Predicting retweet count based on sentiment analysis of historical tweets

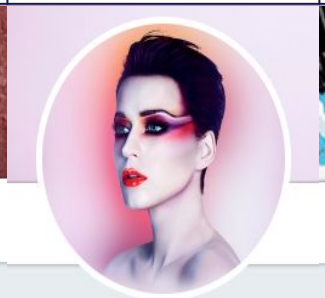
BY
Alex Szilagy [A20324479]
Animesh Patni [A20403240]
Chandana Ravindra Prasad [A20406271]

INTRODUCTION

Problem Statement : Analyzing the possible correlation between the retweet count and the sentiment of a certain tweet.

Proposed Solution : Predicting the retweet count based on the sentiment values of a tweet that is classified among positive, negative and neutral.

Aim : To study, observe, and implement various machine learning algorithms to find the ones that fit the needs of the problem the best way.

Politics	Sports	Authors	Artists	Company	CEO
					
Barack Obama @BarackObama	**LeBron James** @KingJames	**J.K. Rowling** @jk_rowling	**KATY PERRY** @katyperry	**Pepsi™** @pepsi	**Elon Musk** @elonmusk
					
Hillary Clinton @HillaryClinton	**Serena Williams** @serenawilliams	**John Green** @johngreen	**Justin Bieber** @justinbieber	**SpaceX** @SpaceX	**Mary Barra** @mtbarra

DATA

Total number of tweets collected : 38400 (3200*12)

Number of tweets manually tagged : 1200 (100*12)

Name	Id	Text	Retweet count	Follower Count

Fig : Format of the extracted data from twitter using Tweepy

Training set : 75% of the data

Testing set : 25% of the data

SENTIMENT ANALYSIS

Logistic Regression : Provides best results when the target variable is categorical.

Support Vector Machines : Provides maximum margin classification.

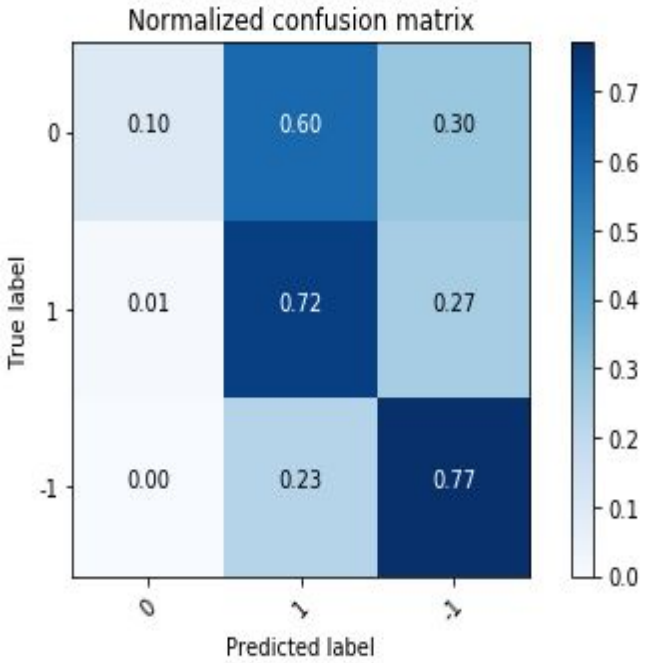
Naive Bayes : Provides improved results as the classifier learns with every new information fed.

Text Blob : Used for the preliminary polarity classification.

Vader : Used to get the compound polarity score which acts as a single measure of polarity (combines positive, negative, and neutral polarity scores).

Manual Tagging : Used to obtain a baseline estimation to compare the other results against with.

