TEAM XXXXXX

General Assembly

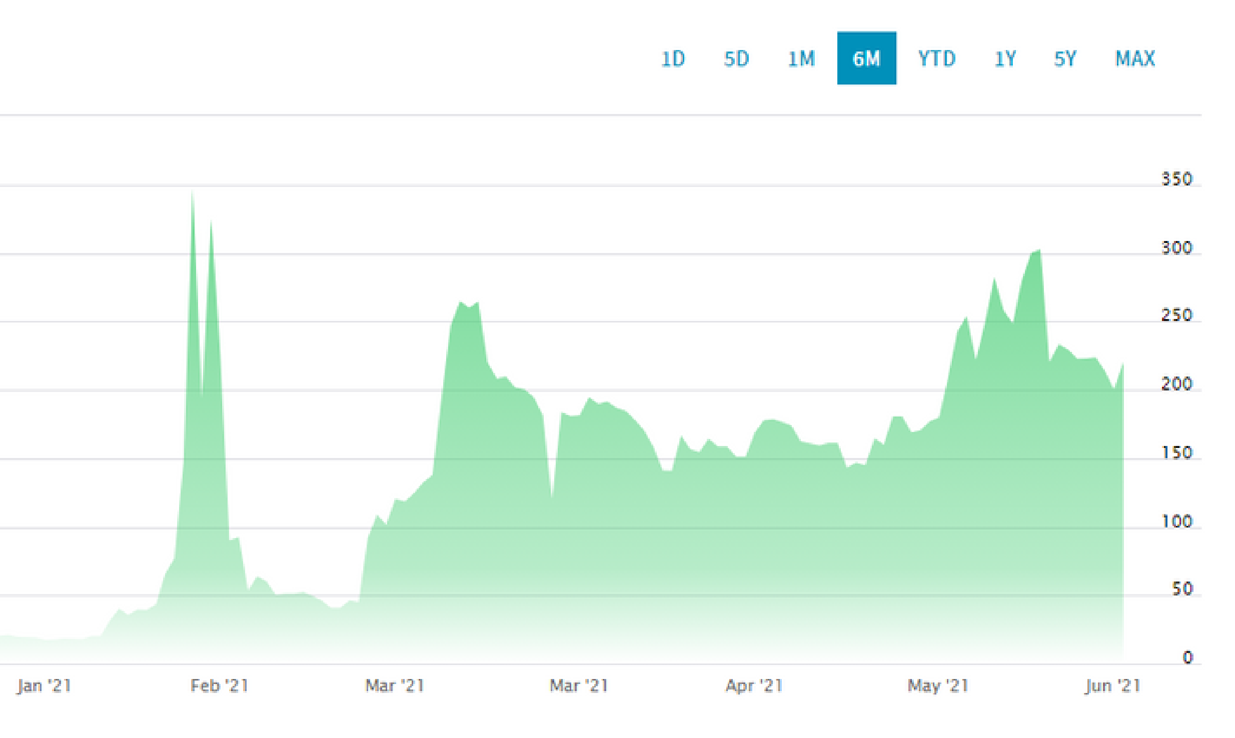
Project 3 - Subreddit Classification

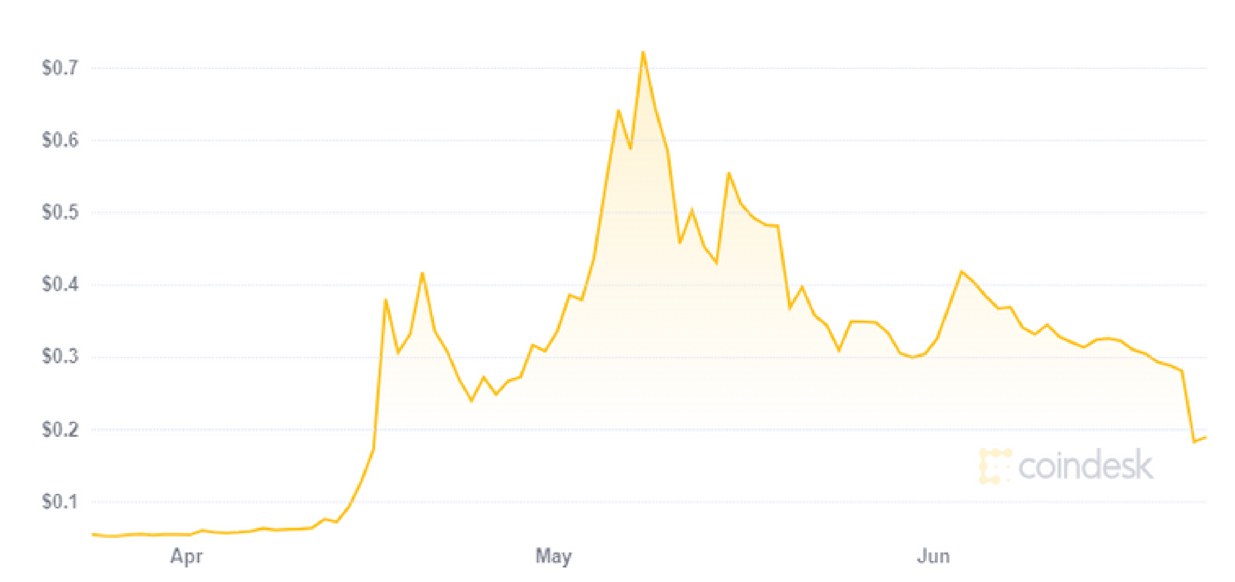
Background

- 2021 GameStop & Dogecoin craze

- Huge increases in price in a short period of time as a result of subreddit activity and influence of prominent figures like Elon Musk.
  - GameStop (21 Jan - 02 Feb 2021):
    - 39 USD to 348 USD (+900%)
  - Dogecoin (26 Apr - 05 May 2021):
    - 0.01 USD to 0.8 USD (+80,000%)

- Prices continue to fluctuate wildly today





Problem Statement

- Retail investors looking to capitalize on the recent craze in GME & Dogecoin
  - Price movements depend on like-minded investors who make decisions based on Reddit/other social media platforms

- Can we use NLP to build classification models based on GME & Dogecoin subreddits to determine the more popular asset?
  - Invest in the more popular asset accordingly
  - Non-quantitative investing perspective vs. traditional financial analysis

