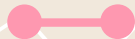




# Subreddit Classifier

r/investing or  
r/cryptocurrency!?

Team 2  
Clara, Luka, Shu Kai,  
Kang Yang, Wee Cheng





# TABLE OF CONTENTS

**01** Background &  
Problem Statement

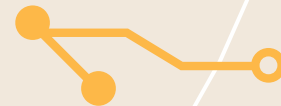
**02** Preprocessing &  
NLP

**03** Exploratory Data  
Analysis

**04** Model 1: Naive Bayes  
Classifier

**05** Model 2: Logistic  
Regression

**06** Conclusion



01

# Background & Problem Statement





# Chosen Subreddits

## **r/investing**

- <https://www.reddit.com/r/investing/>

## **r/CryptoCurrency**

- <https://www.reddit.com/r/CryptoCurrency/>





# Background

We are a new investment company which has two main trading desks :

- one for **traditional securities** and,
- another for **cryptocurrency**





# Problem Statement

- **Automate** the **monitoring** of reddit posts
- Identify **new leads & hot trends**
- **Filter** such information to the **specific** trading desks



02

# Preprocessing & NLP



# Scraping, Preprocessing & NLP



**Additional Data Cleaning**

Preprocessing



**Scraping**



**Tokenizing, Initial  
Dropping of  
Stopwords, &  
Lemmatizing**

NLP





# Scraping










- Extracted 5k unique posts before 25 Oct 2021, GMT+8 2359
- Drop empty posts
- Drop duplicated posts
- Drop nulls

Removed post

 Posted by u/volant007 35 minutes ago

**Vote** **Best FREE app to track stocks** 

 **Sorry, this post has been removed by the moderators of r/investing.**  
Moderators remove posts from feeds for a variety of reasons, including keeping communities safe, civil, and true to their purpose.

 **1 Comments**  **Award**  **Share**  **Save**  **Hide**  **Report** 100% Upvoted

Example of a post that has been removed, hence resulting in an empty selftext & post.

# Preprocessing

## Additional Data Cleaning

- URLs

URL



Posted by u/indivinvest 5 hours ago

31



### Should you invest in IPOs?

Hey guys, if you have ever considered investing in IPOs, or companies that have gone IPO in the recent years, then this post is for you. I did a bit of a quick research into this recently. But first, a quick recap of what Warren Buffet says about investing in IPOs:

- Warren Buffet has always been very vocal about his views on IPOs. His thought process is that "*there are always better businesses to buy than IPOs*". He always compares IPOs with other alternatives in the market, and so far, he has not found a strong justification for investing in an IPO versus another solid business at a good price.
- Buffet also dislikes the fact in IPOs, there are commissions that go to the stock salesmen, remember that IPOs are executed by investment banks so there is a push to sell IPOs like any other product, He likes to buy stocks where "*no one is making any money on the sale*".

I'd like to share with you some stats that I put together. Bear in mind the following:

- I took the [full list](#) of IPOs in the last 20 years from iposcoop.com.
- Filtered only companies that went public between years 2000 and 2015, so that at a minimum we have at least 5-6 years of post IPO performance.
- Calculated annualized performance using stock price from the first day close of IPO and current stock's price at the time this video was made.
- Performance from companies that are not currently trading were excluded.
- List [here](#) for your reference.

Example of a post that has a URL.

# Preprocessing






## Additional Data Cleaning

- HTML Special Entities & Terms
  - E.g. Codes behind symbols '&amp;', '&gt;'
  - E.g. Non-Breaking Spaces like '#x200B;', '#xa0;'
- Digits
- Special Characters
- Remove characters beyond Basic Multilingual Plane (BMP) of Unicode
  - Example of characters beyond the BMP would be 'å', and emojis.



## Tokenizing, Initial Dropping of Stopwords, & Lemmatizing

↑ 148  
↓

 r/CryptoCurrency · Posted by  u/Dwez1337 Uranium  8.4k · 4 hours ago

**Reminder: Jim Cramer Called for Selling Ethereum a Week before it Rallied to All Time High — Also did the Same for Bitcoin before it Doubled in Price**

MARKETS

A week ago Jim Cramer has predicted that Ethereum has reached its top, and might fall from its then price point.

