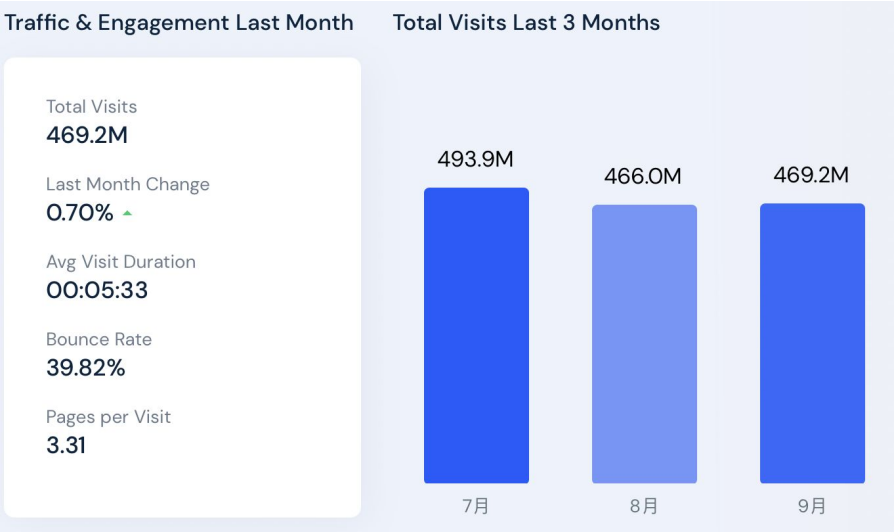
News Headlines Classification

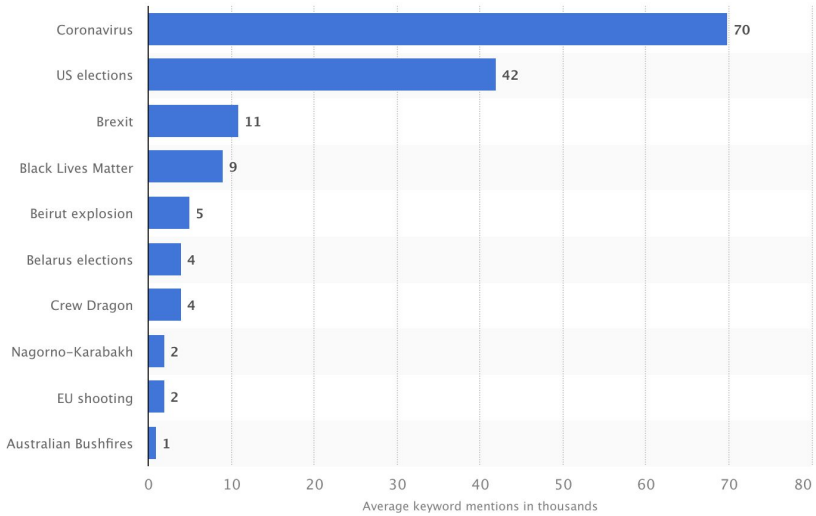
ANLY-580

Yilin Yang, Huiting Song, Shiyu Wang, Tianyi Xu

Introduction

News happen every second. People need them.

