Natural Language Processing on reddit posts

Edward Hiah

Content

- Problem Statement
- Scraping API and Cleaning
- Exploratory Data Analysis
- Modeling Process
- Evaluation
- Summary

Problem Statement

Reddit is a network of communities where users can share news content or comment on each other posting.

As there are over 1.5 million subreddits on reddit. Given that users can post on anything on any of the subreddits, Moderator will have difficulties to visually going into all posting to ensure that postings are relevant to the respective subreddit.

The objective of this project is to use machine learning to create a classification model and see how well that can it distinguish the postings and re-classify the posts to the respective subreddit or remove it. As such I picked two closely-related subreddits for the challenge.

cryptocurrency and bitcoin

The Data Scrapping

- Bitcoin : 9998 rows 80 variables / columns

4	I happen to be a crypto enthusiast and I've be...	Bitcoin
6	YungSnxw's\nCrypto Exchange Price's\n\n100$ - ...	Bitcoin
12	For 5 years, I've been dabbling in sum purchas...	Bitcoin
13	Figured I share a couple of observations regar...	Bitcoin
19	Sentiment on Wall Street is that the bear mark...	Bitcoin
...	...	...
9985	Some people are saying bitcoin is a store of e...	Bitcoin

- CryptoCurrency : 9955 rows 82 variables / columns

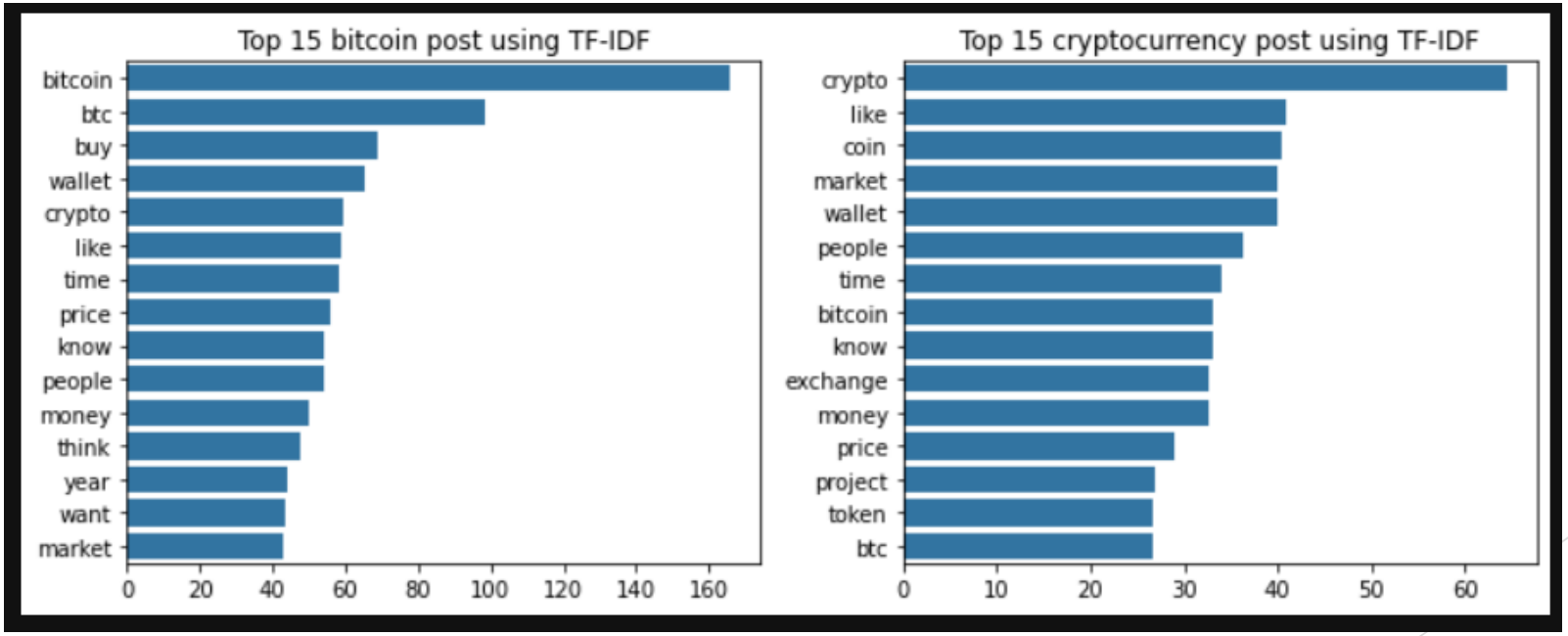
1	Hi, I'm a young guy in his 20's and I think I ...	CryptoCurrency
3	I have Coinbase but was fortunate enough to g...	CryptoCurrency
8	my friend owes me money and has money in cryp...	CryptoCurrency
9	\nAre you bullish or bearish in the coming wee...	CryptoCurrency
12	**Welcome to the Daily General Discussion thre...	CryptoCurrency
...	...	...
9936	So I've decided that I need to take the storag...	CryptoCurrency
9938	My PNL and ROE are offset?\n\nSo i have been b...	CryptoCurrency
9940	DCA = dollar cost averaging. That means buying...	CryptoCurrency

