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, there are more choosable categories. 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.

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

5.1 Collect the Dataset

To solve the problem, we decided to create our own dataset. We will 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. The category of each post collected corresponds to one of the categories of the following set: { entertainment, politics, travel, parenting, business, sports }. 4000 posts will be collected per category, and a total of 24000 posts will be collected. The collected dataset will be shared through GitHub.

Column	Description
title	title of post
content	content of post
category	[entertainment | politics | travel | parenting | business | sports]

Table 1 : Dataset Description.

5.2 Process of crawling

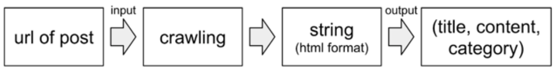

Figure 1: A figure with a caption that runs for more than one line.

Using urllib and bs4 packages, all posts are stored as one row in the csv file through the following process Figure 1. Extract the title and content information of the post through the Beautiful Soup of the bs4 package.

5.3 Dataset Description

One row consists of three columns and the dataset has a total of 24000 rows.

5.4 News Category Dataset

- Hugging Face, News Category Dataset

  It is used as a dataset to compare with our model. This dataset will be pre-processed in the same format as the previous dataset and will be compared to our model by learning another model in the same way.

6 Method

6.1 Pre-processing

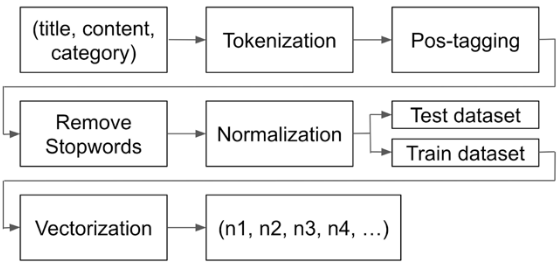

Figure 2 : Pre-processing process.

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, do lemmatization 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 dataset is used to learn the naive bayes classifier provided by the scikit learn package.

6.3 Using Classifier

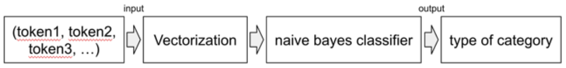

Figure 3 : How to use trained classifier.

The pre-processed test dataset 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 Evaluation

7.1 Evaluation Method

The classifier will be measured Accuracy, Precision, Recall, and F-score using the pre-processed test dataset. Our target accuracy is 0.7 or higher.

8 References

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

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

- Hugging Face, New Category Dataset

  https://huggingface.co/datasets/Fraser/news-category-dataset

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

  https://arxiv.org/abs/1410.5329