DSI 18 Project 3

Subreddit Classification

By: Benjamin Lee

Introduction - Problem Statement

As a Data Analyst in a consultation firm (Data Insights Pte Ltd), the firm was recently **engaged by Sibeh Rich Bank** to **predict** whether if a particular post is **savings related** as the bank wants to better engage their customers on savings based on ground sentiments. A **classification model** will be devised and evaluated based on accuracy score.

Introduction - Summary

In this study two subreddit (r/SavingMoney and r/Investing) were examined. Both topics **revolves around** the idea of preparing for the future with the emphasis of **money**. However, while being **similar in nature** where money is the center of gravity, the utilization is **different in concept** as one emphasize the importance of saving while another shares the idea of growing wealth through investment. The goal of this project is therefore to try and figure out how distinct these concepts are from one to another.

End Product

S/N	Model	Vectorizer	Feature	Train	Test	TN	FP	FN	TP
1	Naive Bayes (Multinomial NB)	cvec	title	0.9089	0.8192	131	20	27	82
2	Naive Bayes (Multinomial NB)	cvec	post	0.9669	0.9462	141	10	4	105
3	Naive Bayes (Multinomial NB)	cvec	combine	0.9669	0.9462	141	10	4	105
4	Naive Bayes (Multinomial NB)	tvec	title	0.9172	0.8115	129	22	27	82
5	**Naive Bayes (Multinomial NB)**	**tvec**	**post**	**0.9867**	**0.95**	**144**	**7**	**6**	**103**
6	Naive Bayes (Multinomial NB)	tvec	combine	0.9867	0.95	144	7	6	103
7	K Nearest Neighbors (KNN)	cvec	title	1.0	0.5615	40	111	3	106
8	K Nearest Neighbors (KNN)	cvec	post	1.0	0.5923	47	104	2	107
9	K Nearest Neighbors (KNN)	cvec	combine	1.0	0.5923	47	104	2	107
10	K Nearest Neighbors (KNN)	tvec	title	0.8294	0.7730	117	34	25	84
11	K Nearest Neighbors (KNN)	tvec	post	1.0	0.9346	142	9	7	102
12	K Nearest Neighbors (KNN)	tvec	combine	1.0	0.9346	142	9	7	102
13	Logistic Regression (LR)	cvec	title	0.9354	0.8153	137	14	34	75
14	Logistic Regression (LR)	cvec	post	0.9933	0.9230	136	15	5	104
15	Logistic Regression (LR)	cvec	combine	0.9933	0.9230	136	15	5	104
16	Logistic Regression (LR)	tvec	title	0.9586	0.8038	133	18	33	76
17	Logistic Regression (LR)	tvec	post	0.9983	0.9153	140	11	11	98
18	Logistic Regression (LR)	tvec	combine	0.9983	0.9153	140	11	11	98
19	Random Forest (RF)	cvec	title	0.9668	0.7615	119	32	32	77
20	Random Forest (RF)	cvec	post	1.0	0.9230	134	17	7	102
21	Random Forest (RF)	cvec	combine	1.0	0.8923	135	16	6	103
22	Random Forest (RF)	tvec	title	0.9668	0.7807	115	36	20	89
23	Random Forest (RF)	tvec	post	1.0	0.8961	134	17	6	103
24	Random Forest (RF)	tvec	combine	1.0	0.9153	134	17	6	103

Steps

- Create function
- Scrape subreddit
- Feature Selection

Scrapping

- Model Tuning (NB, KNN, LR, RF) with vectorizer (cvec & Tfid)
- Features (title, post, combined)