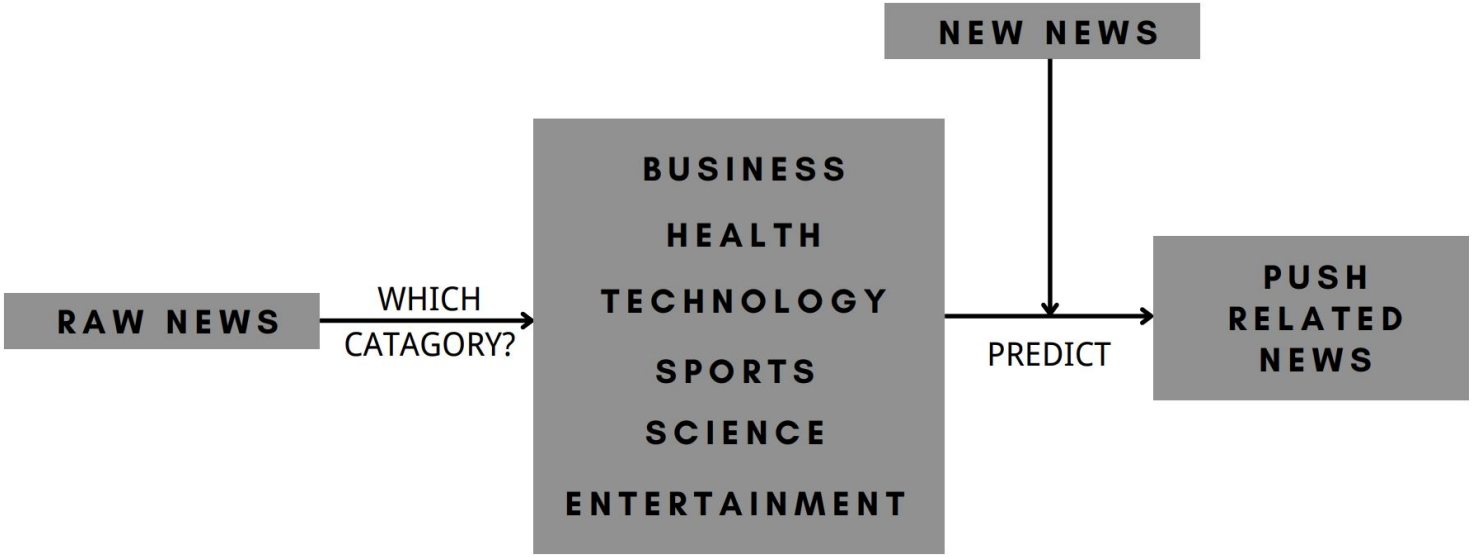




Project Purpose

1.Classify the news

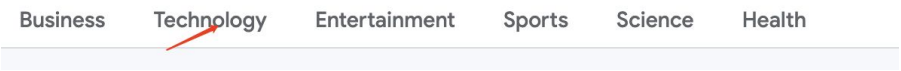
2.Prediction and News Push



Data Collection

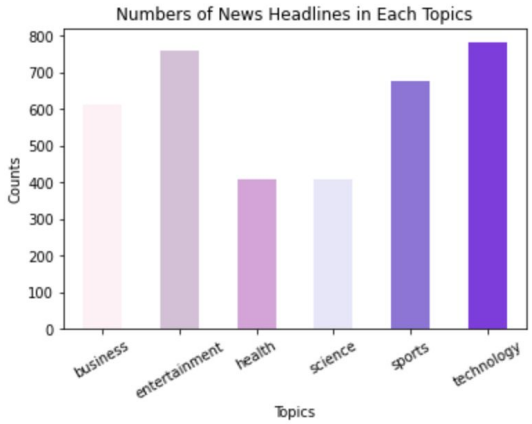
Crawling data from Google News - 3600+ latest new in six topics.

Implement by Python Package Selenium.

