Partisan bias in political news reporting during election periods in Taiwan and the US

PPOL 565 Data Science II: Presentation

Sharon Chuang 04/25/2023

Background

- The 2020 US presidential election was a highly contested and polarizing event, exposing deep political divisions within the country
- The 2020 Taiwan presidential election was characterized by a polarized political climate, influenced by the 2019 Hong Kong protests and the values of democracy and the China challenge.
- Taiwan's polarization is largely driven by issues related to **national identity** and **cross-strait relation**s with China,
- the US's polarization is more focused on issues related to **race, immigration,** and **economic** inequality
- Media bias in news coverage played a significant role in influencing the election outcome.
- → Examining **the role of news bias in shaping public opinion** is essential in understanding the election outcomes in both countries

MEDIA · **Published** April 18, 2023 5:37pm EDT

FOX News Media, Dominion Voting Systems reach agreement over defamation lawsuit

Dominion Voting Systems filed $1.6B lawsuit against network in 2021

By **Joseph A. Wulfsohn | Fox News**

Background : 2020 presidential election

Joe Biden

Democratic

Donald Trump

Republican

Tsai (蔡英文)

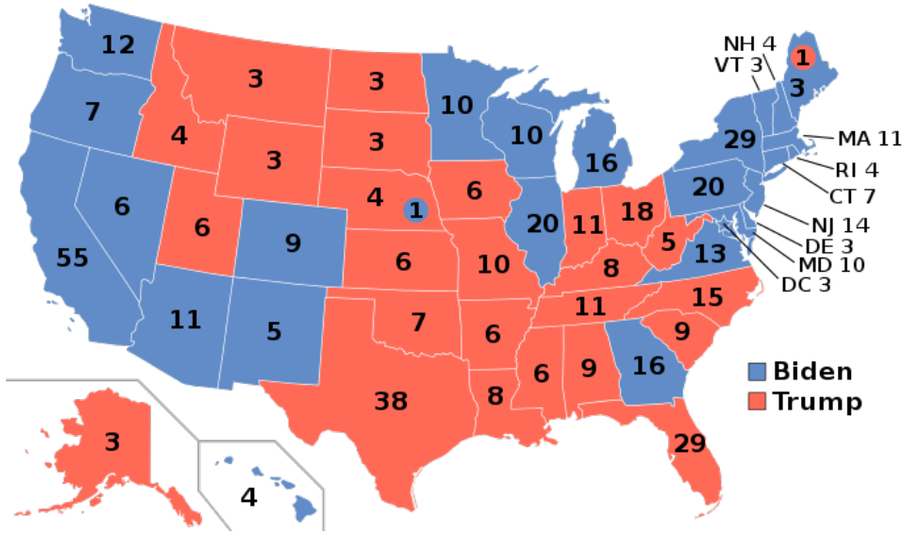
DPP

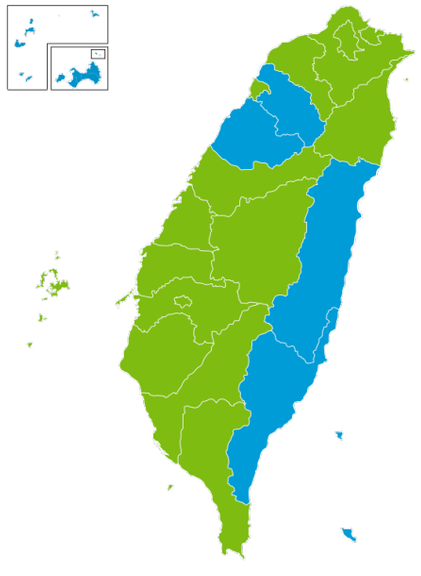
(the liberal party)

Han (韓國瑜)

KMT

(the conservative)





Research questions

1. To what extent do conservative and liberal media outlets exhibit news media bias in their coverage of the presidential election?

2. How does this bias compare to publicly funded non-profit news media?

3. What are the most likely biased keywords in news coverage of the presidential election?

4. Which country exhibits more media bias, Taiwan or the United States?

My Hypothesis:

