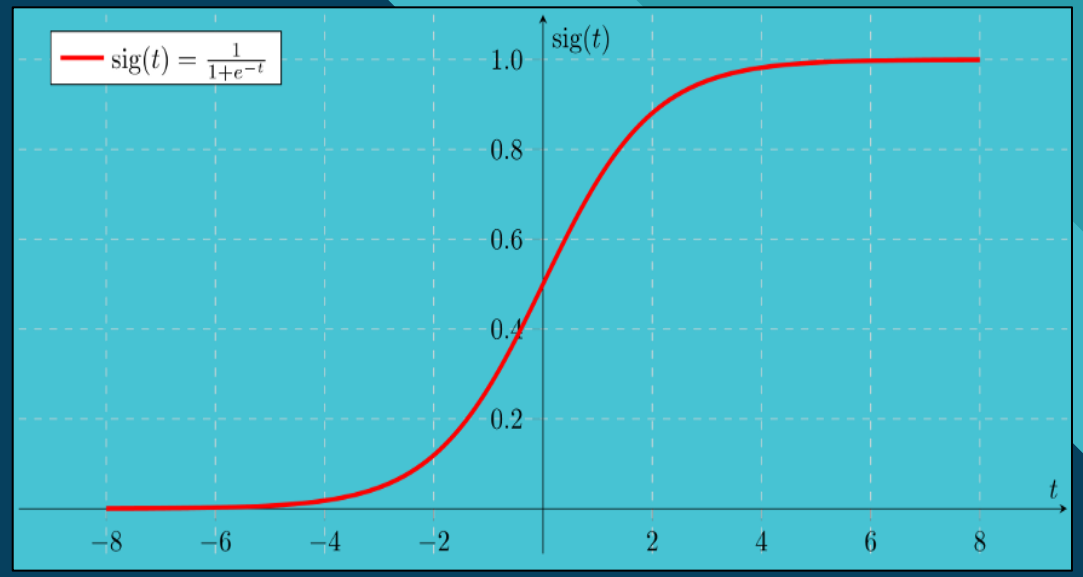
Gradient Boosting Models

- Works on an additive design process, where each step is an attempt at an improvement on the previous step.
  - Use errors from layer x to adjust the weights of values for trees in layer x + 1. Some trees are given less weight, some more.
  - Process repeats until accuracy no longer improves after weight adjustments.
- Top Kaggle Model – XGBoost Regressor
  - Easy to use, setup, and adjust.
  - Multi-faceted, can cover a wide variety of areas.
  - Dwivedi 2020 & Morde 2019



