Reddit API & NLP

DSI - 13 Project_3

Julian Broche - East Coast cohort

I'm a moderator for a popular Finance related subreddit....I'm lazy, so I built a script/model to detect non-related posts and automatically remove it from my precious forum.

Problem Statement

Could I build an NLP model that reduces tinfoil posts on my assigned subreddit with high degree of accuracy?

Let's find out.....

Methodology

- PRAW Data Sourcing
- NLP Text to Vectors
 - Count Vec
 - TFIDF Vec
- Models Classification
 - Logistic Regression
 - Random Forest
 - SVM



Quantitative Findings

SubReddit Stocks ~885k

WORDs in Stonks

startbeat guyworkcashindex 100 data clickstate past best short sector 2020 say future deal recent article tradelower number increase right trump bounds bounds contact trump deal recent article tradelower number increase right trump bought eps service current october potential product q3 stimulus etfgood nio money going looking helpportfolio coronavirus question techlong techlong techlong techlong business way

share yearsale price company termearnings newstock

makerevenue investment friday doing gain

investorestimateknow time²⁰option energy day expected fund hold friday doing invest invest friday doing marketjust averagenews buying marketjust weektradingelection high today buying marketjust weektradingelection high

bank post positive tesla value return chart apple million really better release thing need lot 10 come biglevel