Model Tuning

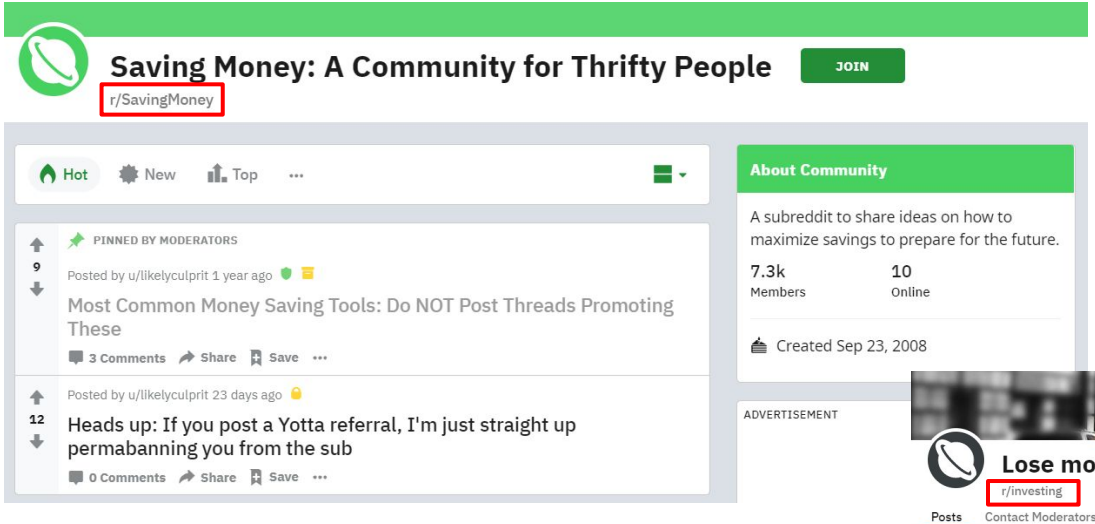
EDA and Data Cleaning

- Top 50 word(s) in title & body
- Preprocessing (remove html, keywords, numerical values and stop words)
- Merging of Data

Model Selection

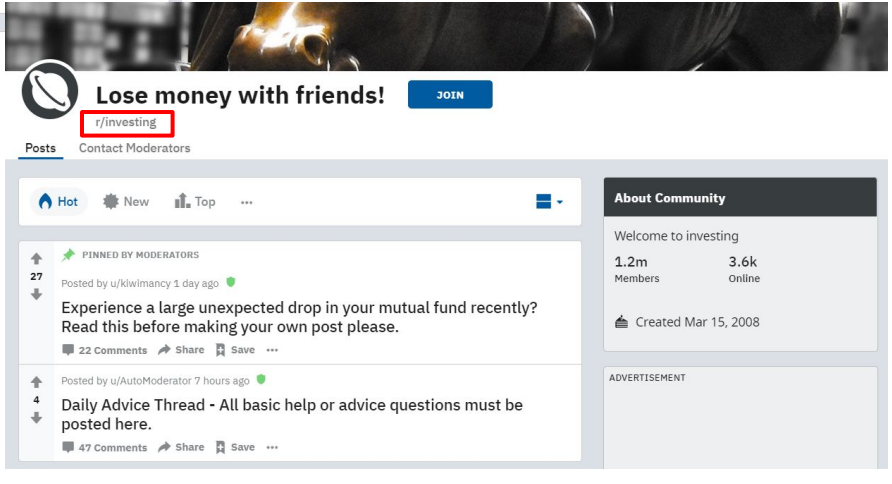
- Comparing accuracy of all models between train and test.
- Find best model

Scraping



- Credit cards
- Savings Account
- Savings Challenge

- Stock Market
- Investment Strategy
- Investment platform/news



Scraping - Data (r/SavingMoney)

```
saving.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 748 entries, 0 to 747
Columns: 107 entries, approved_at_utc to media_metadata
dtypes: bool(26), float64(30), int64(10), object(41)
memory usage: 492.5+ KB
```

```
# Based on the scrapping, some post was scrapped twice.
# Apart from that, may be advertisement so there will repost of with similar title or body.
save_dup = saving[(saving.duplicated(subset = ['title']) == True) & (saving.duplicated(subset = ['selftext']) == True)]\
['title'].value_counts()

print(len(save_dup))
save_dup.head()
```

```
59

My opinion on Yotta savings                                  1
How to Save Money to Have More Money to Spend                1
The Ultimate Guide To Saving Money In Your Home              1
Welp, I love saving money So I had to make a video about it  1
ISO advice                                                   1
Name: title, dtype: int64
```

```
saving.drop_duplicates(subset=['title', 'selftext'], inplace = True)
```

```
saving.shape
```

```
(689, 107)
```

