Classifying Star Wars Posts on Reddit

Patrick Zabriskie

February 19th, 2016

Introduction

Posts of similar themes can be found in many different subreddits on Reddit's website. One can, for example, find Star Wars-themed posts in a variety of subreddits, some of which are Star Wars-specific, some of which are not. With this in mind, the question arises of whether or not it is possible to determine the specific subreddit of a post, given the content of the post itself. Such classification would have multiple useful applications for Reddit and other companies.

First, it would allow Reddit to make subreddit recommendations to its users. For example, if someone's Star Wars posts in a particular subreddit are similar in content to other users' posts in another subreddit, then Reddit could advise that user to switch subreddits for a better experience.

Second, it would assist Reddit in determining ways to advertise to current users. For example, if all of someone's Star Wars-themed posts occurred in a specific subreddit, then Reddit might focus Star Wars advertising (for Star Wars-themed games, collectibles, movie tickets, etc.) within that subreddit.

Third, this method of classification assists Reddit and other companies in performing general market research. Beyond the realm of Star Wars-related content, being able to classify a Reddit post can lead into key insights into certain demographics. If, in the world of politics, Reddit posts could be classified to Republican or Democrat subreddits based on issues discussed within those posts, then this classification could be a way to determine which issues are most discussed (and in what way) amongst different political parties. A similar analysis could be made between different religions, or people from different countries, in determining what is most pertinent for their discussions on Reddit. And of course, this can be extended to other social media sites outside of Reddit, or other websites where user posts are common.

This project will focus on the classification of Star Wars posts into two different subreddits: movies and Star Wars. It will do so by attempting to separate Reddit posts based on certain words in certain categories that could appear within Star Wars-themed posts. There are two main steps in this project. First, a Support Vector Machine (SVM) approach will be used to create a linear classifier for the data and determine the order of importance for categories of words in classification. Following this, rule mining will then be implemented, using the most important category of words, for further analysis.

Searching and Cleaning

The data for this project comes via the website www.kaggle.com, which currently hosts Reddit data from May of 2015. By writing a script on this website, 55000 Reddit posts from the Star Wars and movies subreddits were extracted and stored as a Microsoft Excel spreadsheet. A larger number of posts would have been extracted, but Kaggle encounters memory limitations at around 55000 Reddit posts.

Once the data was stored and downloaded, a word search was performed on the Reddit posts. A dictionary of words, grouped into categories, was defined; and a tally was kept of how many times a word from a particular category appeared in a Reddit post; these tallies were stored as a spreadsheet, with rows corresponding to the individual Reddit posts and columns corresponding to word categories[1]. A spreadsheet holding the subreddit type for each of these Reddit bodies was also created. The dictionary, with categories, used for this search can be seen below.

Table 1. Star Wars Dictionary.

Cast and Crew	Creatures	Factions	Main Characters	Planets and Locations
mark hamill	jawa	galactic senate	luke skywalker	tatooine
harrison ford	womp rat	the republic	obi-wan kenobi	dagobah
carrie fisher	sarlacc	rebel	han solo	hoth
george lucas	ewok	stormtrooper	chewbacca	coruscant
alec guinness	gungan	clone trooper	darth vader	naboo
john williams	hutt	empire	princess leia	alderaan
ewan mcgregor	rancor	separatist	anakin	cloud city
hayden christensen	wookie			endor
natalie portman	droid			mos eisley
				yavin

Plot Points	Secondary Characters	Ships and Weapons	Titles
the force	r2-d2	light saber	star wars
jedi	c-3po	blaster	new hope
the dark side	palpatine	x-wing	the empire strikes back
the good side	darth maul	tie fighter	return of the jedi
sith	sidious	at-at	phantom menace
trench run	dooku	at-st	attack of the clones
	yoda	death star	revenge of the sith
	boba fett	millennium falcon	special edition
	jabba		the force awakens
	lando		despecialized edition
	jar jar binks		
	mace windu		

[1] An example of this spreadsheet can be found in Appendix I.

