Classification of online post categories based on community tendencies

20190052 Geon Kim

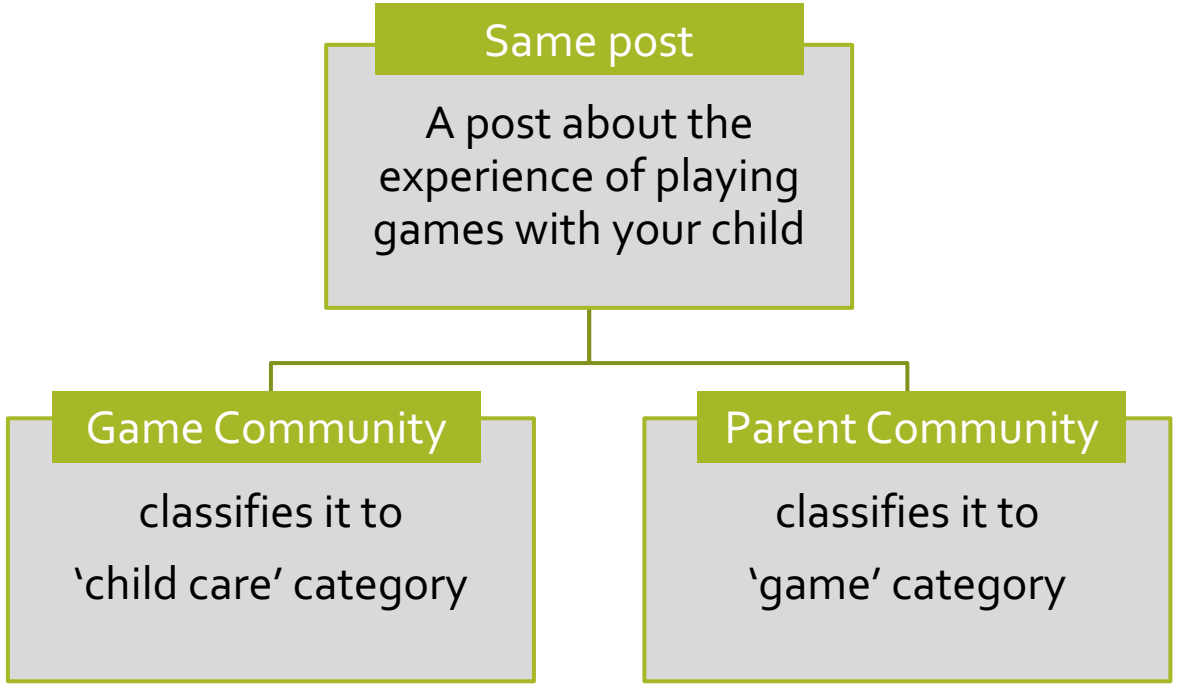
20190703 Geonha Hwang

Introduction – Motivation

- Community users read and write the post by selecting the appropriate category
- If the post are not properly categorized, it will cause dissatisfaction for community users and lead to community decline in the long run
- Therefore, category recommendation system is needed for community
- Can rapid-growing online communities use dataset from other communities to create automatic category classifier?

Introduction – Hypothesis

- Each online community has a tendency to influence text classification



Introduction – Problems

- Create a classifier
  - trained with posts categorized by the community

- Create a classifier
  - trained with posts categorized by another community
  - to compare previous classifier for proving whether our hypothesis is correct

- Create a classifier
  - trained with combined posts from the previous two communities
  - to show that our test is not related to the size of the dataset