In the following week Ethereum has rallied to a new all time high of over \$4,600, making Cramer predict exactly the opposite of what has happened

He has done the same for Bitcoin, called for selling it around the \$30,000 price point, in the following weeks it rallied to over \$60,000

(source on cramer selling) (source2)

59 Comments Share Save Hide Report 95% Upvoted

Words that  
do not add  
value to  
meaning

Symbols,  
Punctuations,  
Digits

# Example of Lemmatizing

**Original** websites changing regularly

---

**Lemmatizer** website changing regularly



03

# Explorative Data Analysis



# Text Corpus



r/investing · Posted by u/DarthTrader357 1 day ago

0



## Can someone explain to me how margin actually is useful?

I'm struggling to justify the use of margin, particularly in naked puts. I learn best by doing, but before doing I need to learn, so I'm in a bit of a catch-22 with margin.

The problem I'm having is visualizing the risk-reward to using it. I can take on substantially more risk with cash than I can take on with margin. And I can substantially concentrate my trades more with cash than with margin. (That is to say, the more concentrated in higher volatility, the more my margin buying power just begins to look like cash-only).

So inevitably it looks like I can outperform with cash more than I can with margin.

Thus the question, how is margin actually useful?

For simplicity, if I want to 2x my buying power, but I have to reduce my return by 1/2x, then I have effectively accomplished nothing. Correct? So for margin to be useful I have to be able to both maintain a certain rate of return AND increase the buying power. Yes?

I have some other advanced questions about house surplus, equity and journaling times, but I think first I just need to see someone justify the use of margin in the first place.

Another problem with a margin call that I have is that it removes my ability to choose when I lose. Now the brokerage gets to choose when I lose, and they don't care if 2 weeks later the trade reverses and I would be a complete winner.

That alone is a very big hindrance when working a long-game high risk trade.

Maybe someone can comment a little on that as well.

