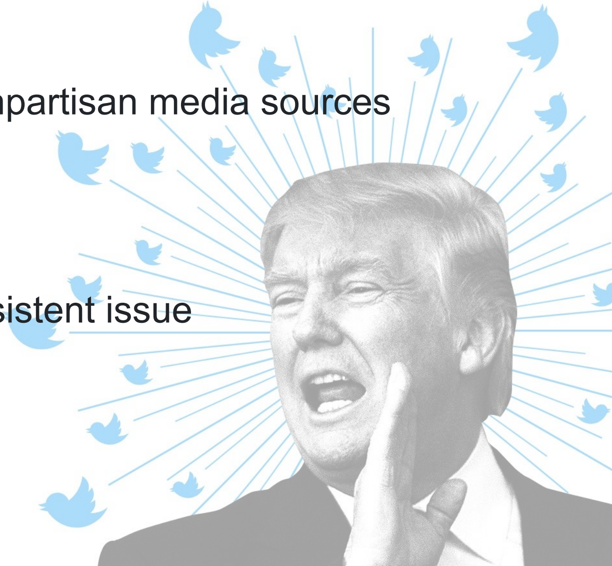
Predicting political affiliation based on tweets

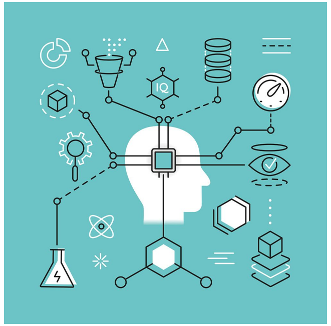
Cooper Chia, Evan Yip, and Walker Azam

Motivation and Background

- How to politicians and candidates interact with the public/voters?
  - Twitter!!
- How can an automated algorithm help inform voters?
  - Gauging potential political bias
  - Political leaning of ‘non-partial’ users.
  - Deciphering political biases of self-proclaimed nonpartisan media sources
  - New politicians
- Why now?
  - 2020 is an election year
  - Fake news through twitter ‘bot-accounts’ as a persistent issue



Research Questions

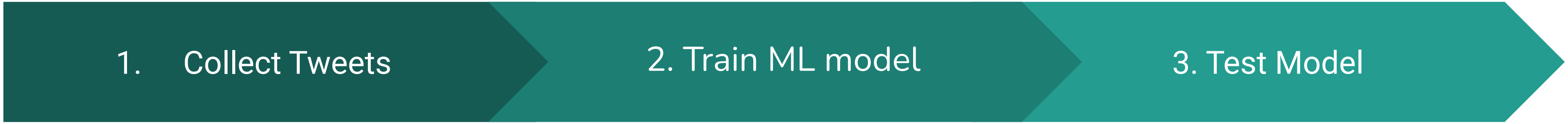


1. Can we use ML to predict political affiliation based on tweets?
   a. Yes, our model was able to predict with high levels of accuracy. ~95%
2. What are the limitations of the Naive Bayes model?
   a. The assumption of independent occurrences of each word in a tweet.
3. What is the best way to visualize the accuracy of our model?
   a. Bar charts and confusion matrix proved to provide a useful visualization of our model.
4. Given the results/accuracy of our model, is the model a suitable tool for determining the political sentiment of other public figures and non-politicians?
   a. Yes, to an extent.

Methodology

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B	= events
$P(A\|B)$	= probability of A given B is true
$P(B\|A)$	= probability of B given A is true
$P(A), P(B)$	= the independent probabilities of A and B



1. Collect Tweets

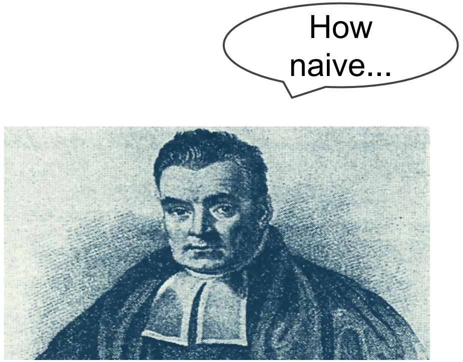
2. Train ML model

3. Test Model

- Existing Data (Kaggle)
- Web-scraping (Beautiful Soup)