Logistic Regressions

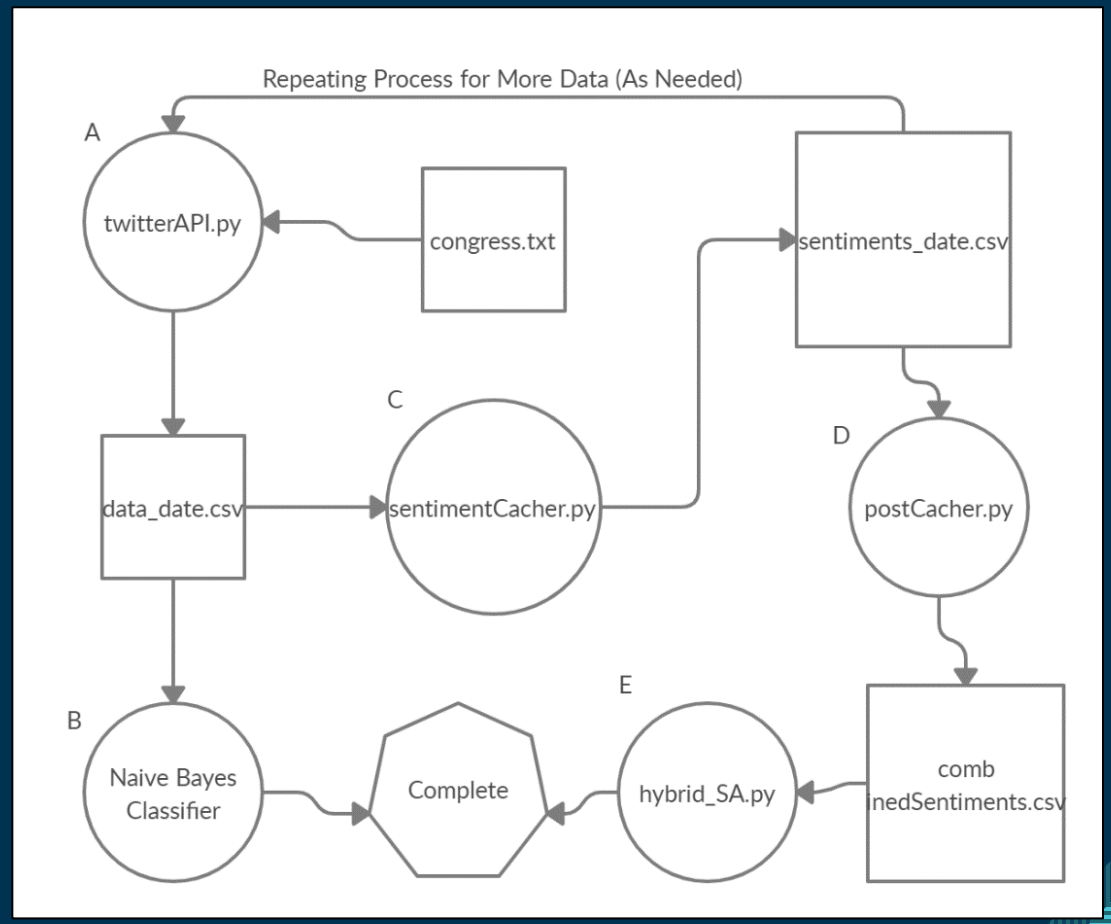
- Similar to Naïve Bayes:
  - Takes in a number of features and determines the probability of the label occurring given that set of features.
- Use a complex formula to adjust weights and coefficients associated with specific features (Maximum Likelihood Estimation). Plug in data and get a probability.
  - Convert this probability into a binary decision to get classification.
- Pranckevičius & Marcinkevičius Text Classification Study
  - Beat every other tested supervised learning model by at least 10%.
  - Great results in other ML competitions and studies.



Methods & Implementation

Pipeline Part 1.

- Before discussing methodology and implementation of various models, we do a brief discussion on the pipeline.
- 1st Major Component – *twitterAPI.py*
  - Uses an authorized Twitter development account to pull data on ~500 different users.
    - Congressional figures and other high-level politicians (cabinet members, president, etc.).
  - Gets five tweets per user, and loads these into a DataFrame and CSV file, both of which can be loaded into other models and files (Naïve Bayes, Hybrid ABSA/ML, etc.).



Naïve Bayes

- In terms of our specific problem, our model uses an equation:
  - $P(Party|Feature) = \frac{P(Feature|Party) * P(Party)}{P(Feature)}$
  - Likelihood of a given tweet being of a certain party given any number of features that are associated with the party.
- Implementation
  - Relies on built in Naïve Bayes modeling tools from the TextBlob Python library, which was used in other studies (Hasan et. al, 2018).
  - Analyzes words usage and bag of words/vectorization results to compute party prediction.

```
def chunk(data, mode, classificationS):
    length = len(data);
    curPos = 0;
    classifier = None;
    if classificationS is not None:
        classifier = classificationS;

    if mode == "train":
        while curPos <= length:
            if curPos == 0:
                d = data[0:50]
                for i in d:
                    print(i)
                classifer = NaiveBayesClassifier(data);
                curPos = 50;
            else:
                if curPos + 50 >= length:
                    classifer.update(data[curPos:length]);
                else:
                    classifier.update(data[curPos:curPos + 50])
                curPos = curPos + 50;
                time.sleep(2);
        return classifier;

    elif mode == 'test':
        listOfAccs = [];
        while curPos <= length:
            if curPos + 50 >= length:
                listOfAccs.append(classifier.accuracy(data[curPos:length]));
            else:
                listOfAccs.append(classifier.accuracy(data[curPos:curPos + 50]));
            curPos = curPos + 50;
            time.sleep(2);
        return listOfAccs;
```

Sentiment Analysis

- Goal is to cover as many current events and ongoing political debates as possible.
  - Key is picking topics that have some easily definable “Republican” or conservative side and “Democratic” or liberal side.
- Eight Chosen “Key” Topics
  - Economics
  - Police
  - Foreign Policy & Immigration
  - Presidency and Elections
  - Military
  - Abortion
  - Health/Healthcare

Sentiment Analysis Cont.