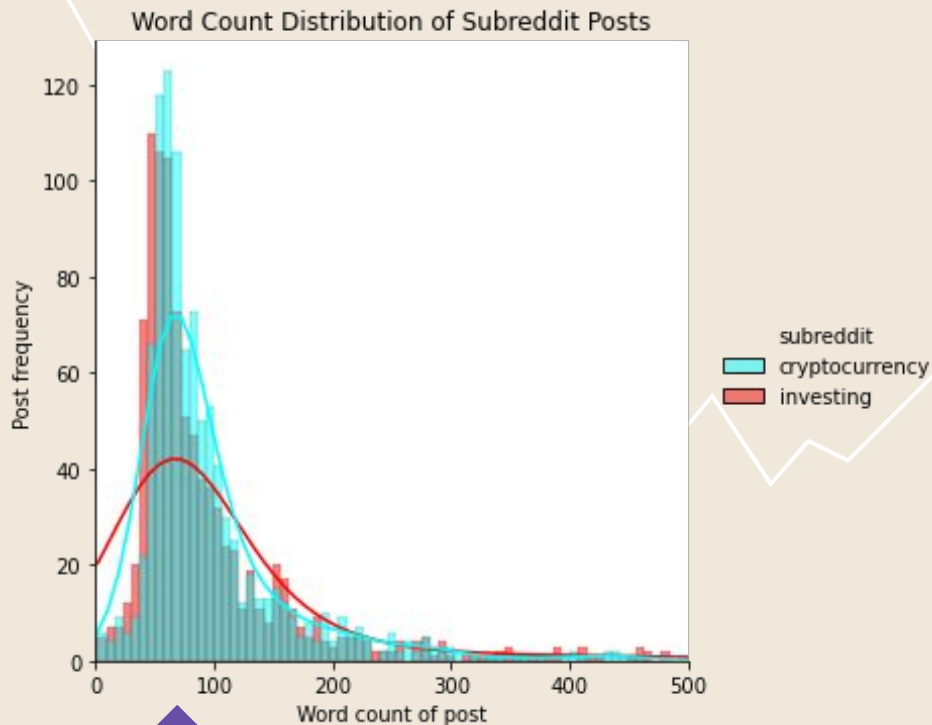
Thanks

Title

+

Body





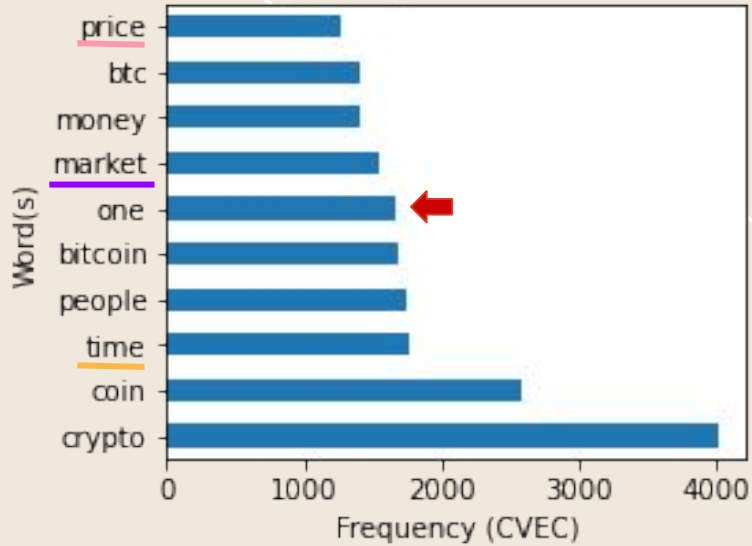
# Word Vectorization

- Count Vectorization
- TF-IDF Vectorization
- N-gram Frequency

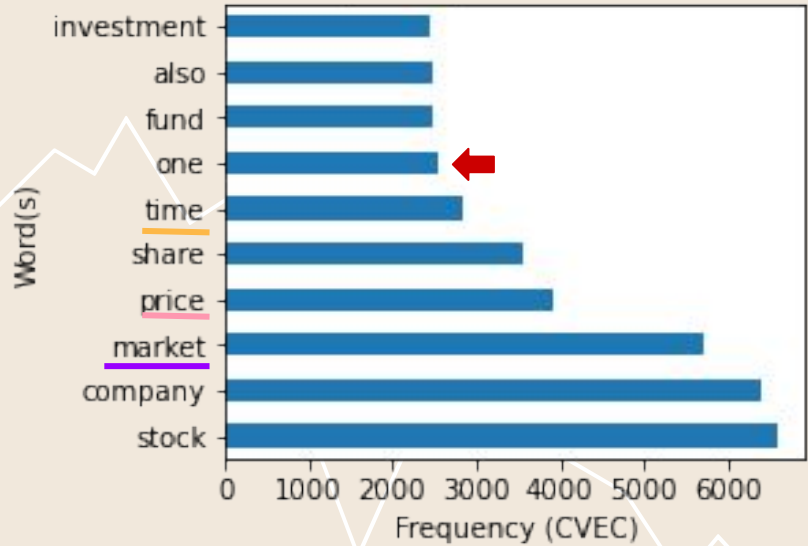


# Unigram (CVEC)

Top 10 Word(s) in r/cryptocurrency  
(CVEC) - Lemmetizer



Top 10 Word(s) in r/investing  
(CVEC) - Lemmetizer

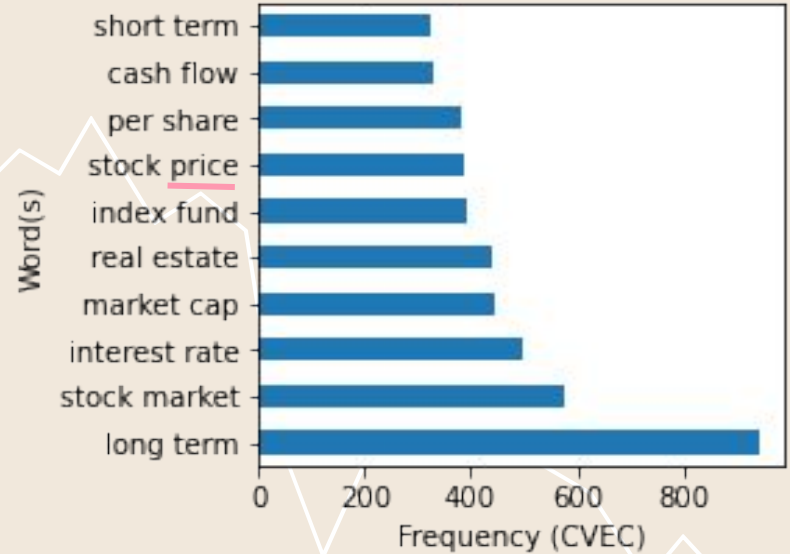


# Bigram (CVEC)

Top 10 Word(s) in r/cryptocurrency  
(CVEC) - Lemmetizer



Top 10 Word(s) in r/investing  
(CVEC) - Lemmetizer



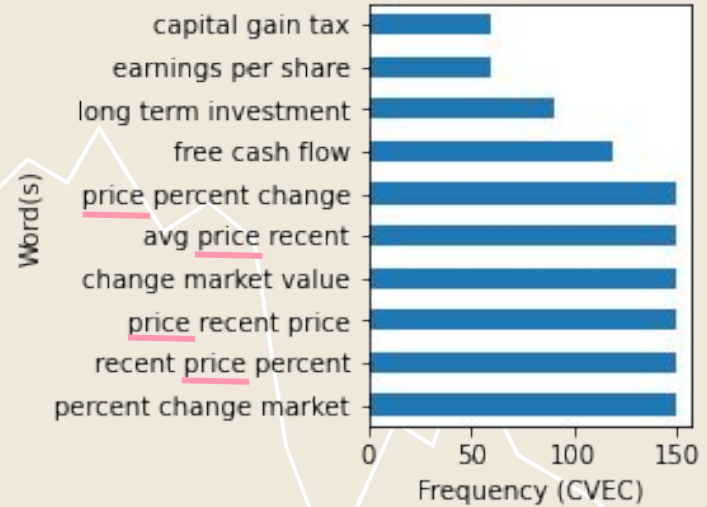
# Trigram (CVEC)



