Patrikios George
January 2019

Sentiment Analysis on Directors

Sentiment Analysis on Directors using Twitter API

Introduction:

The aim of this personal project is to examine how the twitters' users feel about three of the best movie directors. In particular, 1500 tweets were extracted using "Alfred Hitchcock", "Martin Scorsese" and "Steven Spielberg" as keywords.

Methodology:

1. A developer's account was created on Twitter in order to obtain the necessary "keys" that provide access to Twitter's API
2. After obtaining the "keys", programming took place on python. As it shown on the script provided, a 'for loop' was created in order to extract the tweets and store them into a database. Fisrty, a connection was created to Sqlite. Secondly, a table "directors'_tweets" was created in order to be filled by the tweets. Thirdly, there is "thetwitterextractor". It is a "for loop" which uses the twitter's cursor from 'tweepy' so that we can extract tweets. Before storing every tweet in the database, sentiment analysis takes place using "TextBlob" library.

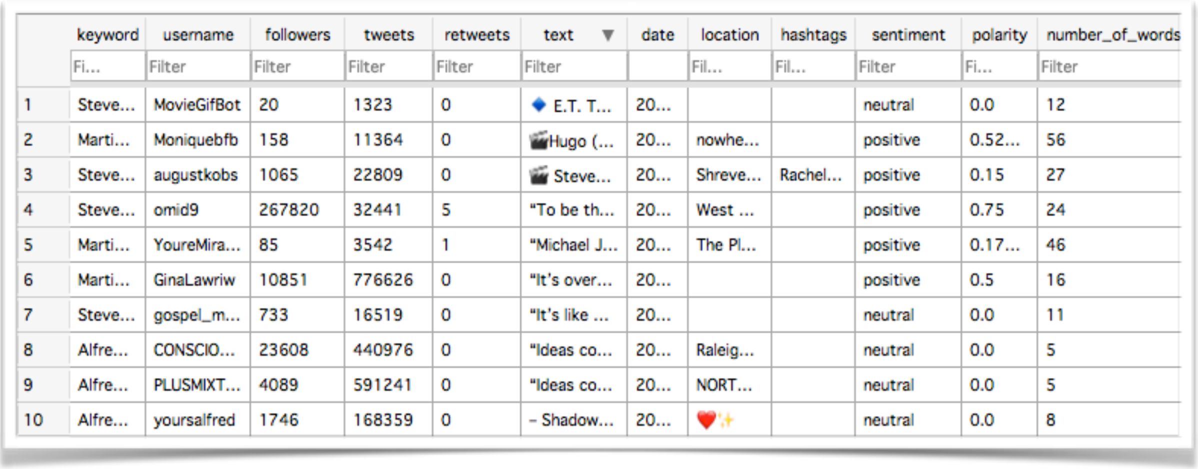
For every tweet we extract the following data:

A. User account (e.g. gpatrick)
B. Number of account's followers
C. Number of account's tweets
D. Number of tweet's retweets
E. Tweet text
F. Date
G. Location
H. Hashtags
I. Sentiment
J. Polarity
K. Number of words in the tweet

3. Then, the extraction of the directors' tweets followed. For every director, we extracted 500 tweets for 10 days (50 tweets daily between the 10th and the 19th of January 2019).
4. Finally, "pandas" and "matplotlib" libraries used to plot the extracted data.

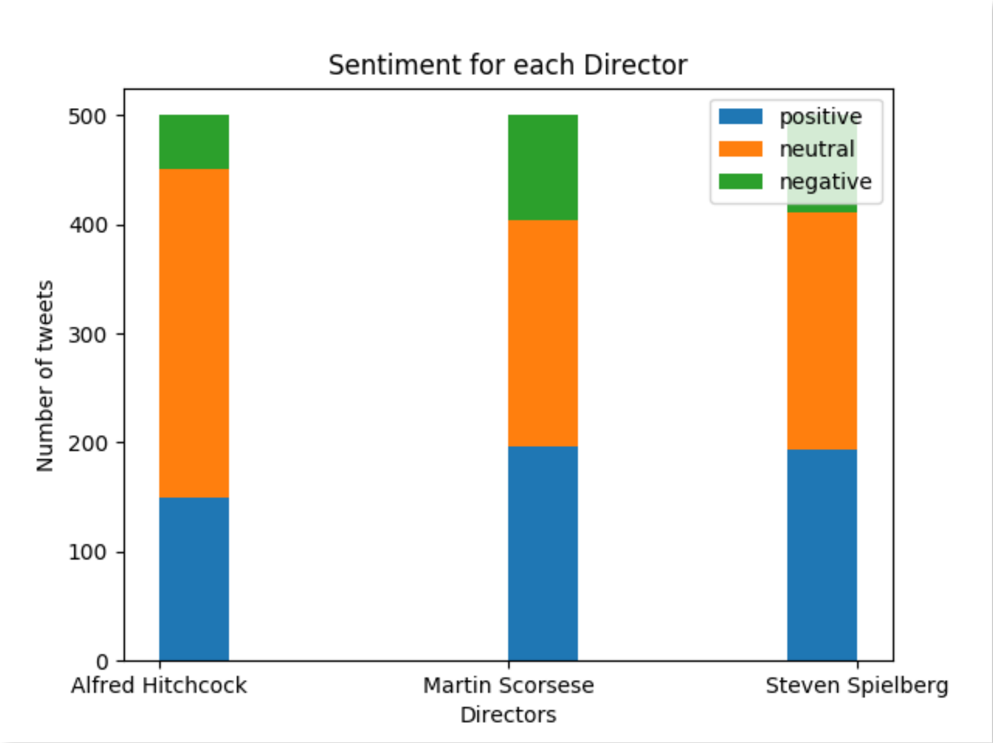
Results:

