Using Machine Learning to Predict a Popular Reddit Meme

David M. Arnold
Data Scientist
April 11th, 2021
Orion99da@gmail.com

Problem Statement

This is project is a study of using machine learning algorithms and neural networks to determine if using the tabular data and image data from a collection of memes from Reddit could accurately predict whether or not the meme is a popular meme based upon user upvotes.

Social media is probably the greatest form of communication in the past 20 years. Prior the advent of the social media age, people communicated and got their news through other means, call conversations, emails, sending letters, watching the news on television and reading online articles. Social media has grown considerably since the late 1990s, with AOL instant messenger and My Space gave birth to Facebook, Twitter, 4Chan and Reddit.

A major phenomenon that is commonly used in social media are memes, which is usually an image, gif or short video that are often humorous, but can be political and evoking an opinion or sentiment. Certain groups of people refer to the 2016 election between Trump and Hillary Clinton as the "Meme War Election" and many of them jokingly refer to themselves as "Meme War Veterans".

While this project focuses on the memes and the upvotes of just small subset of Reddit users, this project could open the potential to study how various groups of individuals feel and express themselves through memes.

Data Background

The dataset used in this project comes from the website Kaggle and the dataset being used is located here: https://www.kaggle.com/sayangoswami/reddit-memes-dataset?select=db.json. The data files is a zip file of 3,200 meme images in jpg, jpeg and png form. The second file included is a json database file that contains all the data relating to the actual post on Reddit.

The information provided in the json data is: title/description, date posted, author name, post unique ID number, upvotes, downvotes, image link, thumbnail image link, thumbnail image width and height.

Data EDA and Analysis

The json data was converted from the semi-structed json form into tabular data, the columns being the categories listed above. The data was then checked for null/missing values and duplicate values.
The data itself needed to be converted into numerical form in order to be utilized by the machine learning algorithms. The title column was vectorized into 600 keywords by removing common words such as "the, had, was, etc" and punctuation.

The date posted column was made into columns of year posted, month posted, day of the week and if the post was made on a weekend.

The links to the images were not needed and discarded. The unique ID number was just a number of the meme image name, this was used for sorting the memes into "Good" meme and

"Bad" meme folders and then discarded. The pixel height and width were made into a thumbnail image area column and then the height and width columns discarded.

The downvotes column contained all zero values, so it was discarded. The upvotes column was split along the median value of 24,000 votes. Anything greater than 24,000 was considered a "Good" meme, anything below 24,000 was considered a "Bad" meme. This was done because I wanted to approach this project from a categorical perspective, this will also be the target variable of the project.

Now that all the features are in proper numerical form and the target variable, "Good" meme is identified, the data can be trained on the machine learning models.

Modelling Overview and Results

To ensure that the ML algorithms were properly trained and tuned to the data, I separated the dataset into three sections: Training data, Validation data and Testing data. The ML algorithms should never be used on the test data until you are confident that the data has been scaled and focusing on the correct number of features using dimensionality reduction, or PCA (principle component analysis). Once the model has been trained and scored using these techniques, then the test data can be introduced to the model for accuracy scoring.

The dataset was trained and scored on three different ML algorithms to determine which model performed best. The models used are: Logistic Regression, Decision Trees and K-Nearest Neighbors (KNN). Data scaling and dimensionality reduction were used on the Logistic Regression and KNN models, but these are not necessary when using the Decision Trees model.

Model Accuracy Scores:

1. Logistic Regression: 66.4%
2. Decision Trees: 65.2%
3. K-Nearest Neighbors: 59.%

Based on the results of the data ML models, the Logistic Regression model performed the highest, but the Decision Trees model performed similarly well. The results of the KNN model were the lowest. Since the models all scored in the 59-66% accuracy range, the small size of the dataset, only 3,200 datapoints was a contributing factor and perhaps the models would be benefitted from additional data.

The image data, or the meme images themselves were trained and validation using an image classification Convolutional Neural Network or (CNN). The images were segregated into classes of "Good" memes and "Bad" memes, as mentioned based on their upvote score of greater than or less than 24,000 upvotes.

The CNN model utilized a pre-trained model named 'VGG16', which is a image classification model that has been trained on millions of images. If your goal was to train a model for instance

on pictures of food or animals, this model is pre-trained and would help your model determine different types of food or animals with a high degree of accuracy.

There are additional aspects of the model that can also be adjusted: the train/validation split of the image data, the learning rate of the model, the model optimizer, the batch size of the images being introduced to the model, the pixel size or image resolution and the number of epochs or iterations the model runs through during the training/learning process.

I tried the model on the images settling on a train/validation split of 90% / 10% split, using the 'Adam' optimizer, also tried using the 'SVD' optimizer, a batch size of 20 images and the model was trained on the images at 64 pixels and 128 pixels. Additional the model augments each image, or distorts the images and looks at them from varying angles to learn more from each image.

The model was able to achieve validation accuracy scores ~70% for both the 64 pixel and 128 pixel models. The models were then given 5 examples of "Good" memes and "Bad" memes. The 64 pixel model actually performed slightly better when predicting both sets of data. The higher resolution model seemed to develop a "Good" meme bias towards an meme that had a person in the images, regardless if it was a "Good" or "Bad" meme. I also introduced several newer memes from the internet just to see what the model would predict. The meme of a lady yelling at a confused cat was a "Good" meme while thumbs up Harold was a "Bad" meme.

Further considerations I would consider for the model process would be to see if there were more sources of meme data from Reddit. Additionally for this dataset, I would see if separating the 2014-2015 data from the 2016-2018 data would improve the accuracy of the models. The early year data had fewer data points but much higher upvote counts. One other consideration would be to adjust the size of the Train-Validation-Test set sizes.

Summary

Social Media is a dominant platform for communication in the digital world. If you're a celebrity, a politician, a corporation or even a government agency, most likely you have a social media account on Facebook and Twitter. Social Media is about expressing yourself, your brand and your opinions.

A major part of communication on Social Media is through the creation and sharing of memes, whether they are intended to be funny, express an opinion on current events and influence a social movement. People on Twitter follow accounts like Wendy's and KFC (Spain) because they often share popular memes, this attracts a large crowd of younger followers.

Using machine learning, memes can be studied to determine what appeals to different audiences in order to improve communication and reach larger audiences while giving a voice to not only the largest social influencers, but rather the smallest individual.

Appendices

Project Data Source: https://www.kaggle.com/sayangoswami/reddit-memes-dataset?select=db.json