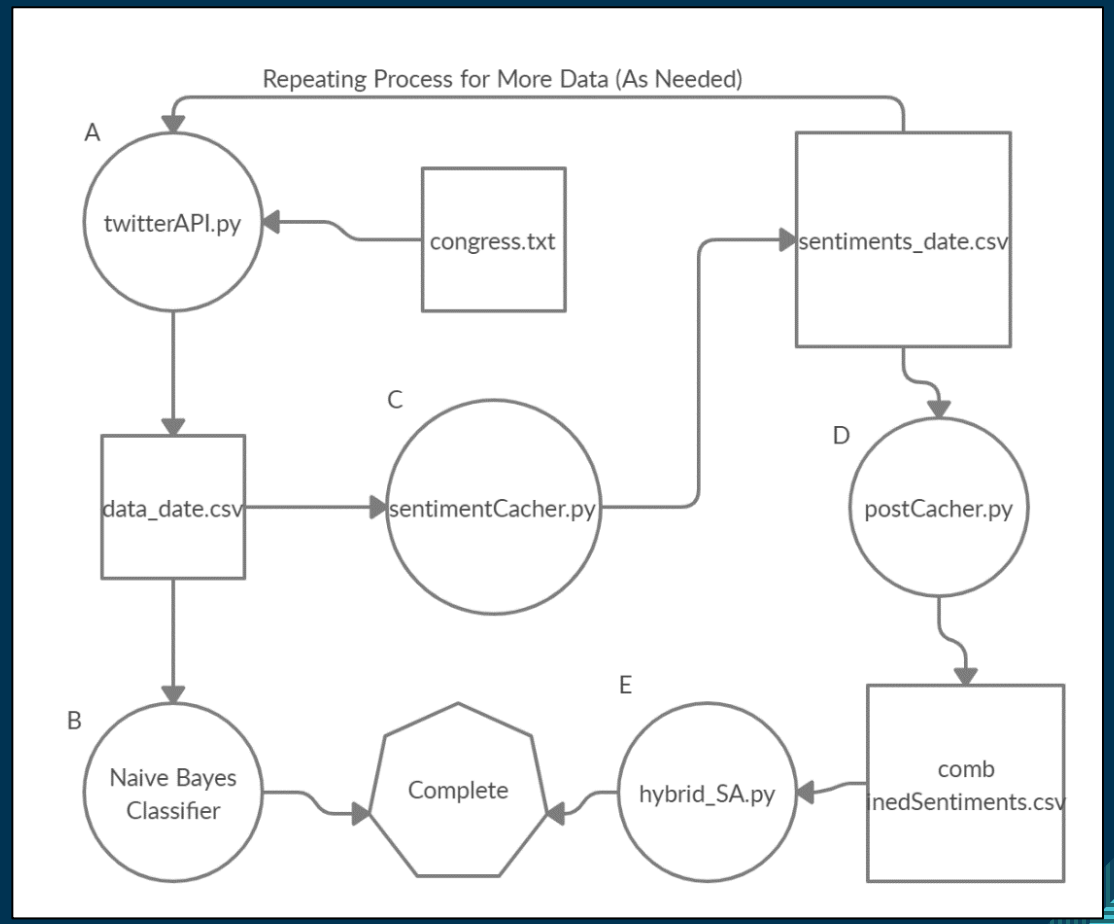
- Each topic is split into several individual aspects.
  - In economics, we want to gather sentiment on several ideas – Chinese trade, GDP, the stock market, natural resources, trade, and much more.
- Implementation
  - Each tweet is parsed and tested on each aspect.
    - If a tweet contains a word or phrase we test for, we analyze sentiment. If not, we say the tweet is “neutral” towards that specific topic.
  - Built on library used in previous studies solely for ABSA.
  - Used a caching system to split parsing process among four computers, dramatically reducing time.

```
results = nlp(row['tweet'], aspects=needToRun);
for term in needToRun:
    if term in row['tweet'].lower():
        if results[term].sentiment == absa.Sentiment.negative:
            data.at[index, term] = 1
        elif results[term].sentiment == absa.Sentiment.positive:
            data.at[index, term] = 2
    else:
        data.at[index, term] = 0
```

```
topics = {
    'economics': ['consumption', 'commerce', 'economics', 'economic', 'trade', 'gdp', 'China', 'investment',
        'stock market', 'stocks', 'stock', 'goods', 'financial', 'fiscal', 'economical', 'profitable',
        'economy', 'efficent', 'finance', 'monetary', 'management', 'economist', 'macroeconomics',
        'microeconomics', 'protectionism', 'resources', 'real value', 'nominal value', 'capital', 'markets'],
    'police': ['blm', 'defund', 'police', 'militiarization', 'police officer', 'sheriff', 'crime',
        'fatal', 'shooting', 'abuse of power', 'line of duty', 'protect', 'protecting', 'patrol',
        'patrolling', 'law enforcement', 'riot', 'looting', 'arrest', 'racism', 'law', 'black lives matter'],
    'foreign_policy': ['imperialism', 'occupation', 'un', 'united nations', 'united', 'hrc',
        'who', 'free trade', 'anarchy', 'nationalism', 'foreign', 'china', 'russia', 'cuba',
        'multinational', 'regional', 'trade', 'international', 'commerce', 'alien', 'refugee',
        'border', 'ambassador', 'israel', 'pakistan', 'terrorism'],
    "immigration": ['alien', 'wall', 'emigration', 'immigration', 'migration', 'illegal alien',
        'naturalization', 'visa', 'citizenship', 'refugee', 'welfare', 'family reunification', 'border',
        'immigrants', 'enforcement', 'seperation', 'asylum', 'sanctuary city', 'sancuary cities'],
    "president": ['trump', 'president', 'genius', 'smart', 'incompatent', 'idiot', 'leadership',
        'cabinet', 'election', 'biden', 'elect', 'harris', 'joe biden', 'donald trump', 'government',
        'corrupt', 'russia', 'head of state', 'presidency'],
    "military": ['military', 'armed forces', 'air force', 'coast guard', 'national guard', 'army',
        'navy', 'marines', 'combat', 'forces', 'invasion', 'occupation', 'overseas', 'over seas',
        'foreign policy', 'defense', 'intelligence', 'military intelligence', 'militaristic', 'militia',
        'peacekeeping', 'occupy', 'regiment', 'noncombatant', 'naval'],
    "abortion": ['abortion', 'birth control', 'contraceptives', 'condoms', 'abortion laws', 'feticide',
        'abortion clinic', 'pro-choice', 'pro choice', 'prochoice', 'abortion pill', 'trimester',
        'first trimester', 'planned parenthood'],
    'health': ['healthcare', 'ppe', 'personal protective equipment', 'covid', 'health care', 'health', 'coronavirus', 'covid-19']
}
```