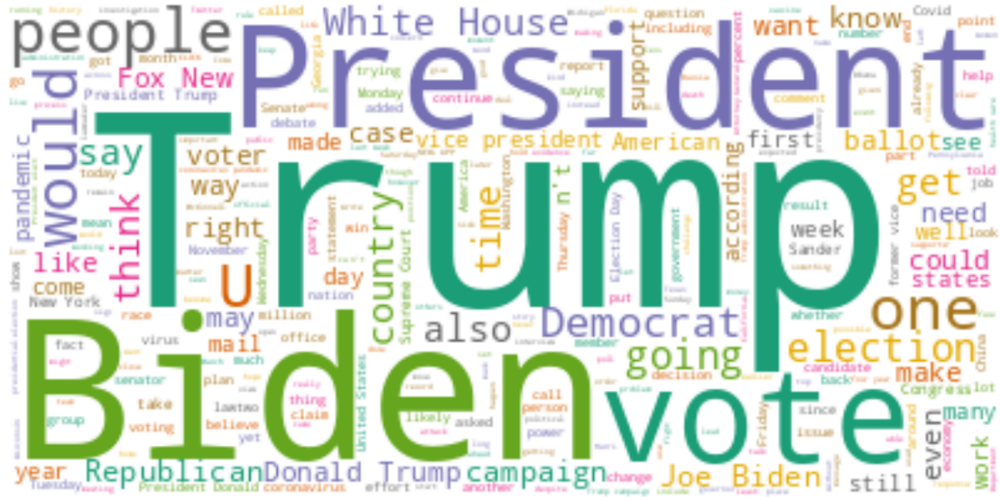
the performance of classification models is good
→ Easier to find difference between news sources

Data sources

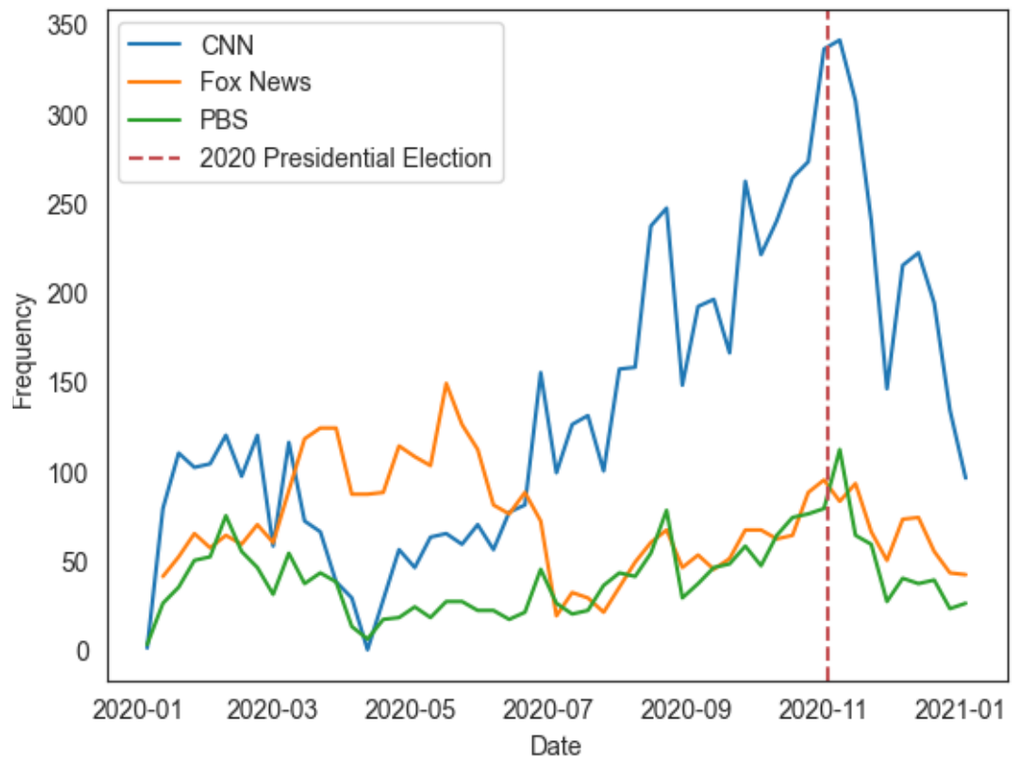
News for the US

- "NELA-GT-2020" dataset from Harvard Dataverse
  - a large multi-labeled news dataset for the study of misinformation in news articles
  - contains nearly 1.8M news articles from 519 sources collected between 01/01/2020 and 12/31/2020
    - has a subset collection of election news articles from 403 different media.
- I chose the news articles from **FOX news, CNN,** and **PBS,**
  - 7663 observations from CNN
  - 3926 observations from Fox news,
  - 2280 observations from PBS

