Project 3 - Classifying Reddit Posts

The primary objective of this project is to utilize natural language processing to classify which subreddit a post comes from. All models evaluated in this repo used post titles as the only text. A secondary objective is scraping data for use. This repo uses Pushshift API to grab raw data to be transformed into ready-to-use data.

Problem Statement:

Build a model that will accurately predict whether a given post is from one of two subreddits. Subreddits evaluated in this repo are:

1. LifeProTips
2. UnethicalLifeProTips

Summary:

We start out establishing a baseline to compare future models to. Our baseline accuracy was 50.3% where we simply predict the majority class, LifeProTips, for all posts. We evaluated 4 separate model types and performed a gridsearch across various different parameters for each type. Below are the models considered and their highest accuracy achieved when only text is considered without lemmatization.

1. Logistic Regression (.832792)
2. Extra Trees Classifier (.821834)
3. Random Forest Classifier (.819508)
4. KNN (.767857)

When lemmatization is included, our best model is Logistic Regression achieving an accuracy of .833198. The same score was achieved when including average length of words in the title and number of words in the title as additional features.

Areas for Improvement:

There are a couple of areas that could prove to be beneficial when evaluating this problem but we were did not implement as of yet.

1. Using the body of the text or even comments could be considered as in some cases titles are very short or simple and the body contains the majority of relevant text. And in the case of 'request' type posts the comments are an important part of the post as a whole.
2. Sentiment analysis could prove to be beneficial given the nature of these subreddits.

Note: Ultimately these posts from random people on the internet and them being posted to one subreddit or the other does not truly indicate whether something is ethical or not. Models created here do not indicate ethicality, simply whether post belongs to one subreddit or another.