Hybrid Approach (Regressions and Boosting Trees)

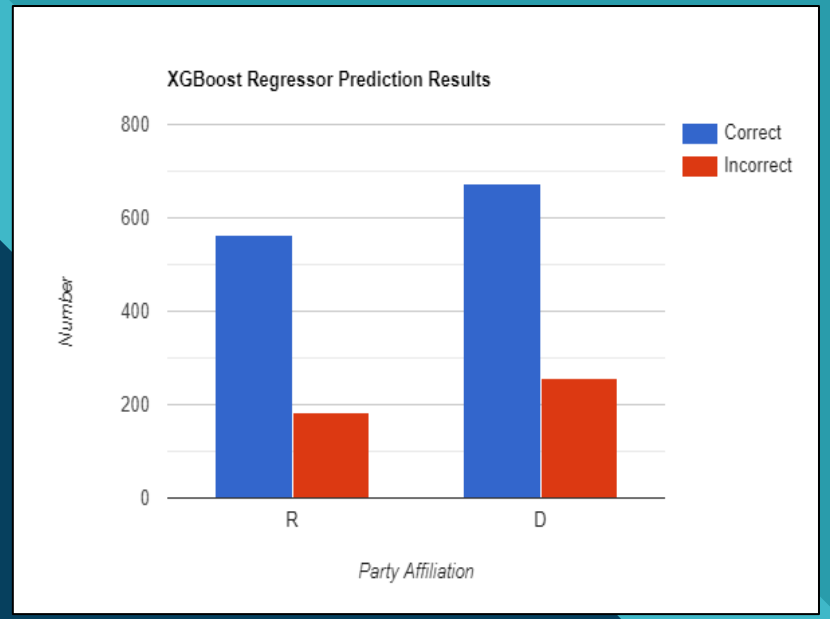
- Data is loaded in and passed through a vectorizer, with tweets being turned into “bags of words.”
- These results are concatenated with the data from the sentiment analysis chain and combined into training and testing datasets.
- Models loaded from the XGBoost library, which has the “best” tuned XGB Regression model, and Sci-Kit Learn, the most popular and well known Python library for data science.

