US Presidential Election 2020

with

Economics and Twitter

ECON4140

Yam Tsz Chung

1155122291

Motivation

- Continue with the last paper
- Result of 2020 election was out
- A close result between two candidates

Motivation

- Will the models used in last paper perform also well in 2020 election?
- Try to understand social media influence

Target Goals

- Using multiple economic variables to predict the county vote result in 2020
  - With the models used in last paper (Logit, KNN and Decision Tree)
- Build a classifier to identify tweets opinion towards Trump and Biden
- Predict swing state result with social media
  - Can social media reflect state voting result?

Data collection

- Update economic variables with 2019 record
  - Data from US government Census Bureau
  - Including employment, education, population and poverty
  - 3111 x 51 data

- Social media data
  - Using Twitterscraper to collect twitter data
  - Period from 1 Oct to 8 Nov (one month before election to the end of election)
  - 1st set keywords: #DonaldTrump & #Trump
  - 2nd set keywords: #JoeBiden & #Biden
  - Around 1,800,000 tweets in total

Data Processing – Economic variables

- Perform LDA(Linear discriminant analysis) on the economic variables
  - Dimensionality reduction on dataset
  - Normalize the inputs
  - Evaluate models that use LDA projection as input
- Assign Republican win as “1” and Democrat win as “0”
  - High autocorrelation in election, so the result in 2016 will be used as a dummy variable in the model
  - Swing counties as “1”, unchanged as “0”

Data Processing – Text-Data

- Prepare Tweets dataset
  - Cleaning tweets (to lower case, remove@#hyperlink, etc.)
  - Detect tweets in language using **PYCLD2 library**
  - Focus on **English** tweets as to keep the model more simple
  - Remove stopwords and stem with **NLTK**
  - Tokenize the tweets
  - Using **TextBlob** to get polarity

Model - Economic variables
~Setting~

- Predict voting result:
  - Split the dataset with testing size = 0.25 (778 counties for testing)
  - Best model: **Logistic Regression** with C = 1

- Find swing counties:
  - Add swing counties in 2016 election as dummy variable
  - Split the dataset with testing size = 0.3 (934 counties for testing)
  - Best model:
    - Decision Tree with {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 2}

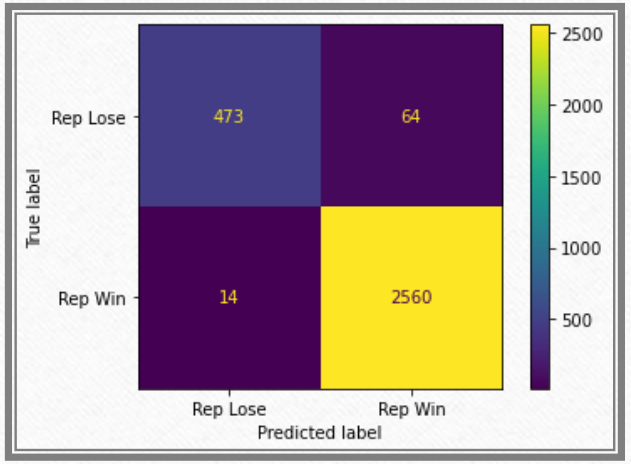
Model - Economic variables ~Result~

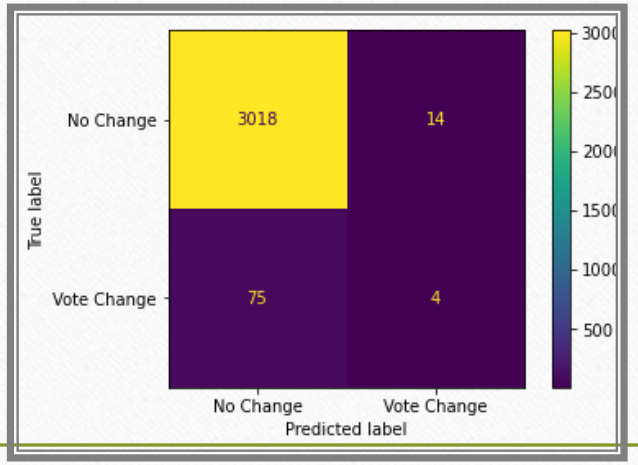
Predict voting result on county level:

