Project Proposal
CS 5785 - Applied Machine Learning
Professor Kuleshov

James Chu (qjc2)

October 3, 2022

Project Title: Reddit Post Classification
Category: Application of machine learning to a practical problem

Reddit is an online community where users are able to submit content to the site under user-created boards called subreddits. As Reddit grows its user base, the number of subreddit communities will increase, and each subreddit will become increasingly narrow in its scope of discussions (e.g. posts for "Hyundai cars" will spin up a new subreddit, rather than being aggregated under a generic "car" subreddit). Therefore, Reddit users will find it increasingly difficult to locate the appropriate subreddit to post their content. This project will develop a classifier that will aim to identify the most appropriate subreddit aligned to the content of a user-generated post. As a byproduct, we'll also be able to identify the topics of each subreddit by revealing the most commonly used words in that subrreddit's posts.

For this project, we'll be using this labelled data set from Kaggle with 1.03M rows containing the title, content, and subreddit of each post. As the data set is extremely large, we'll plan to only use a subset of the data for training and development. The bag of words model will be used to vectorize the title and content of each post, and standard pre-processing steps will be applied (e.g. lemmatization, stemming, stripping punctuation and stop words). We'll then leverage several supervised learning algorithms such as Naive Bayes, Logistic Regression, and Nearest Neighbors on the frequencies of specific words to perform the multi-class classification. The goal is to determine the classifier that resulted in the highest F1-score.

The following experiments will be performed and assessed against the F1-score:

a) Adding regularization terms with varying regularization strengths

b) For Naive Bayes, experimenting with different values of alpha for Laplace smoothing

c) Applying different threshold values for occurrences of words (i.e. `min_df`)

d) Binary vs. non-binary feature vectors

e) Using the n-gram model when vectorizing the text

f) Applying the supervised learning techniques on the title of the post by itself, the content of the post by itself, and both the title and post

content combined

g) For Nearest Neighbors, experimenting with different values of k