Scrapped: 748

Duplicates: 59

Entries: 689

Scraping - Data (r/Investing)

```
investment.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 745 entries, 0 to 744
Columns: 103 entries, approved_at_utc to author_cakeday
dtypes: bool(26), float64(31), int64(10), object(36)
memory usage: 467.2+ KB
```

```
#checking for duplicates
in_dup = investment[(investment.duplicated(subset = ['title']) == True) & (investment.duplicated(subset = ['selftext']) == T
['title'].value_counts()

print(len(in_dup))
in_dup.head()
```

```
223

Daily Advice Thread - All basic help or advice questions must be posted here.                                    18
Anti-capitalist forms of investing?                                                                                1
My elderly parents wants to put some of their savings into the stock market rather than just gaining interest in banks    1
Thoughts on £CEY Centamin                                                                                          1
Realistically what is ARKK 5-year and 10-year realistic return?                                                    1
Name: title, dtype: int64
```

```
investment.drop_duplicates(subset=['title', 'selftext'], inplace = True)
```

```
investment.shape
```

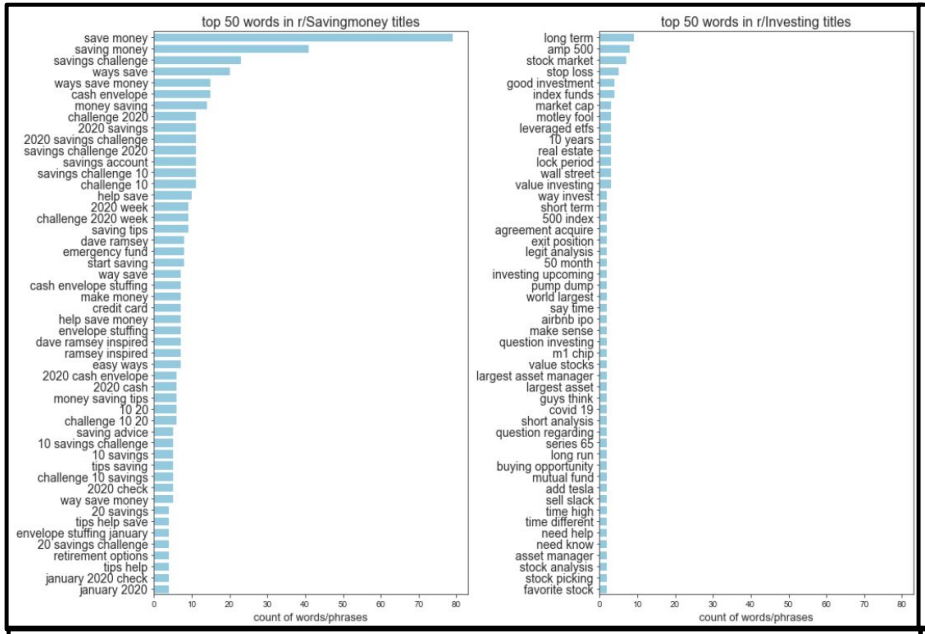
```
(505, 103)
```

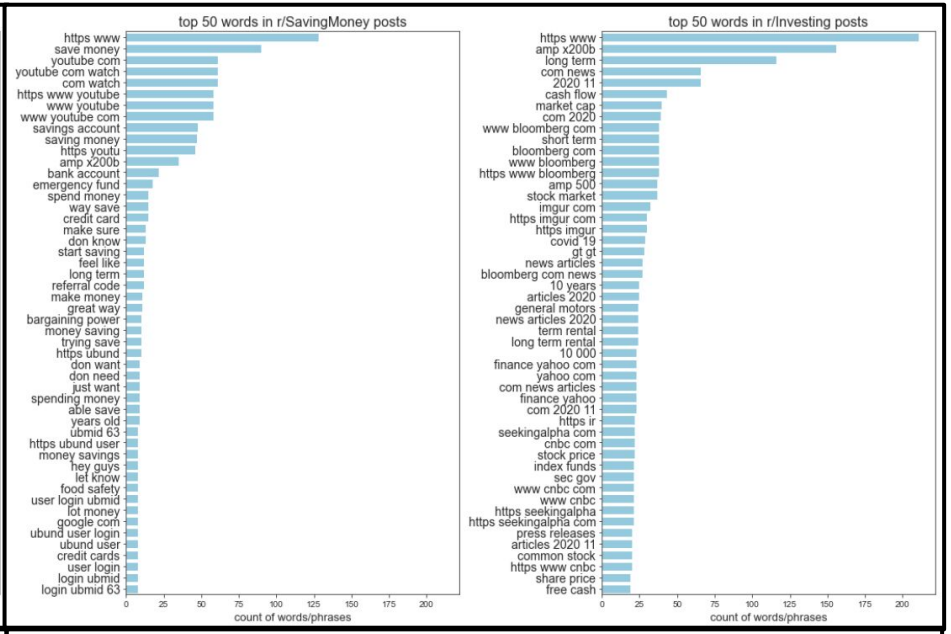
Scrapped: 745

Duplicates: 240

Entries: 505

EDA - Top Word(s)





SavingMoney: save money, saving money

Investing: long term, amp500, stock market

SavingMoney: https www, save money

Investing: https www, amp x200b, long term

Preprocessing

```
saving[saving['distinguished'] == 'moderator']
```

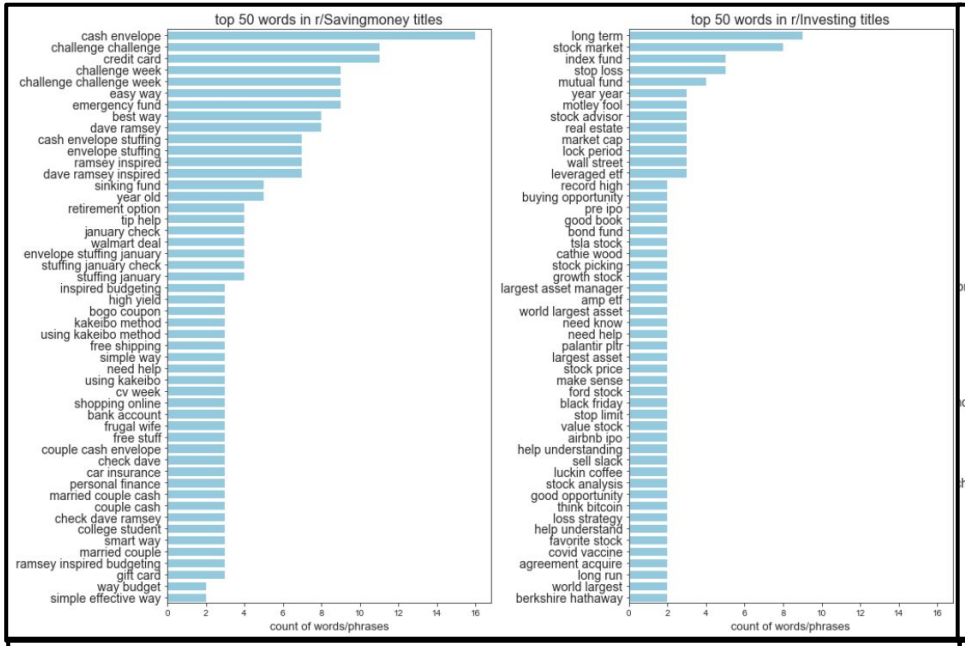
	subreddit	id	distinguished	title	selftext
0	SavingMoney	calpl0	moderator	Most Common Money Saving Tools: Do NOT Post Th...	In order to minimize the constant referral pos...