- F1 score 0.98 in testing set
- 0.97 in predicting all counties
- Predicted 3033 counties correctly
- The model performed quite well like last paper

Find swing counties:

- Fail to identify swing counties in training
- Only identify **4** correctly in all counties
- Model not performing well in both stage
- Same as last paper, economic variables are not able to identify swing counties



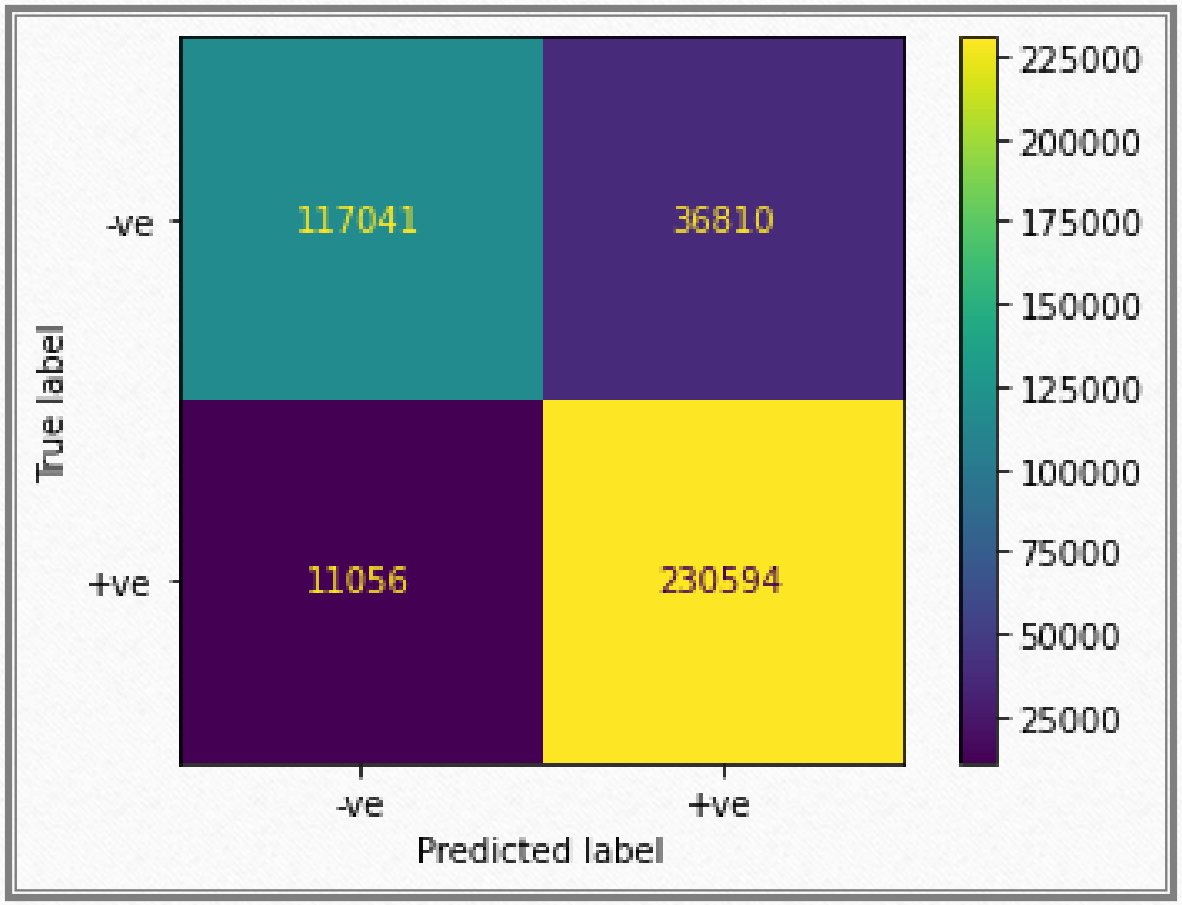


Model – Text-Data
~Setting~

- Sentiment Analysis with TextBlob:
  - Building a **Multinomial Naive Bayes Classifier** to classify positive and negative tweets
  - Vectorize tweets with **TF-IDT**