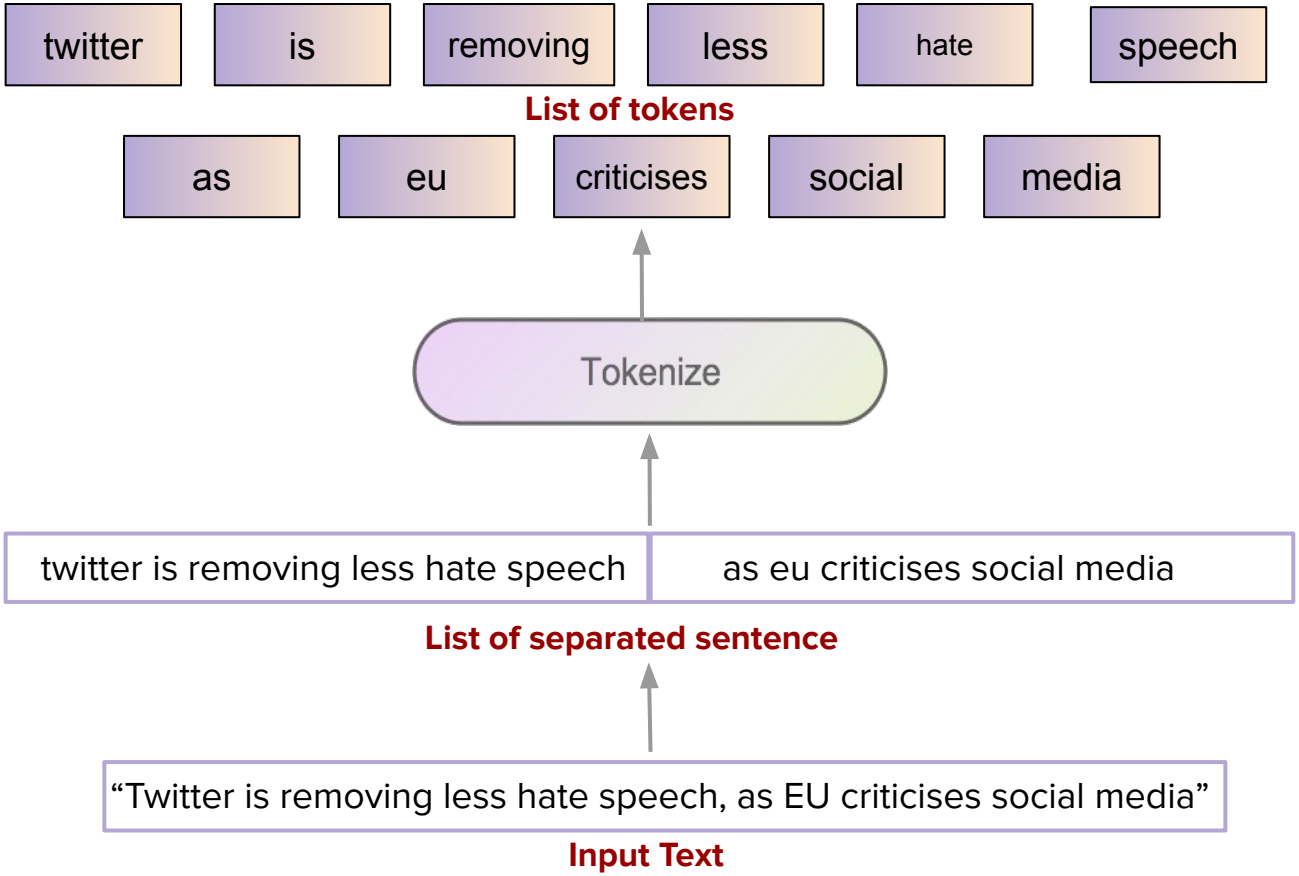


news_label

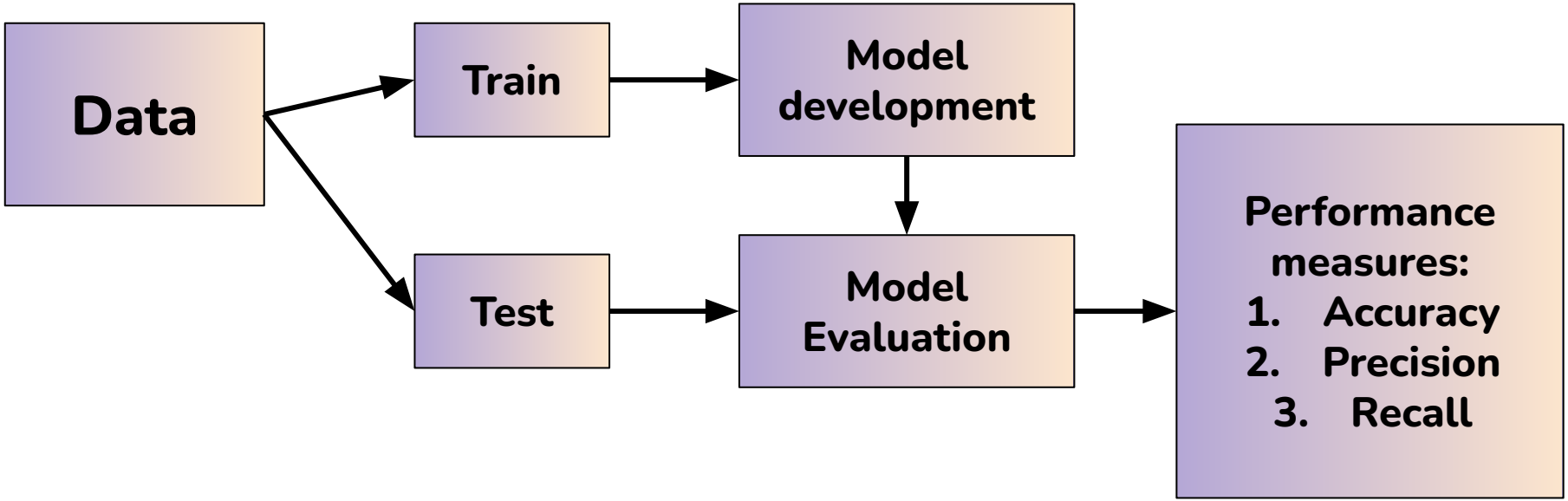
	title	topic	sub_topic	topic_label
0	Holiday shopping returned to a lower-key normal this Black Friday	business	Latest	0
1	Elon Musk says Twitter will re-launch its verification program next week	business	Latest	0
2	Musk says Twitter will launch blue check subscription next week	business	Latest	0
3	Twitter relaunching Verified, with manual authentication checks	business	Latest	0
4	Twitter Will 'Tentatively' Relaunch Paid Verification System Next Friday: Musk	business	Latest	0
5	Elon Musk says Twitter's verified service with colors to start next week	business	Latest	0
6	24 Cheap Doodads Available at Amazon's Black Friday Sale	business	Latest	0
7	215+ Best Black Friday Deals of 2022	business	Latest	0
8	Best Black Friday deals at all-time low price: Apple Watch, Roomba	business	Latest	0
9	Black Friday discounts aren't over: Amazon just dropped 9 fantastic new deals	business	Latest	0
10	The best Black Friday tech deals for 2022: discounts on TVs, laptops, smartwatches and more	business	Latest	0
11	US bans Chinese telecom devices, citing 'national security'	business	Latest	0
12	US FCC bans sales, import of Chinese tech from Huawei, ZTE	business	Latest	0
13	U.S. Expands Bans of Chinese Security Cameras, Network Equipment	business	Latest	0
14	U.S. bans Huawei, ZTE equipment sales citing national security risk	business	Latest	0
15	FCC bans U.S. sales of Huawei and ZTE equipment over national security concerns	business	Latest	0
16	Stocks close mixed on holiday-shortened trading day	business	Latest	0
17	Dow closes more than 150 points higher. Stocks notch gains for holiday week	business	Latest	0
18	Stocks Finish Mixed in Shortened Trading Day	business	Latest	0