The following picture is a screenshot of the table created after extracting 1500 tweets for the three directors. The “keyword” attribute was created in order to query the results for each director.

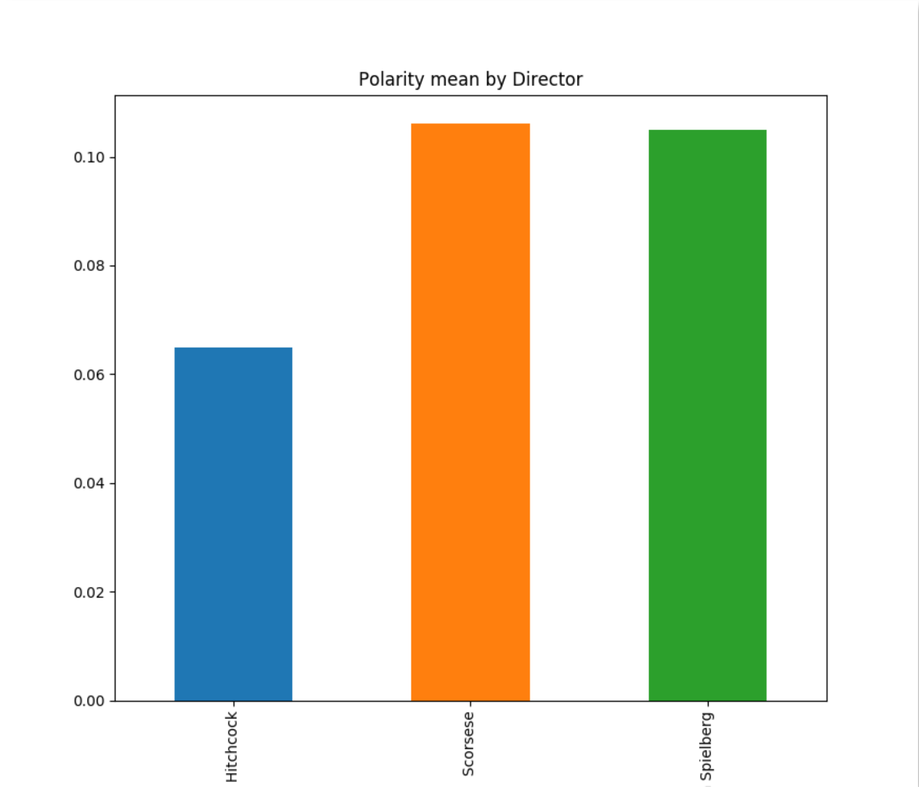


	keyword	username	followers	tweets	retweets	text ▼	date	location	hashtags	sentiment	polarity	number_of_words
	Fi...	Filter	Filter	Filter	Filter	Filter		Fil...	Fil...	Filter	Fi...	Filter
1	Steve...	MovieGifBot	20	1323	0	🔷 E.T. T...	20...			neutral	0.0	12
2	Marti...	Moniquebfb	158	11364	0	🎬Hugo (...	20...	nowhe...		positive	0.52...	56
3	Steve...	augustkobs	1065	22809	0	🎬 Steve...	20...	Shreve...	Rachel...	positive	0.15	27
4	Steve...	omid9	267820	32441	5	“To be th...	20...	West ...		positive	0.75	24
5	Marti...	YoureMira...	85	3542	1	“Michael J...	20...	The Pl...		positive	0.17...	46
6	Marti...	GinaLawriw	10851	776626	0	“It's over...	20...			positive	0.5	16
7	Steve...	gospel_m...	733	16519	0	“It's like ...	20...			neutral	0.0	11
8	Alfre...	CONSCIO...	23608	440976	0	“Ideas co...	20...	Raleig...		neutral	0.0	5
9	Alfre...	PLUSMIXT...	4089	591241	0	“Ideas co...	20...	NORT...		neutral	0.0	5
10	Alfre...	yoursalfred	1746	168359	0	– Shadow...	20...	❤️✨		neutral	0.0	8

