John Flanigan

December 19, 2019

Reddit Sports Subreddit Analysis

Summary

Social media, online news, and content aggregation websites have changed the way people consume media since the invention of the World Wide Web. One of the largest websites of this type is Reddit, which was founded in 2005. It has since grown to become the 5th most visited website in the United States and the 13th in the world[1]. Reddit is organized into user-created boards called “subreddits” where users can share links, text posts, or media and discuss in a threaded comment section. The subreddits are user created but generally focus on a single topic (e.g. a hobby, a specific video game, or a type of content).

Some of the more popular subreddits on Reddit are sports related; the general sports subreddit is the 26th most popular on the site and the NBA subreddit is the 68th most popular.[2] All United States major leagues and their teams have subreddits. This means there is a large amount of content, particularly text content, generated on Reddit by users. This content is relatively accessible and can be retrieved through the Reddit Application Programming Interface (API) or by scraping the website. The goal of this project is to collect data from various sports subreddits and search for trends within the data. Because the majority of the content on Reddit are text comments, text analysis tools and practices will be used to analyze the data. If trends or patterns are identified in the data, the results can be used in a variety of ways. Therefore, there is potential for either commercial or academic funding depending on how the results will be used. Although there is potential for commercial or academic applications of the results, the primary focus of this paper will be on commercial uses.

Related Work

There has been work done on subreddit analysis previously. One paper, CS 229: r/Classifier - Subreddit Text Classification, attempts to create a classifier to match posts to subreddits[3]. This is very similar to the goal of this project; one significant difference, however, is that the CS 229 paper analyzes 12 subreddits that are on different subjects. They chose subreddits to analyze based on the number of text posts on the subreddits. The paper also only examines the original post and does not analyze comments when creating the classifier.

Implementation

Manual Prework

The first step of this project was to identify the subreddits to analyze. American sports teams' subreddits were chosen because they likely have a number of similarities with a just few defining features. They are also spread out geographically meaning a classifier might be able to pick up on regional language. Another advantage of sports subreddits is that they every team has a subreddit. The first step of this project was to identify all of the subreddits to analyze. This was a manual process that first required retrieving a list of all teams for each major American sports league (NFL, MLB, NHL, and NBA). Then a search engine was used to find the subreddit for each league. The league subreddit contained links to all individual team subreddits. The team subreddit names were then recorded in a Jupyter notebook for later use.

Data Gathering

Once the subreddits had been identified, the next step was to gather data for each subreddit. Initially, the Reddit API was used to gather data[4]. The Reddit API is relatively easy to consume and provides useful features such as retrieving data in JSON format instead of HTML but does have some challenges. For example, not all comments are retrieved immediately. The response from the initial request only contains text for top level comments. Other comments are only linked. Additional requests must be made to retrieve the text contents. Because this data gathering was more complicated than originally anticipated, alternative approaches were explored.

The easiest way to retrieve data was using the Python Reddit API Wrapper (PRAW)[5]. This package provided an easy-to-use Python interface to consume the Reddit API; the PRAW documentation even included an example on how to perform comment extraction. In order to have enough data to analyze, the top 100 posts (by user votes) for each subreddit were retrieved and then all comments on these posts were stored in a list. A challenge encountered at this step was that the Reddit API is rate limited, meaning only a certain number of requests can be made per minute. Because of this, the data gathering took longer than expected and had to run for about 7 hours.

Model Development

Once the data had been retrieved, the next step was to transform the data into something that a model can be trained on. Two approaches to text analysis were used, bag-of-words and Word2vec. To create the bag-of-words, first the number of occurrences of each word were counted and any word appearing fewer than 200 times was filtered out. This was to shrink the dataset from the 118,110 unique words down to 5,003 words. Once the most used words were identified, a bag-of-words was created for each post by counting the number of occurrences of each popular word in the post's comments. These bags-of-words were then used to create a Naïve Bayes and a Logistic Regression model using the scikit-learn package. Three models of each were created: one to classify the city, one to classify the sport, and one to classify the team.

Another set of models were created using Word2vec. The Word2vec implementation used came from the Gensim package[6] and used a pretrained embedding trained on a Google News corpus[7, 8]. Word vector averaging was applied to the pre-filtered comment text to create data to train and test the models. This time, only Logistic Regression models were created for each of the three labels to predict.

Final Product

The results of these models demonstrate that labels can be correctly applied to subreddits based on comments on posts. Overall, the bag-of-words model outperformed the Word2vec model significantly. Below are the overall accuracies. More data on the classifier's performance can be found in the Jupyter notebook.

Model Accuracies on Test Data

Pre-processing	**Model Type**	**City**	**Sport**	**Team**
Bag-of-words	**Naïve Bayes**	0.505	0.941	0.550
Bag-of-words	**Logistic Regression**	0.556	0.920	0.530
Word2vec	**Logistic Regression**	0.092	0.823	0.049

The Naïve Bayes and Logistic Regression models resulted in similar accuracies. Naïve Bayes was better than Logistic Regression at classifying the sport and team while the Logistic Regression was better at classifying the city. The high accuracy of the sports classifier compared to the city and team classifiers indicate that there is the language used to talk about sports is more different than the language used in different cities. This makes sense because of how many words there are that are specific to sports. For example, touchdown would indicate comments are discussing football and bat would be used when discussing baseball. Surprisingly, the Word2vec had a significantly lower accuracy across all three categories and particularly struggled with city and team classifications. This is likely because the bag-of-words models were able to heavily weight keywords that only apply to certain sports or cities. Additional work could be done to refine these models and develop more sophisticated data pre-processing.

Work Plan

These results show that it is possible to predict certain labels based on the comments of a post. Based on these preliminary results, additional work could be put in to refine the models to give better results and expand the analysis beyond sports subreddits. These machine learning techniques could also be expanded to compare subreddits and identify similarities between subreddits. Subreddit comparisons would have a number of applications. A user could be given recommendations of new subreddits to explore based on the user's interactions with similar subreddits. It could also be a useful tool for advertisers to target their ads. Advertisers can specify what subreddits they want to run their ads on so similarity data could be used to suggest additional subreddits to run ads on based on past successes. There are a number of valuable applications that this analysis can be applied if this work is continued.

Sources

1. reddit.com Competitive Analysis, Marketing Mix and Traffic. https://www.alexa.com/siteinfo/reddit.com
2. Subreddit Stats. https://subredditstats.com/
3. Giel, A., NeCamp, J., Kader, H. CS 229: r/Classifier - Subreddit Text Classification. http://cs229.stanford.edu/proj2014/Andrew%20Giel,Jon%20NeCamp,HussainKader,rClassifier.pdf
4. API Documentation. https://www.reddit.com/dev/api/
5. PRAW: The Python Reddit API Wrapper. https://praw.readthedocs.io/en/latest/
6. Řehůřek, R. Genism: Topic Modelling for Humans. https://radimrehurek.com/gensim/
7. GoogleNews-vectors-negative300.bin.gz. https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

8. Li, S. Multi-Class Text Classification Model Comparison and Selection. https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568