Cleaning

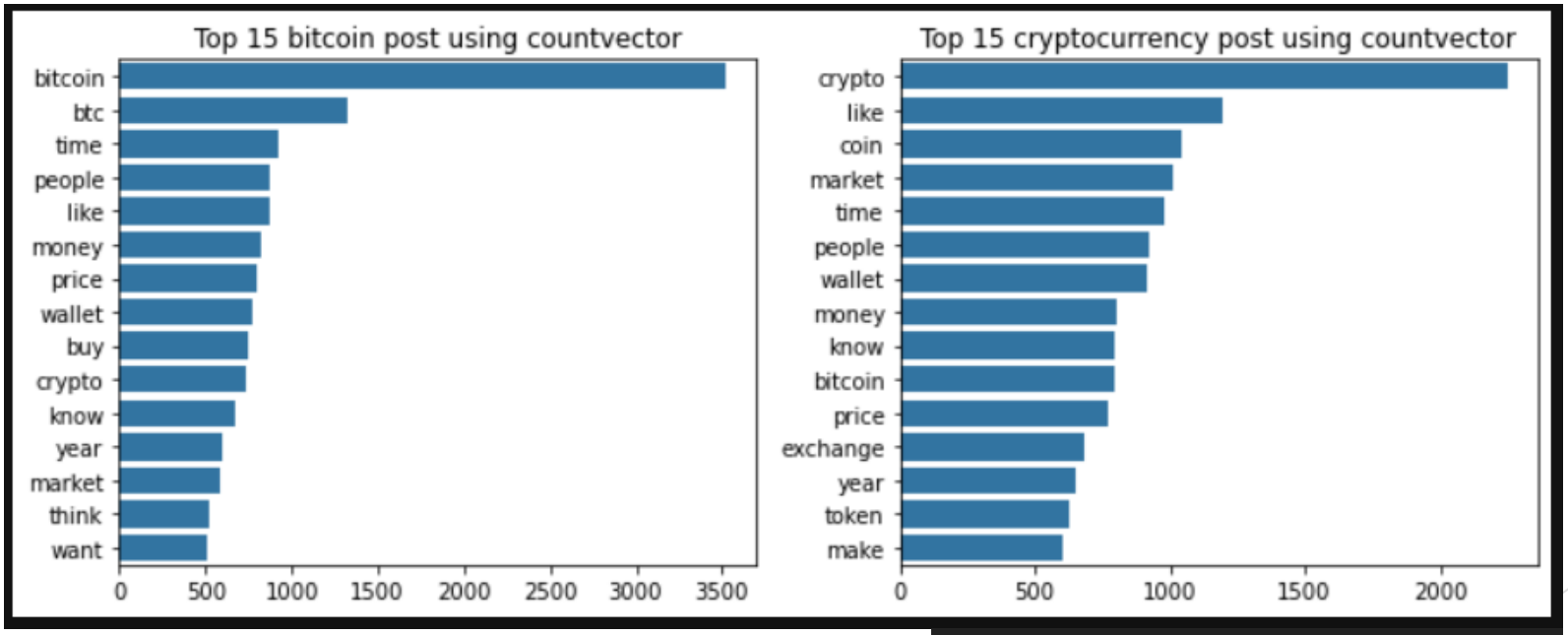
- Removing unneeded columns
- Removing duplicates posts
- Removing [deleted] & [removed]
- Reformatting subreddit into binary post
- Drop the null post
- Cleaning the self text