Stakeholders

1

Alternative Investors

Use sentiment to understand popularity of asset

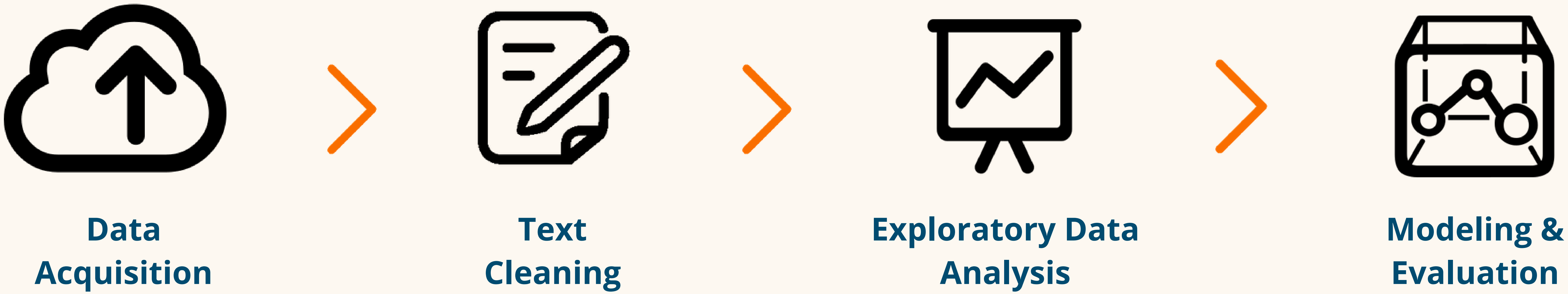
2

Reddit Company

Subreddit Moderators & Community

reddit

Workflow



Data Acquisition

Text Cleaning

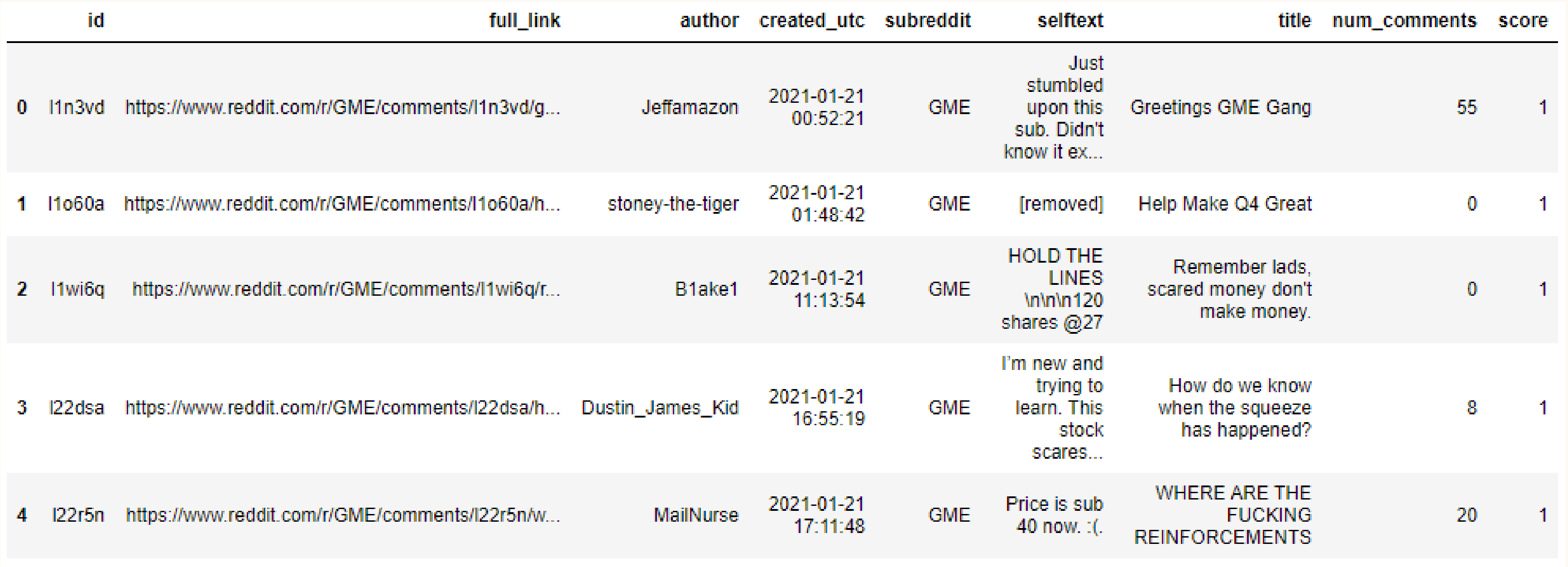
Exploratory Data Analysis

Modeling & Evaluation

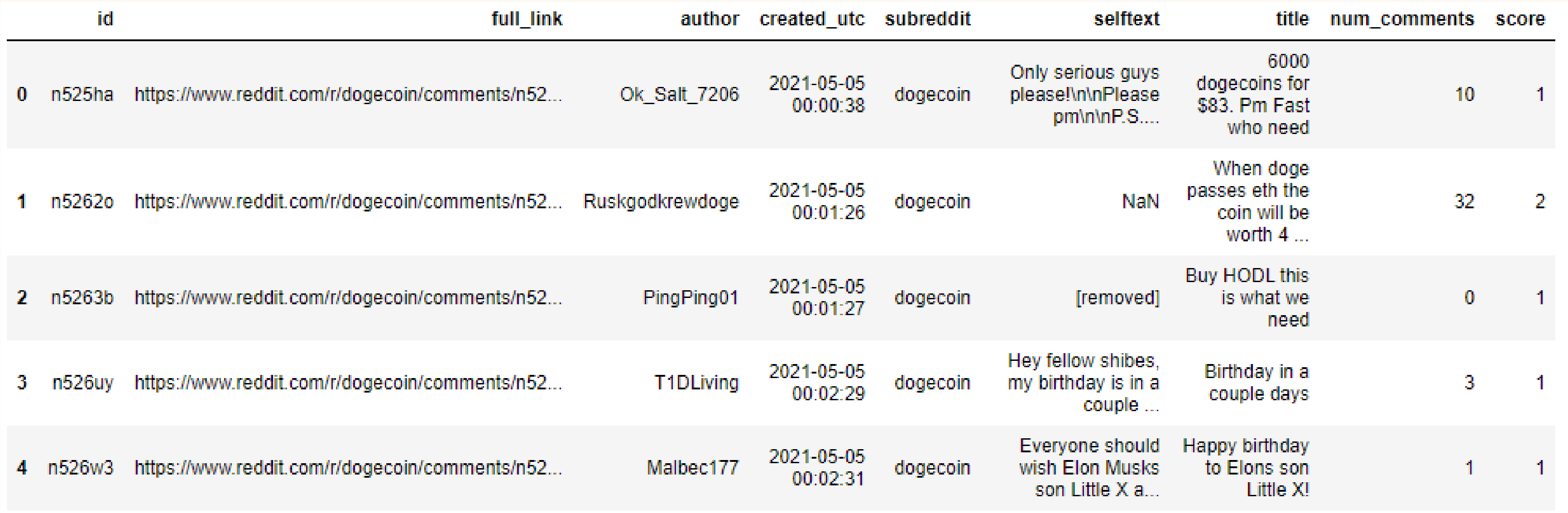
Data Acquisition

Collected posts from r/GME - GameStop Stock subreddit and r/dogecoin - Dogecoin subreddit using Pushshift's API

- Data collected from:
  - GME: 21 Jan 2021 - 2 February 2021
  - Dogecoin: 5 May 2021
- There are about 6,805 posts from GME subreddit and 10,750 posts from Dogecoin subreddit
- The title and content of the posts are used as features for EDA and modelling



	id	full_link	author	created_utc	subreddit	selftext	title	num_comments	score
0	l1n3vd	https://www.reddit.com/r/GME/comments/l1n3vd/g...	Jeffamazon	2021-01-21 00:52:21	GME	Just stumbled upon this sub. Didn't know it ex...	Greetings GME Gang	55	1
1	l1o60a	https://www.reddit.com/r/GME/comments/l1o60a/h...	stoney-the-tiger	2021-01-21 01:48:42	GME	[removed]	Help Make Q4 Great	0	1
2	l1wi6q	https://www.reddit.com/r/GME/comments/l1wi6q/r...	B1ake1	2021-01-21 11:13:54	GME	HOLD THE LINES \n\n\n120 shares @27	Remember lads, scared money don't make money.	0	1