Word Cloud of the US dataset



Weekly frequency by sources of the US dataset

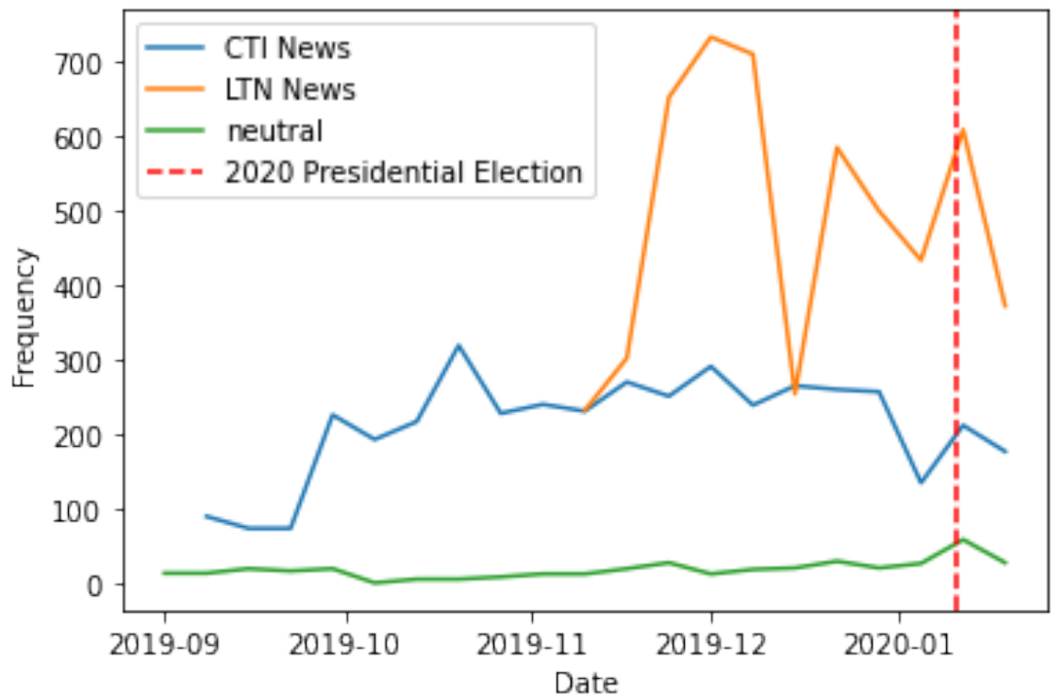


Data sources

News for the Taiwan

- Without organized dataset,
  I **scraped news articles** related to election in Taiwan from a popular Taiwanese news websites
- The Presidential election happened in January 11, 2020. I collected **news articles from September 1st, 2019 to January 18, 2020**
- I chose the news articles from CTI news(中天新聞), LTN news(自由時報), and neutral media collection: CNA (中央社)and The Reporter(報導者)
  - 5387 observations from LTN
  - 4269 observations from CTI
  - 420 observations from CNA and The Reporter

Weekly frequency by sources of the Taiwan dataset



Data sources

Target and Features

- Both datasets contain the same variables: title, release date, content, one hot encoded keyword
- Created one-hot coded keyword records providing a more structured and easy-to-use representation of the keyword information.
- **Content as feature and Source as target**
- All content text will be preprocess using tf-idf
  and Mandarin news articles will processed with word segmentation module

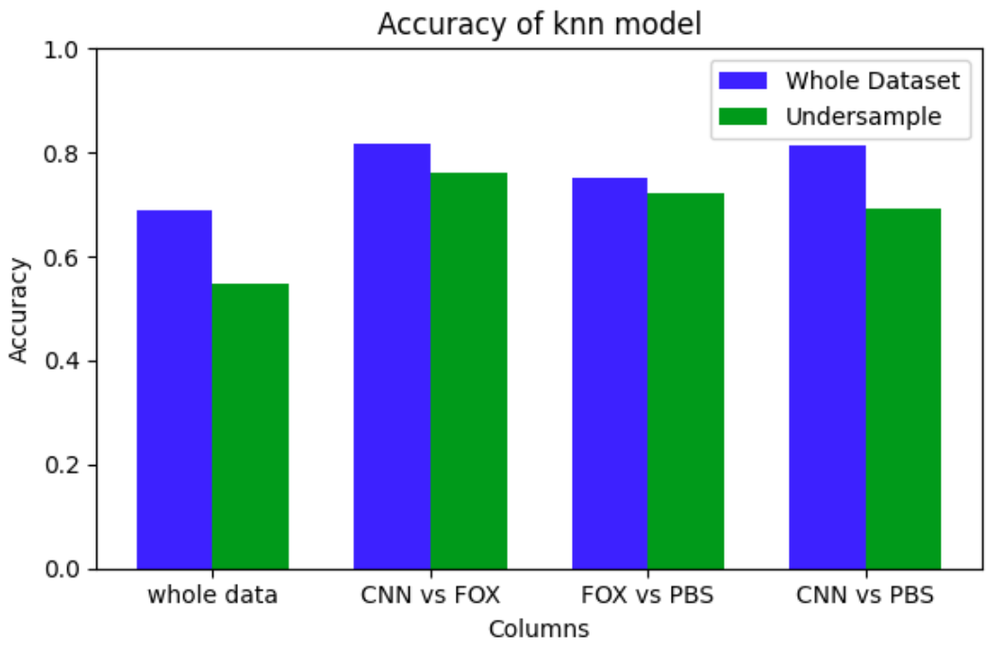
content (feature)	source (target)
MADISON Wis Wisconsin finished a recount of its presidential results on Sunday confirming Democrat Joe Biden s victory over President ….	pbs
Those things that are required that the President has directed us to do in compliance with the decision that the GSA made yesterday I ll do all of those things Pompeo said …	cnn
DEARBORN Mich The Democratic presidential primary is down to two major candidates and it shows Former Vice President Joe Biden and Vermont Sen Bernie …	pbs

