Classification of online post categories based on community tendencies

20190052 Geon Kim, 20190703 Geonha Hwang

1 Introduction

Community users read and write by selecting the appropriate category to communicate with other users about the topic they are interested in. If community writings are not properly categorized, this will cause dissatisfaction with community users and lead to community decline in the long run.

Therefore, it is important to provide clear categories so that users can choose the appropriate categories. Since, as time goes on, the online community not only grows rapidly, but also the tendency and categories of the community are diversifying. a text can have multiple themes, and there are more categories to choose. This makes it difficult for users to select categories.

So, in terms of the community system, it will be increasingly necessary to recommend appropriate categories to users. The important point at this time is that even if the text is the same, it can be classified into different categories depending on the tendency of the community. For example, suppose there is a post about the experience of playing games with your child. If the 'game' and 'childcare' categories are the same within the parent community and the game community, it would be appropriate to classify them as 'game' categories in the former and 'childcare' categories in the latter. This is because it is natural that childcare content is basically included in the parent community post and game content is basically included in the game community post.

Therefore, the purpose of this study is to organize a system that classifies the user posts in consideration of the tendency of the community, rather than simply classifying them into the most relevant categories.

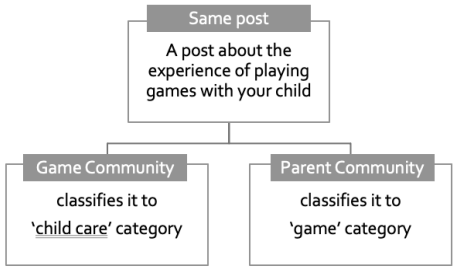


Figure 1: Example about community tendency

2 Problem Definition

2.1 Problem Description

If you simply classify the content of the post without considering the tendency of the community, it is classified into the most relevant category.

However, as in the example of introduction, there are features that become less important according to the tendency of the community. Therefore, it is a wrong approach to train a classification model on random text that is independent of the community.

In order to create a classification model considering the propensity of the community, this study creates a model based on the accumulated and already classified posts of the community.

2.2 What community we have to use

Posts from any community cannot be used as data. There should be many posts accumulated in each category enough to be used for model training. This is because, as the category is selected by the user who wrote the post, there will be noise in which the category is incorrectly selected.

3 Objective

Present a model that recommends a new post category based on the online community's accumulated categorized posts, which reflects the tendency of the community.

4 Related Work

- Ankita Dhar et al., Text categorization: past and present, 2020

  This paper summarizes the methods of text classification.

- Sebastian Raschka, Naive Bayes and Text Classification I, 2014

  This paper describes the method of Naive Bayes Classification.

5 Datasets

To solve the problem, we decide to use two datasets. And they have the same categories { business, entertainment, parenting, politics, sports, travel }.

5.1 Crawled reddit dataset

We do crawling the reddit (https://www.reddit.com) and place it in one row per one post. Each post has a title, content and corresponding category. 1000 posts will be collected per category, and a total of 6000 posts will be collected. The collected dataset is shared through GitHub.

5.2 Process of crawling

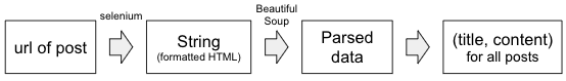


Figure 2: Process of crawling

Using selenium and bs4 packages, all posts are stored as one row in the csv file through the following process in Figure 1. We extract the title and content information of the post.

5.3 Dataset Description

One row consists of six columns and the dataset has a total of 6000 rows.

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

Table 1: Description of reddit dataset

5.4 News Category Dataset

- Kaggle News Category Dataset

It has 200k news headlines obtained from HuffPost and it is shared in kaggle. It is used as a dataset to compare with our model. This dataset is pre-processed in the same format as the previous dataset and is compared to our model by learning another model in the same way.

5.5 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

Table 2: Description of news dataset

6 Method

6.1 Pre-processing

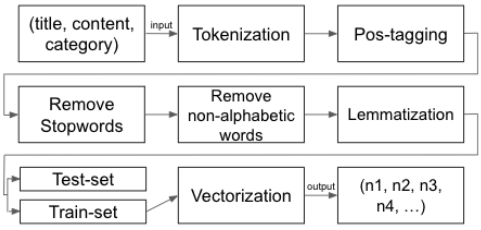


Figure 3: Pre-processing process

In dataset, all posts are tokenized using the nltk's sentence segmenter of and word tokenizer. And each token is tagged with a part-of-speech tagger of nltk. After tagging, tokens with stopword are removed using nltk's stopword corpus. Using WordNet lemmatizer, we lemmatize the tokens. And each post will be divided into a set for test and a set for train at a ratio of 3:1 in a random order. Finally, the train-set will be vectorized using CountVectorizer of keras.

6.2 Train Classifier

The pre-processed train-set is used to learn the naive bayes classifier provided by the scikit learn package. We train three naive bayes classifier. They are based on reddit dataset, based on news dataset and based on a dataset that be combined with reddit and news datasets.

6.3 Using Classifier

Figure 4: How to use trained classifier

The pre-processed test-set can be entered into the learned classifier. The input value(list of tokens) is entered into the classifier after vectorization based on the train dataset tokens.

