Project 3 - Web APIs and NLP Classification

Marc
13 Aug 22

Executive Summary

1. Background information
2. Data Gathering
3. Data Cleaning
4. Exploratory Data Analysis
5. Modelling
6. Results & conclusion

Background Information

Data has been gathered from Reddit. A social news aggregation, content rating, and discussion website. Registered users submit content to the site such as links, text posts, images, and videos, which are then voted up or down by other members.

Method of scraping for data

- We'll be using Pushshift's API to scrap subreddits. Pushshift's API provides enhanced functionality and search capabilities for searching Reddit comments and submissions.
- Documentation may be found at https://github.com/pushshift/api

Outcome

Using NLP, we'll train a classification model to predict which subreddit a given post came from.

Data Gathering

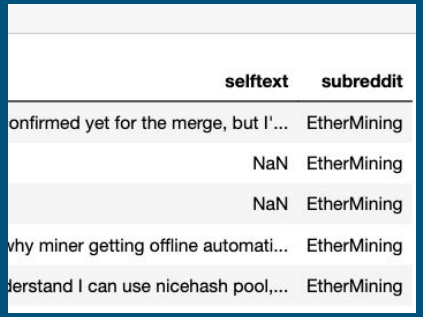
Two subreddits have been chosen. They are:

1. Cryptocurrency

2. EtherMining

Scraped 2,000 post per subreddit with the use of Pushshift's API

Data Cleaning



	selftext	subreddit
	onfirmed yet for the merge, but I'...	EtherMining
	NaN	EtherMining
	NaN	EtherMining
	why miner getting offline automati...	EtherMining
	derstand I can use nicehash pool,...	EtherMining

Replaced Nan with empty string

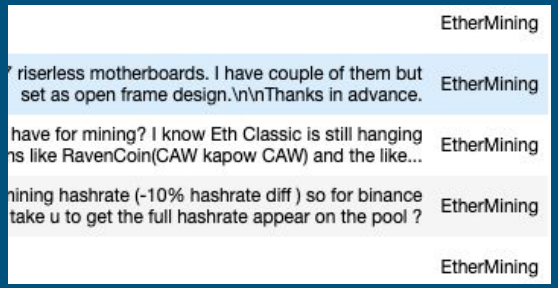
	title	selftext	subreddit
	t's your exit strategy?	[removed]	CryptoCurrency
	Your new version 2 h...	[removed]	CryptoCurrency
	t is your exit strategy?	[removed]	CryptoCurrency
	are offered a way ac...		CryptoCurrency
	forex trading robot		CryptoCurrency

Replaced removed selftext with empty string

	title	selftext	subreddit
	is it dead	[deleted]	EtherMining
	merge event. Eth...	[deleted]	CryptoCurrency
	he law is on our s...	[deleted]	CryptoCurrency
	Up Being "Rug Pull"	[deleted]	CryptoCurrency
	es Over His Pro-D...	[deleted]	CryptoCurrency

Replaced deleted selftext with empty string

Data Cleaning

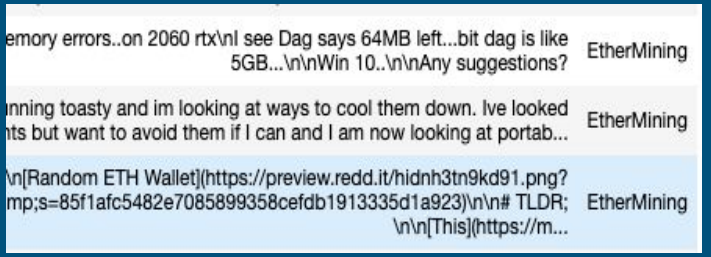


	EtherMining
riserless motherboards. I have couple of them but set as open frame design.\n\nThanks in advance.	EtherMining
have for mining? I know Eth Classic is still hanging ns like RavenCoin(CAW kapow CAW) and the like...	EtherMining
ining hashrate (-10% hashrate diff) so for binance take u to get the full hashrate appear on the pool ?	EtherMining
	EtherMining

Remove newline characters (\n)

