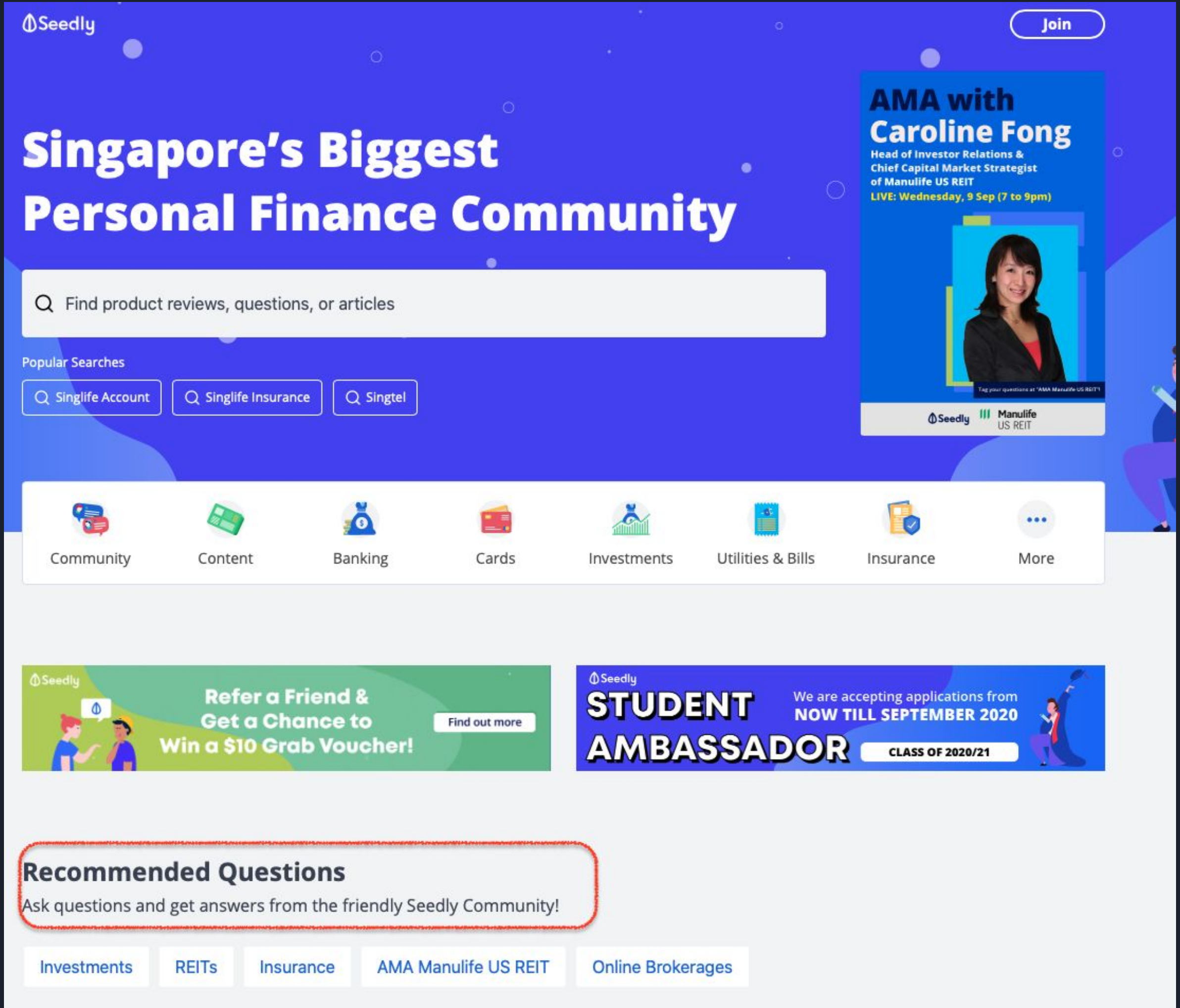
Seedly Text Classifier

Prakash N
Data Analyst @ Seedly.sg
github.com/Prakzter



Seedly

Join

Singapore's Biggest Personal Finance Community

Find product reviews, questions, or articles

Popular Searches

Singlife Account | Singlife Insurance | Singtel

AMA with Caroline Fong

Head of Investor Relations & Chief Capital Market Strategist of Manulife US REIT

LIVE: Wednesday, 9 Sep (7 to 9pm)

Tag your questions at "AMA Manulife US REIT"

Seedly | Manulife US REIT

Community | Content | Banking | Cards | Investments | Utilities & Bills | Insurance | More

Seedly

Refer a Friend & Get a Chance to Win a $10 Grab Voucher!