Sample features and targets of the US data

Models used in the Project

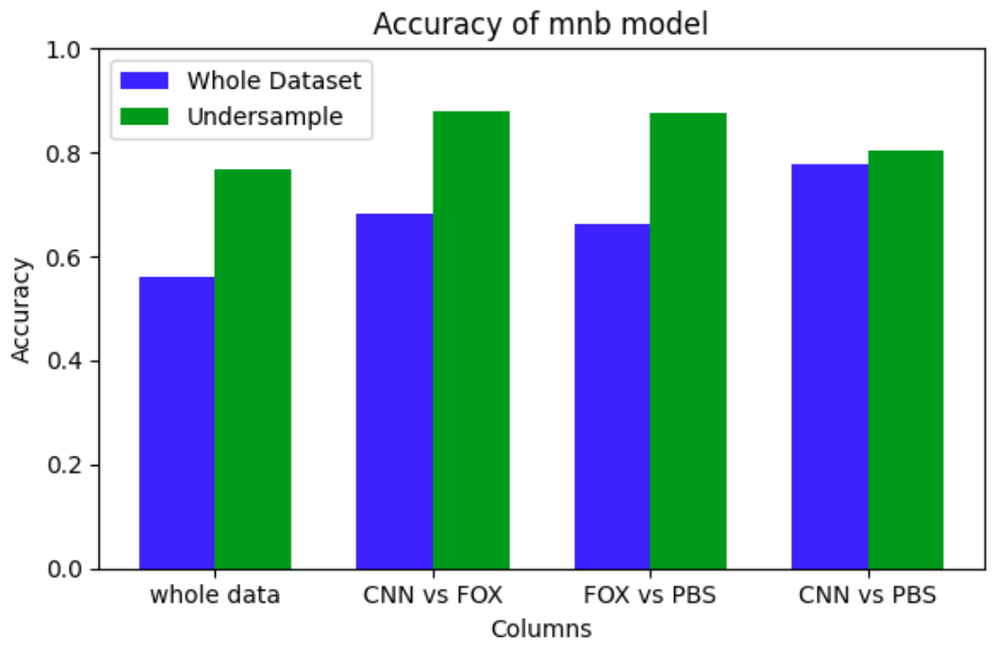
K-Nearest Neighbor (KNN)

- non-parametric
- simple to interpretable



Multinomial Naive Bayes (MNB)

- Parametric
- simple to implement, fast
- handle both numerical and categorical features
- Balances data have a better accuracy