Exploratory Data Analysis

- 15 Top words between the two subreddits

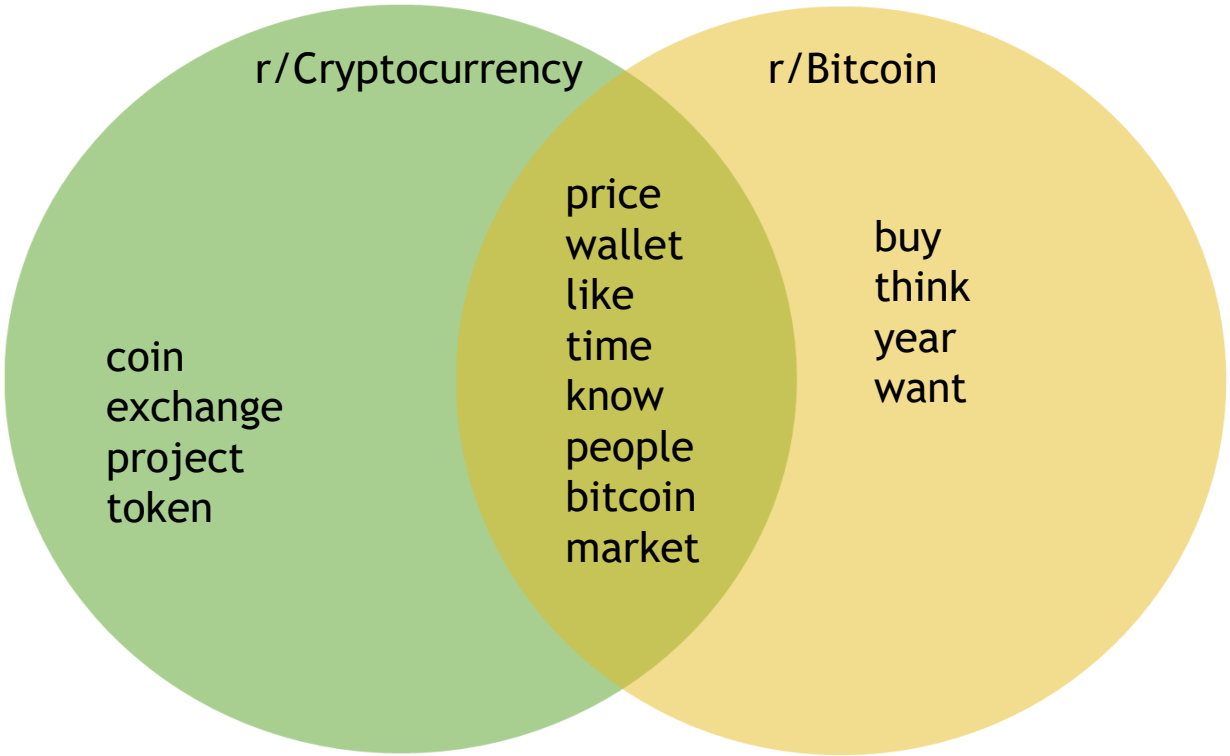


Exploratory Data Analysis




```
'bitcoin','crypto','btc'
```

Common words in the top 15 words



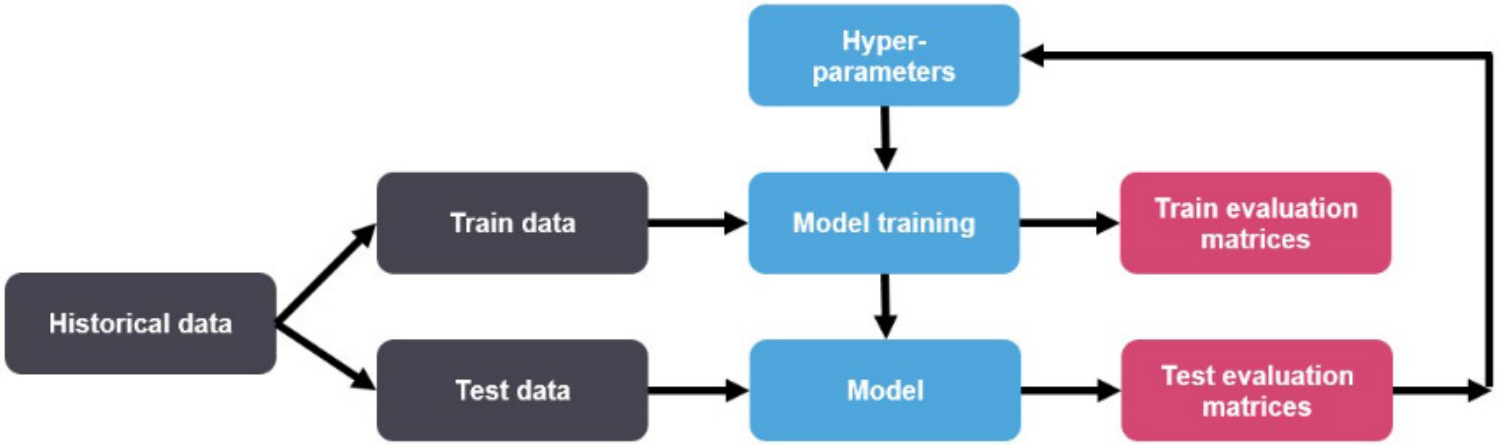
Modeling

- Naïve Bayes
- Naive Bayes **predict the tag of a text**. They calculate the probability of each tag for a given text and then output the tag with the highest one

- Logistic Regression
- Logistic regression is the base- line supervised machine learning algorithm for classification.
- As a Sentiment classification
- Has close relation with neural networks

Hyperparameter

```
#hyperparameters:
params = {
    'cvec__max_features': [1_000, 2_000, 3_000],
    'cvec__min_df': [2, 3],
    'cvec__max_df': [.9, .95],
    'cvec__ngram_range': [(1,1), (1,2)],
    'lr__C' : [0.1,1.0,10],
}
gs = GridSearchCV(pipe,
                  param_grid=params,
                  cv=5)
# Fit GridSearch to training data.
gs.fit(X_train, y_train)
```



- Guesswork is necessary to specify the min and max values for each hyperparameter.
- Searching for the best hyper-parameter can be tedious, hence search algorithms like grid search.

