

Classification of Reddit Text Posts

Pradeep Kumar Srinivasan
Purdue University
sriniv68@purdue.edu

Mohammad Haseeb
Purdue University
mhaseeb@purdue.edu

ABSTRACT

In this project, we worked on the problem of large text classification, 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 used two traditional machine learning algorithms and one deep learning algorithm to learn to predict the class (i.e., the subreddit) given the title and body of a post. We evaluated the performance of each model using the Precision-at-K metric. We conducted additional experiments like adding features for sentiment analysis and readability level of a post to see if that increases the performance of our models.

Keywords

NLP; large text classification; multi-class classification

1. INTRODUCTION

Text classification with few classes, such as in sentiment analysis, has been well-studied [2] with state-of-the-art techniques like LSTM. However, those techniques do not always work as well in scenarios with many classes [3]. Another issue with training on many-class datasets is that they have a very large number of labels (such as WikiLSHTC-325K dataset [6], which has 325K labels) and are sparse - most labels in the above dataset have less than 100 examples [4].

We aim to study this problem by using a dataset that has a large number of classes but also a large number of examples per class. To this end, we used the Reddit Self-Post Classification Task (RSPCT) dataset and trained three models and evaluated their performance.

We think this is a useful problem to solve because it has not been adequately studied, to the best of our knowledge, and because this could pave the way to future work on such many-class datasets.

2. LITERATURE REVIEW

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

???

3. DATASET

Reddit is a popular social link aggregation, web content rating, and discussion website. Reddit's users submit posts, which are then voted up or down by other members. This allows the highest rated posts to gain the most attention. Reddit is organized into "subreddits", which are sub-forums on Reddit dedicated to a specific topic, for example r/geography, r/gameofthrones, r/MachineLearning.

There are two main types of posts a user can submit on Reddit: a simple URL link and a self-post. A self-post consists of a topic and a body of text and is generally much longer in length, thereby, providing more information to Redditors. Moreover, most of the self-posts talk closely about the subreddit that they were posted in, which makes for a good classification task.

We obtained the Reddit Self-Post Classification Task (RSPCT) dataset from Kaggle [1]. It 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. The data consists of 1.013M self-posts, posted from 1013 subreddits (1000 examples per class). For each post, its title, body and the subreddit that it belongs to are provided.

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 [7]) with 1000 classes and around 1400 examples per class. This potential is what made us find this project interesting.

4. PREPROCESSING

As mentioned in the previous section, the RSPCT dataset contains real self-posts from Reddit. This means that the text in these posts may not always conform to the grammar, spelling, vocabulary and other nuances of the English language. For this reason, we carried out a series of preprocessing steps.

Because this is a text classification problem and most of our models used the bag-of-words approach, we first lowercased all the text in the subreddit, title and body of each post. We noticed that many of the posts contained special HTML encoded text and tags like >, &, <, <lb>, <tab> etc. We used regular expression matching to replace these tags with a blank space. We also removed punctuation characters like !, ", #, \$, %, &, (, [, *, + etc. such that we were only left with words at the end of this step.

Finally, we used Python's NLTK's PorterScanner to replace all words with their stem words. This resulted in words like "extremely" and "extreme" to be replaced with "extrem". *Figure __ shows the whole process for pre-processing one self-post*

5. EXPLORATORY DATA ANALYSIS

5.1 Distribution of Characters in Post and Title

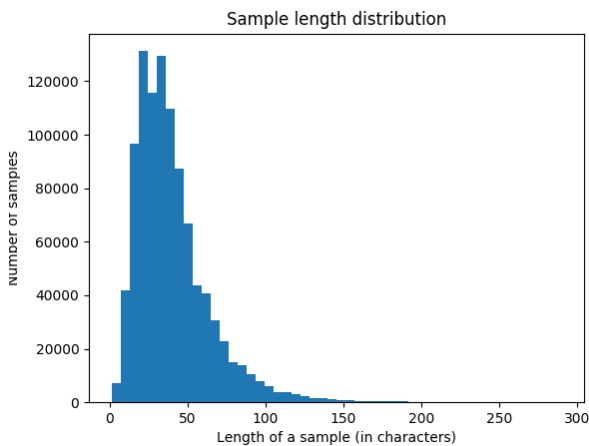


Figure 1: Sample Distribution for Length of Post's Title (in characters)