	EtherMining
Lettme explain, i bougt these 2 3080s LHR at the , etc. they had really bad temps on their vrams...	EtherMining
B;\n\nhttps://preview.redd.it/nisj1uupk9d91.png?ba6c1c230507e6e14cbd0a8fa9c83882ab7e9fd1	EtherMining
	EtherMining
d to mine ETC? Vitalik Buterin did Shout Out ETC Ethereum "non-altered" blockchain. Was he se...	EtherMining

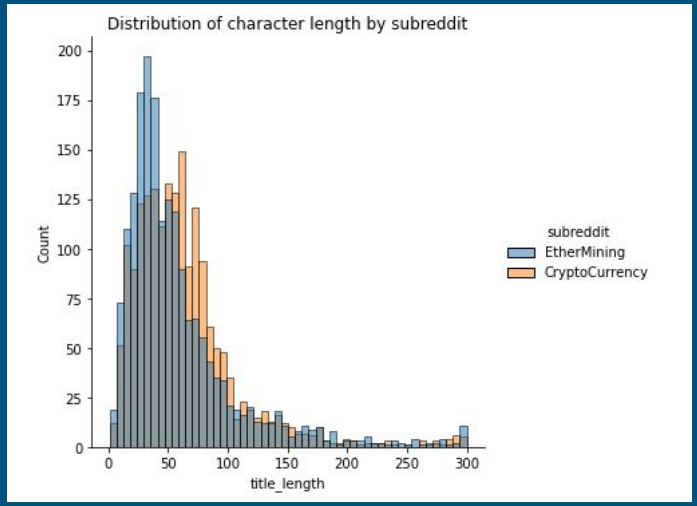
Remove URL links using RegEx

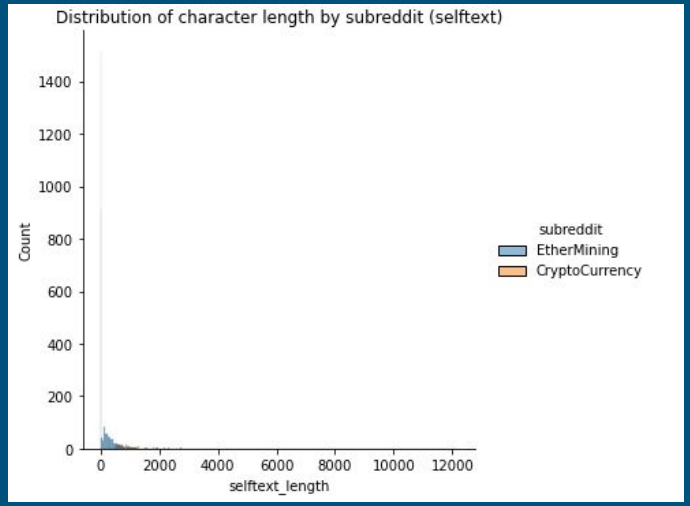


emory errors..on 2060 rtx\nI see Dag says 64MB left...bit dag is like 5GB...\n\nWin 10..\n\nAny suggestions?	EtherMining
nning toasty and im looking at ways to cool them down. Ive looked nts but want to avoid them if I can and I am now looking at portab...	EtherMining
\n[Random ETH Wallet](https://preview.redd.it/hidnh3tn9kd91.png?mp;s=85f1afc5482e7085899358cefdb1913335d1a923)\n\n# TLDR; \n\n[This](https://m...	EtherMining