Top 10 Word(s) in r/cryptocurrency  
(CVEC) - Lemmetizer



Top 10 Word(s) in r/investing  
(CVEC) - Lemmetizer



# Additional Stopwords



## Market

'bear market'  
'market cap'

'low market cap'  
'change market value'  
'market cap coin'

## Time

'time high'

'new time high'

## Price

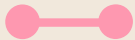
'stock price'

'avg recent price'  
'price recent price'  
'recent price percent'

## Long term

'Short term'

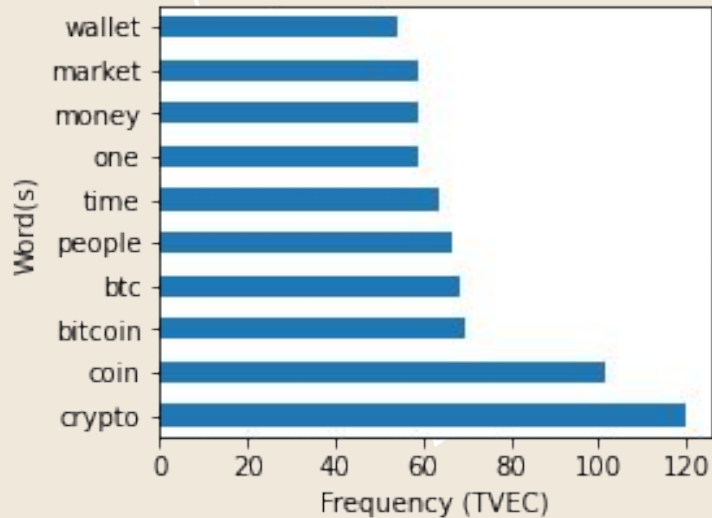
'Long term investment'



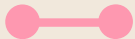
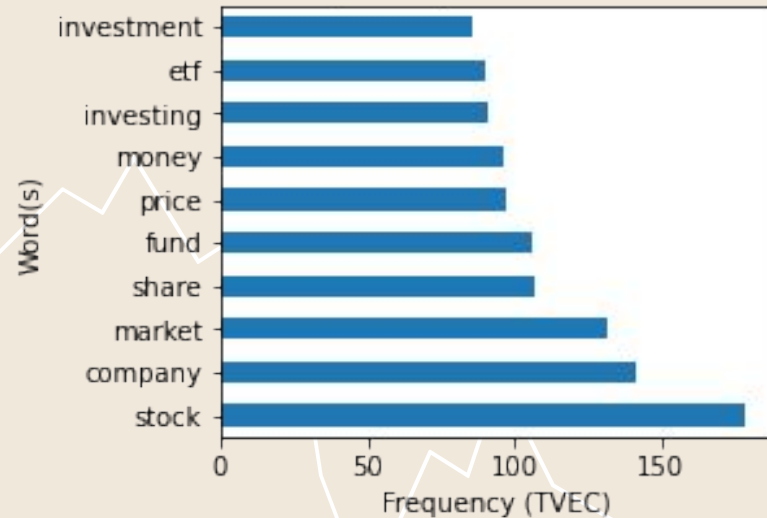
# Unigram (TVEC)