Find out more

Seedly

STUDENT AMBASSADOR

We are accepting applications from NOW TILL SEPTEMBER 2020

CLASS OF 2020/21

Recommended Questions

Ask questions and get answers from the friendly Seedly Community!

Investments | REITs | Insurance | AMA Manulife US REIT | Online Brokerages

Problem Statement

Questions coming into the discussion section needs to be tagged into the respective sections, as soon as readers post them up, and with minimal errors

Our Solution

Create a text classifier app that can distinguish a post between the top 2 categories:
Real Estate Investment and Stocks Investment

Building our corpus using relevant subreddits

Stocks Investment

r/Stocks
630 posts
50.5%

Real Estate Investment

r/RealEstate
618 posts
49.5%

* Initial 818 posts
Filter: < 30 char

EDA & Data cleaning

- Dropped null values
- Removed http links, non-word characters, digits
- Lemmatized across various POS tags:
  - Verb (v)
  - Noun (n)
  - Proverb (r)
  - Adjective (a)

Stop,words!

- Created custom stopword list on top of NLTK stopwords
  - Domain names
    - stocks , real estate, property, investment
  - Words in high frequency across both datasets
    - buy, just, new, know, year
  - Just plain gibberish
    - ve, amp, gt, don, http

ALL CLEANED UP

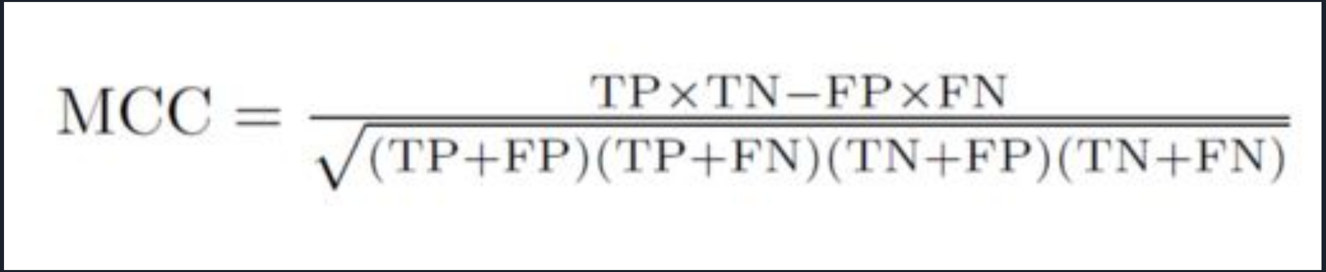
NOW LET'S MODEL!!

Pipeline it!

- Created 6 pipeline permutations using 2 transformers and 3 classifiers:
  - Transformers: CountVectoriser and TF-IDF Vectoriser
  - Classifiers: Logistic Regression, Multinomial Naive Bayes and Support Vector Classifier
- Ran model on default parameters then fine tuned hyperparameters using GridSearch