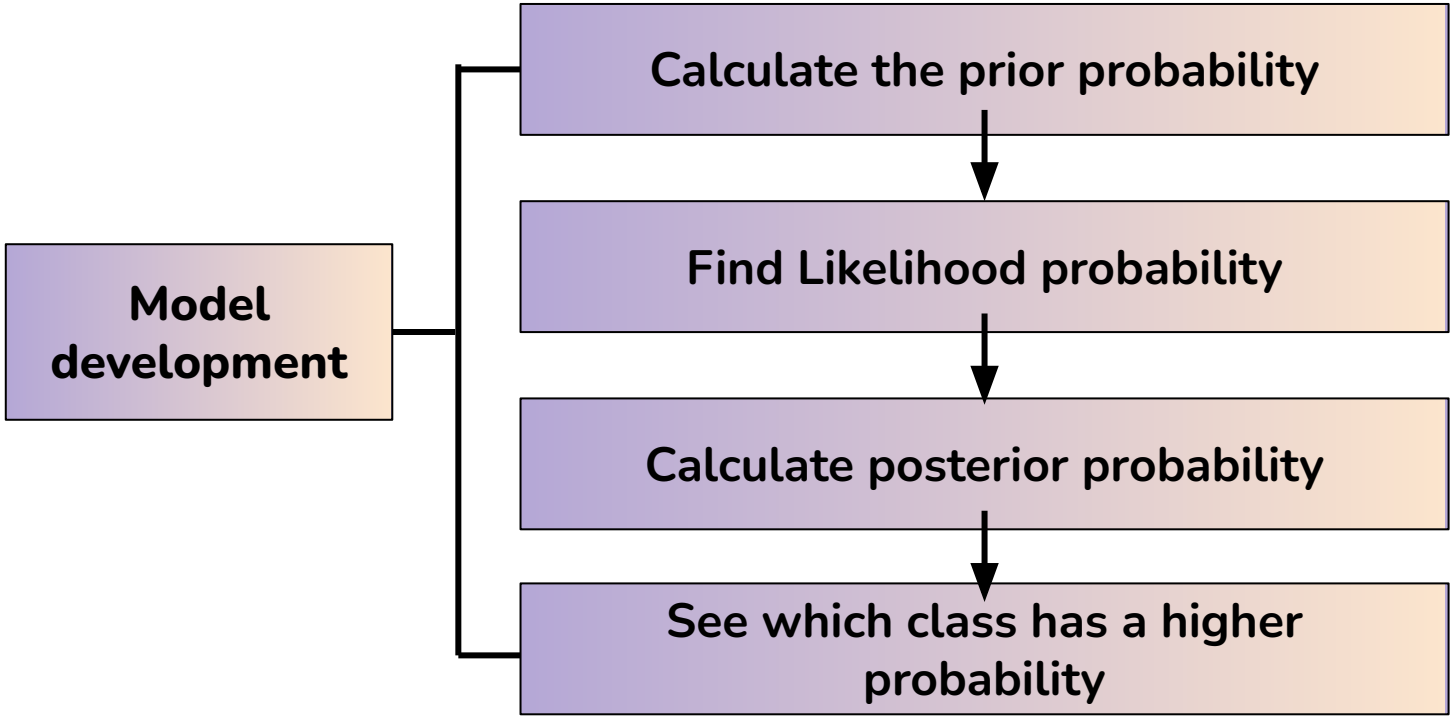
Data Processing: Tokenization



Naïve Bayes Classification



Naïve Bayes Classification



Naïve Bayes: Optimization

Laplace smoothing

$$P(w'|positive) = \frac{\text{number of reviews with w' and y = positive} + \alpha}{\text{N} + \alpha * K}$$

Solving the zero probability problem in Naive Bayes algorithm

```
lambda_a = 0.6
```

Remove words with low frequencies

- The dataset is sufficient enough
- Overfit

- Low frequencies words can only be served as noise and decrease the accuracy

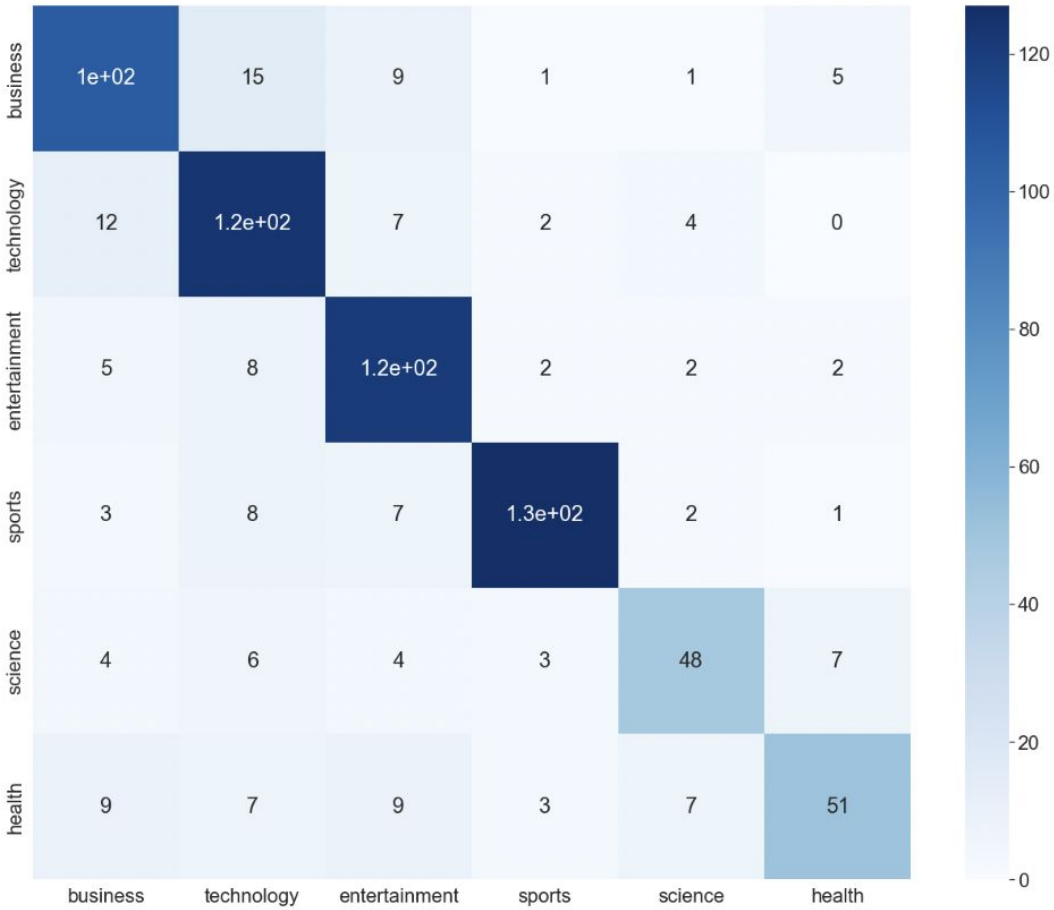
Naïve Bayes: Final results

```
print(classification_report(gold_label, pre_label))
```