Top 10 Word(s) in r/cryptocurrency  
(TVEC) - Lemmetizer



Top 10 Word(s) in r/investing  
(TVEC) - Lemmetizer

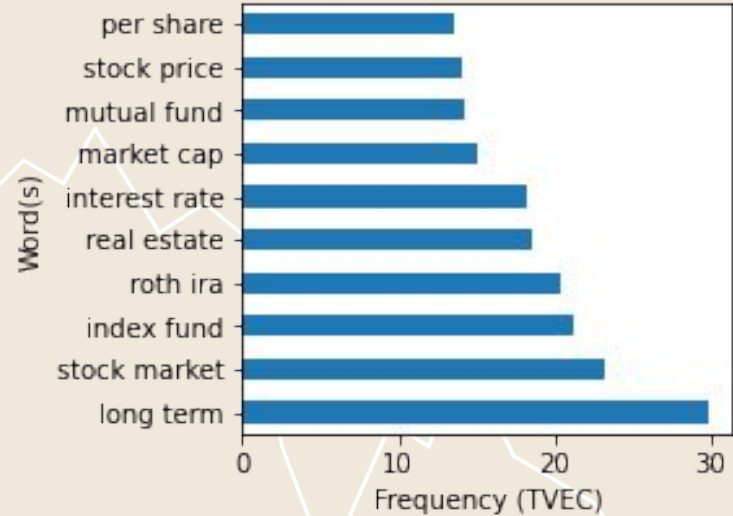


# Bigram (TVEC)

Top 10 Word(s) in r/cryptocurrency  
(TVEC) - Lemmetizer



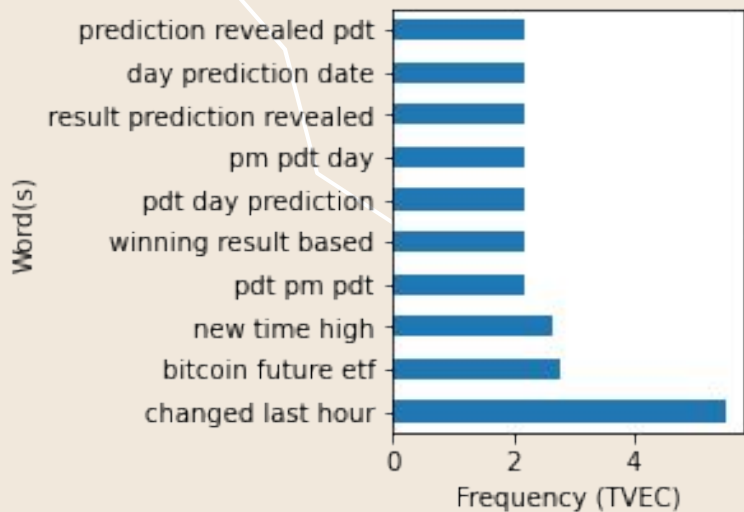
Top 10 Word(s) in r/investing  
(TVEC) - Lemmetizer



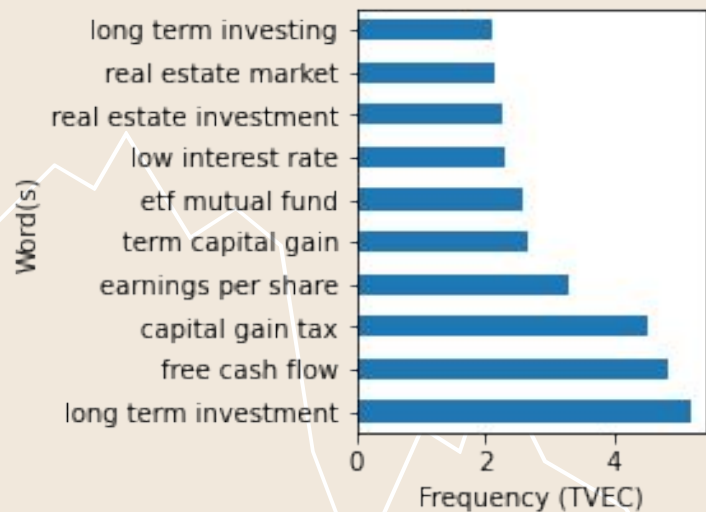
# Trigram (TVEC)



Top 10 Word(s) in r/cryptocurrency  
(TVEC) - Lemmetizer



Top 10 Word(s) in r/investing  
(TVEC) - Lemmetizer











# Initial Insights

#1



Repetition of  
words

#2



Common words in  
both subreddits

#3



Presence of  
distinctive  
keywords

04

# Naive Bayes



# Choice of Vectorizers and Hyperparameters


**Count Vectorizer**

**TF-IDF Vectorizer**

**Hyperparameters**

**Max features=**  
**2000-5000**  
**ngram range= (1,1), (1,2) ,**  
**(1,3)**

# Choice of model and Hyperparameters



**Multinomial Naive Bayes (CVec)**

**Multinomial Naive Bayes (TD-IDF)**



**Hyperparameters**

**CV = 5, 10, 15, 20**



# CVEC vs TF-IDF Naive Bayes



## CVEC Naive Bayes Classifier

## TF-IDF Naive Bayes Classifier

Train Accuracy

94.4%

95.2%

Test Accuracy

93.0%

93.7 %

Sensitivity

(Hit rate for correctly classifying r/investing)