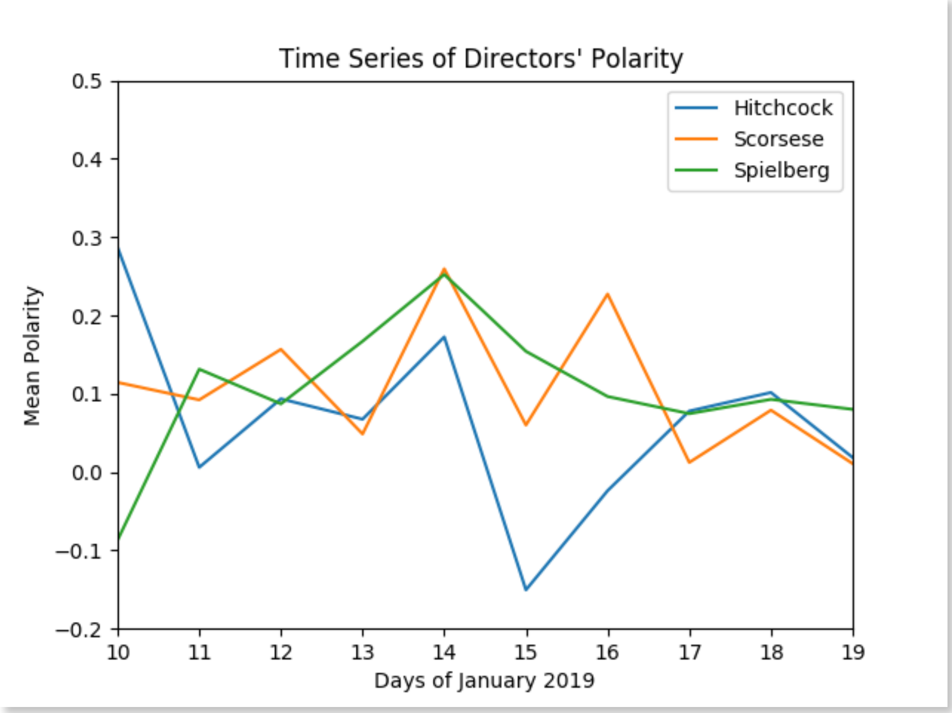
The following section presents the plots that created using “pandas” and “matplotlib”. More plots are available in the python script provided:



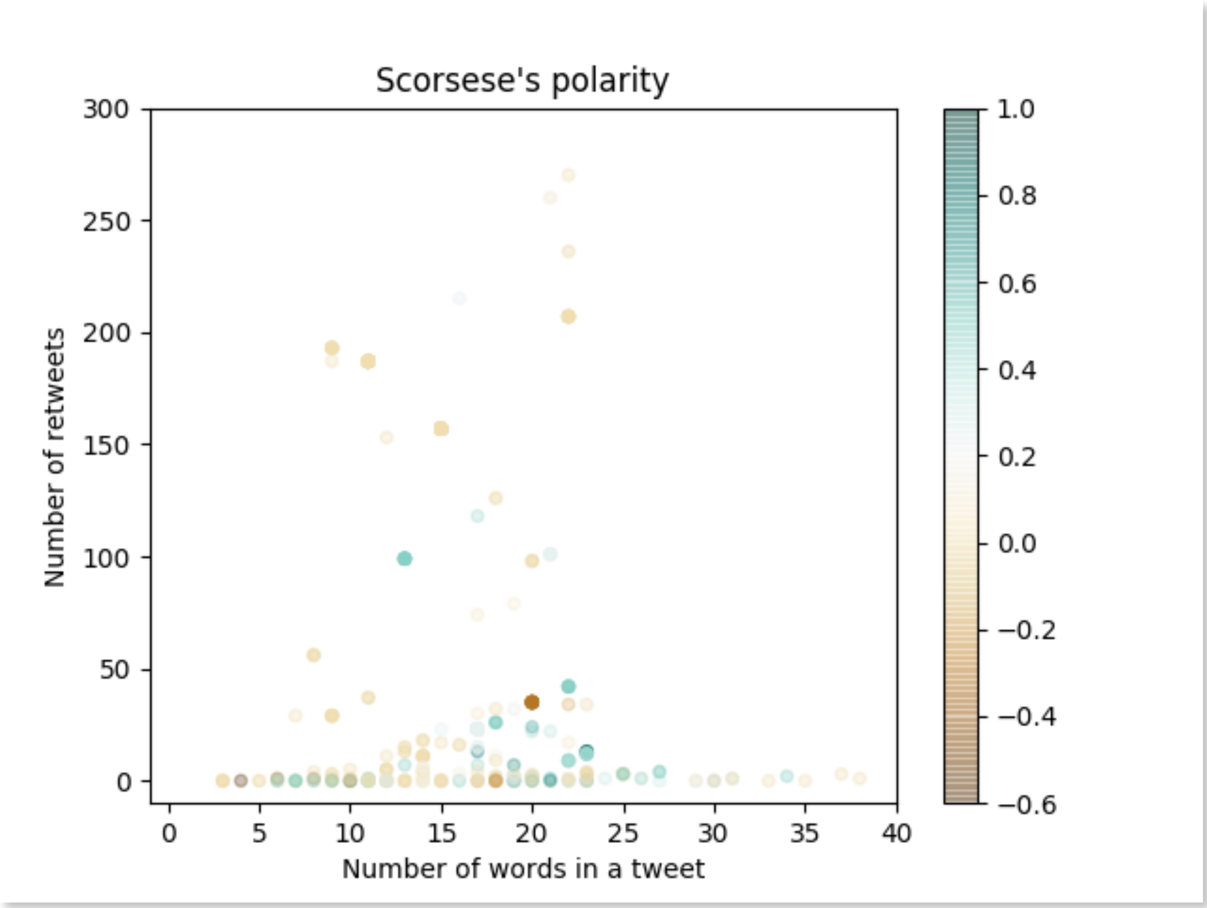
At first glance, the number of negative tweets for all three directors is relatively low comparing it to the other two sentiments. As it is shown, Hitchcock has the less positive tweets (149) in contrast to Scorsese who has the largest number (197) and Spielberg (193). Regarding the neural number of tweets, they scored 301, 206, 218 respectively.



