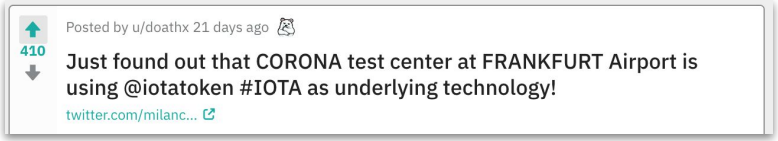
Subreddit predictions: Classifying subreddit posts between IOTA and IOTAmarkets subreddits using NLP

January 28, 2020

Prepared by: Noah Zuckerman

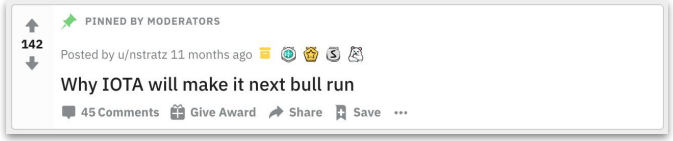
Selected two subreddit within the IOTA cryptocurrency community

IOTA



Community focused on development of the technology

IOTA Markets



Community focused on price speculation of the crypto token

Build a classification model to accurately predict whether a post was originally posted in IOTA or IOTAmarkets

Subreddit post raw data

10K rows of data

	subreddit	
0	Iota	**This posting was written(translate
1	Iota	
2	Iota	
3	Iota	Can someone explain this:\n\n“An atta
4	Iota	

Most frequently occuring words

r/IOTA
r/Iota

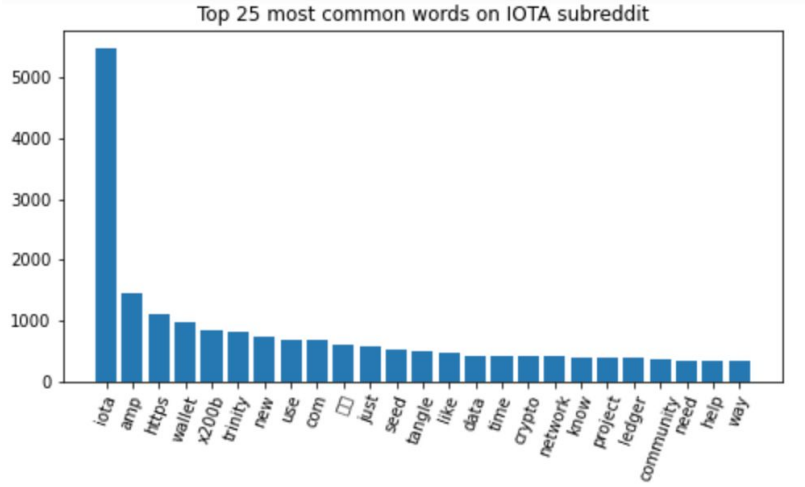
Word	Frequency
iota	5503
amp	1458
https	1108
wallet	975
x200b	831
trinity	824
new	7474

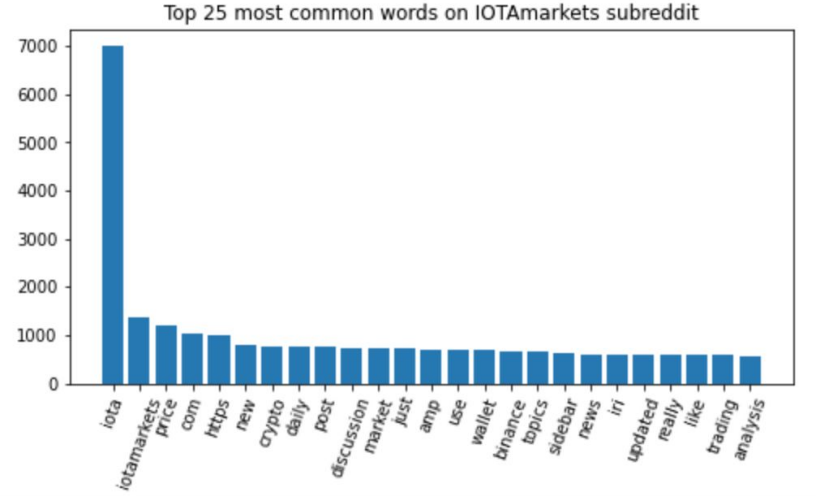
r/IOTAmarkets
r/IOTAmarkets

Word	Frequency
iota	6992
iotamarkets	1370
price	1210
com	1017
https	1009
new	790
crypto	767

The model predictions will be evaluated based on overall accuracy of the model

Across both subreddits, there are many overlapping keywords included in posts making classification more challenging





Overwhelmingly and unsurprisingly the word Iota comes up the most frequently across both subreddits

Model building methodology

Data cleaning

- Drop nulls, removed and deleted posts

EDA and Feature Engineering

- Investigate word frequency across corpuses
- Label encode **iota as 1** and **iotamarkets as 0**

Model Preparation & Preprocessing

- Evaluate baseline accuracy **~50.1%**
- Test/Train split
- Transform text columns into features using Countvectorizer