- Naive Bayes Algorithm
  - Bayes’ Theorem
  - Assumption of Independence
  - Multiple training/test splits

- Test on Kaggle data and scraped data



Thomas Bayes

Results

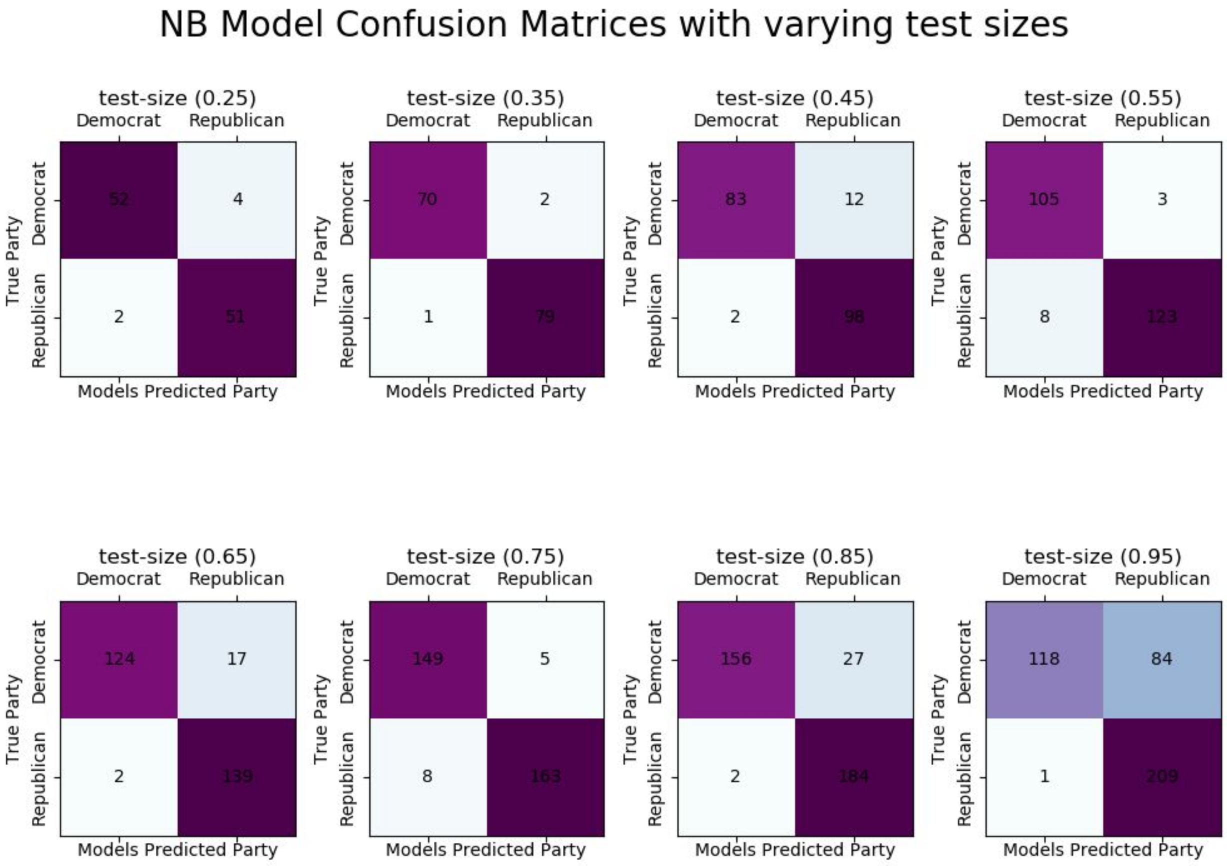
Figure 1: Confusion matrices of the Naive Bayes model applied to test data. High accuracy models correspond to darker values along the diagonal.