Datasets

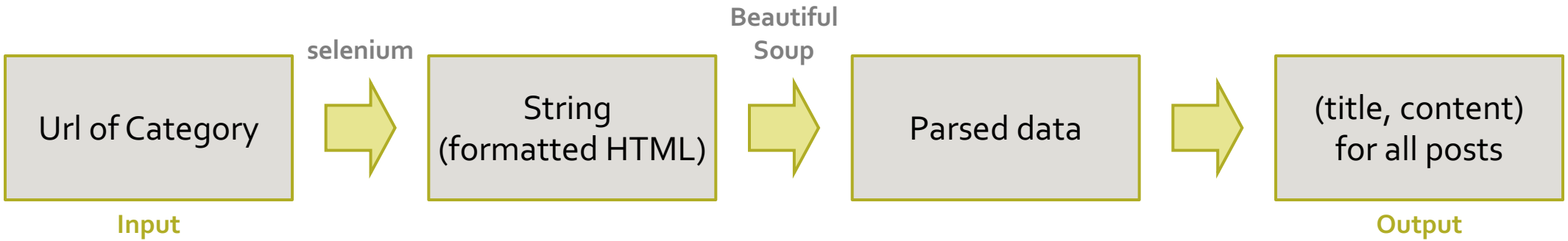
- We use two datasets
  - Crawled reddit dataset
  - News category dataset

- The two datasets have the same categories
  - business, entertainment, parenting, politics, sports, travel

Datasets – Crawled reddit dataset

- Reddit
  - American social news aggregation, content rating, and discussion website
  - https://www.reddit.com/

- We do crawling the reddit site for all categories



Datasets – Crawled reddit dataset

- Dataset description

Column name	Possible value	Description
title	string	Title of post
content	string	Content of post
link	string	If a link exists in the post, it is true or false
image	True | False	If a image exists in the post, it is true or false
video	True | False	If a video exists in the post, it is true or false
category	entertainment | politics | travel | parenting | business | sports	Category of post

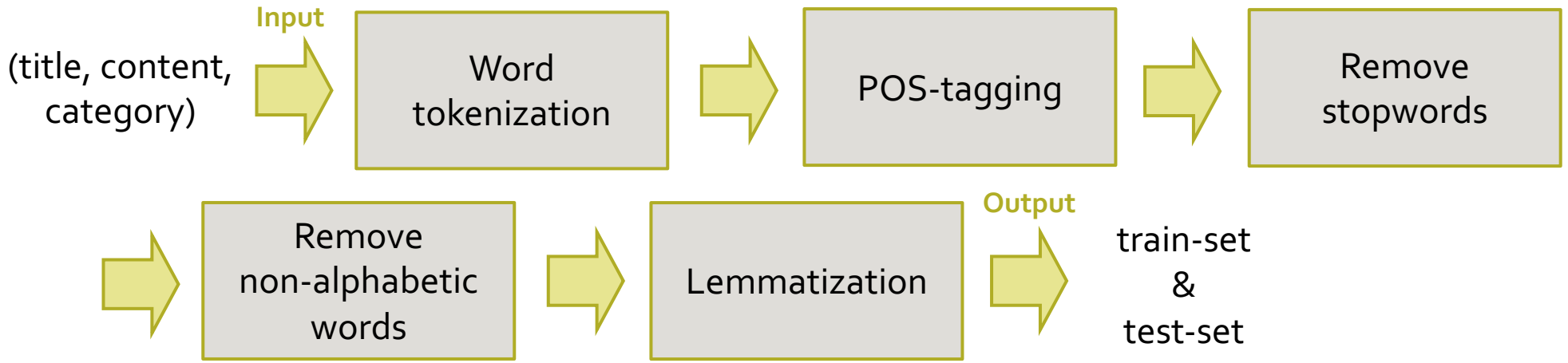
- It is uploaded to ' https://github.com/14KGun/CS372-Community-Category-Classification/tree/main/dataset'

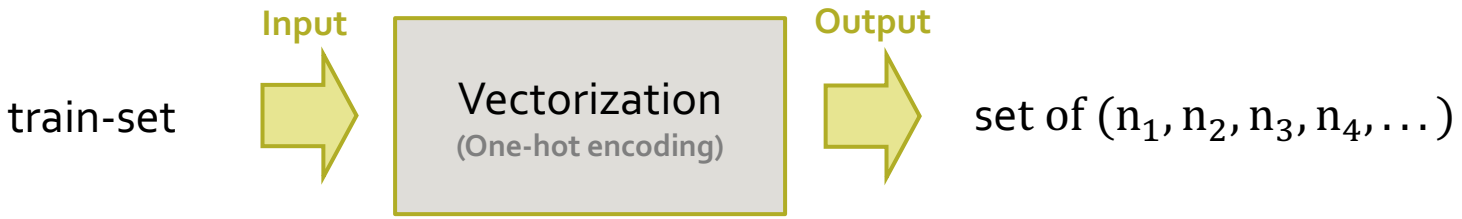
Datasets – News category dataset

- Kaggle News Category Dataset
  - 200k news headlines from the year 2012 to 2018 obtained from HuffPost
  - https://www.kaggle.com/datasets/rmisra/news-category-dataset
- Dataset description