Model Evaluation

- Create Pipe with CountVectorizer and then various Algorithms
- Define pipe params
- Gridsearch fit and score

Production Model Selection

- Feature engineer combined ‘features’ column including selftext and title
- Fine tune top 3 performing models using both cvec and tvec
- Hypertune two best performing models to optimize accuracy while decreasing model variability
- Apply bagging to final model

Conclusions and Recommendations

Model Evaluation - Round 1

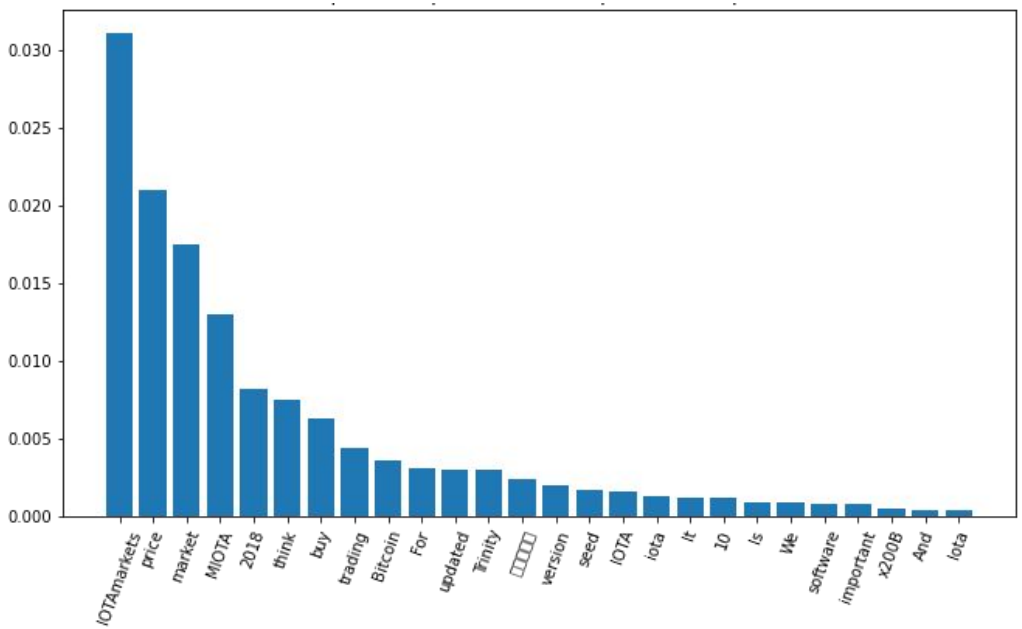
Model Name	Best Score	Train Score	Test Score	Notable Parameters
NB - Multinomial *(Title)*	0.722	0.824	0.728	Cvec max features: 5000 Cvec stop words: English Cvec lowercase: False Cvec ngram range: (1,1)
KNN *(Title)*	0.651	0.760	0.653	
Logistic Regression *(Title)*	0.705	0.867	0.717	Cvec max features: 5000 Cvec stop words: None Cvec lowercase: False Cvec ngram range: (1,1)
Decision Trees *(Title)*	0.624	0.634	0.621	
Random Forests *(Title)*				Cvec max features: 5000 Cvec stop words: None Cvec lowercase: True Cvec ngram range: (1,2)

Model Evaluation - Round 2

Model Name	Best Score	Train Score	Test Score	Notable Parameters
NB - Multinomial (TfidfVectorizer)	0.740	0.847	0.752	
NB - Bernoulli (CountVectorizer)	0.831	0.865	0.812	
Logistic Regression (CountVectorizer)	0.897	0.954	0.901	C: 0.5 Cvec max features: 10,000 Cvec max df: 0.9 Cvec strip accents: unicode
Random Forests (TfidfVectorizer)	0.923	0.953	0.899	Max Depth: None Min Samples Leaf: 2 Min Samples Split: 3 N estimators: 150 Tvec max features: 5000 Tvec max df: 0.98 Tvec strip accents: Ascii

