Lana Palmer | November 15, 2020 | lana@lanapalmer.com

Bot or Not? Bot Classification on Reddit

Introduction

While the word ‘bot’ may evoke images of artificially intelligent robots run amok, most bots found on online message boards are simple software programs written to execute commands and automate mundane tasks. Bots on Reddit range from the benign, such as u/JImmyButler, a bot that replies to users with compliments; the administrative, like the many moderator bots used to help subreddit ‘mods’; and the simply annoying, such as bots that spam URLS.

While these 'friendly' bots may seem somewhat innocuous, the growing number of bots on online platforms provides cause for serious concern. Social media platforms Facebook and Twitter are currently setting new precedents for free speech vs misinformation, and the proliferation of bots and 'inauthentic' accounts only exacerbate these concerns. Left unchecked, bots can pose a threat to public health during a pandemic, the democratic process during an election, and the stock market in the midst of a global recession.

In this project, I will develop a classifier that can predict whether a comment on the Reddit platform was made by a bot or a human user.

Data Collection

I used the Pushshift API to collect Reddit comments from both bots and non-bots. In order to ensure the authenticity of my non-bot data set, I verified that the accounts were human users by reading several of their posts and comments. My non-bot data set includes a mix of users, from celebrities u/janellemonae, popular Redditors u/dickfromaccounting, and a selection of regular users.

To gather bots, I used Beautiful Soup to scrape a list of known bots on Reddit.

For both bots and non-bots, I collected the following features:

- Author: The Redditor username, i.e. PresidentObama
- Comment: The raw comment, which may contain emojis and links.
- Subreddit: The subdirectory the comment was posted on.

- Score: The total upvotes/downvotes, which may be a negative number
- Time: The epoch time of the comment.
- Flair: A special designation that can be awarded by moderators in each subreddit.

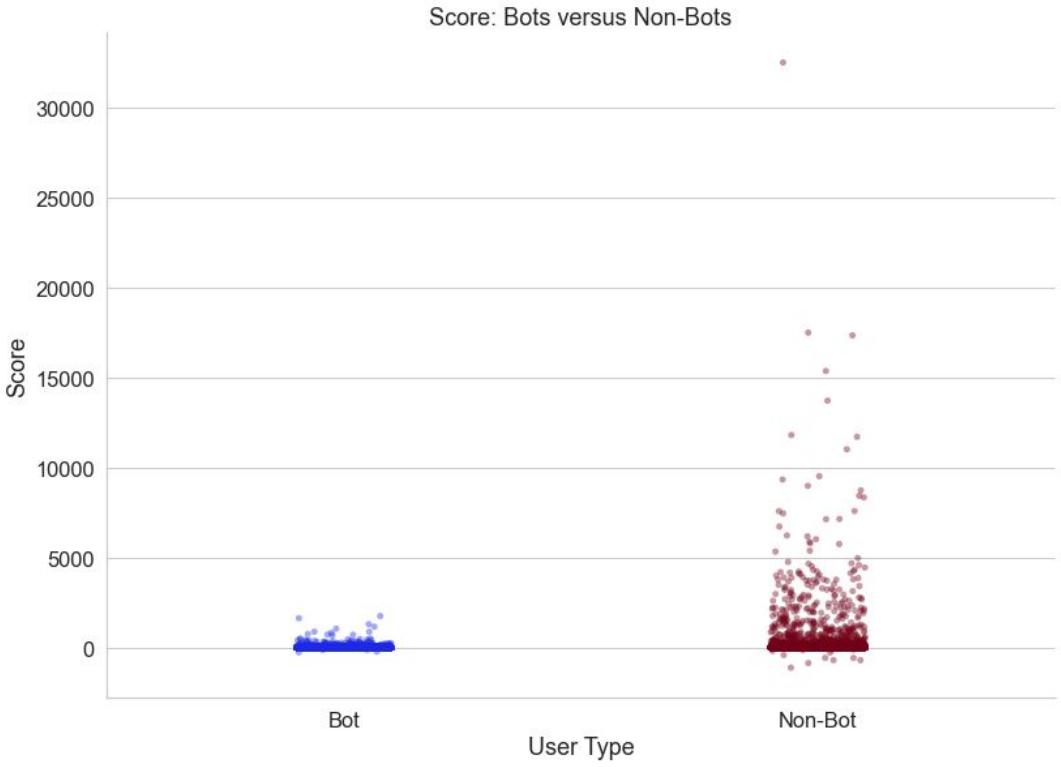
The data did not require extensive cleaning, as null values were excluded from the API call. I converted Epoch time to datetime, and added 'Class' feature to specify the comments as from Bots or Non-Bots, and merged the two data frames.