Column name	Description
category	Category article belongs to
headline	Headline of the article
authors	Person authored the article
link	Link to the post
short_description	Short description of the article
date	Date the article was published

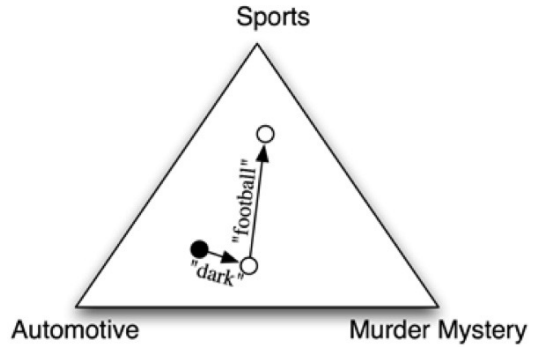
Pre-processing





Train Classifier

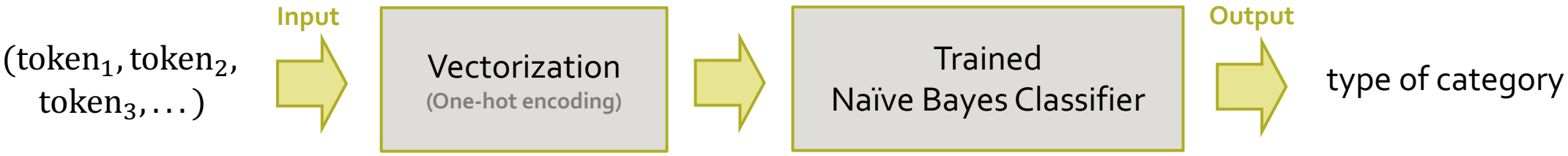
- Naïve Bayes Classifier
  - Probabilistic machine learning model based on the Bayes theorem
  - Bayes Theorem : $P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$



- Using two train-sets, we train three classifier
  - Classifier based on reddit dataset
  - Classifier based on news dataset
  - Classifier based on a dataset that be combined with reddit and news datasets

Test

- Process to test trained classifier

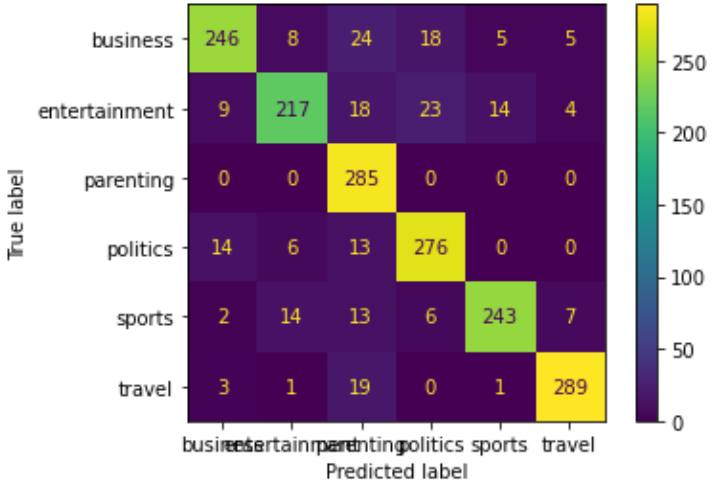


- Test the classifiers in five ways.