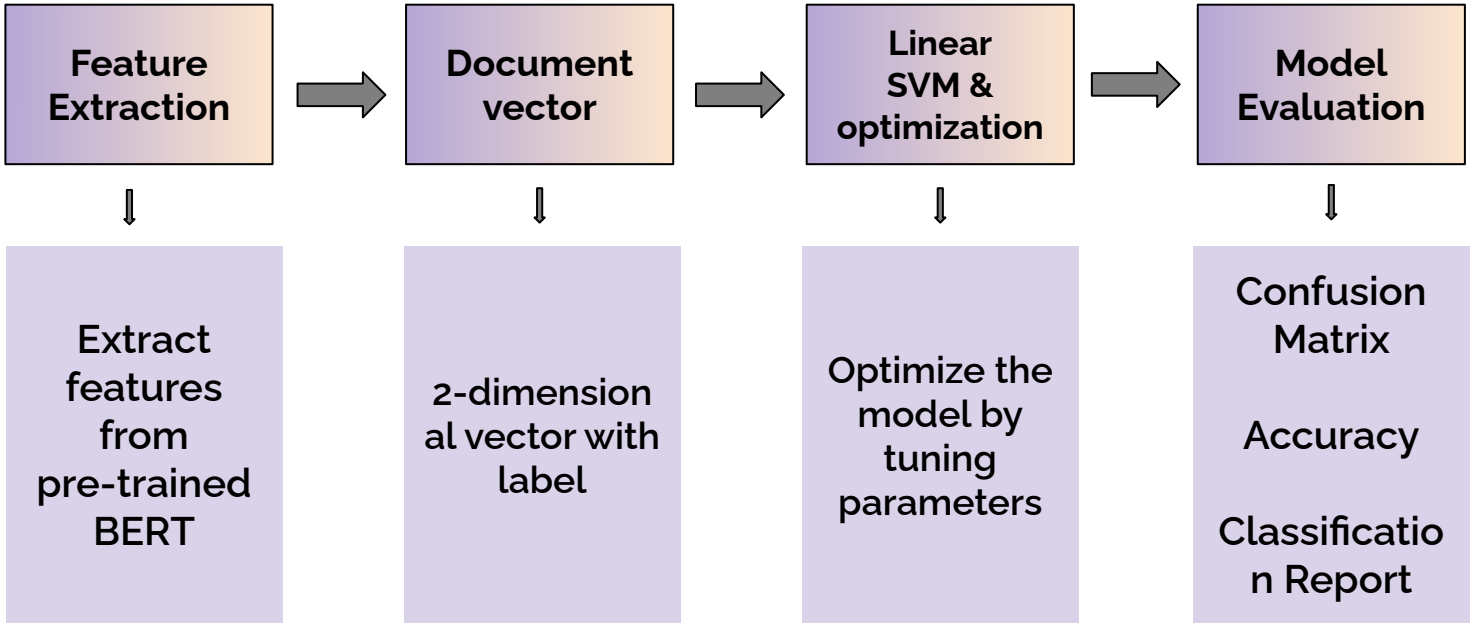
✓ 0.3s

	precision	recall	f1-score	support
0	0.76	0.77	0.77	136
1	0.74	0.83	0.78	149
2	0.77	0.86	0.81	140
3	0.92	0.86	0.89	148
4	0.75	0.67	0.71	72
5	0.77	0.59	0.67	86
accuracy			0.79	731
macro avg	0.79	0.76	0.77	731
weighted avg	0.79	0.79	0.79	731