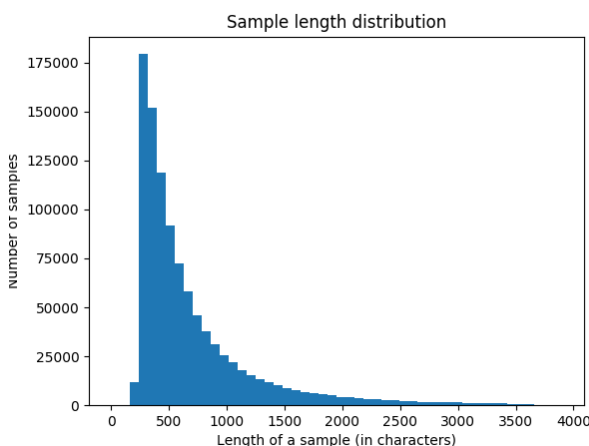


Figure 2: Sample Distribution for Length of Post's Body (in characters)

5.2 Feature Importance

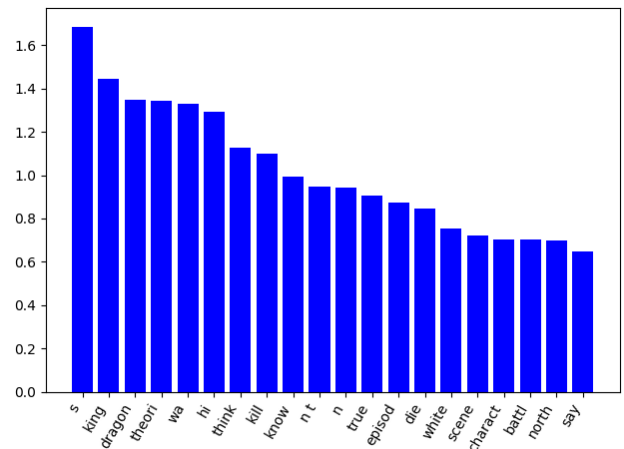


Figure 3: Important Features for subreddit r/gameofthrones

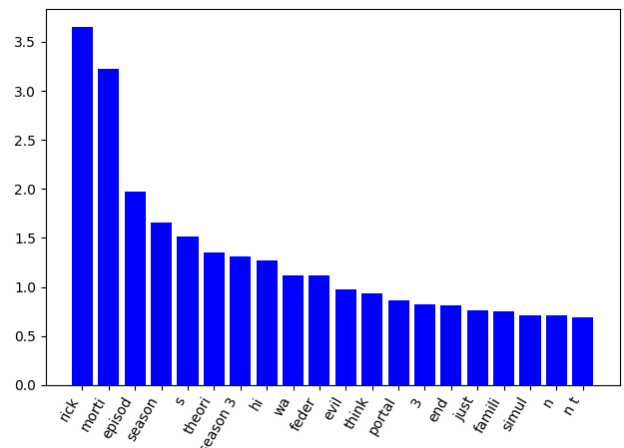


Figure 4: Important Features for subreddit r/rickandmorty

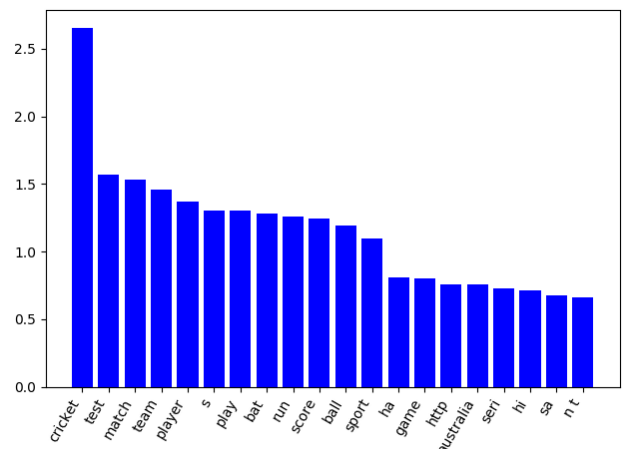


Figure 5: Important Features for subreddit r/cricket

5.3 TSNE

???