Exploratory Data Analysis

My initial EDA was focused on quantitative analysis of existing features:

1. Score

The median for both Bots and Non-Bots was 1, which is unsurprising given that all Reddit comments have a default score of 1. However, the mean and standard deviation differed substantially: Bots had a mean score of 3.59 and a standard deviation of 27.54, while non-bots had a mean of 91.16 and a standard deviation of 687. Clearly, users were engaging more with non-bots than bots, with both upvotes (positive) and downvotes (negative).



2. Unique Subreddits

The percentage of unique subreddits posted in by bots and non-bots is similar, at 9.5% for Bots and 10.6% for Non-Bots

3. Time Span

Feature Engineering

After the exploratory data analysis stage, I decided to use Natural Language Processing techniques to create features from the raw comment text.

* Amount of flair: The count of 'flair' for the each comment
* Emoji Count: The number of emojis contained in the comment.
* Clean text: All lowercase, and remove special characters.
* Comment Length: The total number of words.
* Average Word Length.

Using the textstat library, I created these more sophisticated features:

* Lexicon Count
* Sentence Count
* Readability Score
* Syllable Count

After creating dummies for categorical features and splitting into training and testing groups, I created bigrams (the 1000 most common two-word phrases) using the train group, and used these to transform the test group.

At the completion of feature engineering, my dataset included a total of 1010 features.

Algorithms and Machine Learning

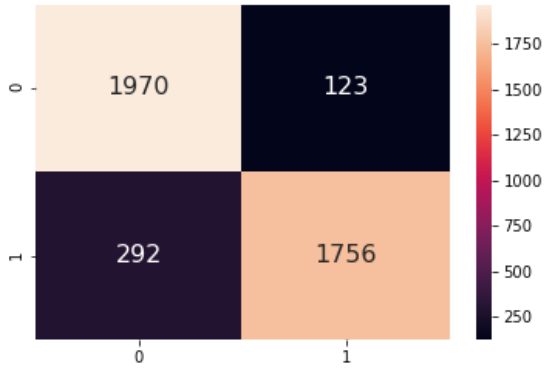
I selected Logistic Regression and Gradient Boosting algorithms for my initial model.

The Logistic Regression model had an accuracy rate of 85% on the test set. After optimizing features using Grid Search, the Gradient Boosting model had an accuracy rate of 90% on the test set. I opted to use the Gradient Boosting model for the remainder of my analysis.

Model Evaluation

Accuracy rate on the test set of 4141 comments was 90%, with the following breakdown:

True Positives (Accurately classified bot comments as bots): 1970
True Negatives (Accurately classified non-bot comments as non-bots): 1756
False Positives (Classified non-bot comments as bots): 123
False Negatives (Classified bot comments as non-bots): 292



For bot comments, precision rate was .87 and recall was .93
For non-bot comments, precision was .93 and recall was .86

	precision	recall	f1-score	support
0	0.87	0.94	0.90	2093
1	0.93	0.86	0.89	2048
accuracy			0.90	4141
macro avg	0.90	0.90	0.90	4141
weighted avg	0.90	0.90	0.90	4141

