Project 3:

Classifying Reddit Posts from Trump and Biden Subreddits using NLP Classification

Samuel Koh

Agenda

1. Problem Statement
2. Preprocessing
3. Modeling
4. Interpretation of Results
5. Future Work

Problem Statement

- Purpose of the project: To identify the political preferences and attitudes of users of Trump and Biden subreddits using NLP classification techniques
- Importance of the project: Shedding light on broader trends in American politics and evaluating our skills in NLP and machine learning.

Scraping, Preprocessing

- Scraping
  - Users might include the details of their post in the title
  - We have combined title and body of the post
  - Criterion: number of characters in title > 60

- Preprocessing
  - We have removed the following:
    - Html codes
    - Urls
    - Emails
    - Emoticons
    - Noise (#)
    - Single letters
    - Words that are part of Trump's or Biden's names

Modeling

- Challenges: Perpetual overfitting

- Methods:
  - Removing words that occur often but have low impact
  - Selecting a model and attempt to tune hyperparameters

- Estimators:
  - Multinomial Naive Bayes (NB)
  - Logistic Regression (LR)
  - AdaBoostClassifier with DecisionTreeClassifier (ADA)

- Transformers:
  - Count Vectorizer
  - Tfidf Vectorizer

Modeling: Method 1

We started with a baseline of having 56% of data belonging to Trump's reddit.

Removing words that occur often but have low impact:

- First:
  - Train: 0.91
  - Test: 0.79
- Second:
  - Train: 0.93
  - Test: 0.78
- Third:
  - Train: 0.94
  - Test: 0.75

Modeling: Method 2

We compared NB, LR and ADA, each of which are coupled with both count vectorizer and tfidf vectorizer

Found that (ADA, Count Vectorizer) had the least amount of overfitting:

- Train: 0.8
- Test: 0.71

Scores after tweaking hyper parameters:

- Train: 0.96
- Test: 0.68

Modeling

Final choice:

- Transformer: tfidf Vectorizer
- Estimator: Linear Regression
- Train: 0.91
- Test: 0.79

Interpretation of Results

For r/Trump:

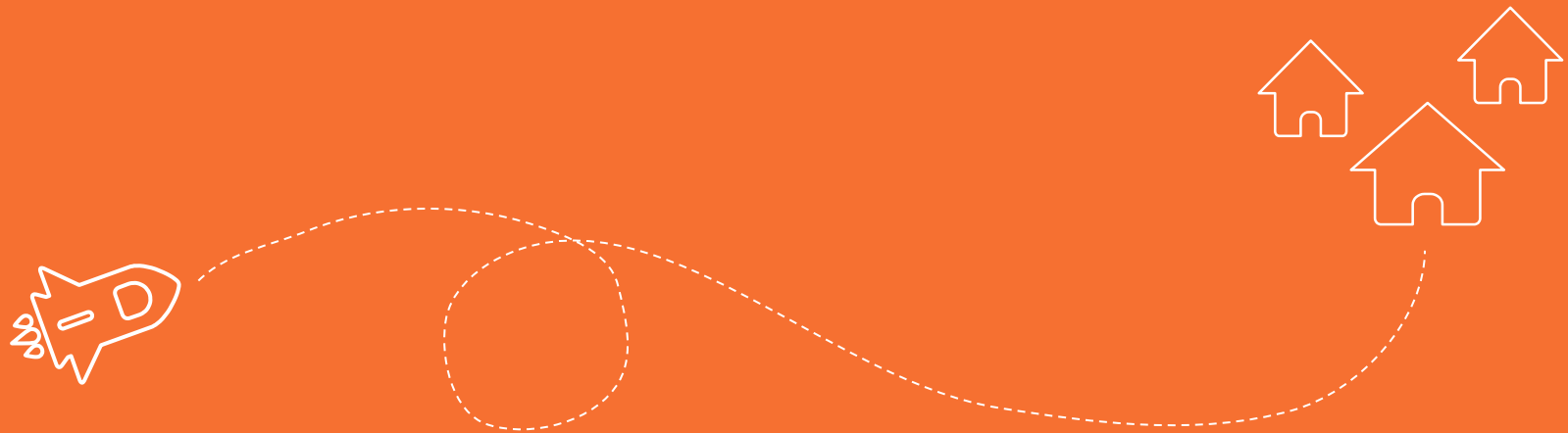
feature_name	coefficient
covid	1.307774
hunter	1.296712
conservative	1.286011
tucker	1.245308
medium	1.241109

For r/JoeBiden:

feature_name	coefficient
union	-1.456752
midterm	-1.477989
administration	-1.587333
republican	-2.105995
president	-2.238303

Future Work

- To improve the overfitting problem
- Include sentiment analysis once we have a proper working model
    - Might be possible to manipulate public opinion



Thank you