Key Metric used

Matthews Correlation Coefficient



$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

** Better than precision or recall as it factors in True Negatives as well to give a more holistic metric

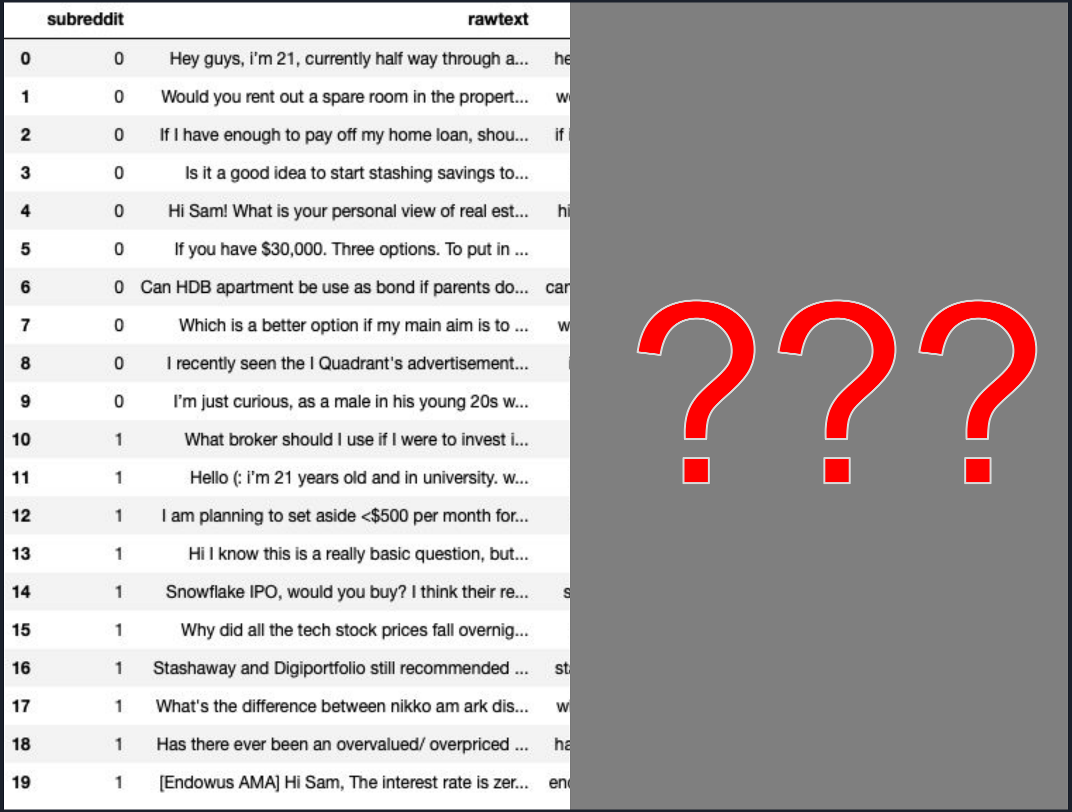
Consolidated model results using default hyperparameters

Transformer/Classifier	MCC	Misclassified RealEstate	Misclassified Stocks
CV & Log Reg	0.8883	9	5
TF & Log Reg	0.9050	9	3
CV & Multinomial NB	0.8974	3	10
TF & Multinomial NB	**0.9211**	**2**	**8**
CV & SVC	0.8558	18	1
TF & SVC	0.8974	10	3

Consolidated model results after tuning hyperparameters

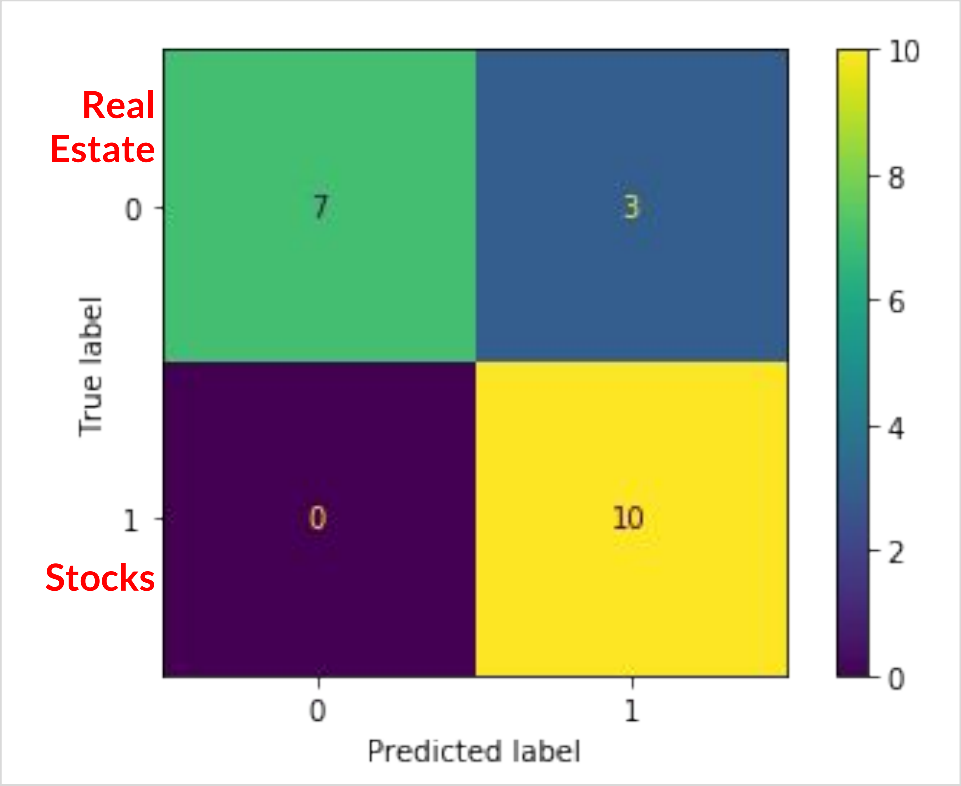
Transformer/Classifier	MCC	Misclassified RealEstate	Misclassified Stocks
CV & Log Reg	0.8807	10	5
TF & Log Reg	0.8983	11	2
CV & Multinomial NB	0.9120	5	6
TF & Multinomial NB	**0.9280**	**5**	**4**
CV & SVC	0.8702	16	1
TF & SVC	0.9122	7	4