Strengths:

- Quantitative values
- Standard visualization

Weaknesses:

- Harder to interpret



Results

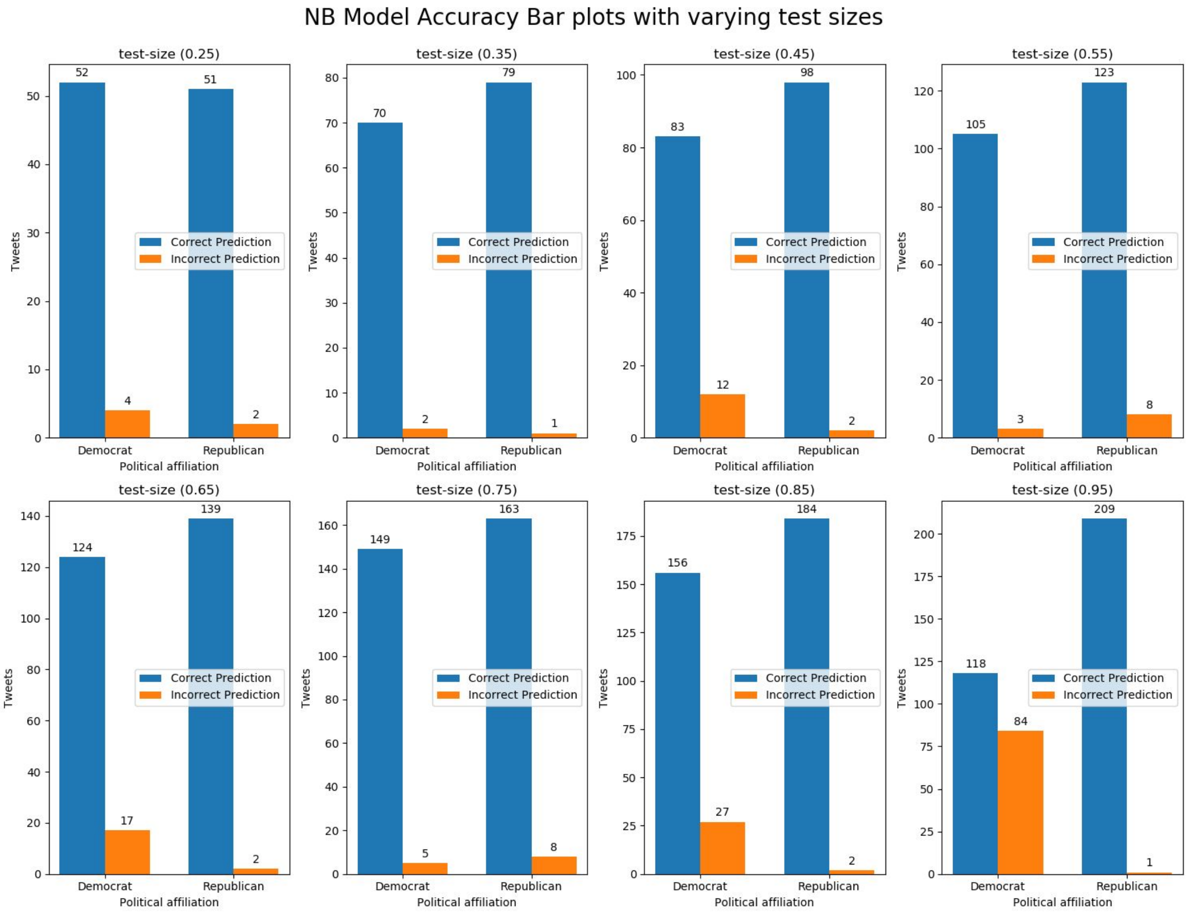
Figure 2: Bar plots of the accuracy of the Naive Bayes model at various test sizes.