After creating this tally, only 3355 out of the original 55000 Reddit posts contained any of the words in the dictionary. As a simplification for this project[2], all but these 3355 were eliminated from consideration for future steps.

Support Vector Machine Implementation

Once properly arranged and formatted, Support Vector Machine (SVM) classification was implemented to create a classifier for the data. SVM views data as points in some *n*-dimensional space. After being given sample data to train and hone its classification approach on, SVM attempts to classify future data by dividing its *n*-dimensional space into regions. Data that falls within a particular region is designated to be one class.

The benefit of SVM is that it is a relatively efficient and robust way to separate comparatively large sets of data. It often takes mere seconds to run on a computer, and given sufficient training samples, it can achieve relatively high classification accuracy. An additional benefit is there are multiple types of SVM to choose from, each using different techniques to classify. These include RBF SVM, polynomial SVM (of many different degrees), and Linear SVM[3].

For this project, though, Linear SVM has particularly strong appeal: it assigns special values, or weights, to each variable it works with. The magnitude of these weights has a distinct relationship to the importance of a variable. The larger the magnitude of a weight, the more significant its associated variable is in classification. By running Linear SVM with the aforementioned category tally and subreddit-type spreadsheets, it is possible to glean insights into which category of words was most important in classification; and, therefore, which words might serve as a clue to whether or not a Reddit post belongs in the Star Wars or movies subreddit.

A script was written to run Linear SVM on the column tally and subreddit-type data. The script ran Linear SVM ten thousand times; each time SVM was trained on a randomly chosen third of the entries in these spreadsheets and tested on the remaining two-thirds to determine the accuracy of the classifier. After all iterations, the average and standard deviation of the accuracy was calculated. It can be see in the table below.

1. Accuracy Mean	2. Accuracy Standard Deviation	3. Proportion of Star Wars Subreddits	4. Proportion of Movies Subreddits
0.665	0.016	0.604	0.396

Table 2. Linear SVM Accuracy Statistics (approximate)

Table 2 shows the accuracy of the classifier as about 66.5%. The actual proportion of Star Wars subreddits is about 60.4%, and the proportion of movies subreddits is about 39.6%. In other words, Linear SVM is performing better than what would be expected from random

[2] A project of larger scope would, of course, include these other Reddit posts in some capacity.
[3] Graphs showing different types of SVM separating two-dimensional data, as well as tables comparing their accuracy on the Reddit data can be found in Appendix II.

guessing on this data set, which is promising. Training on larger data sets would likely increase accuracy.

A bar graph showing the average magnitudes of the weights calculated from Linear SVM, along with the standard deviation of the weights and the standard deviation of the original categories[4], is depicted in Figure 1 below.