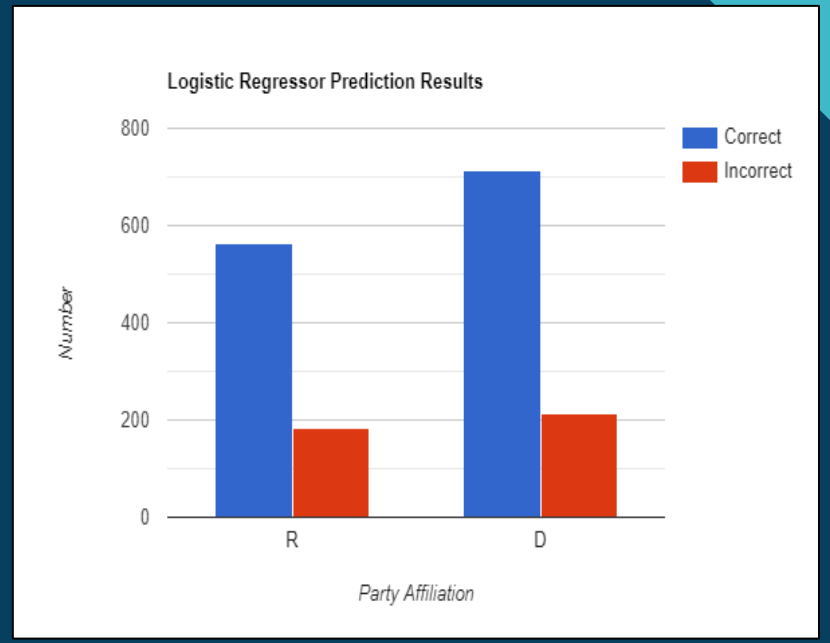


Results & Thoughts

Results

- Naïve Bayes
  - Smaller training and testing dataset than other models.
  - ~70% accuracy based on tweet text and party alone.
- XGBoost Regressor
  - Larger training dataset by ~3,500 tweets, but slightly smaller (~700) testing dataset.
  - 73.77% Accuracy Level
- Logistic Regressor
  - Same datasets as XGBoost model.
  - 76.21% Accuracy Level
  - Higher accuracy a result of better predictions when looking at Democrats, while both models had similar accuracies for Republicans.