SVM Results:



	precision	recall	f1-score	support
0	0.67	0.10	0.17	20
1	0.74	0.72	0.73	160
-1	0.66	0.77	0.71	122
avg / total	0.70	0.70	0.69	302

Accuracy:70.2%
Precision:68.9%
Recall:53.2%
F1:53.9%

SENTIMENT ANALYSIS RESULTS

Text	Manual Tag	Text Blob	Vader	SVM	LR	NB
This tweet wouldn’t have happened five years ago. How have we let this *proudly* racist rats crawl out of our national wood.	-1	1	0.064	0	-1	-1
People with red hair are *less responsive* to anaesthetic.	0	-1	0.298	-1	1	0
\xf0\x9f\x8c\x9e\xe2\x9d\xa4\xef\xb8\x8f\xf0\x9f\x8c\x99 https://t.co/iZv7sjjHzj	0	0	0.0	1	0	0

PREDICTIVE ANALYSIS

Random Forest Regression :

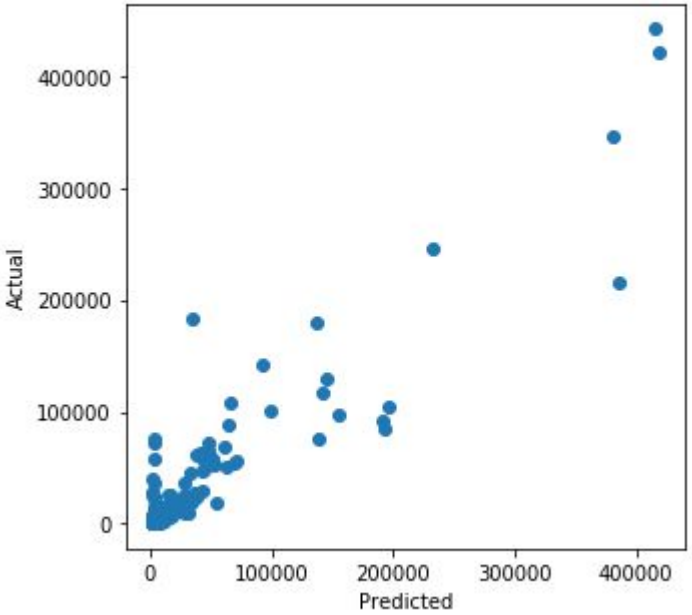
- Works well with non-contiguous data.
- Finds the best split randomly.
- More sensitive to outliers.

Linear Regression :

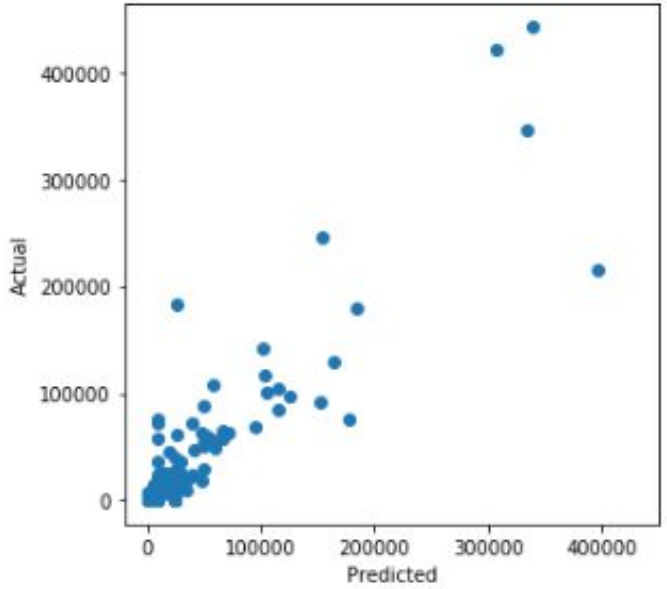
- Works well when there is a linear relationship between target variable and explanatory variable.
- Requires more input data to provide higher accuracy results.