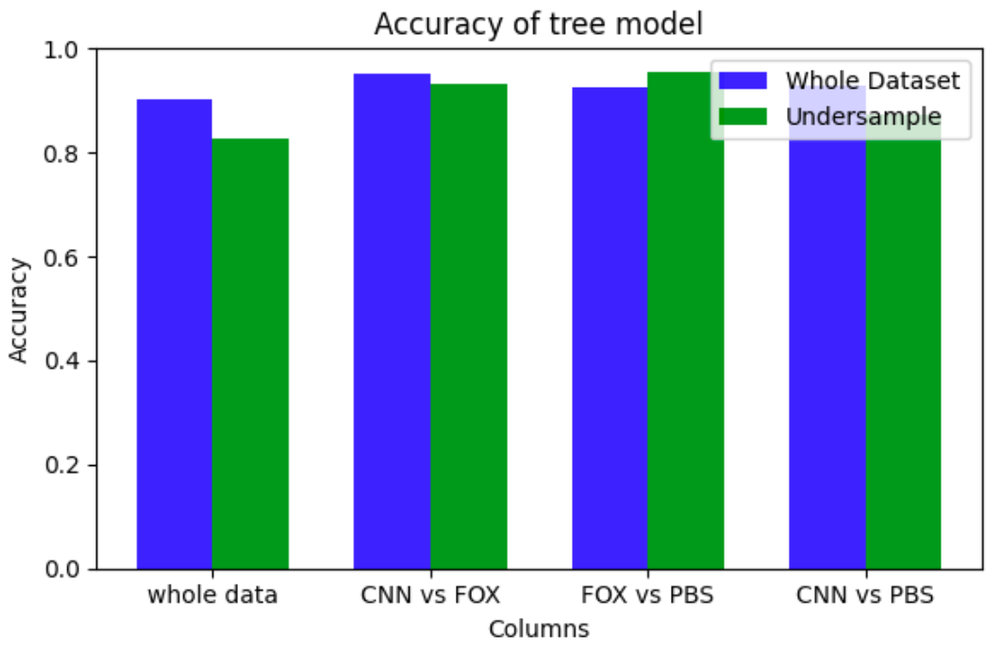
Models used in the Project

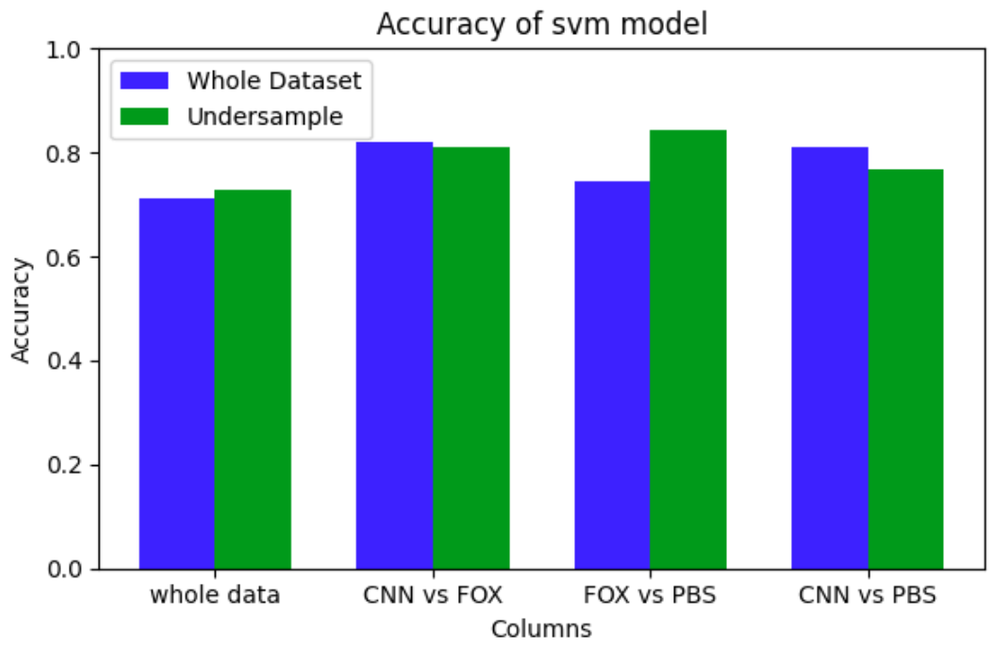
Decision Tree

- non-parametric
- Easy to interpret, but there's only one features in the data
- have the best performance

Support Vector Machine (SVM)