  - **Drop** neutral statements(polarity == 0) to simplify the model
    - Total Tweets left for Trump = 395501, Biden = 302741
  - Split the dataset with testing size = 0.33 (Trump = 130516, Biden = 99905)
  - Assign "1" for polarity > 0 and "-1" for polarity < 0
  - Parameter to tune: Alpha : [0.01, 0.1, 0.3, 0.5, 1.0, 10.0]
  - Apply 5-fold Cross-Vaildation to train and test the models

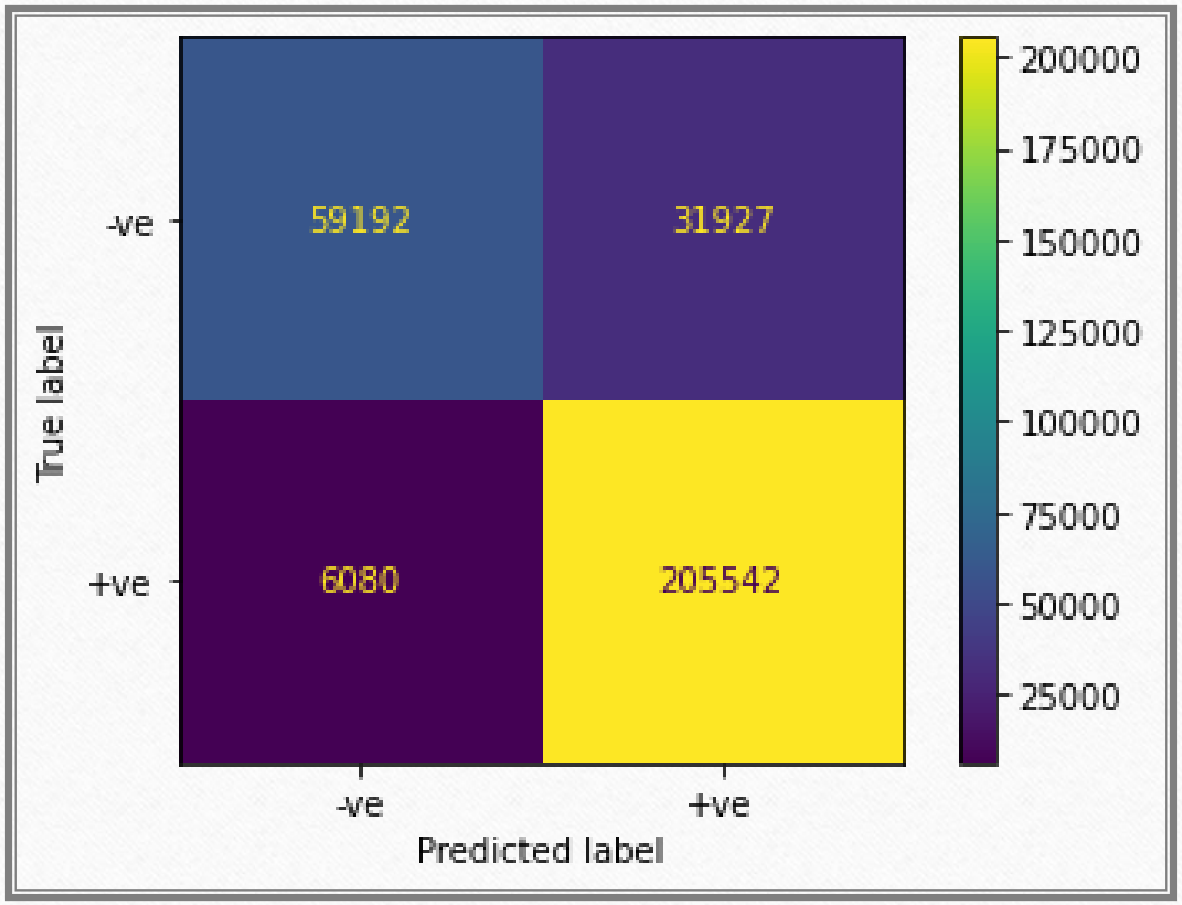
Model - Text-Data ~Result~

- Trump tweets:
  - best parameters: alpha = 0.1
  - F1-score in testing set: **0.82**
  - Overall performance in all tweets:
    - F1-score at **0.88**