	Test - 1	Test - 2	Test - 3	Test - 4	Test - 5
Classifier	Reddit	News	Both	Reddit	News
Test-set	Reddit	News	Both	News	Reddit

Test Results – Using reddit test-set, test reddit classifier

- Test the classifier trained with reddit train-set, using reddit test-set



[Confusion matrix]

train-set size	4158
test-set size	1782
accuracy	**0.87**

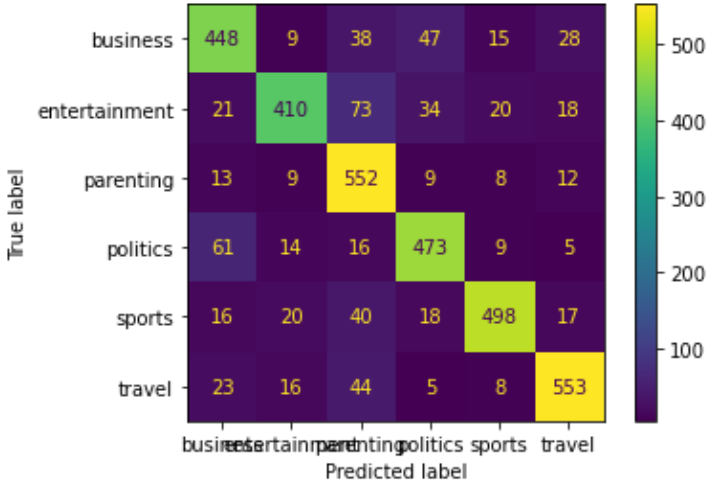
[Accuracy]

	precision	recall	f1-score	support
business	0.90	0.80	0.85	306
entertainment	0.88	0.76	0.82	285
parenting	0.77	1.00	0.87	285
politics	0.85	0.89	0.87	309
sports	0.92	0.85	0.89	285
travel	0.95	0.92	0.94	313

[Report by category]

Test Results – Using news test-set, test news classifier

- Test the classifier trained with news train-set, using news test-set



[Confusion matrix]

train-set size	8400
test-set size	3600
accuracy	**0.81**

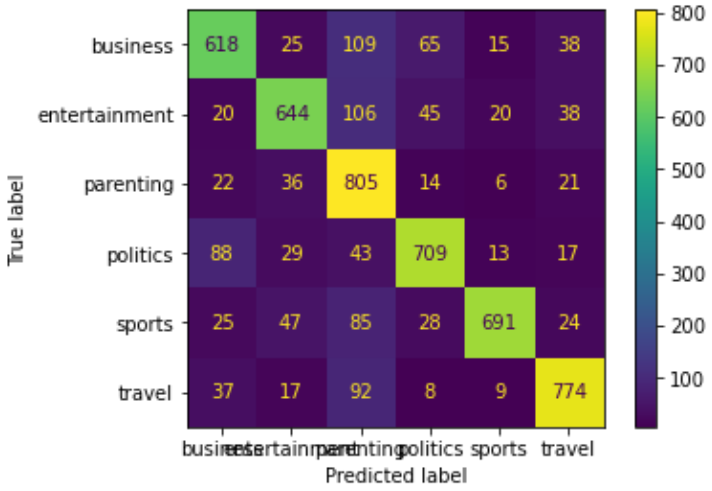
[Accuracy]

	precision	recall	f1-score	support
business	0.77	0.77	0.77	585
entertainment	0.86	0.71	0.78	576
parenting	0.72	0.92	0.81	603
politics	0.81	0.82	0.81	578
sports	0.89	0.82	0.85	609
travel	0.87	0.85	0.86	649

[Report by category]

Test Results – Using combined test-set, test classifier trained combined-set

- Test the classifier trained with combined train-set, using combined test-set



[Confusion matrix]

train-set size	12558
test-set size	5382
accuracy	**0.79**

[Accuracy]