- non-parametric
- works well with both linear and non-linear datasets

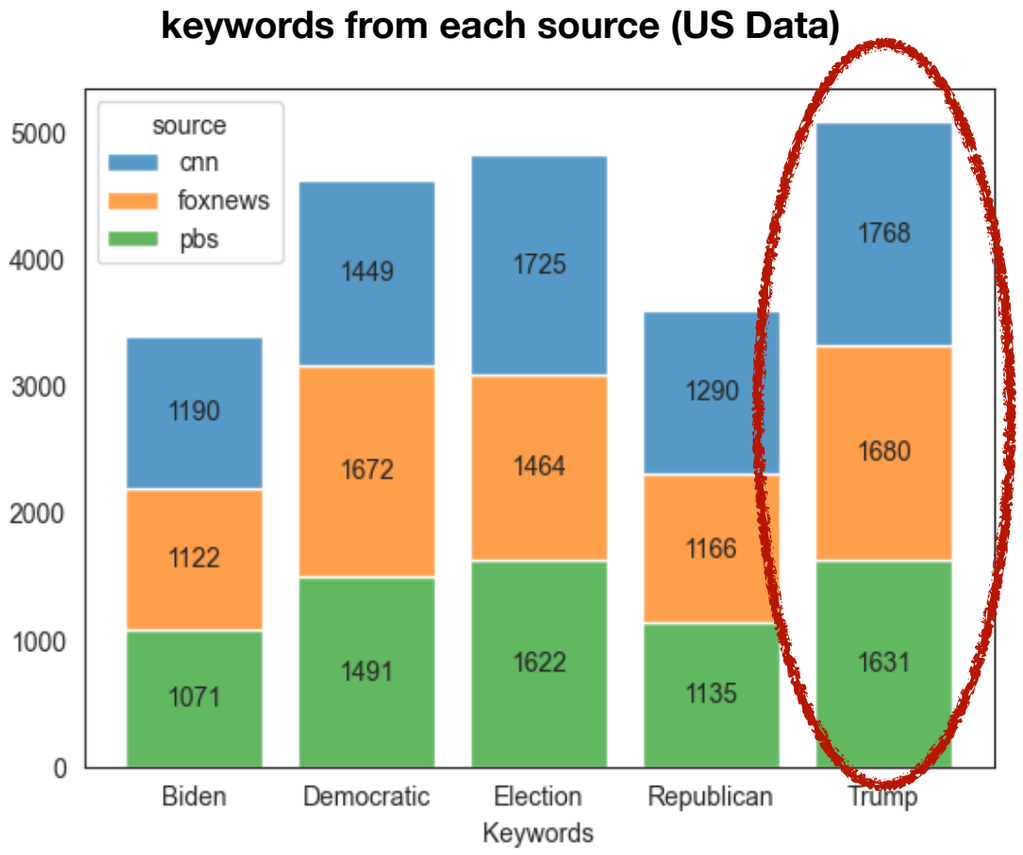


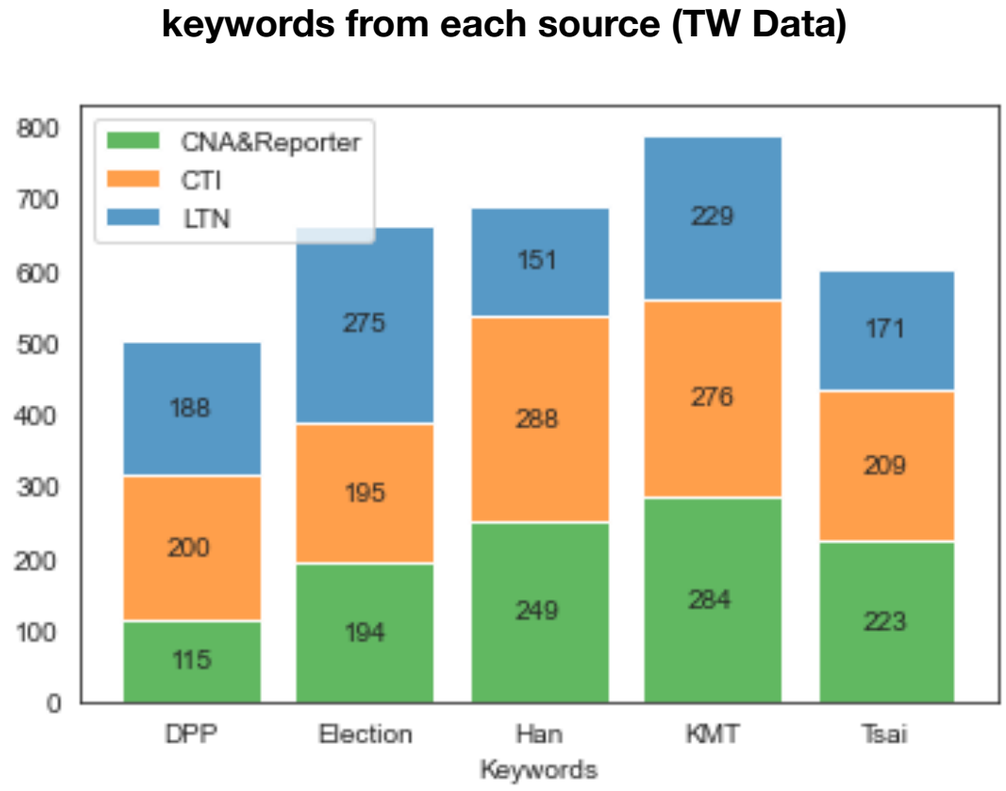


Evaluation and interpretation

Data

- The controversial candidates attracted more attention in both election
- In the US, the media don’t have preference on specific keywords
- In Taiwan, the neutral media use
  - more keyword “election”, less keywords of candidates’ name
  - 2 other media focus more on candidates