Now let’s start the real ‘test’ with actual Seedly.sg posts



	subreddit	rawtext	
0	0	Hey guys, i’m 21, currently half way through a...	he
1	0	Would you rent out a spare room in the propert...	w
2	0	If I have enough to pay off my home loan, shou...	if i
3	0	Is it a good idea to start stashing savings to...	
4	0	Hi Sam! What is your personal view of real est...	hi
5	0	If you have $30,000. Three options. To put in ...	
6	0	Can HDB apartment be use as bond if parents do...	car
7	0	Which is a better option if my main aim is to ...	w
8	0	I recently seen the I Quadrant's advertisement...	
9	0	I’m just curious, as a male in his young 20s w...	
10	1	What broker should I use if I were to invest i...	
11	1	Hello (: i’m 21 years old and in university. w...	
12	1	I am planning to set aside <$500 per month for...	
13	1	Hi I know this is a really basic question, but...	
14	1	Snowflake IPO, would you buy? I think their re...	s
15	1	Why did all the tech stock prices fall overnig...	
16	1	Stashaway and Digiportfolio still recommended ...	st
17	1	What's the difference between nikko am ark dis...	w
18	1	Has there ever been an overvalued/ overpriced ...	ha
19	1	[Endowus AMA] Hi Sam, The interest rate is zer...	en

Stocks were great but.....



Real Estate

Stocks

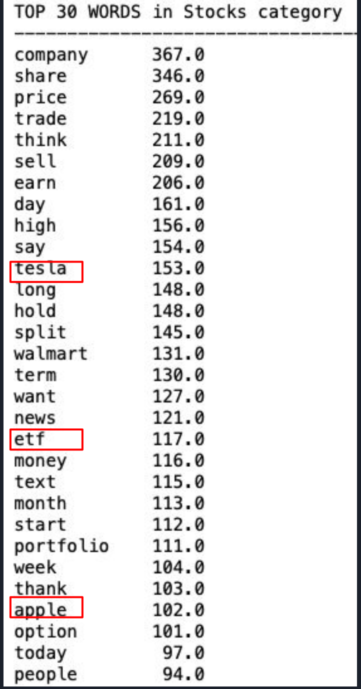
All 10 stock posts were correctly classified but there was some noise with the real estate posts, with 3 of them classified as being related to Stocks.

MCC for blind test: 0.7338

MCC from training data: 0.9280

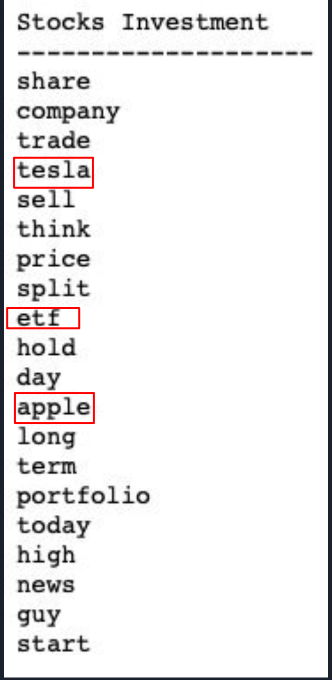
So which words matter and why...