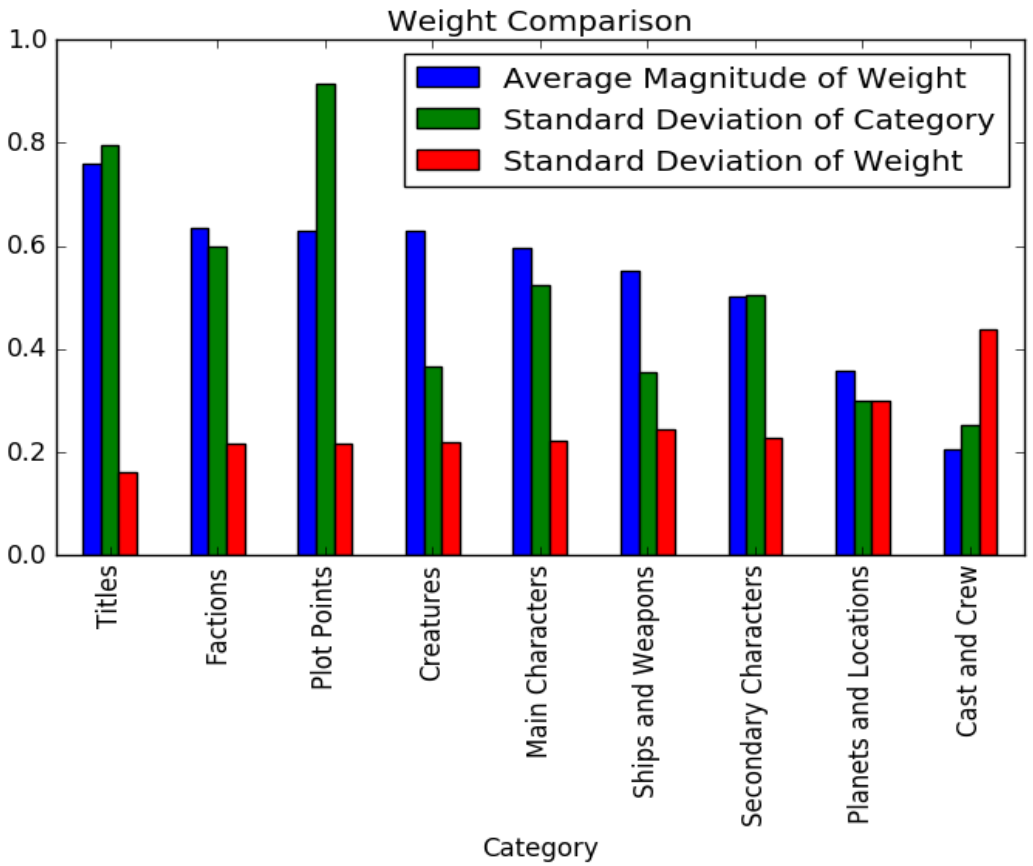

Figure 1. Bar graph showing average weights for each category, the standard deviation of the weights, and the standard deviation of the original category, ordered in decreasing order of average magnitude of weight.

From Figure 1, it can be seen that the "Titles" category had the weight of highest average magnitude[5], suggesting that it is the most important category in classifying Reddit posts as either in the Star Wars or movies subreddit, while "Cast and Crew", with the weight of lowest average magnitude, is least important. A few of the middle categories have very similar weight averages

[4] That is, the standard deviation of the column corresponding to the category in the aforementioned tally Excel spreadsheet.
[5] A table showing precise values for this graph can be found in Appendix III.

and standard deviations of weights. To determine if all of these averages are statistically different from each other, a two-sample t-test was performed on every distinct pair of categories.

A two-sample t-test compares two population[6] averages, using sample averages, the number of samples used to obtain them, and their variances to determine whether or not the population averages are statistically distinct. The result of a t-test is a p-value. The p-value, if smaller than a certain threshold (common thresholds are 0.05 and 0.01), indicates that two averages are statistically different from each other. The resultant p-values between distinct pairs of categorical weight means are displayed in the table below. It should be noted that a zero-value should be interpreted as a number small enough to be rounded to zero by the computer.

Table 3. The p-values from the two-sampled t-test.

Category	Titles	Factions	Plot Points	Creatures	Main Char.	Ships and Weapons	Sec. Char.	Planets
Titles								
Factions	0							
Plot Points	0	0.14						
Creatures	0	0.09	0.82					
Main Characters	0	1.51E-33	2.09E-26	3.72E-25				
Ships and Weapons	0	1.14E-132	1.23E-118	1.21E-115	1.88E-38			
Secondary Characters	0	0	0	0	9.31E-166	2.39E-41		
Planets and Locations	0	0	0	0	0	0	0	
Cast and Crew	0	0	0	0	0	0	0	3.29E-173

Using a threshold of 0.05, it can be seen in Table 4 that the population averages between most pairs of categorical weights are statistically different from each other. However, “Plot Points”, “Factions”, and “Creatures” are not statistically different from each other. This puts the ranking of weights and importance for these categories in question; making it difficult to determine the true order of importance in terms of categories; and it might suggest that new categories should be made.