Remove symbols and special characters using RegEx

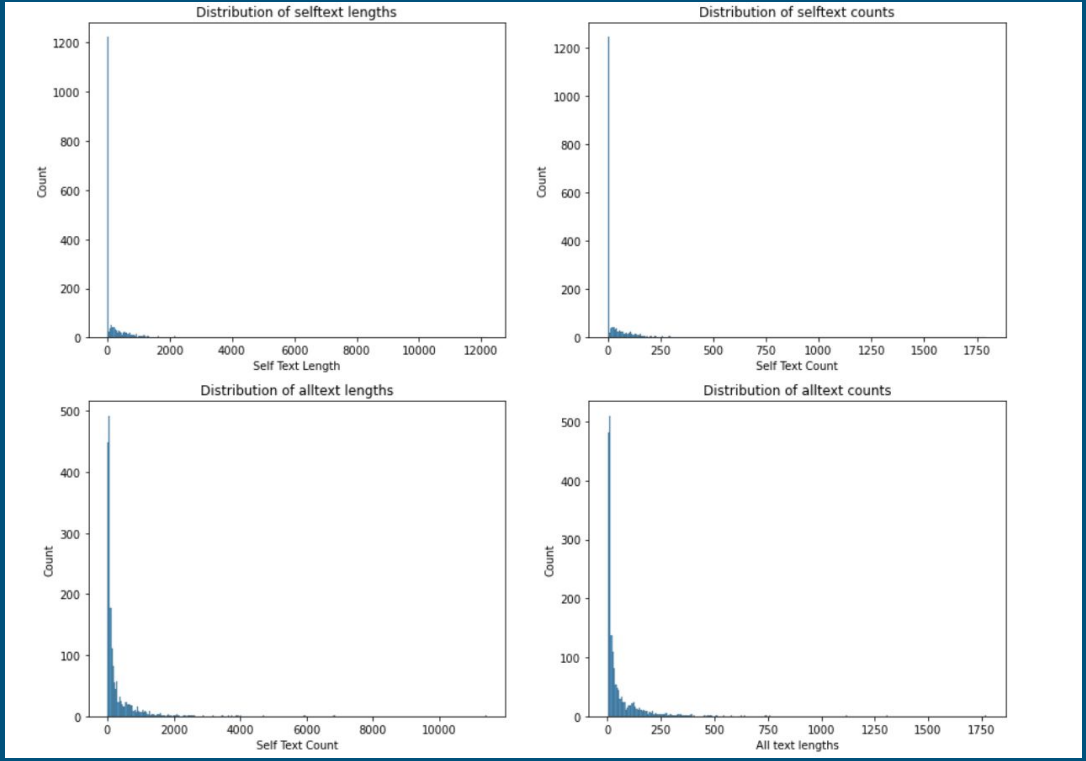
Exploratory Data Analysis





- Character count seems to be maxed out at 300 characters.
- Selftext distribution shows large standard deviation indicating a lot of varian in observed data.

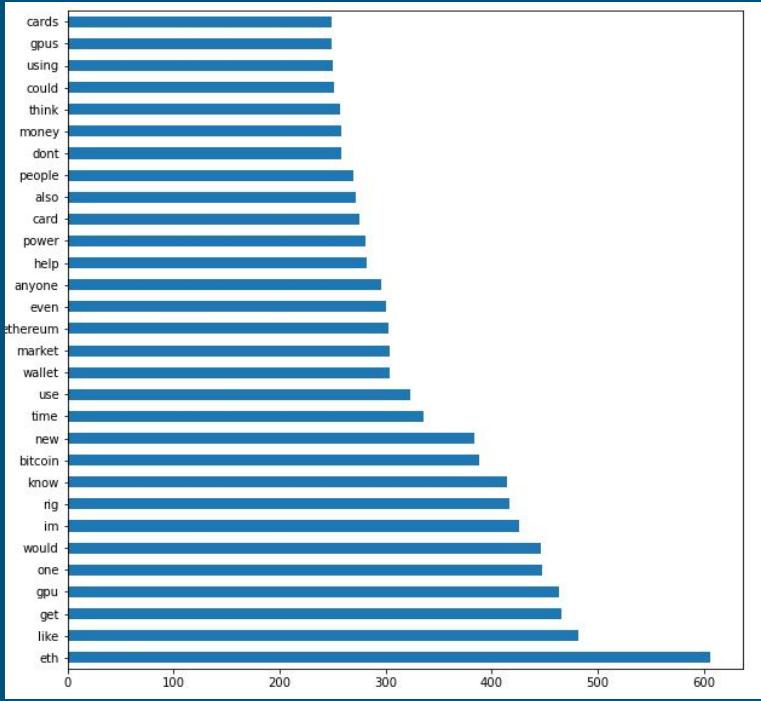
Exploratory Data Analysis



- Highly skewed distribution (Right-skewed)
- Outliers kept for these reasons:
  - We're predicting which subreddit a post originated from.
  - Users may have different preferences to either use the 'title' or 'selftext' portion to fill up their discussion points.
  - By combining both 'title' and 'selftext', we'll have a more reliable dataset (some people may have clickbait titles)
  - Having outliers in terms of length and word counts will probably not affect the model accuracy adversely.