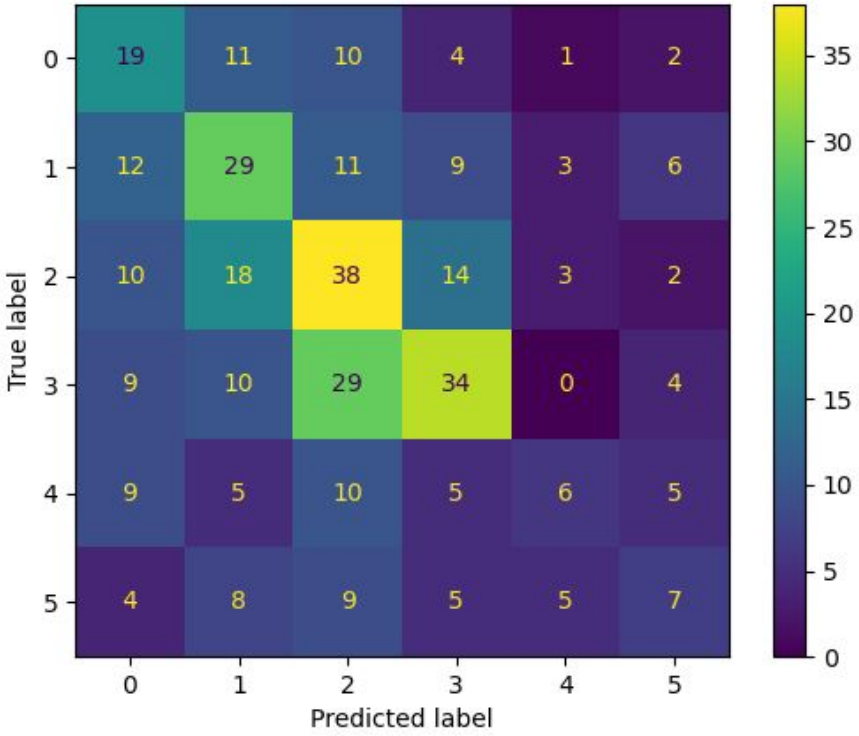
Naïve Bayes: Final results



SVM: Steps to Build a Model



SVM : Result



	precision	recall	f1-score	support
0	0.33	0.29	0.31	59
1	0.40	0.46	0.43	83
2	0.40	0.50	0.44	76
3	0.47	0.39	0.43	77
4	0.21	0.14	0.16	37
5	0.17	0.18	0.17	34
accuracy			0.37	366
macro avg	0.33	0.32	0.32	366
weighted avg	0.36	0.37	0.36	366

SVM : Conclusions

- **Accuracy score is low: only 37% for this model**
  - Feature vector from BERT needs to be fine-tuned
  - SVM could not effectively classify 2-dimensional data
  - **Not using Word2Vec :**Not suitable for SVM because word to vector will make the text vectors too high-dimensional so that SVM could not classify accurately
  - SO, SVM is not a good model for this data set.
- **Advantages**
  - Use pre-trained BERT to extract features
- **Disadvantages**
  - Low accuracy ——> no reference value

BERT : Encoding

```
>>> input-ids
tensor([[ 101, 2421,  112,  188, 3858, 1996, 3776,  106,  102,    0,    0,    0,
             0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
             0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
             0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
             0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
             0,    0,    0,    0]])
```

Input IDs

```
>>> attn_mask
tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

Attention Mask

1 [CLS] stocks finish mixed in shortened trading day [SEP] [PAD] [PAD] ..... [PAD] 64

Padding to the max length allowed

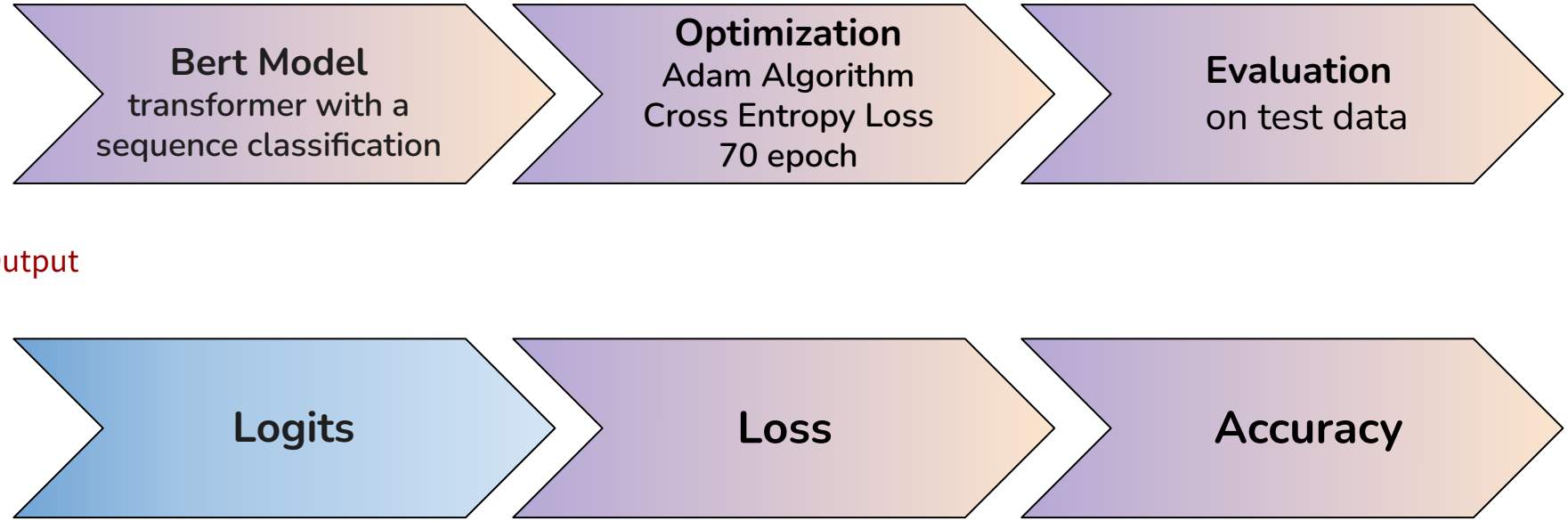
[CLS] stocks finish mixed in shortened trading day [SEP]

Adding [CLS] and [SEP]

stocks finish mixed in shortened trading day

The sentence to encode

BERT : Modeling



BERT : Result

```
[epoch 54] train_loss: 0.026  dev_accuracy: 0.860
              precision    recall  f1-score   support