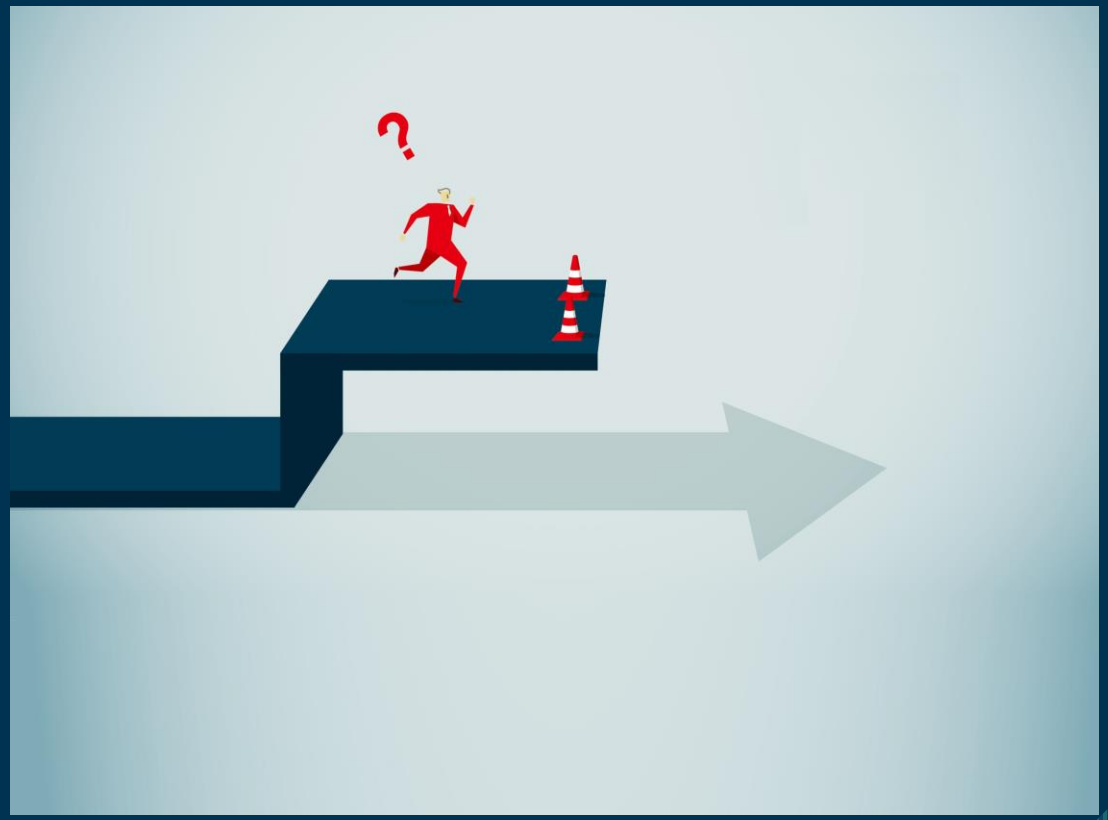
Interface

- Created to allow for duplication of results and testing on custom/newer sets of data.
- Two “Key” Functions:
  - Pipeline
    - Steps through the entire pipeline process.
    - Creates new logistic regressors and boosting models based on this data.
  - Testing
    - Allows users to test on “custom” tweets and sentences and live Twitter accounts.
    - Allows users to test on the models used in the results section or use custom models.

```
(base) C:\School Stuff\HON 450-451\hon450>
```

Limitations

- No clear definition on what is “political ideology” and how you determine it.
  - Liquid vs. Plastic
  - Policy
- How do beliefs coincide with party?
- How do we handle people who have beliefs from both sides of the aisle?
- Does one belief have more weight in one’s political identification than another?



Future Work

- Hashtags
  - Great way of getting data on specific topics/events.
  - Easy to connect tweets and compare sentiments, words/phrases used, etc.
- Language
  - ABSA couldn’t work on words/phrases that aren’t English.
  - Trouble with bag-of-words and vectorization comparisons from English to foreign languages.
  - However, useful in identifying polarity.
- Other Models
  - Unsupervised Learning
  - Hidden Markov Model (HMM) Chains or Conditional Random Fields (CRFs)

Implications & Ideas

Implications

The Good

- Reductions in Campaign Spending
  - 2020 U.S. Presidential Election - $14,000,000,000 in Campaign Spending
  - Reducing search costs/identification costs for users.
- Determining Bots/Fake Users
  - Typically far-right and far-left users.
- Anti-Radicalization & Fake News
  - Determine common phrases, words, and sentiments, use this to remove/monitor account before they get out of hand.

The Bad

- Discrimination
  - Stanford “Gaydar” Study
  - 2017 ProPublica Report
  - Race & Connection to Party
  - Can easily affect “whether or not individuals get a job, get credit, end up in jail, or even experience violence against them.”
- Radicalization
  - Easy to locate members of a specific party or belief system to rally.
  - U.S. Capital Riots / Summer DTP Riots

Solutions: Data Privacy & Protection

- Twitter API Problems
  - Easy to get access too.
  - Easy to download massive amounts of data.
  - No proof of usage.
- Solutions
  - Accounts Set to Default by Private
    - Protects smaller “ordinary” users, no issues for “content creators” and celebrities.
  - Limiting Developer Access
    - Usage Reports
    - Academia Only

Thank You For Your Help in this Complex Process!

Questions?