Word Frequency



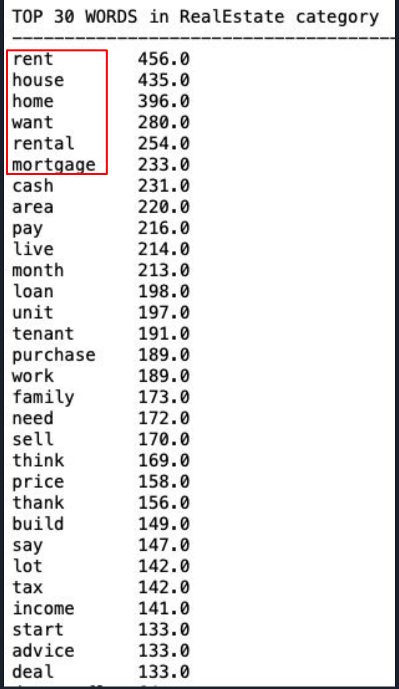
```
TOP 30 WORDS in Stocks category
----------------------------------
company        367.0
share          346.0
price          269.0
trade          219.0
think          211.0
sell           209.0
earn           206.0
day            161.0
high           156.0
say            154.0
tesla          153.0
long           148.0
hold           148.0
split          145.0
walmart        131.0
term           130.0
want           127.0
news           121.0
etf            117.0
money          116.0
text           115.0
month          113.0
start          112.0
portfolio      111.0
week           104.0
thank          103.0
apple          102.0
option         101.0
today           97.0
people          94.0
```

Deterministic words



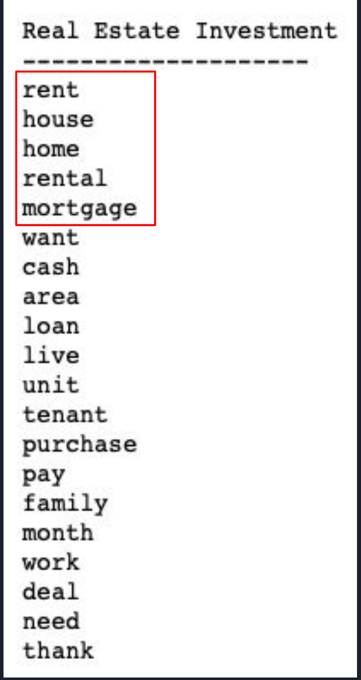
```
Stocks Investment
--------------------
share
company
trade
tesla
sell
think
price
split
etf
hold
day
apple
long
term
portfolio
today
high
news
guy
start
```

Word Frequency



```
TOP 30 WORDS in RealEstate category
------------------------------------
rent           456.0
house          435.0
home           396.0
want           280.0
rental         254.0
mortgage       233.0
cash           231.0
area           220.0
pay            216.0
live           214.0
month          213.0
loan           198.0
unit           197.0
tenant         191.0
purchase       189.0
work           189.0
family         173.0
need           172.0
sell           170.0
think          169.0
price          158.0
thank          156.0
build          149.0
say            147.0
lot            142.0
tax            142.0
income         141.0
start          133.0
advice         133.0
deal           133.0
```

Deterministic words



```
Real Estate Investment
--------------------
rent
house
home
rental
mortgage
want
cash
area
loan
live
unit
tenant
purchase
pay
family
month
work
deal
need
thank
```

Closer look behind the distribution...

The TF-IDF vectoriser was able to give words that were mentioned way fewer more importance in predicting the Stocks class.

However, on the Real Estate front the words with a lower frequency weren’t able to clearly distinguish it from the latter class.

Furthermore, as compared to stocks, Real Estate investment posts within the singapore context would include specific terms like “CPF” or “HDB” or “Condo”, which adding those terms during training would aid in improving the model’s accuracy

DEMO TIME!!!

Let’s test out the app

https://seedy-classifier.herokuapp.com