           0       0.81      0.83      0.82       123
           1       0.82      0.87      0.85       162
           2       0.91      0.86      0.89       148
           3       0.98      0.94      0.96       127
           4       0.86      0.86      0.86        88
           5       0.80      0.80      0.80        83

    accuracy                           0.87       731
   macro avg       0.86      0.86      0.86       731
weighted avg       0.87      0.87      0.87       731

[[102  11   4   0   2   4]
 [ 12 141   4   0   3   2]
 [  6   9 128   3   1   1]
 [  0   1   3 120   0   3]
 [  1   4   0   0  76   7]
 [  5   5   1   0   6  66]]
```

dev_accurate =

acc / (len(dev_loader)*batch_size)

The best accuracy after training on testing set is 0.860

Confusion Matrix

	Correctly Predicted Counts	Error Counts
Business	102	21
Technology	141	21
Entertainment	128	20
Sports	120	7
Science	76	12
Health	66	17

Saved Best model

Naive Bayes

	precision	recall	f1-score	support
0	0.76	0.77	0.77	136
1	0.74	0.83	0.78	149
2	0.77	0.86	0.81	140
3	0.92	0.86	0.89	148
4	0.75	0.67	0.71	72
5	0.77	0.59	0.67	86
accuracy			0.79	731
macro avg	0.79	0.76	0.77	731
weighted avg	0.79	0.79	0.79	731

SVM

	precision	recall	f1-score	support
0	0.33	0.29	0.31	59
1	0.40	0.46	0.43	83
2	0.40	0.50	0.44	76
3	0.47	0.39	0.43	77
4	0.21	0.14	0.16	37
5	0.17	0.18	0.17	34
accuracy			0.37	366
macro avg	0.33	0.32	0.32	366
weighted avg	0.36	0.37	0.36	366

BERT

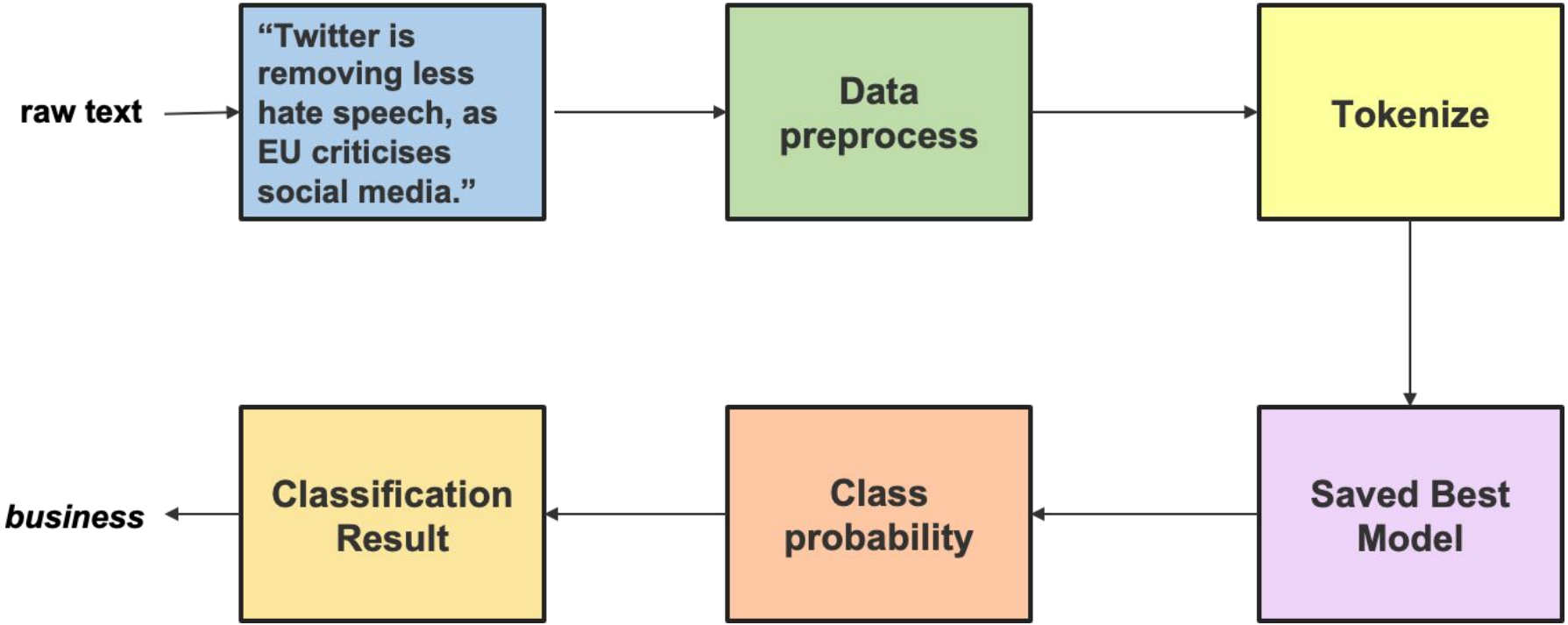
```
[epoch 54] train_loss: 0.026  dev_accuracy: 0.860
              precision    recall  f1-score   support

           0       0.81      0.83      0.82       123
           1       0.82      0.87      0.85       162
           2       0.91      0.86      0.89       148
           3       0.98      0.94      0.96       127
           4       0.86      0.86      0.86        88
           5       0.80      0.80      0.80        83

    accuracy                           0.87       731
   macro avg       0.86      0.86      0.86       731
weighted avg       0.87      0.87      0.87       731