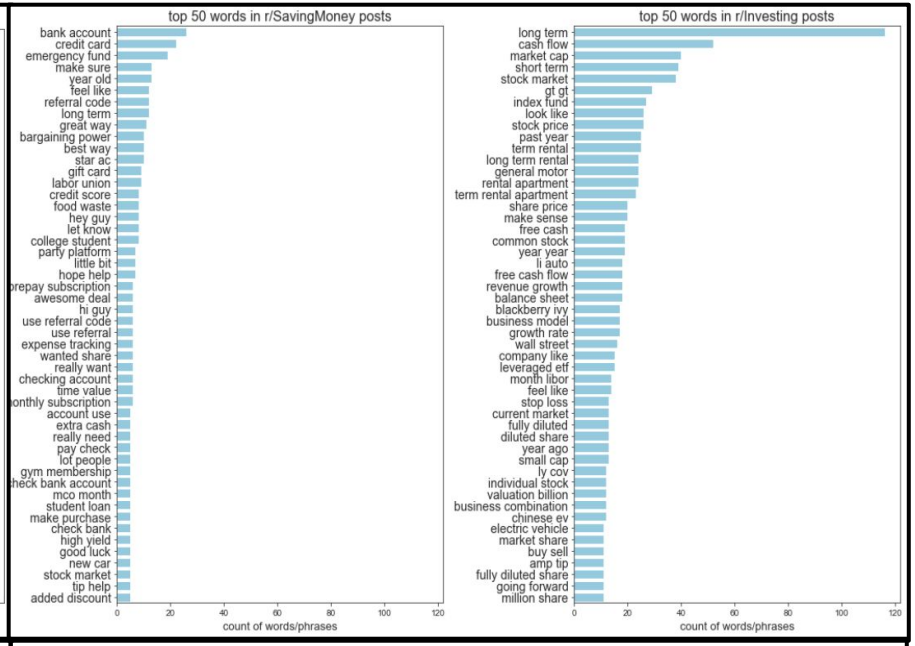
```
investment[investment['distinguished'] == 'moderator']
```

	subreddit	id	distinguished	title	selftext
1	investing	k4jq8t	moderator	Daily Advice Thread - All basic help or advice...	If your question is "I have $10,000, what do I...

- Remove moderator post
- Removed html, numerical values, stop words
- Lemmatize

EDA - Processed Top Word(s)