3	l22dsa	https://www.reddit.com/r/GME/comments/l22dsa/h...	Dustin_James_Kid	2021-01-21 16:55:19	GME	I'm new and trying to learn. This stock scares...	How do we know when the squeeze has happened?	8	1
4	l22r5n	https://www.reddit.com/r/GME/comments/l22r5n/w...	MailNurse	2021-01-21 17:11:48	GME	Price is sub 40 now. :(.	WHERE ARE THE FUCKING REINFORCEMENTS	20	1



	id	full_link	author	created_utc	subreddit	selftext	title	num_comments	score
0	n525ha	https://www.reddit.com/r/dogecoin/comments/n52...	Ok_Salt_7206	2021-05-05 00:00:38	dogecoin	Only serious guys please!\n\nPlease pm\n\nP.S....	6000 dogecoins for $83. Pm Fast who need	10	1
1	n5262o	https://www.reddit.com/r/dogecoin/comments/n52...	Ruskgodkrewdoge	2021-05-05 00:01:26	dogecoin	NaN	When doge passes eth the coin will be worth 4 ...	32	2
2	n5263b	https://www.reddit.com/r/dogecoin/comments/n52...	PingPing01	2021-05-05 00:01:27	dogecoin	[removed]	Buy HODL this is what we need	0	1
3	n526uy	https://www.reddit.com/r/dogecoin/comments/n52...	T1DLiving	2021-05-05 00:02:29	dogecoin	Hey fellow shibes, my birthday is in a couple ...	Birthday in a couple days	3	1
4	n526w3	https://www.reddit.com/r/dogecoin/comments/n52...	Malbec177	2021-05-05 00:02:31	dogecoin	Everyone should wish Elon Musks son Little X a...	Happy birthday to Elons son Little X!	1	1