[[102  11   4   0   2   4]
 [ 12 141   4   0   3   2]
 [  6   9 128   3   1   1]
 [  0   1   3 120   0   3]
 [  1   4   0   0  76   7]
 [  5   5   1   0   6  66]]
```

Prediction


raw text
"Twitter is removing less hate speech, as EU criticises social media."
Data preprocess
Tokenize
Saved Best Model
Class probability
Classification Result
business

Demo

Given any news headlines, our demo will predict the category.

True label:
entertainment

Demo

Given any news headlines, our demo will predict the category.

True label:
technology

Demo

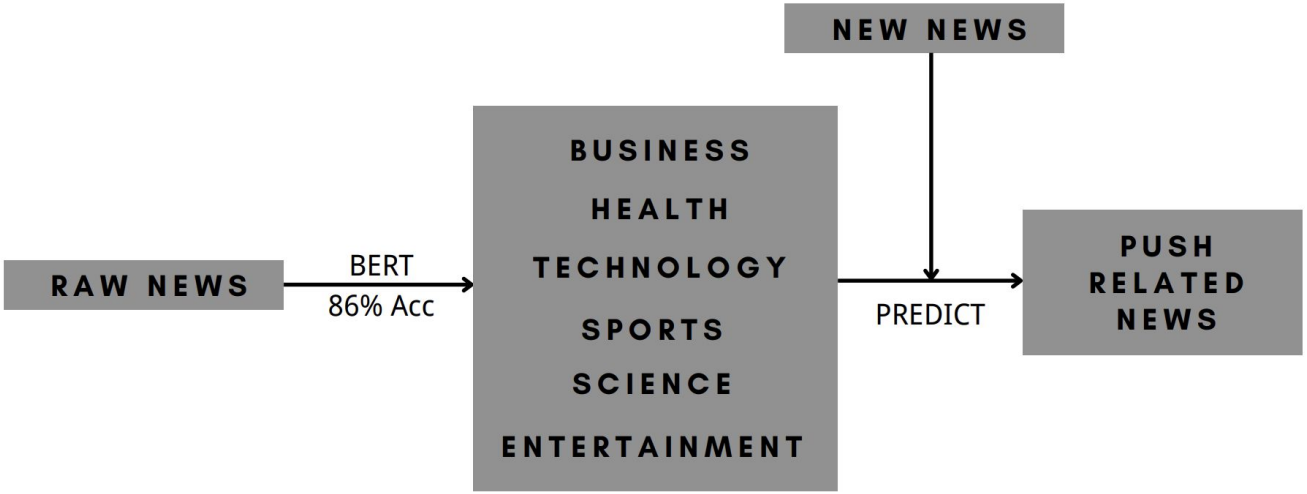
After the label prediction, we could get the related news:

```python
def get_topic(topic):
    website_url = "https://news.google.com/home?hl=en-US&gl=US&ceid=US:en"
    driver = webdriver.Chrome('/Users/xutianyi/Desktop/2022_fall/ANLY580/final project/chromedriver')
    driver.get(website_url)
    search_box = driver.find_element(By.XPATH, "//input[@class='Ax4B8 ZAGvjd']")
    search_box.send_keys(topic)
    search_bottom = driver.find_element(By.XPATH, "//button[@class='gb_rf']")
    search_bottom.click()
    time.sleep(3)
    news = driver.find_elements(By.XPATH, "//a[@class='DY5T1d RZIKme']")
    data = [[elem.text,elem.get_attribute('href')] for elem in news]
    df = pd.DataFrame(data,columns=['title', 'link'])
    return df
```

	title	link
0	The Week in Business: Upheaval in China	https://news.google.com/articles/CBMiV2h0dHBzO...
1	2023 Resolutions For Business Owners	https://news.google.com/articles/CBMiVmh0dHBzO...
2	10 Places to Look for Small Business Grants	https://news.google.com/articles/CBMiZGh0dHBzO...
3	Business Notes for Dec. 4, 2022	https://news.google.com/articles/CBMiQmh0dHBzO...
4	Wildcats take care of business at home against...	https://news.google.com/articles/CBMifWh0dHBzO...
...	...	...
95	Alex Jones has filed for personal bankruptcy	https://news.google.com/articles/CBMiSGh0dHBzO...
96	Brookfield Asset Management Sets Share Ratio f...	https://news.google.com/articles/CBMigQFodHRwc...
97	Rail unions decry Biden's call for Congress to...	https://news.google.com/articles/CBMiTWh0dHBzO...
98	UK bans Chinese surveillance cameras from 'sen...	https://news.google.com/articles/CBMiZGh0dHBzO...
99	The Only Business Idea You Need to Start Makin...	https://news.google.com/articles/CBMibWh0dHBzO...

Conclusion

Best Model - BERT 86% Accuracy



References

1. Adam Algorithm https://www.geeksforgeeks.org/intuition-of-adam-optimizer/
2. Naive Bayes Gandhi, R. (2018, May 17). Naive Bayes classifier. Medium. Retrieved December 5, 2022, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
3. Introduction Picture from: https://www.statista.com/
4. Word Embeddings in NLP and its applications. https://www.kdnuggets.com/2019/02/word-embeddings-nlp-applications.html

Q&A

Thanks for listening!
Any questions?