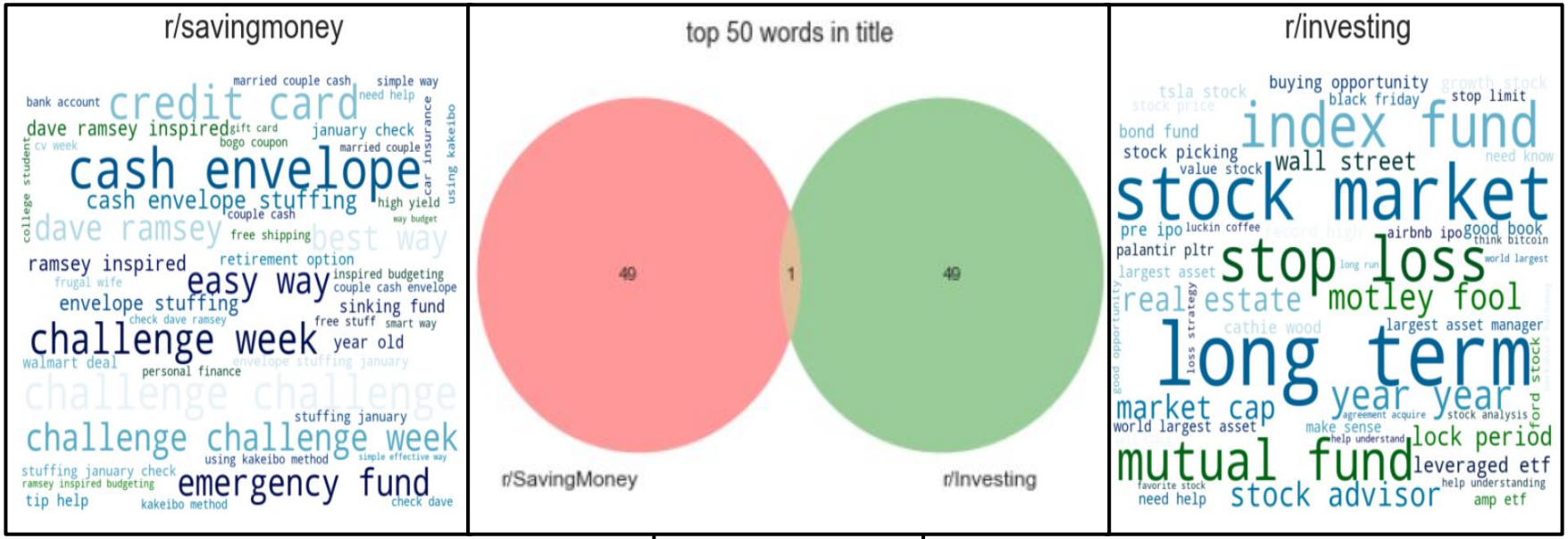
SavingMoney: cash envelope, challenge

Investing: long term, stock market, index fund

SavingMoney: bank account, credit card

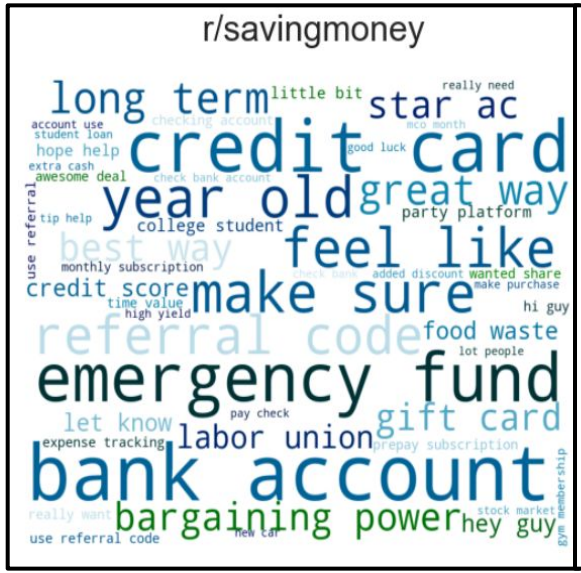
Investing: long term, cash flow

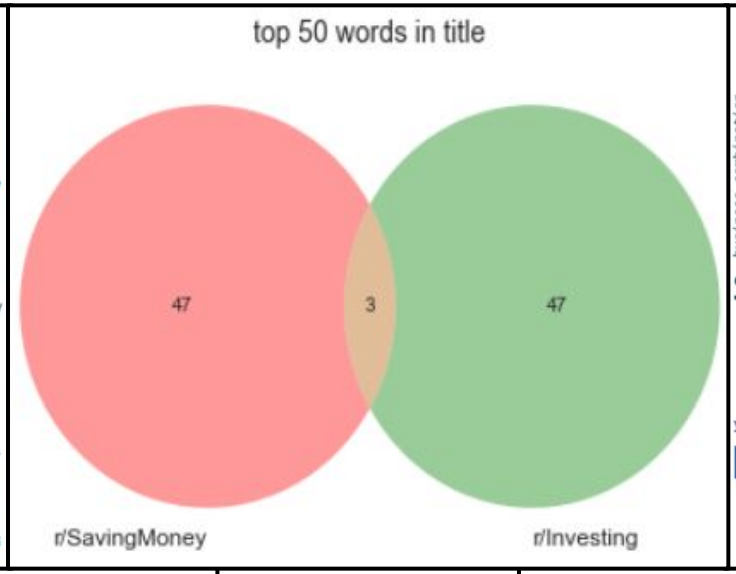
EDA - Processed Top Word(s) Title



1. Best Way

EDA - Processed Top Word(s) Title

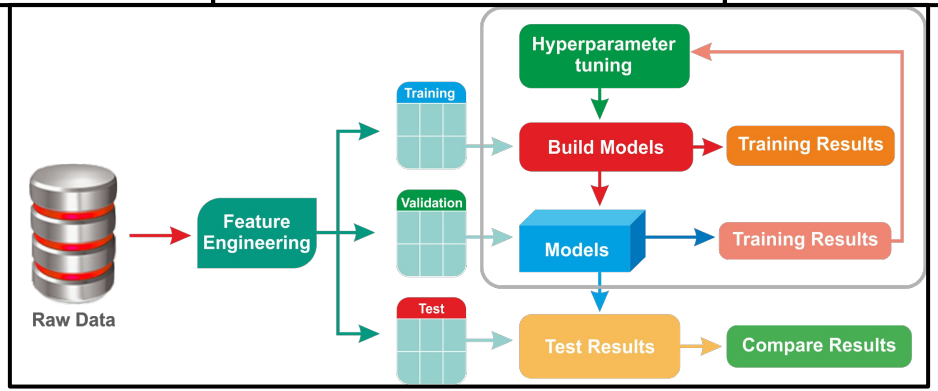




1. Year ago
2. Feel like
3. Long term

Model Tuning

Model:	Vectorizer:	Features:
• Naive Bayes • KNN • Logistic Regression • Random Forest	• Count Vectorizer • Tdif Vectorizer	• Title • Post • Combine (Title + Post)



Model Recommendation

S/N	Model	Vectorizer	Feature	Train	Test	TN	FP	FN	TP
1	Naive Bayes (Multinomial NB)	cvec	title	0.9089	0.8192	131	20	27	82
2	Naive Bayes (Multinomial NB)	cvec	post	0.9669	0.9462	141	10	4	105
3	Naive Bayes (Multinomial NB)	cvec	combine	0.9669	0.9462	141	10	4	105
4	Naive Bayes (Multinomial NB)	tvec	title	0.9172	0.8115	129	22	27	82
5	**Naive Bayes (Multinomial NB)**	**tvec**	**post**	**0.9867**	**0.95**	**144**	**7**	**6**	**103**
6	Naive Bayes (Multinomial NB)	tvec	combine	0.9867	0.95	144	7	6	103