This bar chart demonstrates the polarity mean for every director. It is known, "TextBlob" gives a score-polarity between -1 and 1 regarding the sentiment of a tweet. For neutral tweets the score is 0.0. We can derive from this chart that Scorsese's mean polarity scores up to 0.106 slightly higher than Spielberg's (0.104). Also all three directors have a positive mean polarity since the number of their negative tweets was relatively low as it is shown on the previous chart.



The time series presents the daily mean polarity for each director for our period of interest. According to our data Hitchcock's polarity drops on the 15th of January while Scorsese's peaks on the 14th. Furthermore, Scorsese's daily mean polarity stays above 0 for the whole period series in contrast to Hitchcock's, which has a wavy form. Finally, Spielberg's polarity is the more stable during our period with only a drop below 0.0 on the 10th of January.



This scatterplot demonstrates the tweets of Scorsese regarding the polarity, the number of words in each tweet and the number of retweets. Each dot represents a tweet. The color map illustrates the polarity (e.g. blue means high polarity). We can derive that as the number of words is increasing in a tweet the polarity of the tweet rises also. Specifically, after a tweet exceeds the 15 words the sentiment becomes positive as the colors suggests. Moreover, tweets with 12 to 23 words tend to get more retweets.

Discussion-Findings:

After exhaustively examining the data, a number of findings was raised. It is observed that some attributes contain a relatively large number of null values. Specifically, the hashtag attribute contains 1282 null values. It is a large number comparing to the 1500 tuples that the table contains. Furthermore, capturing the hashtags was done using regular expressions and not the ‘tweet.entities.get(‘hashtags’)’ method provided by “tweepy”. The reason for that is that the above method returns a list of dictionaries which include the hashtag and its position. Using “re” library for extracting the hashtags was easier to manipulate them. Moreover, storing them brought up another issue. Since we do not know the number of the hashtags used in a tweet, we cannot define the number of columns need in order to store them independently. For this reason, hashtags are stored in one attribute separated by commas.

Another significant issue is the high cardinality over the location attribute. Because of that, manipulating this attribute and trying to visualize it is a difficult task. In order for this to happen we need to transform this attribute manually, correcting every tuple.

Another thing worth mentioning is the retweets in the table. Extracting tweets using the ‘cursor’ returns the retweets too. Excluding them is an option and it can be accomplished by setting a regular expression to check if the text begins with ‘RT’. This report incudes the retweets since they have a major role in formulating the sentiment for the directors. More on that, it is noticed that the ‘cursor’ does not give access to the full text to a retweet but only a part of it. For the this reason ‘TextBlob’ may misclassify some tweets.

As it is mentioned above, sentiment analysis was performed using “TextBlob" package. It is noticed that “TextBlob” does not always identify the sentiment for a tweet (e.g. “Martin Scorsese is a legend.” is classified as neutral). In addition, another issue is the misclassification of irony or slang.

To sum up, performing sentiment analysis on Twitter is a hard task since there are a lot of parameters that need to be set. Nevertheless, capturing how people feel is essential for the development of our society by all means.