Reddit posts classification using NLP

Capstone 1- Project Proposal
Springboard Data Science Career Track
Vinod Kumar Varala, August 8th, 2021

Motivation:

Reddit is one of the most used social media to share/discuss about almost everything that is happening to us or around us. Conventionally, with increased user activity, it becomes challenging to filter the posts based on the content for whatever reason may be- either to prevent hateful posts or to just further classify the post to a specific sub-reddit. This project aims to just do the same- to classify incoming stream of posts to their respective sub-reddits using natural language processing (NLP) techniques. Hateful posts can also be screened in a similar fashion but this is outside the scope of this project.

Problem Statement:

To use a very large text corpus (~1M posts) from Reddit to design a machine learning model that would correctly classify a post (or a new post) into its class (sub-reddit) from a predefined set of classes (1013 classes). Here are more details about the various steps needed for this task.

1. Data collection: The dataset is obtained from Kaggle[1]. It consists of 1.013M posts labelled into 1013 classes (about 1000 posts per class). This dataset does not suffer from label sparsity as most datasets of its type do. Kaggle made the effort into selecting a 'good' set of subreddits to minimise overlap in content. For the purpose of this project fewer classes will be chosen.
2. Pre-processing (Data-wrangling): Key steps include formatting, removing stop words, tokenization, lemmatization.
3. Exploratory data analysis (EDA): The goal is to get a set of most likely words that represent each class. Training/testing split (80:20)
4. Modeling: Classification algorithms starting from simple NaiveBayes to more complex FastText. Model parameters are tuned using GridSearchCV.
5. Testing and validation: Once a few models are narrowed down, we test the performance of those on test data and decide the best model to use.
6. Testing on real data: The goodness/accuracy of the model will eventually be tested against a new stream of real data which will be accessed by Reddit API. The goal here is to get the new data into a format that the model can use, by running it through the data cleaning/transformation flow.
7. Final accuracy: Model performance is evaluated based on the accuracy scores.

Some background information on data:

Reddit is divided into various ‘subreddits’ based on the types of posts being submitted, for example r/politics or r/MachineLearning. Subreddits are generally created and moderated by the users themselves, rather than the admins of reddit. Kaggle found from ad-hoc analyses that the large majority of self-posts were talking about the topic that their subreddit implied, suggesting that this may be an interesting task from a machine learning perspective. They downloaded all the self-posts in a two year period (2016/06/01 --- 2018/06/01), and did a number of cleaning steps to try and find posts that were sufficiently detailed with minimal overlap in content.

Challenges:

1. Since there are many classes in the classification task, it is not a straightforward problem to solve.
2. Traditional classification algorithms like LSTM do not always translate seamlessly to the many-class domain.

Citations:

[1] https://www.kaggle.com/mswarbrickjones/reddit-selfposts