7	K Nearest Neighbors (KNN)	cvec	title	1.0	0.5615	40	111	3	106
8	K Nearest Neighbors (KNN)	cvec	post	1.0	0.5923	47	104	2	107
9	K Nearest Neighbors (KNN)	cvec	combine	1.0	0.5923	47	104	2	107
10	K Nearest Neighbors (KNN)	tvec	title	0.8294	0.7730	117	34	25	84
11	K Nearest Neighbors (KNN)	tvec	post	1.0	0.9346	142	9	7	102
12	K Nearest Neighbors (KNN)	tvec	combine	1.0	0.9346	142	9	7	102
13	Logistic Regression (LR)	cvec	title	0.9354	0.8153	137	14	34	75
14	Logistic Regression (LR)	cvec	post	0.9933	0.9230	136	15	5	104
15	Logistic Regression (LR)	cvec	combine	0.9933	0.9230	136	15	5	104
16	Logistic Regression (LR)	tvec	title	0.9586	0.8038	133	18	33	76
17	Logistic Regression (LR)	tvec	post	0.9983	0.9153	140	11	11	98
18	Logistic Regression (LR)	tvec	combine	0.9983	0.9153	140	11	11	98
19	Random Forest (RF)	cvec	title	0.9668	0.7615	119	32	32	77
20	Random Forest (RF)	cvec	post	1.0	0.9230	134	17	7	102
21	Random Forest (RF)	cvec	combine	1.0	0.8923	135	16	6	103
22	Random Forest (RF)	tvec	title	0.9668	0.7807	115	36	20	89
23	Random Forest (RF)	tvec	post	1.0	0.8961	134	17	6	103
24	Random Forest (RF)	tvec	combine	1.0	0.9153	134	17	6	103