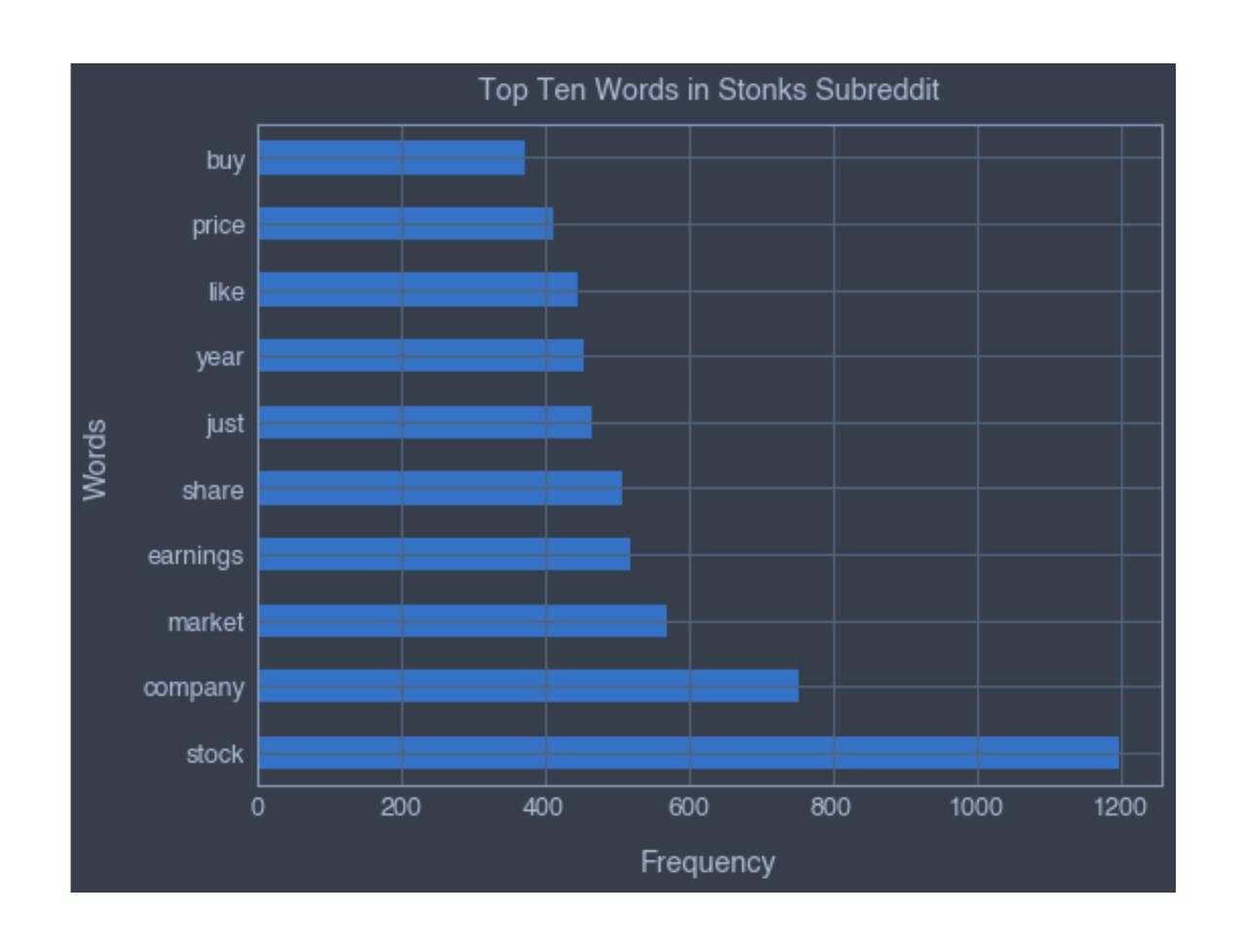
51

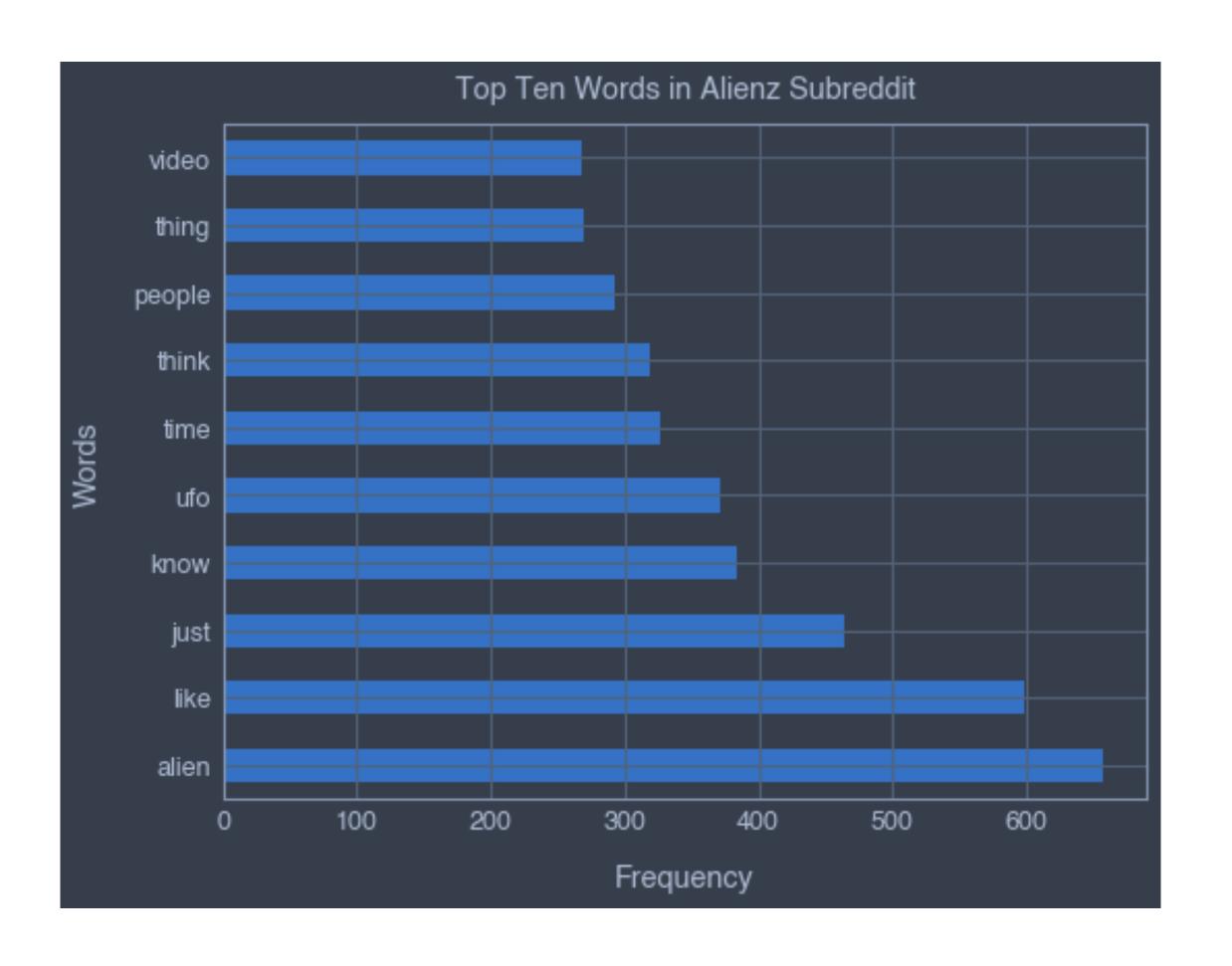
SubReddit Aliens ~229k

Words in Alienz

stateflying grey used fear new shipareasmall agosky heard pretty fact casestrange post reasonpossible placewindow house means leep airtheory airtheory awayencounter interesting outside line of the state of the sta said dayve sawnight evidence actually advanced earth read Want interesting outside contact farworldvideouseextraterrestrial experience timeufo happened report comment question looking thinking footage came make remember right just lot human space think a lien information sure know look people like maybe government yearthing believe wentcome life thought point friend similar different craft little coming picture mind real talking body travel forcetype gotweird sort little coming point forcetype gotweird sort little coming picture mind real talking little coming picture mind real talking sort little coming picture mind real talki