After the tuning

#	Vectorizer	Model	Best Score	Train Score	Test Score (Accuracy)	FN
1	CountVectorizer	Multinomial Naïve Bayes	0.7994	0.8392	0.8021	0.8398
2	TfidfVectorizer	Multinomial Naïve Bayes	0.8296	0.8842	0.8296	0.9089
3	CountVectorizer	Logistic Regression	0.8845	0.9973	0.877	0.9089
4	TfidfVectorizer	Logistic Regression	0.8837	0.9838	0.8735	0.906

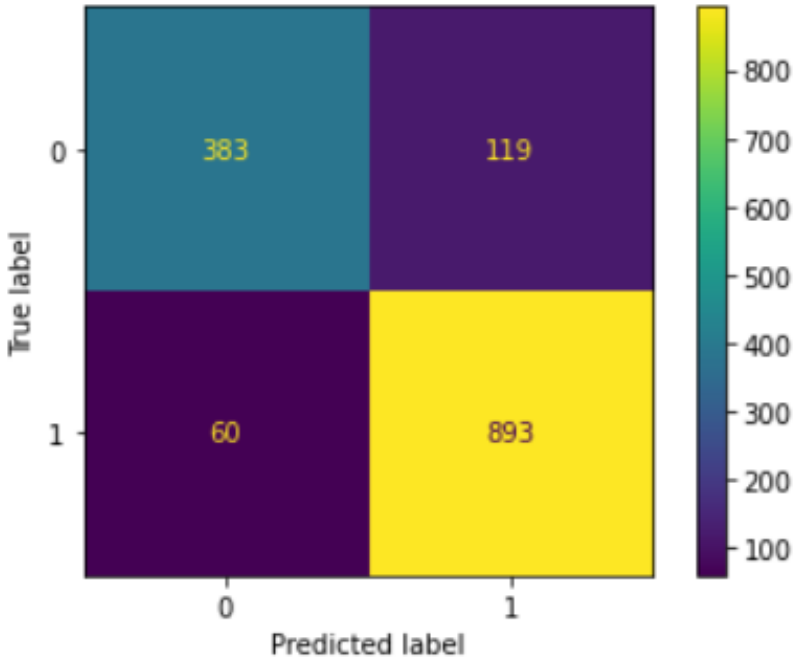
Generalization test error of learning algorithms has two main components:

- Bias: error due to simplifying model assumptions
- Variance: error due to randomness of the training set

Final Model

- **Model Selection: Logistic Regression with CountVectorizer**
- The model exceed the baseline 0.654.
  - True Negatives: 383
  - False Positives: 119
  - False Negatives: 60
  - True Positives: 893

```
Accuracy: 0.877
Misclassification rate: 0.123
Precision: 0.8824
Recall: 0.937
Specificity: 0.7629
f1 score: 0.9089
```



Conclusion & Recommendation

With the prediction we can greatly minimise the need for moderator to manually re-classify the post.

To look into using Random forest modelling :

Does not suffer from overfitting

Get relative feature imporatance