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

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

Text classification with few classes, such as in sentiment analysis, has been well-studied [TODO] with state-of-the-art techniques like LSTM. However, those techniques do not always work as well in scenarios with many classes [TODO]. Another issue with training on many-class datasets is that they have a very large number of labels [TODO] and are sparse - most labels have very few examples [TODO].

Reddit Self-Post Classification Task (RSPCT) is a text corpus containing self-posts (i.e., text posts) from Reddit. RSPCT was collected to help spur research on models that could tackle a large number of classes by ensuring a large population of examples for each class (currently, around 1000 examples for each class) and a large number of classes (around 1000). The aim is to allow for a situation comparable to that in the computer vision community, which was helped by the famous ILSRVC competition (ImageNet [TODO]) with 1000 classes and around 1400 examples per class. This potential is what made us find this project interesting.

TODO: Why is this a useful problem to solve? How is our approach innovative?

2. PLANS FOR EXPLORATORY ANALYSIS

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

3. ALGORITHMS WE AIM TO IMPLEMENT

We aim to implement two traditional classification algorithms and one deep learning algorithm.

First, we plan to use Naive Bayes classification because of its simplicity and past track record in multi-class text classification [TODO]. This will provide us with a good baseline against which to test the other models. Just in case, we will also use a trivial classifier (one that predicts classes at random) as another baseline.

Second, we plan to use logistic regression with a bag-of-words model [TODO] based on word n-grams and character n-grams. Since we are not sure if logistic regression or SVM will be better suited to this problem, we plan to answer this question using our literature review. If prior research suggests SVM performs better on large text classification, we may switch to it.

Lastly, we plan to use Convolutional Neural Networks (CNN) to test whether deep learning models can outperform the traditional models on text classification problems with many classes. Again, we will check if the existing literature recommends other deep learning algorithms for text classification and may switch to something like RNN.

4. PROPOSED EXPERIMENTS

On top of the previous three algorithms, we will experiment with adding a new feature - the sentiment of the text (which we will compute using a pre-trained, off-the-shelf model) - to see if improves performance.

5. EVALUATION OF OUR PROJECT

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

6. TIMELINE

Week 1: We will do exploratory data analysis and study the literature to learn the strengths and weaknesses of past approaches towards large text classification.

Week 2 and Week 3: We will implement and train the different models.

Week 4: We will tune our hyperparameters to get the best performance on the validation sets. We hope to fine-tune variables like learning rate, batch size, and number of epochs.

Week 5: We will spend this week writing our report and presentation (and finishing any work that is left from the previous weeks).

6.1 Citations

Citations to articles [1, 3, 2, 4], conference proceedings [3] or books [6, 5] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [5]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

7. CONCLUSIONS

TODO

8. REFERENCES

[1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[2] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[3] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[4] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[5] L. Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

[6] S. Salas and E. Hille. *Calculus: One and Several Variable*. John Wiley and Sons, New York, 1978.

8.1 References