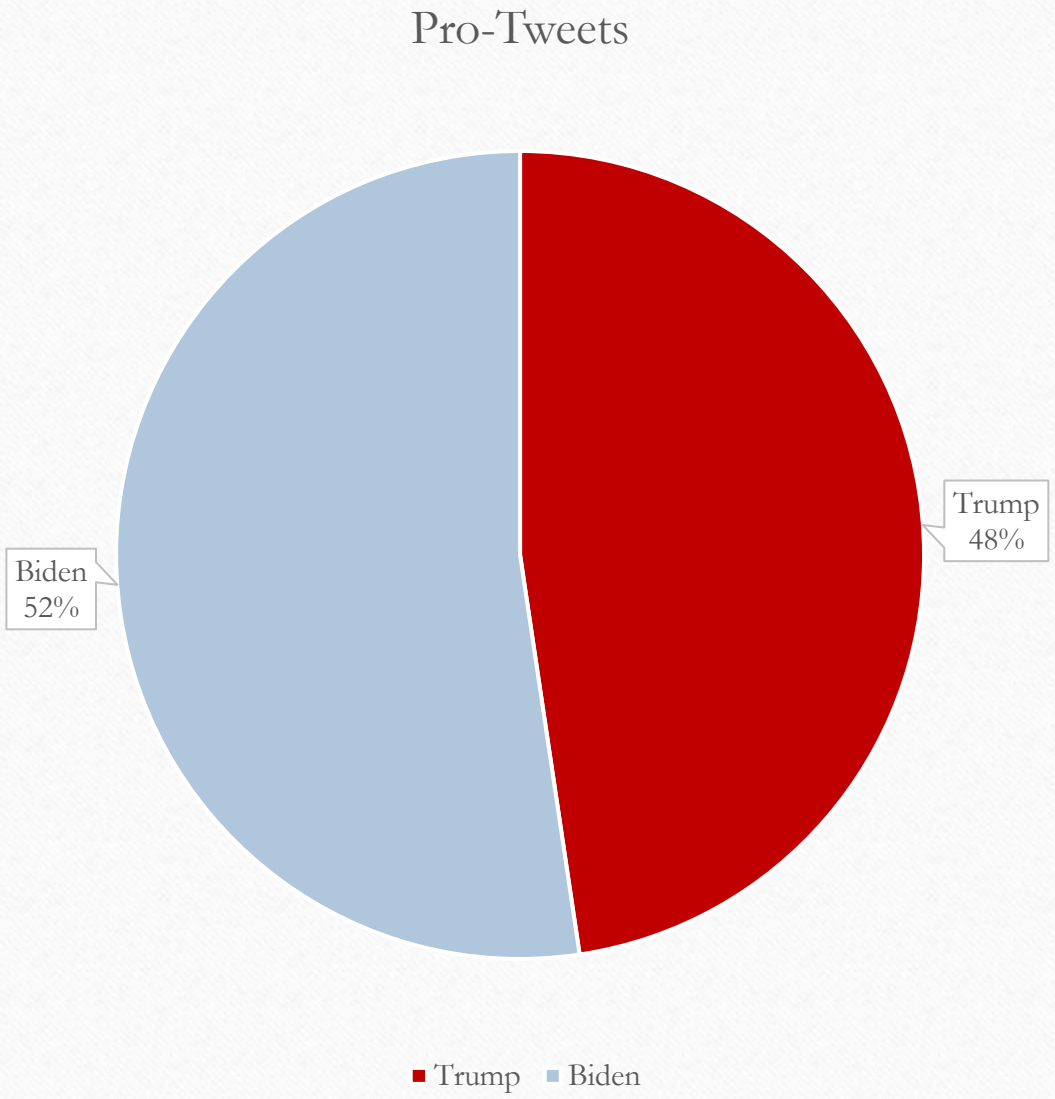
Model - Text-Data ~Result~

- Biden tweets:
  - best parameters: alpha = 0.1
  - F1 score in testing set: **0.83**
  - Overall performance in all tweets:
    - F1-score at **0.87**