Text Cleaning

- Clean characters that stem from (Reddit-specific) markdown formatting
- Convert emoji to text
- Remove HTML tags and links
- Convert all text to lowercase
- Convert slangs/typos to their original word
- Expand abbreviations
- Expand contractions
- Remove special characters
- Remove extra whitespace

The following steps are optional:

- Stopwords removal
- Stemming
- Lemmatization

Clean characters that stem from (Reddit-specific) markdown formatting

- “&amp;#x200B;” (bullet point) , “&gt;” (quote)

Remove HTML tags and links

- Eg. “</h2></html> https://preview.redd.it/”

Convert slangs/typos to its original word

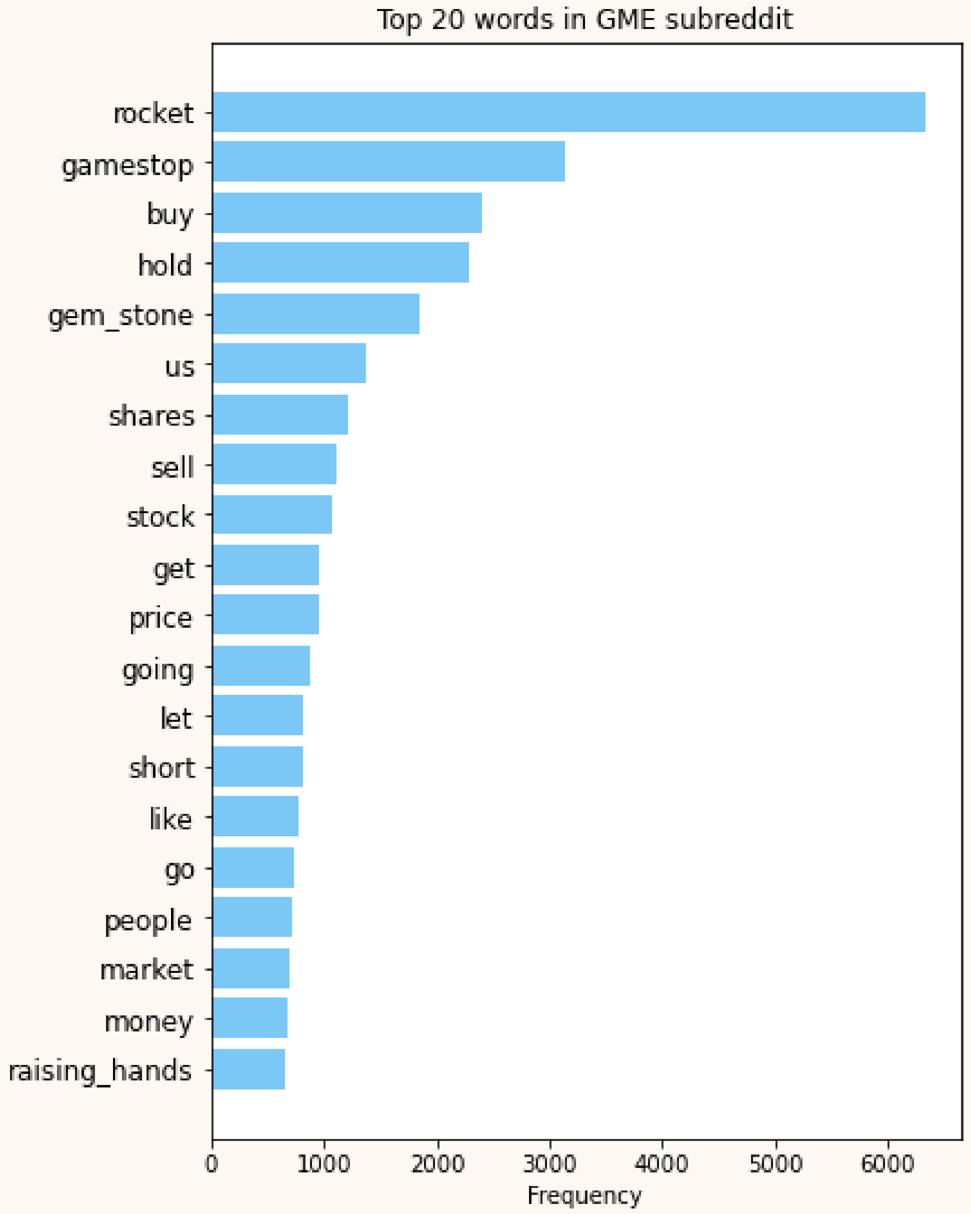
- Eg. “dodgecoin”, “dojehcoin”