In as much as this table shows that the weight means of “Titles” and “Cast and Crew” are statistically distinct from all other category weight means, the results of the t-tests do add evidence for the idea that “Titles” is a more important category than the other categories in this classification, and that “Cast and Crew” is the least important category. In interpreting the importance of the “Titles” category, it is possible that one subreddit mentions the titles of Star Wars films more often than the other, or, at the very least, that one subreddit mentions the titles of specific Star Wars films more than the other, which could give insight into what particular

[6] Population here refers to populations of weights.

sub-topics these subreddits are most interested in discussing. Conversely, a low "Cast and Crew" weight could signify that both subreddits might discuss these values to an equal extent.

Figure 1 also displays the standard deviation of the categorical columns in the tally spreadsheet. Essentially, this gives an indication of how varied the values in different entries of these columns can be. Often, it is the case that categories with lower standard deviations are less important, since, comparatively, their expected change is smaller between instances. Looking at Figure 1, there is a very loose trend of category standard deviations ultimately becoming smaller as the average weights decrease, thus somewhat supporting the idea that the larger expected variations of the columns with higher weights have more significance in classification. However this is not an exact principle; the highest category standard deviation does not belong to "Titles", the category with the highest weight (although it does have the second highest category standard deviation), and in some instances, category standard deviations do increase even as the average magnitudes of weights decrease.

It is important to recognize the current limitations with this SVM approach. The data set is comparatively small; the dictionary being used is far from complete, the organization of categories might need to be changed, and it does not currently account for variants on words or phrases (e.g. users write "Chewie" instead of "Chewbacca") as it should. Given a larger sample of Reddit data, and updates on the dictionary being used, Linear SVM could be honed and refined to become more accurate at classifying posts. With time and data storage restrictions, however, this cannot be done at this stage.

Rule Mining

While Linear SVM can give insight into the importance of words for classifications, it is difficult to infer from it a specific relationship between those words and the classifications. With this in mind, a tangential technique will be performed: rule mining. Rule mining is an implementation of conditional probability. In this instance it will be used to determine an exact statistical relationship between the presence of a word (or words) in a Reddit post and the category of that Reddit post; this relationship can be used to construct an associative rule. Suppose, for example, a possible association that whenever a post contains word *A*, then that post is in category one. This is a rule that can be restated as "*A* implies category one", and the strength of this rule is evaluated in the following way:

$$\text{Strength of rule} = \frac{\#Category\ one\ posts\ that\ contain\ word\ A}{\#\ posts\ that\ contain\ word\ A}$$

This rule is considered meaningful if the strength of the rule is greater than some desired threshold. The benefit of implementing this technique after performing SVM is that now SVM has given an indication of which words are important for classification, so working with these words in rule mining can be more meaningful than merely choosing words randomly; it has also given a possible threshold candidate in the form of its average accuracy.

For simplicity sake, the important words will be limited to the "Titles" category, though future work on this project should include words from other important categories. Running rule mining using words from the "Titles" category yields the following table:

Table 4. Single-word rule mining results**.**

	Star Wars	New Hope	The Empire Strikes Back	Return of the Jedi	Phantom Menace
Number of Movies Posts	655	26	16	18	49
Number of Star Wars Posts	530	44	8	30	37
Word Implies Movies	0.55	0.37	0.67	0.375	0.57
Word Implies Star Wars	0.45	0.63	0.33	0.625	0.43

	Attack of the Clones	Revenge of the Sith	Special Edition	The Force Awakens	Despecialized Edition
Number of Movies Posts	16	29	27	25	4
Number of Star Wars Posts	22	26	15	27	4
Word Implies Movies	0.42	0.53	0.64	0.48	0.5
Word Implies Star Wars	0.58	0.47	0.36	0.52	0.5

If the threshold is set at 60% (the approximate percentage of posts in the data set that are in the Star Wars subreddit), then the rules whose strength surpass this threshold are "Return of the Jedi implies Star Wars", "Special Edition implies movies", "New Hope implies Star Wars", and "The Empire Strikes Back implies movies." In practice, if a computer program attempting to classify Reddit posts was designed to implement these rules, then every time, for example, that it detects the phrase "New Hope" in a Reddit post, it would classify that Reddit post as being in the Star Wars subreddit, and it would be expected to be right approximately 63% of the time. If the threshold is reset to 66.5% (the average accuracy of Linear SVM), then the only rule whose strength meets this threshold is "The Empire Strikes Back implies movies", and if this rule were implemented in a computer program, then that program would be successful in classifying posts containing the phrase "The Empire Strikes Back" approximately 67% of the time.