Model - Text-Data ~Result~

- With the model:
- Pro-Trump tweets = 332,676
- Pro-Biden Tweets = 365,566
- The result saying that twitter users prefer Biden over Trump



Can Twitter Opinion Reflect Voting Result?
~Background~

- Which candidate win in the election mostly depends on few states
- The influence of social media growth a lot in recent years
- When economic variables cannot reflect result, can social media opinion?

Swing states and tipping point states, 2016&2020

2020 election	Margin	2016 election	Margin
New Hampshire	7.35%D	Maine	2.96%D
Minnesota	7.11%D	Nevada	2.42%D
Michigan	2.78%D	Minnesota	1.52%D
Nevada	2.39%D	New Hampshire	0.37%D
Pennsylvania	1.16%D	Michigan	0.23%R
Wisconsin[note 1]	0.63%D	**Pennsylvania**[note 2]	0.72%R
Arizona	0.31%D	**Wisconsin**[note 2]	0.77%R
Georgia	0.24%D	Florida	1.20%R
North Carolina	1.35%R	Arizona	3.55%R
Florida	3.36%R	North Carolina	3.66%R
Texas	5.58%R	Georgia	5.13%R
National	4.45%D	*National*	2.10%D

from wiki

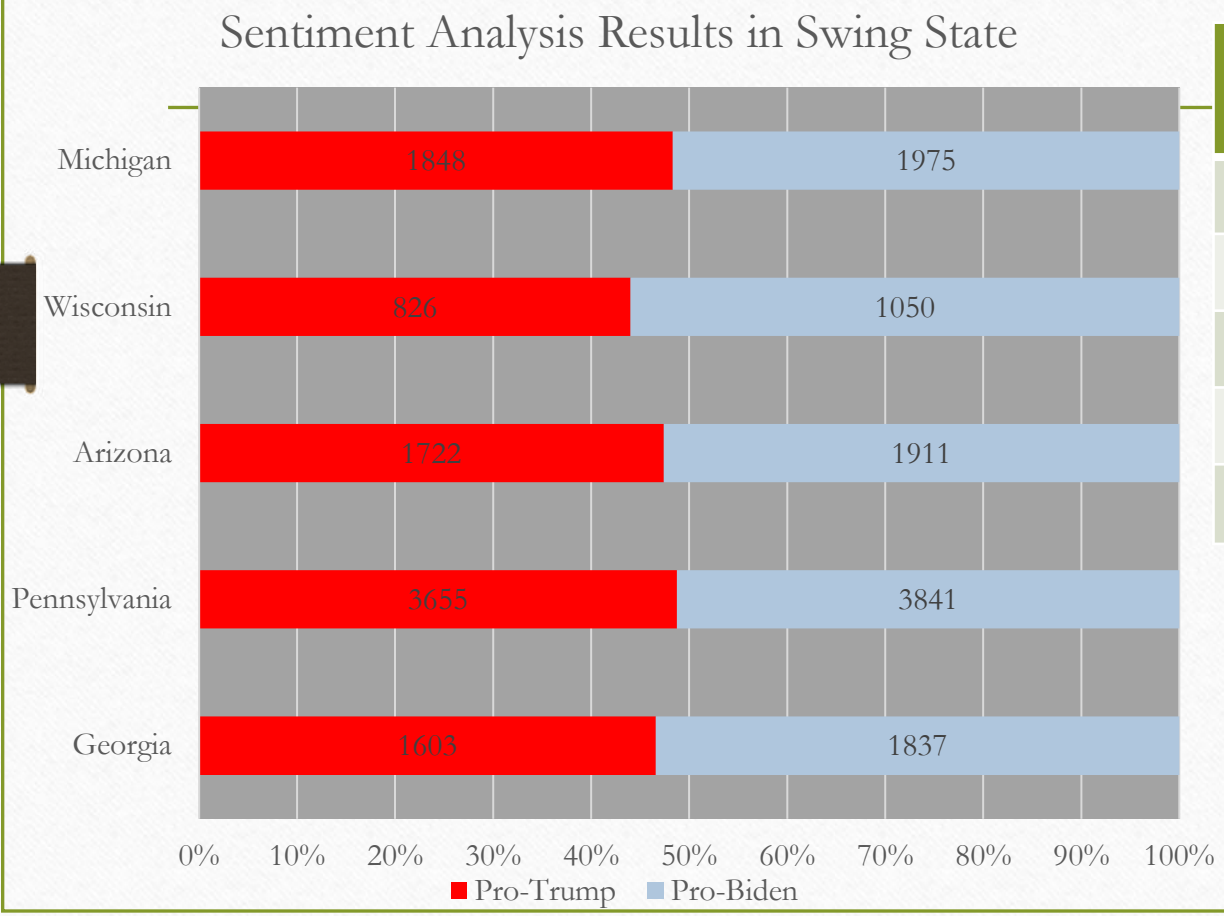
Can Twitter Opinion Reflect Voting Result?
~Setting~