The Naive Bayes and TdifVectorizer model is able predict with an accuracy of 95%

Among all the features, post seems to give the best results in terms of accuracy and computing time

Best estimator: max_df 0.9, max_features=2000, min_df=3, ngram_range(1, 3), stop_words='english, alpha=0.8

Top Correlated Words & Predictions

Top Correlated Words		
S/N	**Saving**	**Investing**
1	money	riskier
2	account	entire
3	make	entry
4	month	environment
5	way	relevant
6	week	established
7	help	estate
8	know	estimated
9	like	enterprise
10	tip	etf

Predictions

	0	1	text
515	0.992489	0.007511	CONTEXT: \nAPPULSE Corp. is a centrifuge manu...
454	0.991243	0.008757	Palantir Technologies Inc. posted its best wee...
673	0.989508	0.010492	Hello all,\n\nI decided I would do a comparabl...
453	0.988205	0.011795	\n\nWall Street is bracing for Tesla Inc's (N...
368	0.987977	0.012023	\nHello all,\n\nI want to highlight a few poin...

	0	1	text
251	0.015176	0.984824	I'm a student and was working a part time job ...
308	0.013565	0.986435	I recently found out about this app called get...
45	0.012238	0.987762	Hi everyone, I'm 21 with a decent paying job f...
77	0.011207	0.988793	If you are anything like me (or the average hu...
188	0.008435	0.991565	Saving up to half your income on one salary re...

Confusion Matrix

Class	Actual Value = 1	Actual Value = 0
Predicted = 1	True Positive 103	False Positive 7
Predicted = 0	False Negative 6	True Negative 144

Misclassification Rate: (FP+FN)/Total = 0.05

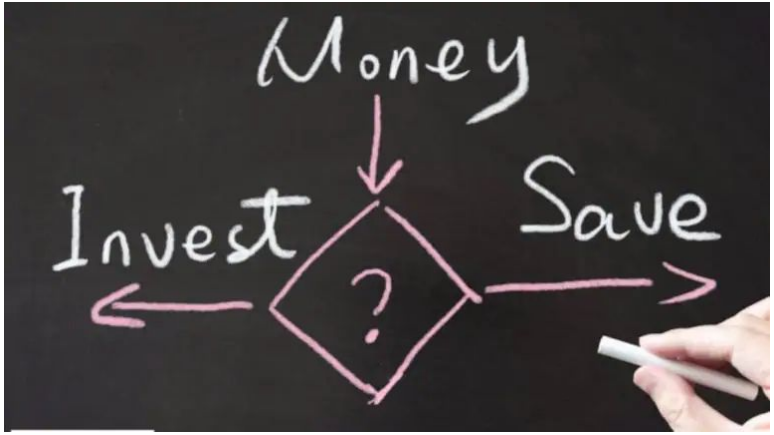
Conclusion

1. In general, both subreddit are quite distinct. Both revolves around money, but concept in utilizing differs. Hence, predicting between saving and investing generally gives a higher accuracy.

2. The Naive Bayes and TdifVectorizer model has the highest accuracy of 95%. ‘Post’ feature give the best results in terms of accuracy and computing time.

3. Improvement:
   a. Reduce misclassification rate by looking into the False Positive and False Negative
   b. Gather more data through scrapping of other subreddit related to Savings and Investment to improve the accuracy of the result

Thank you~



Top Correlated Word