Evaluation and interpretation

Results

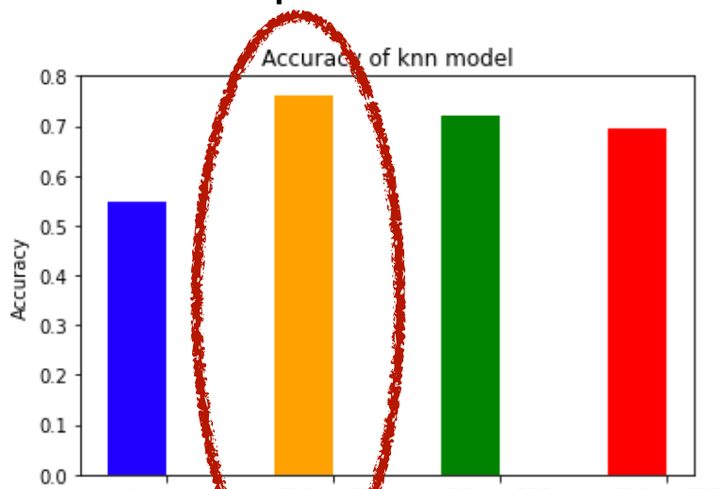
- The complete dataset have 3 different classes, its reasonable the accuracy is lower

- In the US, the liberal v.s conservative model have the best accuracy
- In Taiwan, its more easy to classify other media v.s neutral media

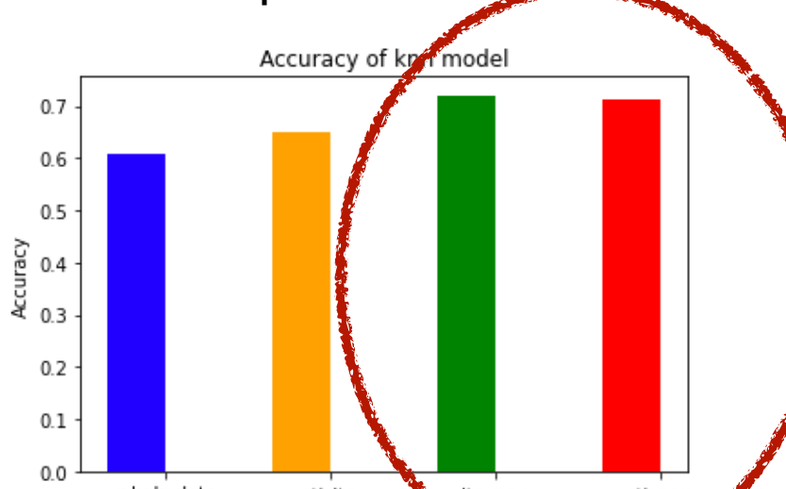
- Model of “election” keyword perform better

- most of the US models have better performance
- But the whole dataset models: Taiwan have a better accuracy