- Group tweets base on states
- Choose states with enough tweets in both swing state and state with large voting margin
- For both tweet dataset, assign "1" for polarity > 0 and "-1" for polarity < 0, drop polarity = 0
- "1" in **Trump set** meaning 1 tweet supporting Trump and "-1" meaning 1 tweet supporting Biden
- "1" in **Biden set** meaning 1 tweet supporting Biden and "-1" meaning 1 tweet supporting Trump
- Compare sentiment percentage with vote result percentage

Twitter Opinion vs Voting Result
~Setting~

- Swing states(2016 → 2020) used:
  - Georgia, Pennsylvania, Arizona, Wisconsin, Michigan
- Stable states used:
  - New York, California, Florida and District of Columbia

Twitter Opinion vs Voting Result
- Result in Swing State

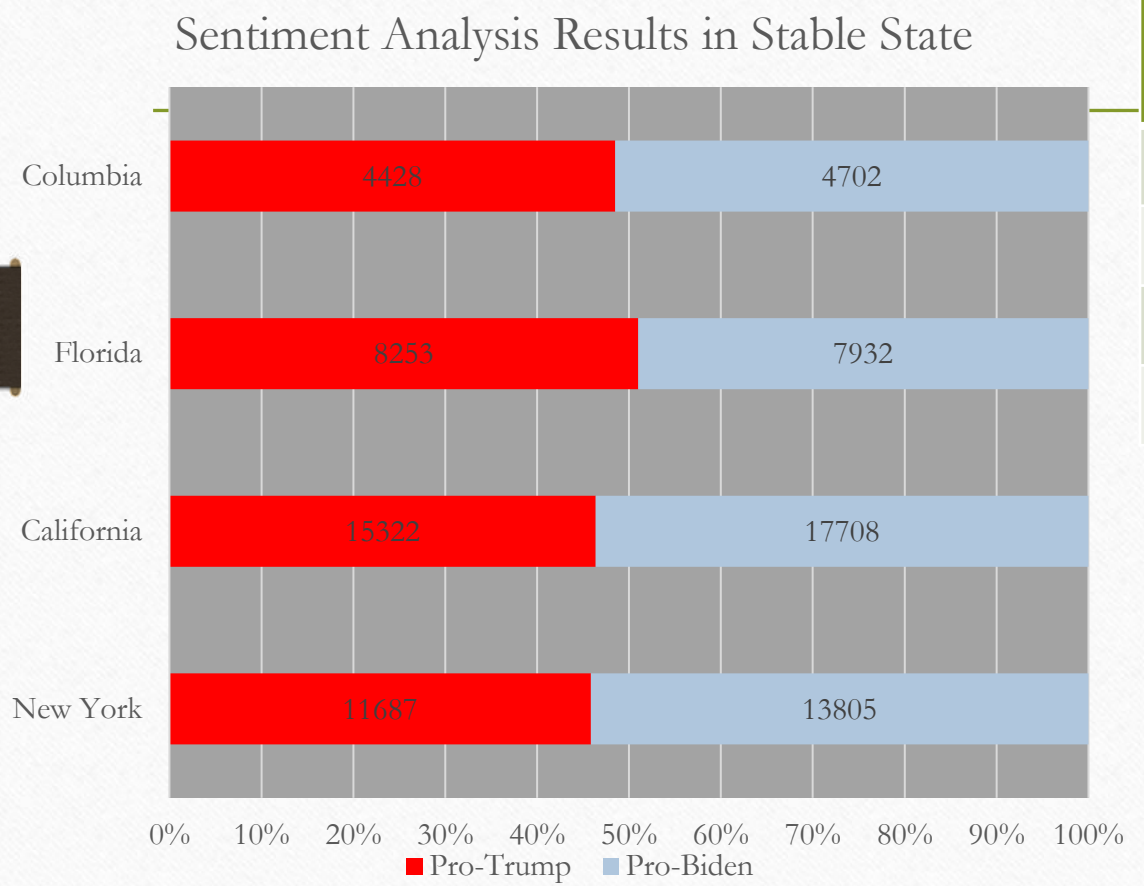


State Name	Sentiment Margin	Voting Margin
Michigan	1.6% (D)	2.78% (D)
Wisconsin	5.0% (D)	0.63% (D)
Arizona	2.6% (D)	0.31% (D)
Pennsylvania	1.2% (D)	1.16% (D)
Georgia	3.0% (D)	0.24% (D)

- Tweet's opinion can reflect voting result in these sates
- But the margins are not accurate enough
- Sentiment over-estimate the voting margin

Twitter Opinion vs Voting Result

~Result in Stable State~



State Name	Sentiment Margin	Voting Margin
Columbia	2% (D)	87.6% (D)
Florida	1% (R)	3.3% (R)
California	3% (D)	13.5% (D)
New York	4% (D)	23.1% (D)

- Tweet's opinion can reflect voting result in these sates
- But the margins are not accurate enough
- Sentiment under-estimate the voting margin by a lot

Conclusions

- Continue from last paper, economic variables **can predict** most of the voting result with **over 90% accuracy**
- But not able to discover swing counties