These rules, to some extent, support the supposition made earlier that certain Star Wars titles are discussed more frequently in one subreddit than the other. The Empire Strikes Back, for example, is seen by many as the best Star Wars film, and therefore might be brought up more frequently in the movies subreddit—where discussing great films occurs—then in the Star Wars subreddit, where, its quality is well-known to the users and is therefore not explicitly discussed. However, with only 24 posts total that mention this specific title, such an inference is hardly certain.

Concomitantly, in addition to the strength of a rule, another key consideration when performing rule mining is the frequency with which a rule may be implemented. For example, "The Special Edition implies movies" has one of the higher strengths of any rule, but the instances where it may be applied—a total of 42 posts out of 3355 posts total—are comparatively few. On the other hand, "Star Wars implies movies" has a strength below the

thresholds of 60% and 67%, but there are more occasions to use it—a total of 1185 posts total. So there may be a trade off between the accuracy of a rule, and the amount of times a rule can be applied.

This level of rule mining only uses one word at a time. Further implementation should involve creating rules involving multiple words, starting with two and then progressing to three, four, etc. Creating rules with multiple words might lead to more reliable rules with higher strengths. Similar to how it would enhance the SVM approach, updating words and adding new words to the dictionary being used for rule mining would be another useful step in this process, as would eliminating words that do not appear frequently enough in Reddit posts, and including more categories than the highest weighted one exclusively.

Conclusion

In summary, this data exercise proved an interesting starting point for this investigation into Reddit post classification. The linear classifier achieved some level of success with the available data, but given a relatively small amount of information to work with, improvements in the process should be made before more definite conclusions are drawn. Rule mining yielded some basic yet intriguing results, but it also needs to be expanded and refined. Time and storage permitting, the following recommendations are made for future work in this area.

- Linear SVM should be trained and run on larger data sets.
- Other forms of SVM should be investigated more thoroughly and their accuracy compared to the accuracy of Linear SVM.
- The dictionary of words should be regularly refined, adding words that are relevant, discarding words that are not, and reorganizing as necessary.
- Rules involving more than one word should be evaluated.
- Other categories should be considered in rule mining.
- Other statistical tests, such as Chi-square tests, should be implemented to give a better understanding of significance/insignificance and independence/dependence between parts of data sets.
- This classification investigation should be expanded to other subject matter and other subreddits.

Appendix I

The table below is an example of the category tally spreadsheet used for this project. Due to space reasons, not every category is seen. The Reddit posts used for this project were given a number, and then a count, partitioned by category, was made of how many words from the dictionary were present in each Reddit post.

Table 5. Sample category tally spreadsheet.

Reddit Post	Cast and Crew	Creatures	Factions	Main Characters	Planets and Locations	Plot Points
0	1	0	1	0	0	1
1	0	0	0	0	0	0
2	0	0	0	0	0	1
3	0	0	0	0	0	0
4	0	2	1	0	0	1
5	0	0	0	0	0	2
6	0	0	0	0	0	1

Appendix II

Many types of SVM classification exist in addition to Linear SVM. These include RBF SVM and Polynomial (of varying degrees) SVM. Shown below is an example of different types of SVM—Linear, RBF, Polynomial (degree 2), and Polynomial (degree 3)—and how they classify two-dimensional data differently. Here, the two dimensional data is made up of points whose coordinates are determined by values from the "Title" category and the "Cast and Crew" category from the category tally spreadsheet. There are red dots and blue dots on these graphs. Points in blue regions are classified as blue points; while points in tan regions are classified as red points.