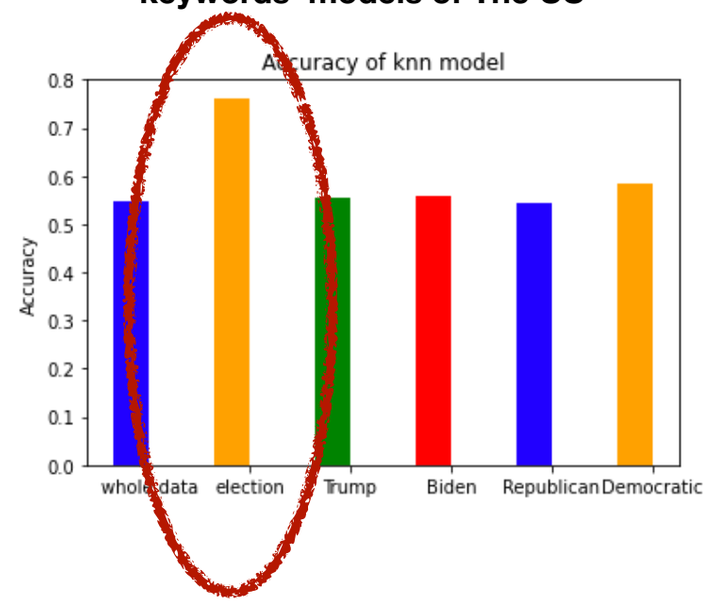
Sources’ comparison models of the US



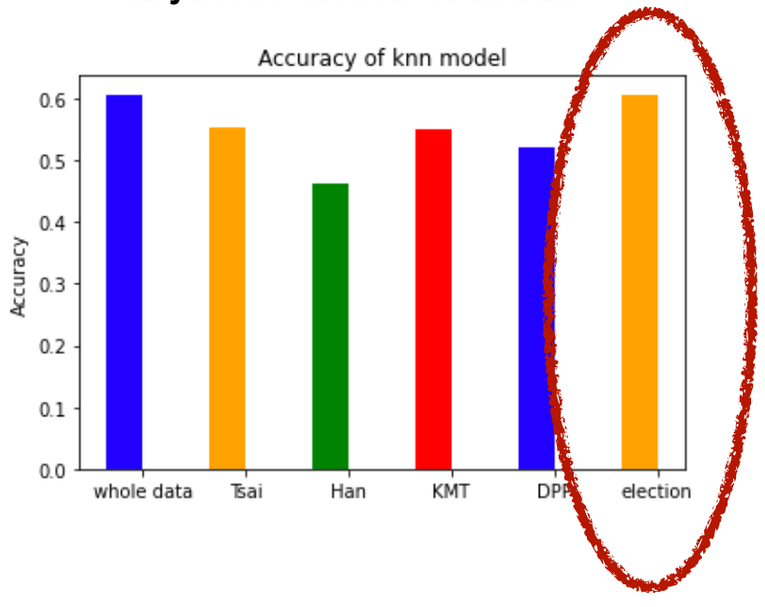
Sources’ comparison models of Taiwan



keywords’ models of The US



keywords’ models of Taiwan

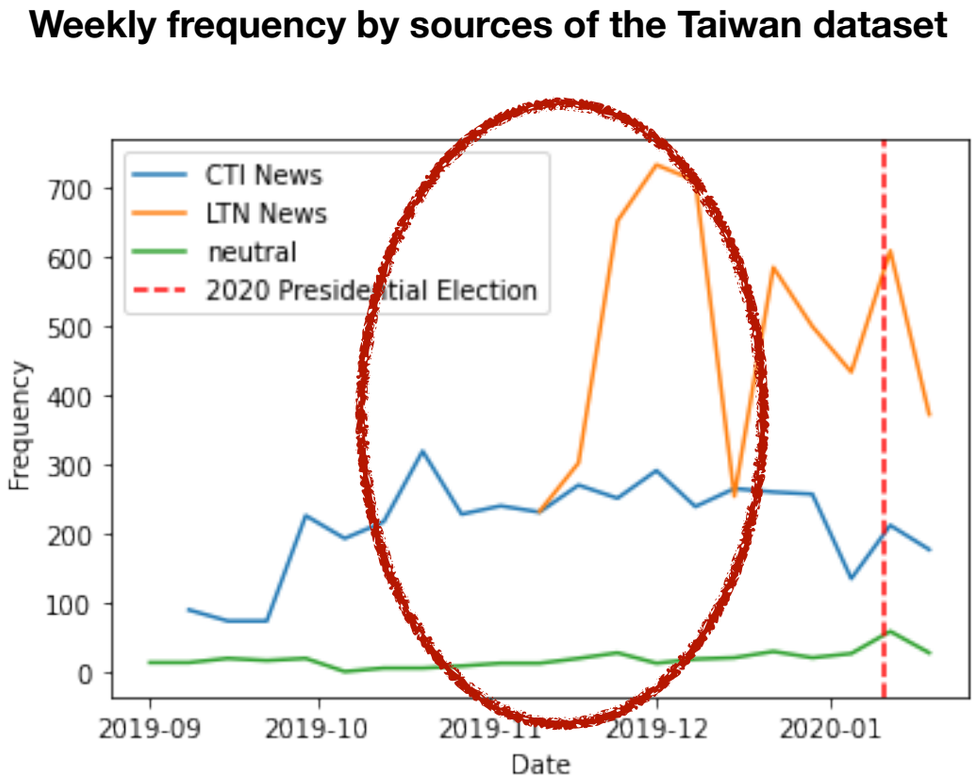


Limitations

- Limited number of media sources used in datasets
- Reliability and accuracy of data may be compromised due to web scraping

- Extremely unbalanced data in Taiwan dataset
  - due to lack of organized dataset :
    Only few media sources in Taiwan considered neutral, and they are rather small media
  - absence of some period data for LTN news:
    did not have data before November due to technical issues while scraping news

- News articles in 2 countries are in different languages:
  this will influence the performance of models



Conclusion

- Decision Tree have the best performance in the news classification

- In the US, the liberal v.s conservative model have the best accuracy
  → the liberal and conservative media is more different
- In Taiwan, its more easy to classify other media v.s neutral media
  → the neutral media is more different

- What else can be done?
  - Evaluate the mean accuracy of different models
  - Use title as the features
  - Add sentiment scores as features