Top 10 words by SubReddit

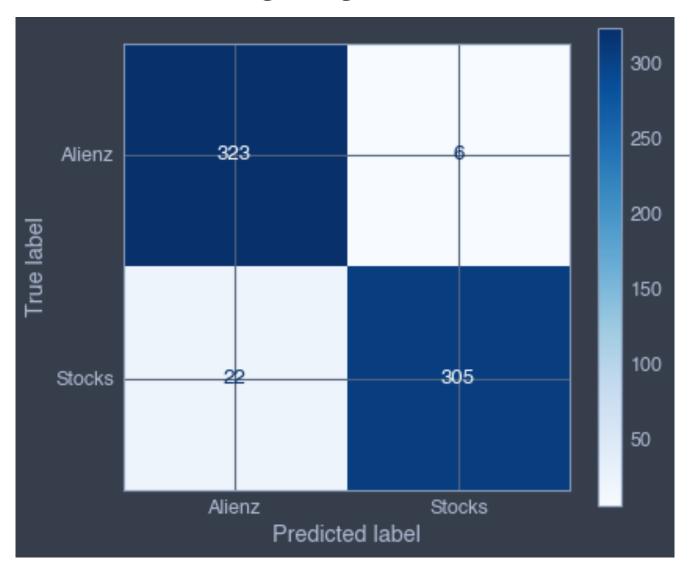




Lets Talk Models

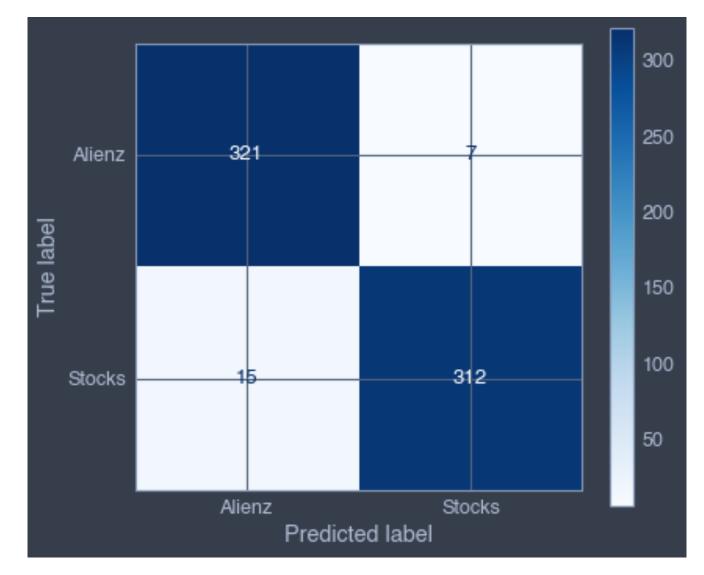
Time to get Confused

Log Reg Matrix



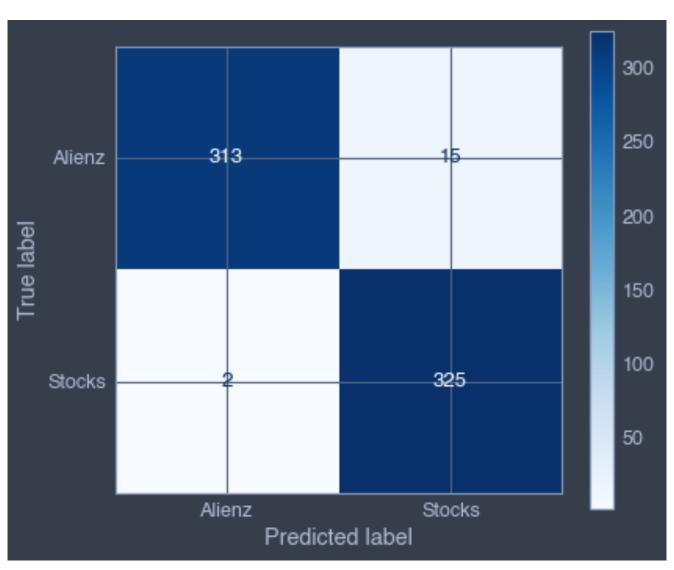
Accuracy Score: 0.957%

Random Forest Matrix



Accuracy Score: 0.966%

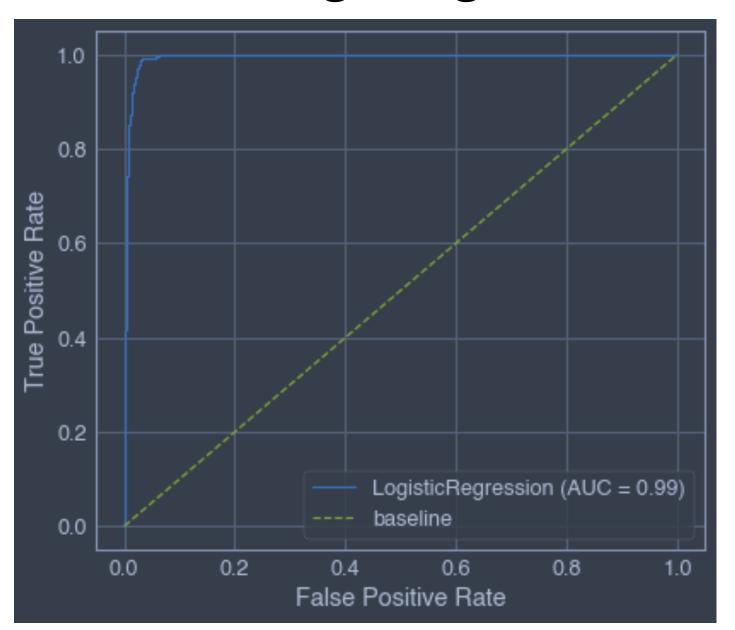
SVM Matrix



Accuracy Score: 0.974%

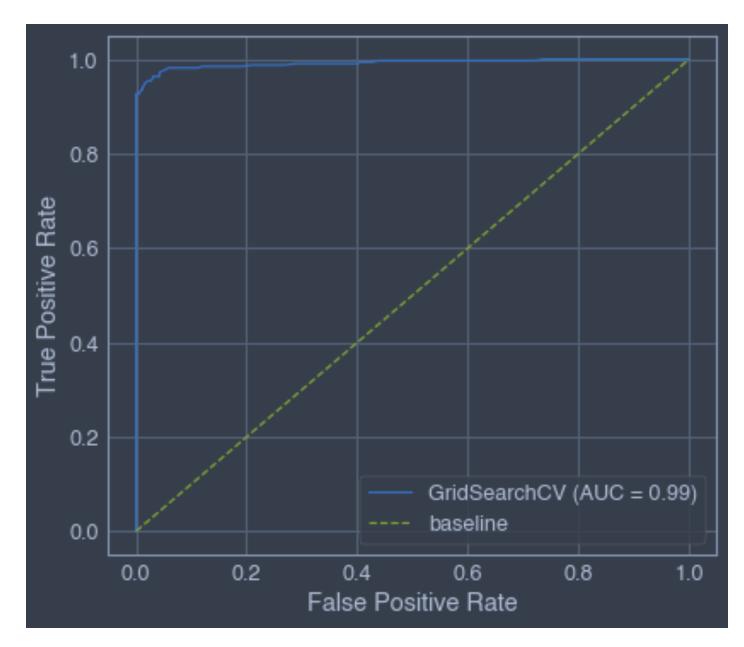
Curves

Log Reg



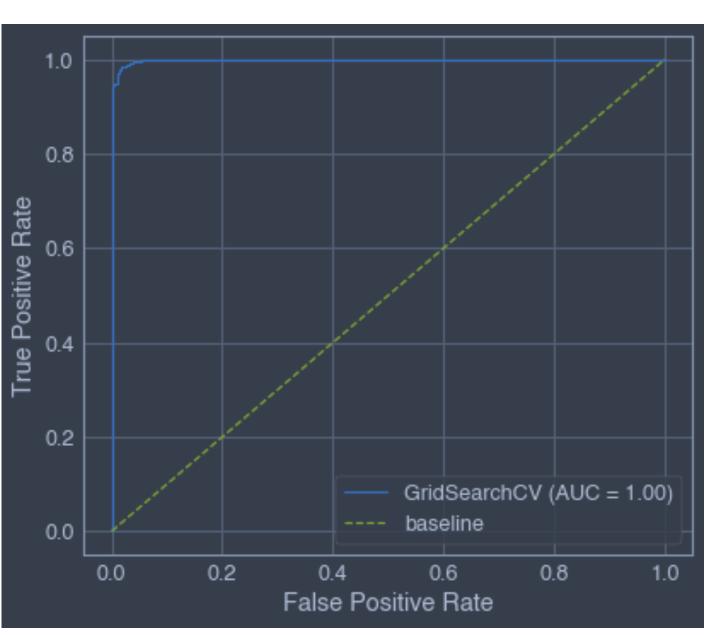
ROC_AUC Score: 0.994%

Random Forest



ROC_AUC Score: 0.991%

SVM



ROC_AUC Score: 0.998%

LogReg vs RF Metrics

Log Reg

Accuracy score 0	.95731707	31707317			
	Actual A	Alienz A	 ctual Stonk	 (S	
Predicted Alienz		323		6	
Predicted Stonks		22	30	5	
pr	ecision	recall	f1-score	support	
0	0.94	0.98	0.96	329	
1	0.98	0.93	0.96	327	
accuracy			0.96	656	
macro avg	0.96	0.96	0.96	656	
weighted avg	0.96	0.96	0.96	656	
ROC_AUC SCORE 0.	994813306	9351105			

Random Forest

	Actual	Alienz A	ctual Stonk	(S	
redicted Alie	enz	321		7	
redicted Stor	nks	15	31	12	
	precision	recall	f1-score	support	
0	0.96	0.98	0.97	328	
1	0.98	0.95	0.97	327	
accuracy			0.97	655	
accuracy macro avg	0.97	0.97	0.97	655	
eighted avg	0.97	0.97	0.97	655	