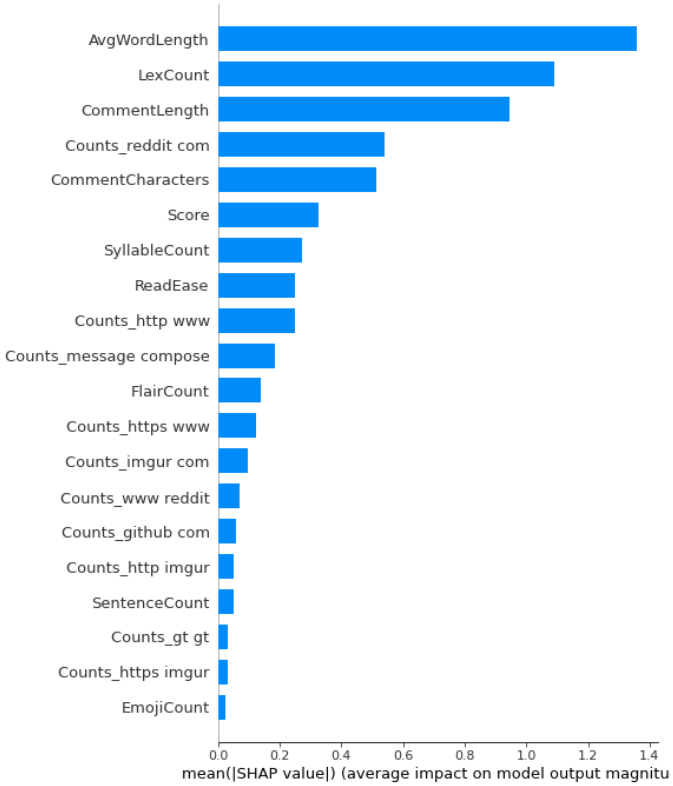
This suggests that the model is more likely to err by classifying bot comments as non-bots than non-bots as bots. In this use case this is likely the best outcome, as unfairly flagging human users as bots would be undesirable to a social network platform.

Feature Importance

I used the SHAP (SHapley Additive exPlanations) library to evaluate feature importance in the gradient boosting model.

As we can see in the bar chart below, the most importance features were:

- Average Word Length
- Comment Length
- Lexicon Count
- Counts_Reddit.com (*The number of times that reddit.com appeared in the comment*)
- Score
- Syllable Count
- Read Ease



Let's dig into individual predictions.

Below is the explanation plot for a comment correctly classified as a non-bot:

Each of the blue features (Score, LexCount, CommentCharacters, etc.) are serving to push this comment away from a positive ('Bot') classification. Below is the original comment:

ID	Author	Comment	Subreddit	Score	Time	Flair	Class	FlairCount	EmojiCount	...
dgiw905	Here_Comes_The_King	what can celebs do to use their voice to support legalization??	trees	421	2017-04-20 10:48:32	None	Non-Bot	0	0	...

By contrast, let's look at a comment that was correctly classified as a bot:

Each of the red features (Score, Comment Length, Flair Count, etc) are serving to push this comment towards a positive ('Bot') classification.

ID	Author	Comment	Subreddit	Score	Time	Flair	Class	FlairCount	EmojiCount	...
g8sxyo6	NoSobStoryBot2	Title|Points|Subreddit|Submitted\n-|-:|-|-:|-\nI'm 36, I just did the thing for the first time ever...|6146|/r/pics|8 hours ago	no_sob_story	1	2020-10-14 08:25:04	RoboCop 2	Bot	1	0	...

The model predicted a .83 probability that this comment was made by a bot.

Conclusion and Caveats

The gradient boosting classifier was able to perform with 90% accuracy on the testing data set, providing evidence that bot comments are distinguishable from non-bots.

A few caveats:

My non-bot user group size is rather small. While I attempted to use a random selection process, I may have introduced bias based on my own opinion of what 'non-bot' user comments look like. Because several of my non-bot users were well-known public figures, I believe the scores for non-bots are skewed higher than they would be for the entire population of Reddit users.

Next Steps

I would love to build a second model which works on a user-basis, rather than on a comment basis. For this model, I would take a rolling average of all features for a user's most recent posts. I believe this model will have a higher accuracy rate, as the originality of Bot posts will be, on average, much lower than for human users.

Credits

Thank you to my Springboard mentor Blake Arensdorf for helping me to hone my skills, and guiding this process with patience and precision. Thank you to Scott Lundberg for the Shap library, Shivam Bansal and Chaitanya Aggarwal for their Textstat library, and to Reddit and Pushshift for providing public access to Reddit data.