Exploratory Data Analysis

List of common words across both subreddits



CountVectorizer N-gram

- Unigram (1- gram)

	frequency	ngram
0	606	eth
1	481	like
2	466	get
3	463	gpu
4	447	one

- Bigram (2 - gram)

	frequency	ngram
0	67	mining rig
1	62	seed phrase
2	60	bear market
3	60	anyone know
4	52	gpu mining

Modelling

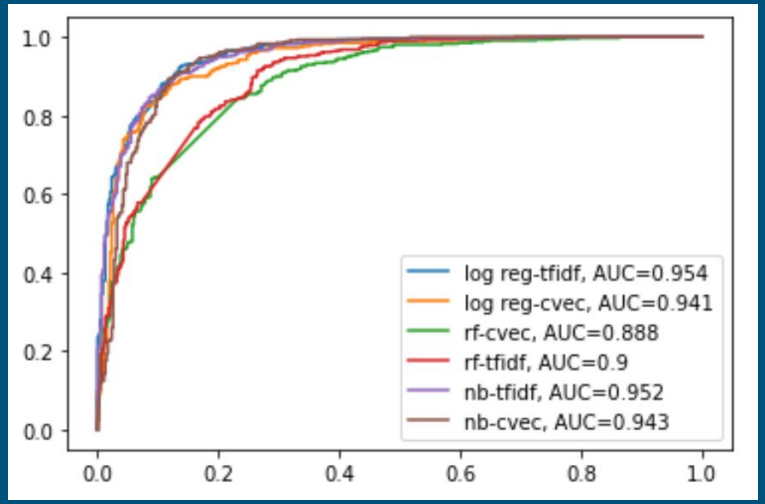
```
#checking for data imbalance
df_combined['subreddit'].value_counts()
```

```
EtherMining        1875
CryptoCurrency     1875
Name: subreddit, dtype: int64
```

- 6 models performed via GridSearchCV
  - MultinomialNB, RandomForest, Logistic Regression with CountVectorizer
  - MultinomialNB, RandomForest, Logistic Regression with tfidfVectorizer

Results

	train_score	test_score	generalisation	f1_score	roc_auc_score
cvec_rf	0.826102	0.777186	5.9214	0.814552	0.8911
tfidf_rf	0.823257	0.764392	7.1503	0.803904	0.9074
cvec_logreg	0.972973	0.861407	11.4665	0.867886	0.9414
tfidf_logreg	0.940256	0.884861	5.8914	0.890244	0.9544
cvec_nb	0.908606	0.885928	2.4960	0.890705	0.9427
tfidf_nb	0.916430	0.882729	3.6774	0.883475	0.9524



Model Evaluation

MultinomialNB with CountVectorizer

- Highest accuracy score
- Best generalization score
- Highest F1 score

Results

- MultinomialNB with CountVectorizer model has the highest accuracy and F1 score as well as the fact that it generalises the best.

- Though the model does not have the highest AUC score, the AUC metric measures performance for the classification problem at various threshold settings.

Downfalls

MultinomialNB nicknamed “naive” due to the fact that the model treats all words equally. Ignores grammar rules and common phrases. A good use case scenario would be separating real content from spam words. (high bais, low variance)

Model	MultinomialNB
Transformer	CountVectorizer
Word processing	Unmodified Text
Max features	5000
Max df	0.85
Min df	2
ngram_range	(1,1)

Conclusion

Though we have chosen the best model to be Multinomial NB with CountVectorizer, there are definitely still things to consider especially in the event of a business setting or when trying to solve a specific problem statement:

1. AUC-ROC curve may not be the only indicator of finding the 'best' model when compared against other models.
2. Different functions may look at different metrics prior to deciding on which model to implement to production. (e.g Marketing/sale may focus on recall , whilst claims functions may want to focus on precision etc)
3. We're dealing with a balanced dataset. For imbalance dataset, class weights and over sampling techniques such as SMOTE may be implemented. F1 score could also come in handy for imbalanced datasets (as its calculated based on Precision and Recall).
4. May consider testing out other models which may perform better for such a task specific to this project.

Thank you!

Q&A