Strengths:

- Conveys scale of accuracy with size
- Highlights differences between Dem vs Rep

Weaknesses:

- Hard to compare relative changes in accuracy



Results

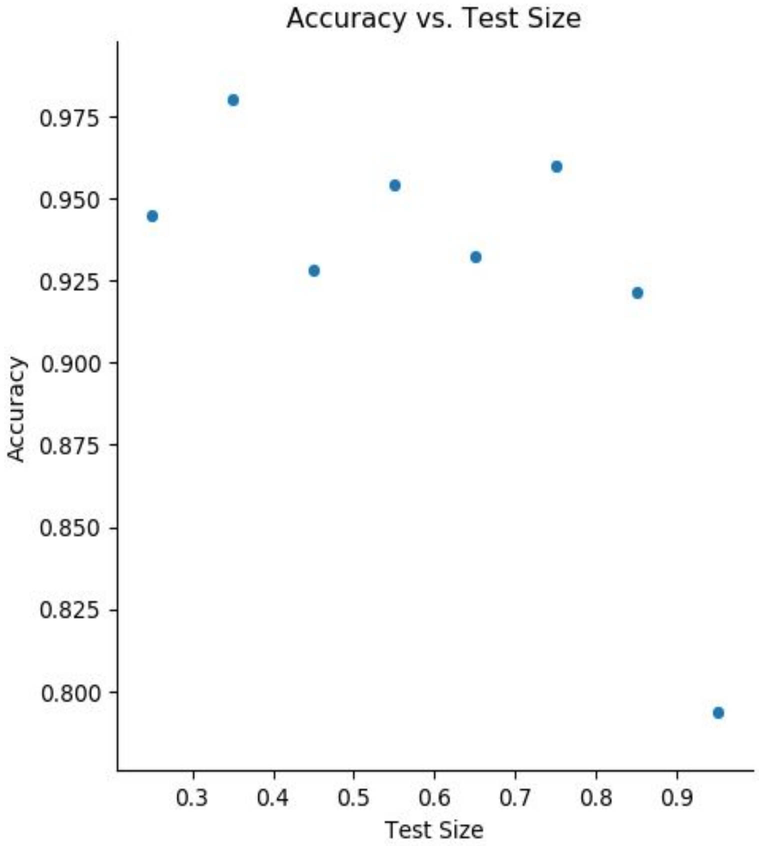
Figure 3: This scatterplot highlights the decrease in prediction accuracy of the model when test size becomes extremely large (ergo, train size is extremely small).

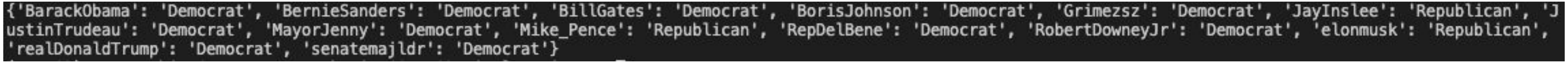
Strengths:

- Emphasizes relationship between test size and accuracy

Weaknesses:

- Loss of information (Democrat vs Republican)





```
{'BarackObama': 'Democrat', 'BernieSanders': 'Democrat', 'BillGates': 'Democrat', 'BorisJohnson': 'Democrat', 'Grimezsz': 'Democrat', 'JayInslee': 'Republican', 'J
ustinTrudeau': 'Democrat', 'MayorJenny': 'Democrat', 'Mike_Pence': 'Republican', 'RepDelBene': 'Democrat', 'RobertDowneyJr': 'Democrat', 'elonmusk': 'Republican',
'realDonaldTrump': 'Democrat', 'senatemajldr': 'Democrat'}
```

Classifications (Scraped Data)

Twitter User	**Prediction**
Barack Obama	Democrat
Jay Inslee	Republican
Donald Trump	Democrat
Elon Musk	Republican
Grimes	Democrat
Bill Gates	Democrat
Robert Downey Jr.	Democrat
Susan Delbene	Democrat

*small sample sizes (~20 tweets)

Jenny Durkan	Democrat
Justin Trudeau	Democrat
Bernie Sanders	Democrat
Mike Pence	Republican
Mitch McConnell	Democrat
Boris Johnson	Democrat

Future work

- Update web-scraping code
- Write Naive Bayes Algorithm ourselves (no Sklearn)
- Different ML models?
- Apply Naive Bayes in other forms of sentiment analysis?