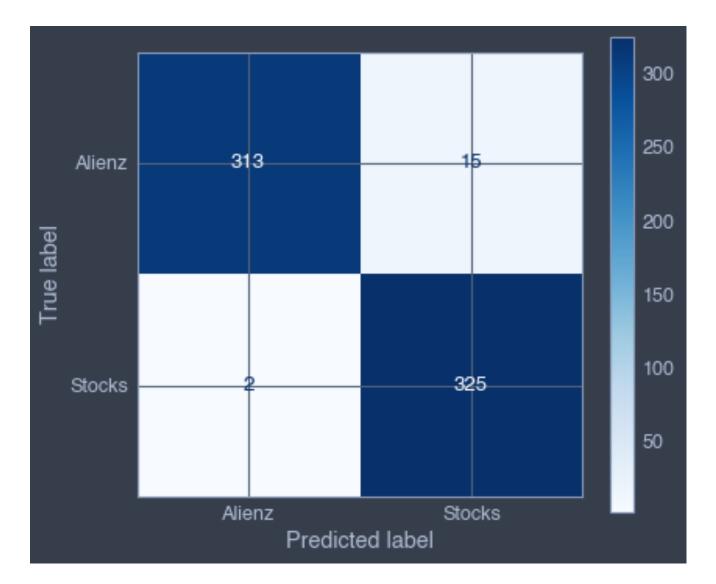
Lets Talk SVM Model

Grid_Best Score: 0.982%

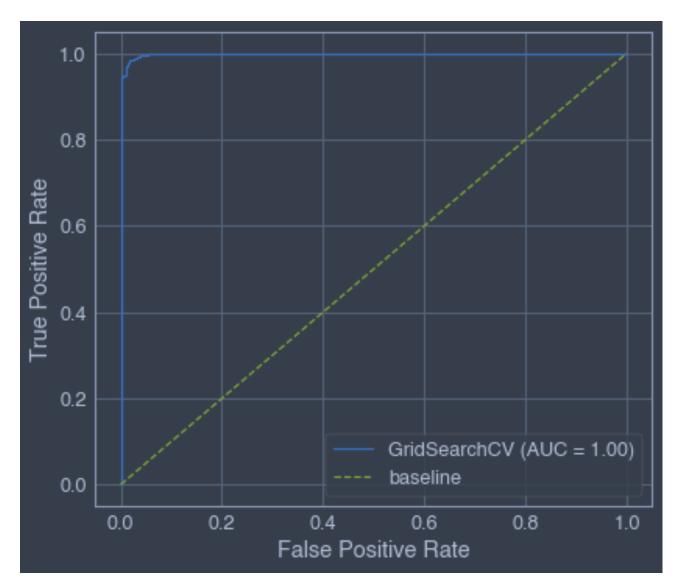
Grid_Best Params: {'C': 10.0, 'gamma': 0.1, 'kernel': 'rbf'}

	Actual	Alienz A	ctual Stonk	S	
Predicted Alier	ız	313	1	.5	
Predicted Stonk	(S	2	325		
į.	recision	recall	f1-score	support	
0	0.99	0.95	0.97	328	
1	0.96	0.99	0.97	327	
1	0.90	0.99	0.97	321	
accuracy			0.97	655	
macro avg	0.97	0.97	0.97	655	
weighted avg	0.97	0.97	0.97	655	

SVM Matrix



ROC_AUC



Recommendation

Don't be posting tin foil talk in the Finance subreddit because it will get swiftly removed.

NOT IN MY HOUSE.

Wrap up Q's

Sources

https://www.reddit.com/r/aliens/

https://www.reddit.com/r/stocks/

Data Visualization - Word Art by Tableau - Julian Broche

Metrics Print Out By Heather Robbins Lesson - NLP