88.4%

91.4%

Specificity

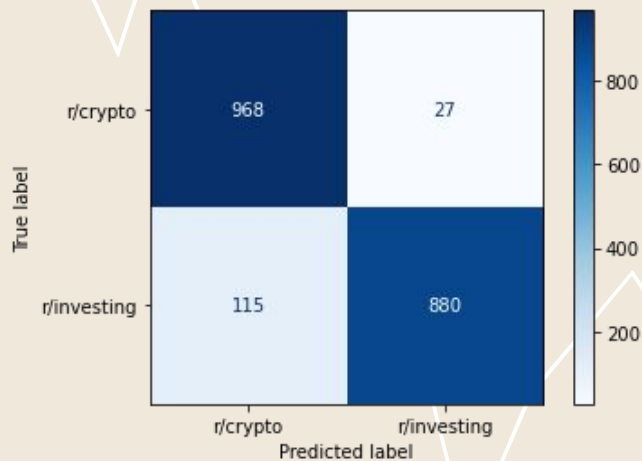
(Hit rate for correctly classifying r/cryptocurrency)

97.3%

96.1%

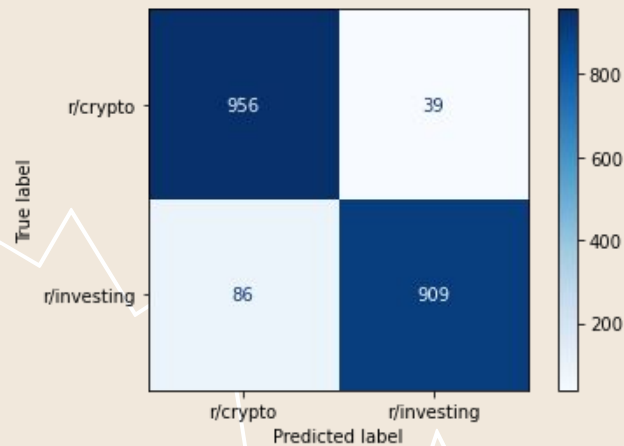


## Multinomial Naive Bayes (CVec)



F1 Score: 92.5%

## Multinomial Naive Bayes (TF-IDF)



F1 Score: 93.6%

# 05 Logistic Regression





# Choice of Vectorizers and Hyperparameters

**Count Vectorizer**

**TF-IDF Vectorizer**

**Hyperparameters**

**Max features =**

**2000-5000**

**ngram range= (1,1), (1,2) ,  
(1,3)**

# Choice of model and Hyperparameters



**Logistic Regression  
(CVec)**

**Logistic Regression  
(TD-IDF)**



**Hyperparameters**

**CV = 5, 10, 15, 20**



# CVEC vs TD-IDF Logistic Regression



## CVEC Logistic Regression

## TF-IDF Logistic Regression

Train Accuracy

97.9%

96.1%

Test Accuracy

93.3%

94.8%

Sensitivity

(Hit rate for correctly classifying  
r/investing)

93.8%

94.4%

Specificity

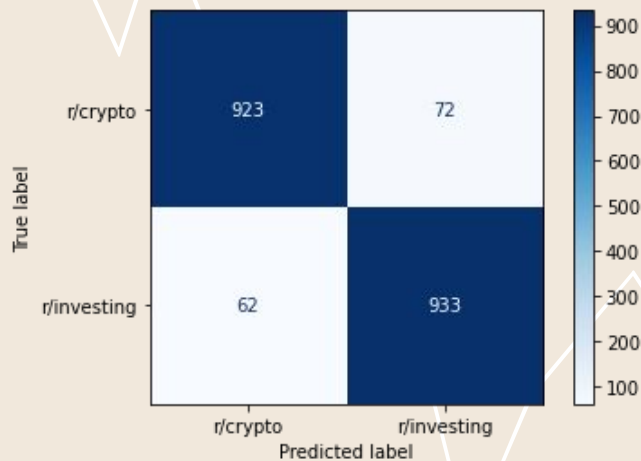
(Hit rate for correctly classifying  
r/cryptocurrency)