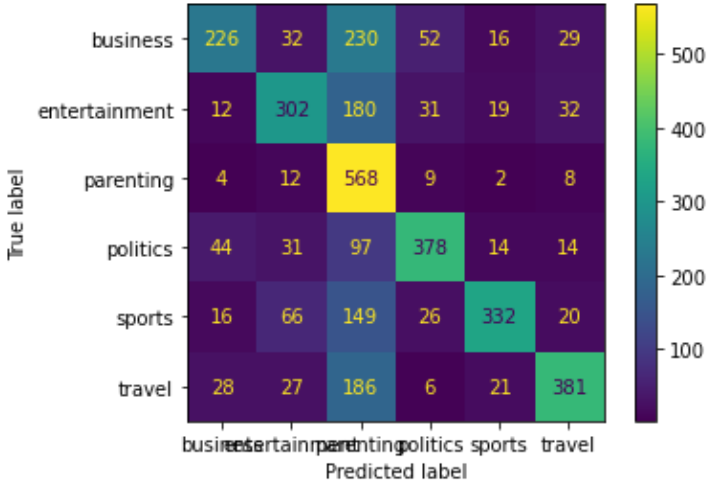
	precision	recall	f1-score	support
business	0.76	0.71	0.74	870
entertainment	0.81	0.74	0.77	873
parenting	0.65	0.89	0.75	904
politics	0.82	0.79	0.80	899
sports	0.92	0.77	0.84	900
travel	0.85	0.83	0.84	937

[Report by category]

- 'combined set' and 'both' mean a dataset combined with reddit and news dataset

Test Results – Using news test-set, test reddit classifier

- Test the classifier trained with reddit train-set, using news test-set



[Confusion matrix]

classifier	reddit
test-set	news
accuracy	**0.61**

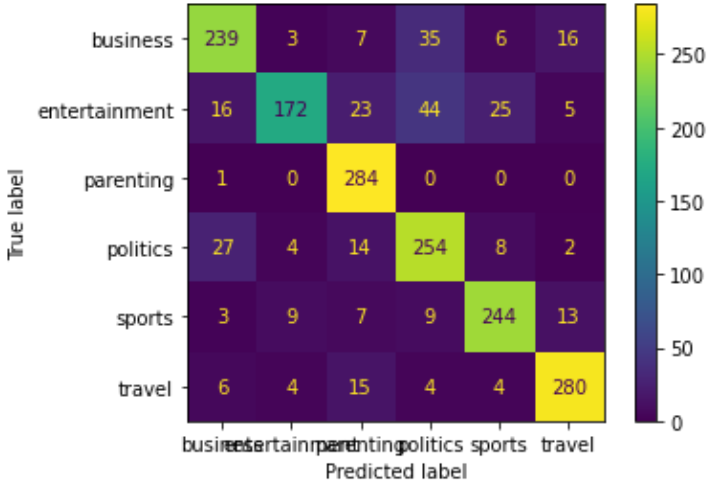
[Accuracy]

	precision	recall	f1-score	support
business	0.68	0.39	0.49	585
entertainment	0.64	0.52	0.58	576
parenting	0.40	0.94	0.56	603
politics	0.75	0.65	0.70	578
sports	0.82	0.55	0.66	609
travel	0.79	0.59	0.67	649

[Report by category]

Test Results – Using reddit test-set, test news classifier

- Test the classifier trained with news train-set, using reddit test-set



[Confusion matrix]

classifier	news
test-set	reddit
accuracy	**0.83**

[Accuracy]

	precision	recall	f1-score	support
business	0.82	0.78	0.80	306
entertainment	0.90	0.60	0.72	285
parenting	0.81	1.00	0.89	285
politics	0.73	0.82	0.78	309
sports	0.85	0.86	0.85	285
travel	0.89	0.89	0.89	313

[Report by category]

Test Results

- Even posts in the same category have different tendencies depending on the community (source of posts)
- Combining the two datasets to train classifier and test it also shows quite high accuracy, but not as much as the case of individual
- If a category classification system is created considering the community that is the source of the post, higher accuracy can be achieved with a much smaller dataset

	Classifier	Test-set	accuracy
Test - 1	Reddit	Reddit	**0.87**
Test - 2	News	News	**0.81**
Test - 3	Both	Both	**0.79**
Test - 4	Reddit	News	**0.61**
Test - 5	News	Reddit	**0.83**

[Summary of each test accuracy]

Thank you

(Github for project : https://github.com/14KGun/CS372-Community-Category-Classification)