Classification of Reddit Text Posts

Pradeep Kumar Srinivasan
Purdue University
sriniv68@purdue.edu

Mohammad Haseeb
Purdue University
mhaseeb@purdue.edu

ABSTRACT

We plan to work on a large text classification problem, specifically the Reddit Self-Post Classification Task (RSPCT). RSPCT is a dataset with around 1000 classes ("subreddits") with around 1000 examples per class, which is unique because most text classification datasets have sparse labels. We plan to use two traditional machine learning algorithms and one deep learning algorithm to learn to predict the class given the title and body of a post.

Keywords

NLP; large text classification; multi-class classification

1. INTRODUCTION

2. PLANS FOR EXPLORATORY ANALYSIS

We will use techniques described in [1].

```
Number of words per sample: 99.0
```

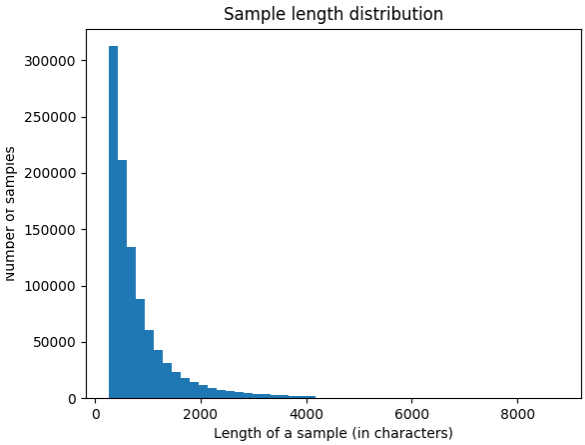

Figure 1: Distribution of sample length (in characters)

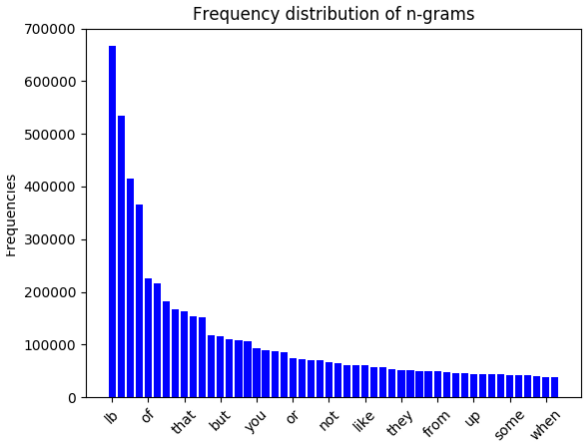

Figure 2: Frequency distribution of n-grams

3. ALGORITHMS WE AIM TO IMPLEMENT

4. EVALUATION OF OUR PROJECT

The key metric for the models will be the Precision-at-K metric [2], since this is a problem with a very large number of classes and we don't expect the correct label to be predicted

DOI: 10.475/123_4

as the top option. We will measure the precision on the top-1, top-3, and top-5 labels. As mentioned before, we will also experiment with adding a new feature for sentiment rating of the text to see if that improves performance. To understand the impact of the title on classification, we will measure performance by training and testing on just the titles of the posts, just the bodies of the posts, and then with both title and body.

In addition, for the deep learning model, we will plot the loss and accuracy curves vs number of epochs. For the traditional models, we will plot the learning curve vs training set size.

5. CONCLUSIONS

6. REFERENCES

[1] Google Developer Guide on exploring text classification data. https://developers.google.com/machine-learning/guides/text-classification/step-2.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.