  - Only **4** counties are correctly predicted
- Multinomial Naive Bayes classifier can classify the sentiment with **nearly 90% accuracy** in twitter text
- Tweet's **opinions can reflect** voting result both in Swing and Stable state
  - But seriously **over / under-estimate** the voting result margin

Limitations

- Hardware limitations
  - Collect tweets required lots of time
- Many tweets have no location record, number of tweets grouped by states reduced a lot compare with original tweets dataset
- Tried to account for COVID influence, but the data is too messy and include a lot of election-unrelated text that affect the model

Suggestions & Future Work

- Collect more keywords to do sentiment analysis
  - In this paper we only use #DonaldTrump, #Trump, #JoeBiden and #Biden
  - With a better computer, we can also collect more keywords
    - But must deal with unrelated tweets that also share the keywords
- Account for a longer period
  - We only use the closest month for this paper (as I think this month will have immediate effect on the election)
  - Can assume earlier month have smaller effect and closer month have more effect on election
- Not enough tweets to show the significant of tweets margin reflection on election margin
- Buy a better computer

Suggestions & Future Work

- My data in this paper cannot fully utilize Neural Network
- I tried using ANN, but the result did not improve by a lot
- If we account for a longer period and more keywords, the model will have a lot of inputs and using Neural Network may have a better result
- Knowledge extraction in Text data
  - Relationship between texts
  - What issues are the most concerned by public

Thank You!!!!!!!!!!