92.8%

95.2%

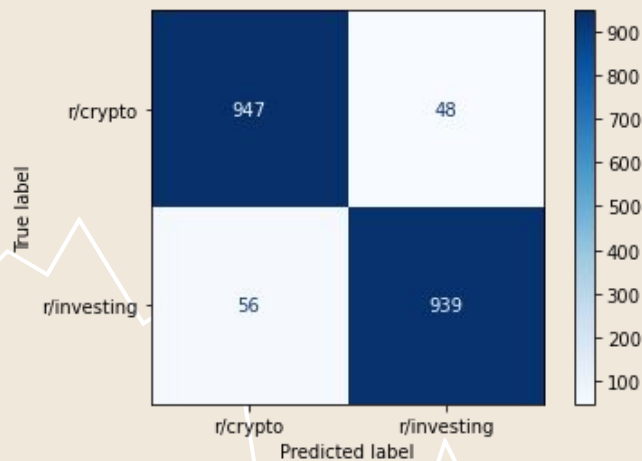


## Logistic Regression (CVec)



F1 Score: 93.3%

## Logistic Regression (TF-IDF)



F1 Score: 94.8%

# Optimal Hyperparameters after GridSearchCV

## Optimized Hyperparameters

**CV = 5**  
**Max Features = 5000**  
**N-Grams = (1,2)**

## Overall Top Performing Model

**Logistic Regression  
(TD-IDF)**

Accuracy

94.8%

Sensitivity

94.4%

Specificity

95.2%

Precision

95.1%

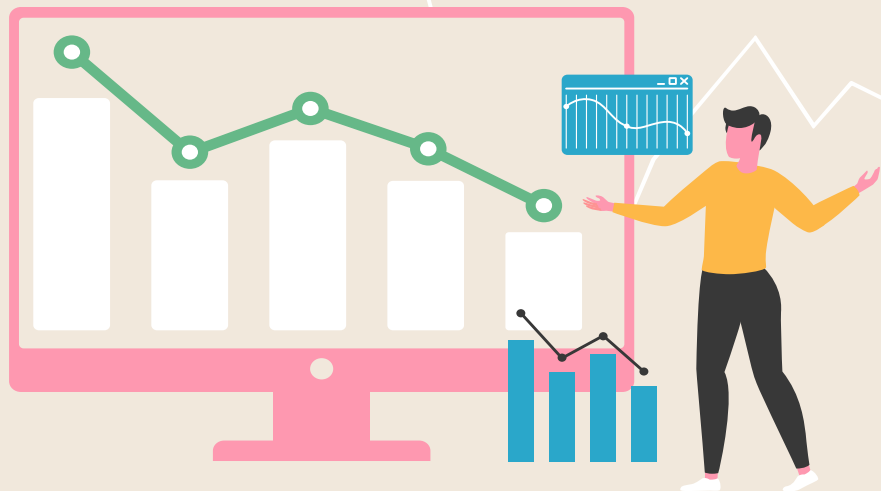
F1 Score

94.8%



06

Conclusion



# Model Comparison Summary



**TF-IDF  
Vectorizer +**

**Accuracy**  
(train/test gap)

**Sensitivity**  
(Hit rate for correctly  
classifying r/investing)

**Specificity**  
(Hit rate for correctly  
classifying r/cryptocurrency)

**Naive Bayes  
Classifier**

**Logistic  
Regression**

Good ~94% (-2%)	Better ~96% (-1%)
Good ~91%	Better ~95%
Better ~96%	Good ~95%



# Final Model Selection

TF-IDF  
Vectorizer +

**Accuracy**  
(train/test gap)

**Sensitivity**  
(Hit rate for correctly  
classifying r/investing)

**Specificity**  
(Hit rate for correctly  
classifying r/cryptocurrency)

Naive Bayes  
Classifier

Good ~94%  
(-2%)

Good ~91%

Better ~96%

Logistic  
Regression

Better ~96%  
(-1%)

Better ~95%

Good ~95%





# Findings / Insights



## Investing market movers?

tesla  
covid  
china

## Crypto is hardly cryptic

Niche subject with distinctive vocabulary and pet phrases



## Top 'coin' mentions

btc  
eth  
shib  
ada

In case you were wondering...

btc : bitcoin  
eth : ethereum  
shib : shiba inu  
ada : cardano





## Conclusion

96%

### Model Metrics

Getting it right 96%  
of the time

Respectable!

5%

### Misclassification

Turns out,  
Machine was correct\*

Human was wrong...  
i.e. 'misposted' into  
the wrong subreddit!