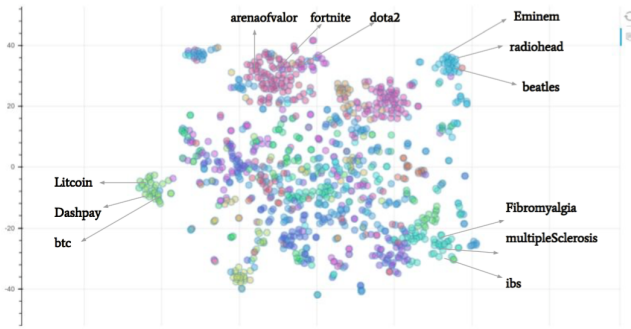


Figure 6: TSNE Plot of All the Self-Posts

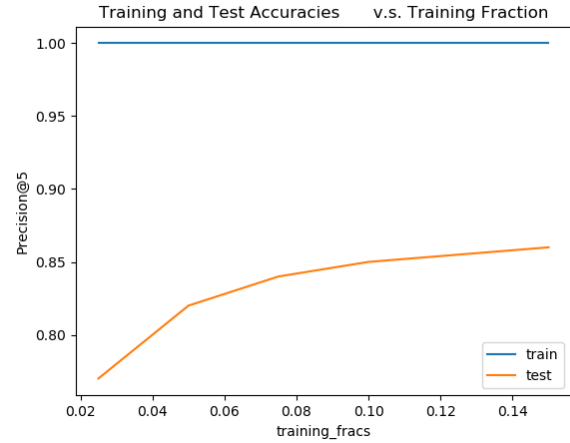


Figure 8: Learning Curve for Logistic Regression

6. ALGORITHMS IMPLEMENTED

6.1 Naive Bayes Classifier (NBC)

6.1.1 Hyperparameter Tuning: α

6.1.2 Performance

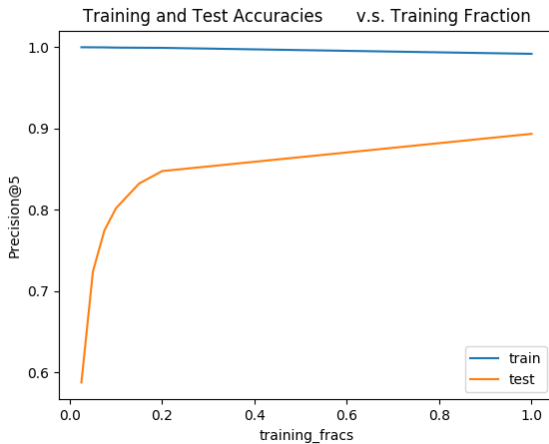


Figure 7: Learning Curve for Naive Bayes Classifier

6.2 Logistic Regression (LR)

6.2.1 Hyperparameter Tuning: C

6.2.2 Performance

6.3 Convolutional Neural Network (CNN)

6.3.1 Hyperparameter Tuning: num_epochs

6.3.2 Performance

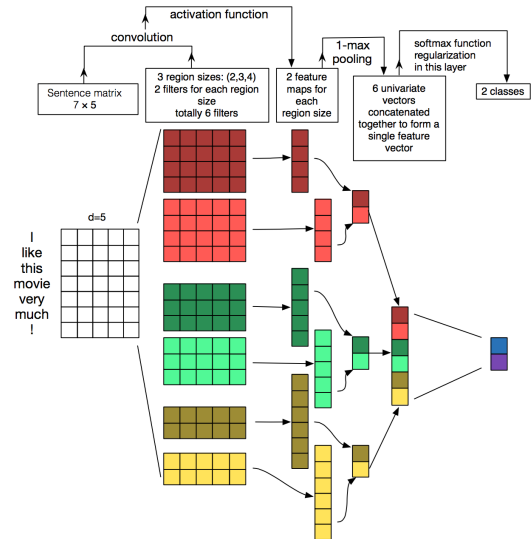


Figure 9: CNN Architecture

7. FURTHER EXPERIMENTS

7.1 Toggle Title and Body

8. EVALUATION OF OUR PROJECT

Compare with the baseline provided in the Kaggle kernel

The key metric for the models will be the Precision-at-K metric [5], 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.

9. CONCLUSIONS

10. TEAM MEMBER CONTRIBUTIONS

11. REFERENCES

- [1] The reddit self-post classification task. <https://www.kaggle.com/mswarbrickjones/reddit-selfposts>.
- [2] A. Agarwa. Sentiment analysis of twitter data. *Association for Computational Linguistics*, 2011.
- [3] R. Ghani. Combining labeled and unlabeled data for multiclass text. 2001.
- [4] M. S. Jones. The reddit self-post classification task (rspct): a highly multiclass dataset for text classification (preprint).
- [5] 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.
- [6] I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari. Lshtc: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*, 2015.
- [7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li. Imagenet large scale visual recognition challenge. *CoRR abs/1409.0575*, 2014.