PREDICTIVE ANALYSIS RESULTS

Accuracy: 0.836453854676
Actual vs Predicted using Linear Regression



Accuracy: 0.814571280092
Actual vs Predicted using Random Forest



PREDICTIVE ANALYSIS RESULTS

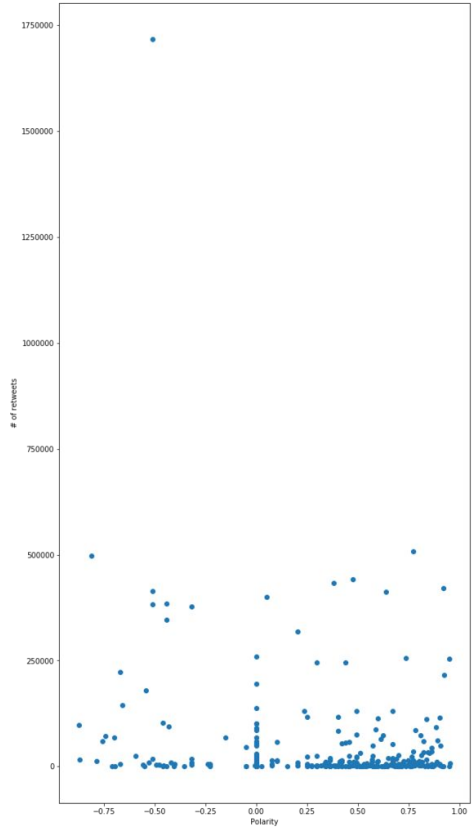
Accuracy: 0.836453854676
Actual vs Predicted using Linear Regression

	Actual Retweets	Predicted Retweet
133	10784	18366.770295
349	1522	2942.175261
1036	0	3146.684905
268	97149	154920.706018
899	3544	10795.812807
60	0	3145.855120
1229	29	960.095713
201	25595	15733.692154
223	10509	31527.818068
866	0	3150.833830
190	8661	15961.606409
457	3051	6045.294216
259	87769	63697.643151
480	64953	48157.708014
1186	16	941.287256

Accuracy: 0.814571280092
Actual vs Predicted using Random Forest

	Actual Retweets	Predicted Retweet
133	10784	21201.400000
349	1522	1409.300000
1036	0	0.000000
268	97149	125618.100000
899	3544	4439.200000
60	0	9490.480912
1229	29	28.400000
201	25595	10364.300000
223	10509	18925.800000
866	0	1.000000
190	8661	9288.800000
457	3051	2779.900000
259	87769	49797.400000
480	64953	66473.500000
1186	16	32.300000

ERROR ANALYSIS - Outlier



Mean values:
Comp_vader - .276
Retweet - 36996.328
Favorite - 150341.071

Outlier Tweet - 1.7M Retweet, 4.6M Favorite

Barack Obama @BarackObama · 12 Aug 2017 1.7M 4.6M

"No one is born hating another person because of the color of his skin or his background or his religion..."

ERROR ANALYSIS

```
[('great', 1.6686141195629294),
 ('happy', 1.526316873470098),
 ('love', 1.4859071489577251),
 ('thank', 1.413918957577895),
 ('proud', 1.3039795020124725),
 ('tesla', 1.2964895701390131),
 ('liftoff', 1.2936327246247472),
 ('congrats', 1.195371054155586),
 ('amazing', 1.0056860545902226),
 ('very', 0.97561523527205185),
 ('excited', 0.96462858954240649),
 ('everyone', 0.92532502062175215),
 ('glad', 0.91155379471269771),
 ('side', 0.9024687323247218),
 ('thanks', 0.89435531568411708),
 ('incredible', 0.87624020482949416),
 ('gm', 0.84787172642345188),
 ('first', 0.83617845240880739),
 ('cutoff', 0.8335967334064549),
 ('gtc39ubc7z', 0.82581452039495118)]
```

PACKAGES USED

1. Tweepy(Data Collection).
2. NLTK(Classifiers).
3. Sklearn(Regression Models)
4. CSV.
5. Vader Sentiment.
6. Pandas.
7. Numpy.