>96%

### Captures topics

Filtering posts  
enables trading  
desks to deep dive  
into #nascent topics  
before they become  
#trending

\*Some of crypto topics posted into r/investing were 'correctly' identified, and would not be out of place in r/cryptocurrency



# Future Improvements / Opportunities



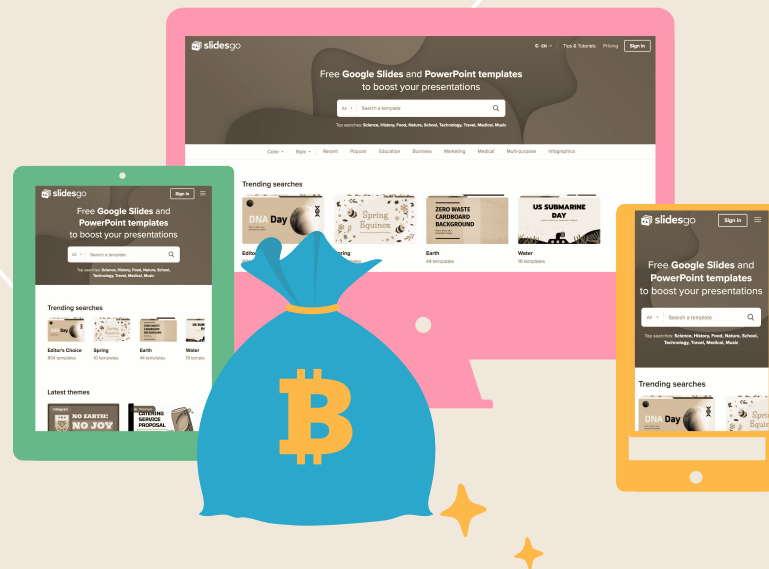
## Explore other Classifier models

Models that are self-learning like Random Forest, Support Vector Machine (SVM), XGBoost, could improve the prediction



## Extend beyond Reddit platform

Consider expanding to social media posts to capture broader web chatter





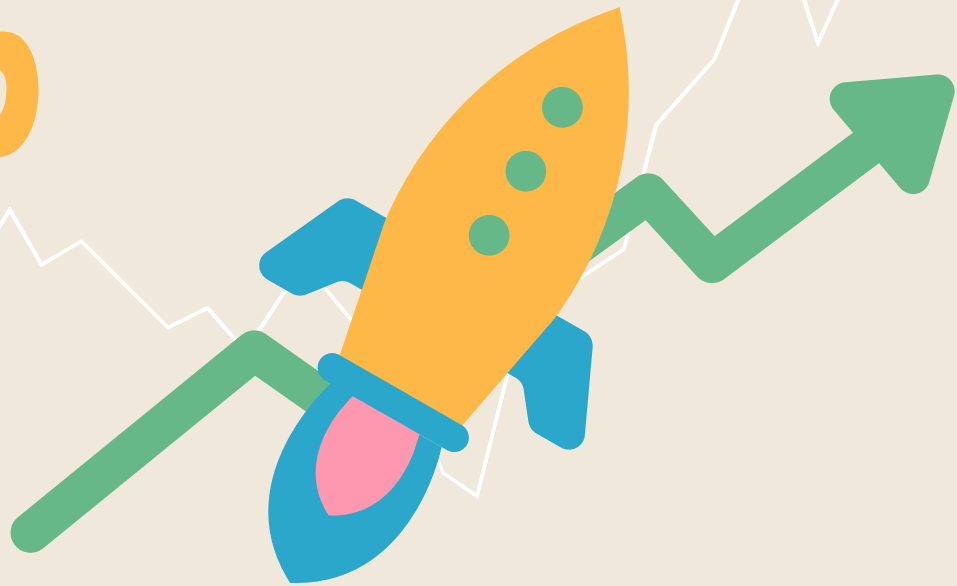
# THANKS!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for the attribution

Back up



# Predictor insights

## r/investing

r/investing Featured Words	Log(Odds)	Odds	Probability
stock	7.52	1839.26	99.95%
company	5.35	210.59	99.53%
share	3.95	51.84	98.11%
investing	3.57	35.46	97.26%
option	2.39	10.96	91.64%
investment	2.29	9.87	90.80%
fund	2.25	9.46	90.44%
investor	2.08	8.04	88.94%
market	2.06	7.84	88.69%
roth	2.01	7.43	88.13%

## r/CryptoCurrency

r/CryptoCurrency Featured Words	Log(Odds)	Odds	Probability
crypto	9.11	9039.02	99.99%
coin	5.94	378.48	99.74%
btc	4.77	117.95	99.16%
bitcoin	4.27	71.52	98.62%
wallet	3.69	40.18	97.57%
moon	3.35	28.39	96.60%
eth	3.17	23.82	95.97%
binance	3.05	21.07	95.47%
project	2.96	19.23	95.06%
cryptocurrency	2.77	15.92	94.09%



# Sneak Preview: XGBoost Pre-tuning



Train Accuracy

Test Accuracy

Sensitivity

Specificity

**TF-IDF** **XGBoost**

99.5%

93.4%

95.0%

90.8%