Interpretation of model results

Logistic Regression Most Predictive Features

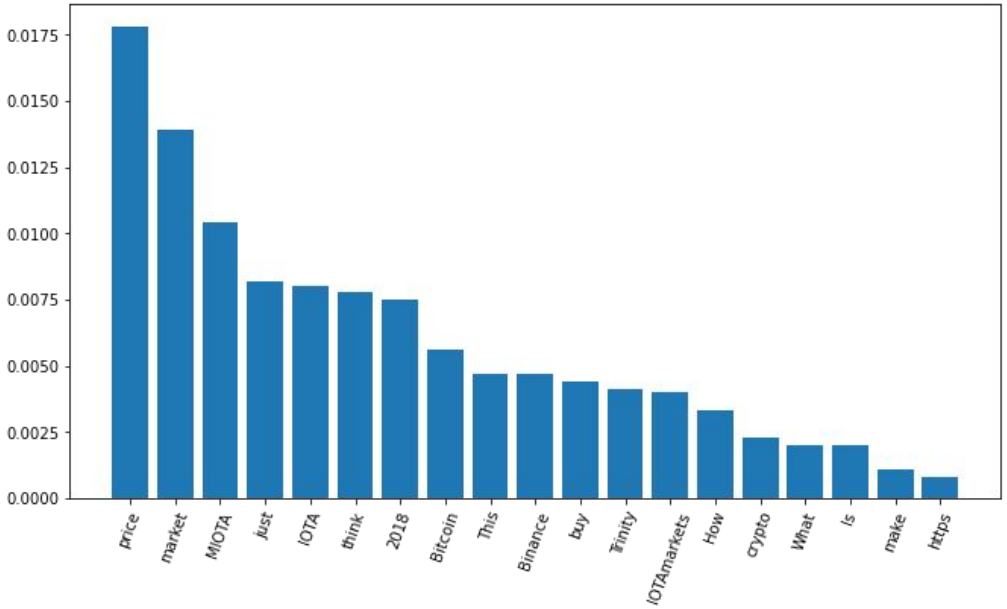



```
IOTAmarkets0.031 +/- 0.002
price    0.021 +/- 0.006
market   0.018 +/- 0.001
MIOTA    0.013 +/- 0.002
2018     0.008 +/- 0.001
think    0.007 +/- 0.000
buy      0.006 +/- 0.001
trading  0.004 +/- 0.002
Bitcoin  0.004 +/- 0.001
For      0.003 +/- 0.000
updated  0.003 +/- 0.001
Trinity  0.003 +/- 0.001
오전     0.002 +/- 0.000
version  0.002 +/- 0.001
seed     0.002 +/- 0.001
IOTA     0.002 +/- 0.001
iota     0.001 +/- 0.000
It       0.001 +/- 0.000
10       0.001 +/- 0.000
Is       0.001 +/- 0.000
We       0.001 +/- 0.000
software0.001 +/- 0.000
important0.001 +/- 0.000
x200B    0.001 +/- 0.000
And      0.000 +/- 0.000
Iota     0.000 +/- 0.000
```

The most predictive features across both models show quite a bit of overlap but also quite a bit of variance as well

Interpretation of model results

Random Forests Most Predictive Features




```
price    0.018 +/- 0.004
market   0.014 +/- 0.001
MIOTA    0.010 +/- 0.001
just     0.008 +/- 0.001
IOTA     0.008 +/- 0.004
think    0.008 +/- 0.001
2018     0.007 +/- 0.001
Bitcoin  0.006 +/- 0.000
This     0.005 +/- 0.001
Binance  0.005 +/- 0.002
buy      0.004 +/- 0.002
Trinity  0.004 +/- 0.001
IOTAmarkets0.004 +/- 0.001
How      0.003 +/- 0.001
crypto   0.002 +/- 0.001
What     0.002 +/- 0.001
Is       0.002 +/- 0.001
make     0.001 +/- 0.000
https    0.001 +/- 0.000
```

The most predictive features across both models show quite a bit of overlap but also quite a bit of variance as well

Model Evaluation - Round 3

Model Name	Best Score	Train Score	Test Score	Notable Parameters
Logistic Regression w/ special chars	0.897	0.954	0.901	
Logistic Regression w/o special chars	0.830	0.856	0.812	
Logistic Regression w/o special chars & lemmatization	0.835	0.855	0.821	
Logistic Regression w/o special chars & stemming	0.838	0.867	0.832	
Random Forests w/ special chars	0.923	0.953	0.899	
Random Forests w/o special chars	0.841	0.866	0.826	
Random Forests w/o special chars & lemmatization	0.845	0.865	0.811	
Random Forests w/o special chars & stemming	0.845	0.877	0.812	

Production Model Selection

Model Name	Best Score	Train Score	Test Score
Logistic Regression	0.921	0.959	0.918
Random Forests	0.911	0.951	0.901

Select Logistic Regression model and hypertuned for optimal results:

96.1%
Train Score

91.9%
Test Score

Apply bagging meta-estimator to selected regression model:

96.3%
Train Score

92.2%
Test Score

93.2%
Sensitivity

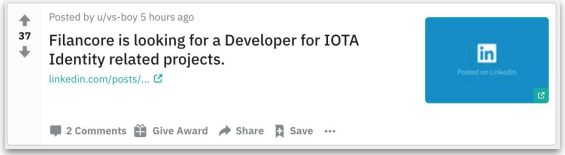
89.9%
Specificity

Recommendations for moderators

1 Consider these words when reviewing posts

- binance
- price
- buy
- mIOTA
- http
- teally
- bitcoin
- just
- 10
- btc
- com
- think
- wallet
- exchange
- crypto
- market

2 Understand this model is only designed to flag



Next steps...

Given additional time:

- Pull more data (historical and layer in comments)
- Optimize for False Negatives (e.g. maximize Sensitivity)
- Investigate tradeoff of using ‘blackbox’ algorithms to further improve performance