7 Test

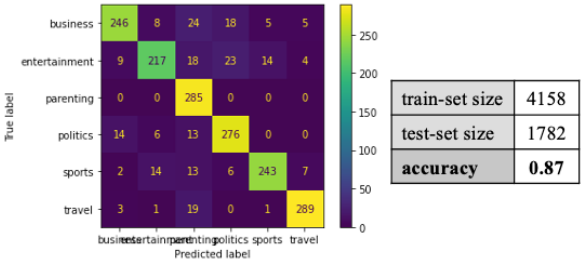
7.1 Test Method

To test three classifiers, we put the test-set into them. And we test three 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

Table 3: Five ways of test

7.2 Test-1, Using reddit test-set, test reddit classifier



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

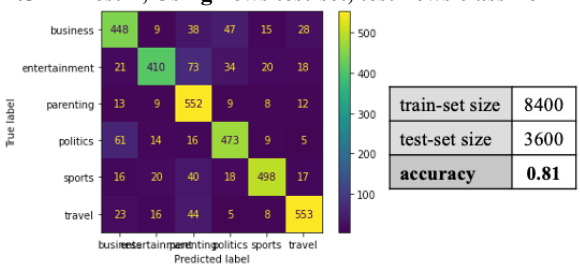
Figure 5: Confusion matrix of Test-1

	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

Table 4: Result of Test-1

This is the result of testing the reddit test-set with a model trained with the reddit trainset.

7.3 Test-2, Using news test-set, test news classifier



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

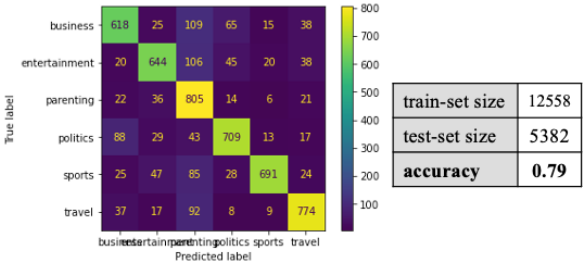
Figure 6: Confusion matrix of Test-2

	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

Table 5: Result of Test-2

This is the result of testing the news test-set with a model trained with the news train-set.

7.4 Test-3, Using combined test-set, test classifier trained combined-set



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

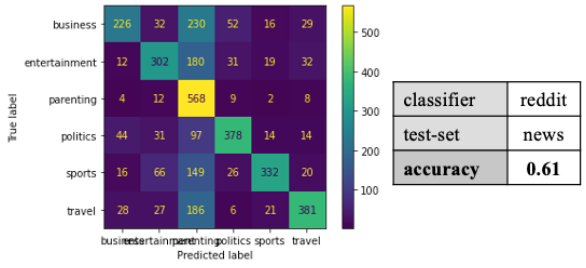
Figure 7: Confusion matrix of Test-3

	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

Table 6: Result of Test-3

This is the result of evaluating the model by merging (mixing) the two data sets and dividing it into train-set and test-set.

7.5 Test-4, Using news test-set, test reddit classifier



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

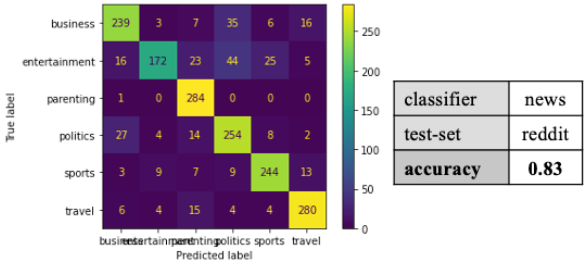
Figure 8: Confusion matrix of Test-4

	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

Table 7: Result of Test-4

This is the result of testing the news test-set with a model trained with the reddit train-set.

7.6 Test-5, Using reddit test-set, test news classifier



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

Figure 9: Confusion matrix of Test-5

	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

Table 8: Result of Test-5

This is the result of testing the reddit test-set with a model trained with the news train-set.

8 Discussion

Test 1 and Test 2 are general model evaluation methods using the same data set to train and evaluate the model. Comparing the results of these two test sets is meaningless.

The first pair of results to be compared is test 1 and test 5, and test 2 and test 4. Each pair is two tests of the same test set. Comparing the two results of each pair, in both cases, the test using the same dataset for train and set has high accuracy. In a way, this seems like a natural result. However, it is important to note that the two datasets used in this study have the same category. Although both datasets are text bundles classified into the same category, the fact that there is a clear difference in accuracy means that each dataset has unique characteristics that cannot be divided by category alone. This analysis is further supported by the results of test 3. Even though the dataset become larger, its accuracy is lower than test 1 and test 2 which use two datasets separately. It means that the two datasets have the same category, but the inherent characteristics are significantly different to the extent that the accuracy decreases when the two datasets are combined. These inherent characteristics is the unique tendency of each community we are talking about. It will blur the tendencies of that particular community that utilizing datasets from other sources of same categories just for using bigger dataset. From these analyzes, we argue that when classifying community posts, only the accumulated posts of the community should be used unconditionally.

Our code and dataset are shared in the GitHub.

- https://github.com/14KGun/CS372-Community-Category-Classification

9 References

- Ankita Dhar et al., Text categorization: past and present, 2020

  https://link.springer.com/article/10.1007/s10462-020-09919-1

- Kaggle New Category Dataset

  https://www.kaggle.com/datasets/rmisra/news-category-dataset

- Sebastian Raschka, Naive Bayes and Text Classification I, 2014

  https://arxiv.org/abs/1410.5329