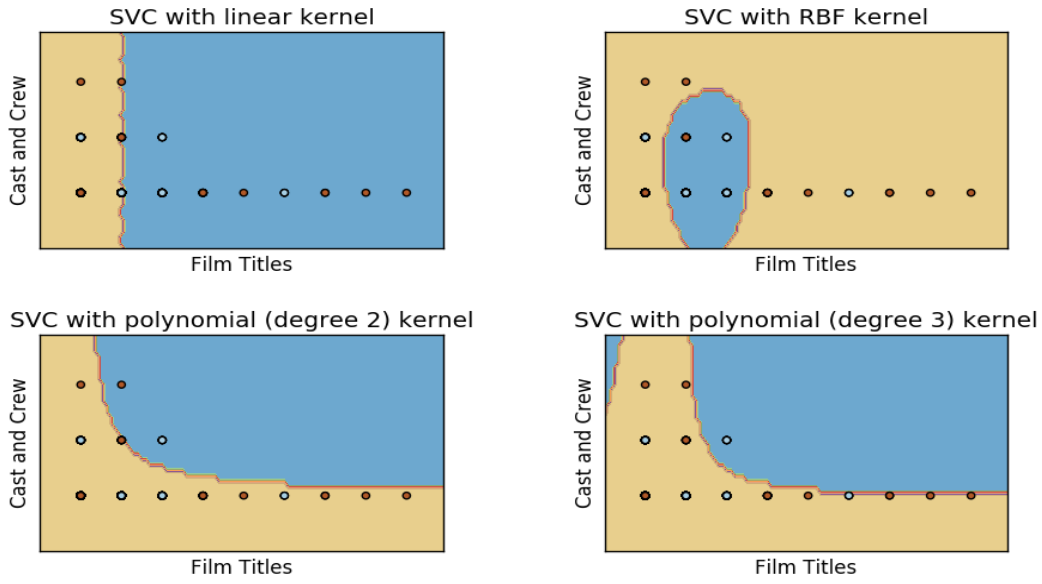

Figure 2. Comparison of different types of SVM.

Additionally, these other types of SVM were tested on the available Reddit data in full, each ten-thousand times. The average accuracies and their standard deviations are given in the table below.

Table 6. Comparison of accuracy of different SVM methods.

1. SVM Method	2. Average Accuracy	3. Accuracy Standard Deviation
Linear SVM	0.665133989	0.016367437
RBF SVM	0.659831127	0.006485004
Polynomial (Degree 2) SVM	0.603747138	0.006691953
Polynomial (Degree 3) SVM	0.602573256	0.00605285

Based on Table 6, Linear SVM appears to be the best classifier, with a Polynomial of degree 3 being the worst. A two-sample t-test was implemented to see whether or not the population averages were significantly different. The p-values from that test can be seen in the table below. Again, values of zero should interpreted to be values that are small enough to be rounded to zero by the computer.

Table 7. The p-values from the two-sampled t-test.

	Linear SVM	RBF SVM	Polynomial (degree 2) SVM
Linear SVM			
RBF SVM	9.3027E-193		
Polynomial (Degree 2) SVM	0	0	
Polynomial (Degree 3) SVM	0	0	1.55447E-38

Using the values from Table 7 and a threshold of 0.05, it can be concluded that all averages are significantly different from each other. And so it can be said that Linear SVM was the best of these four classifiers.

Appendix III

A table showing precise values used for the Figure 1, the Weight Comparison bar graph, is provided below.

Table 8. Precise values of Linear SVM statistics.

1. Category	2. Average Magnitude of Weight	3. Standard Deviation of Weight	4. Standard Deviation of Category
Titles	0.760416846	0.160486235	0.797126476
Factions	0.634142322	0.217403035	0.597835483
Plot Points	0.629789763	0.218179146	0.914991228
Creatures	0.628480821	0.21944223	0.366055518
Main Characters	0.597467659	0.22334002	0.524008554
Ships and Weapons	0.552342495	0.244084462	0.35553359
Secondary Characters	0.503271611	0.227185985	0.50454682
Planets and Locations	0.35935841	0.301189553	0.299891221
Cast and Crew	0.205826127	0.438104894	0.253323273