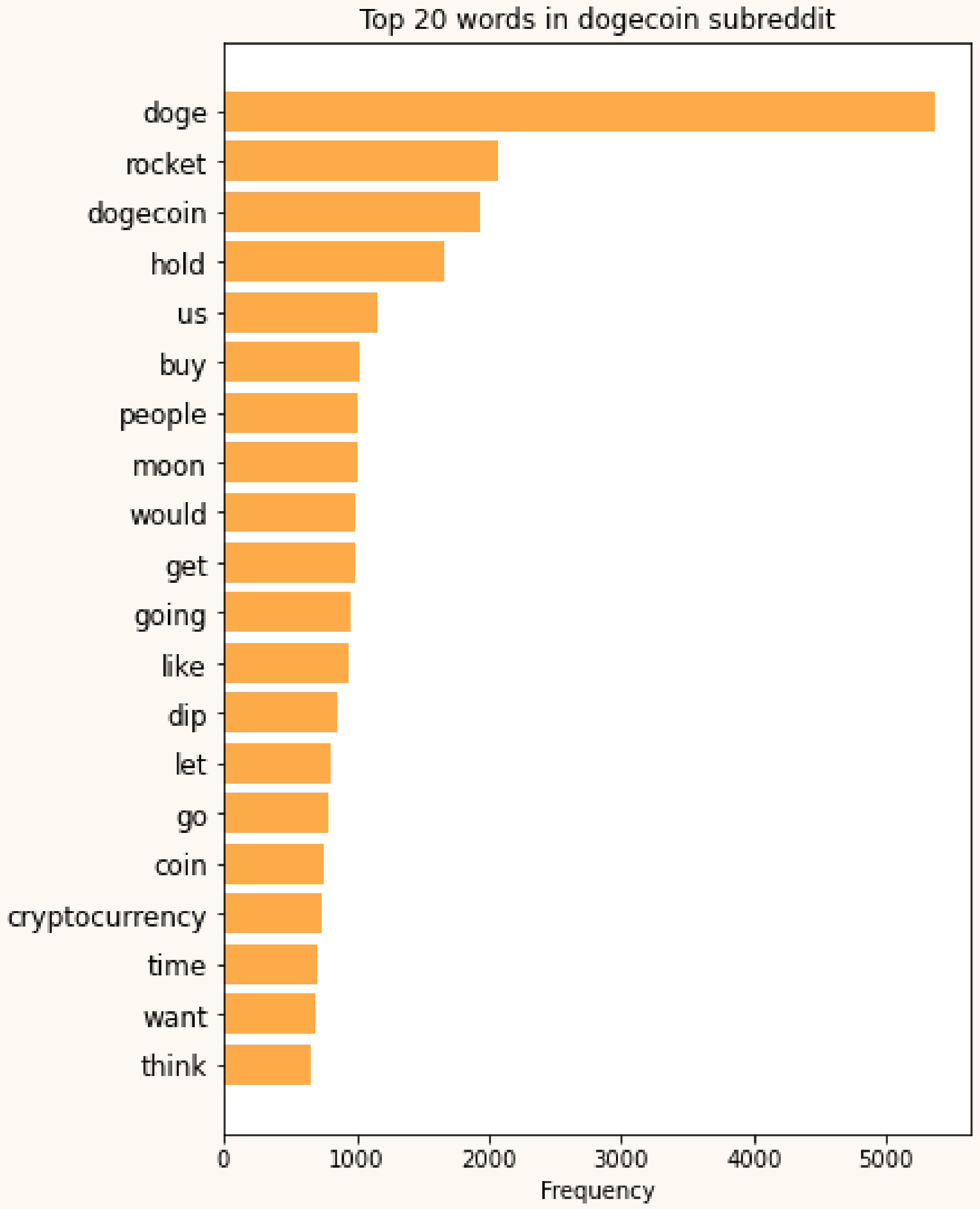
Expand abbreviations

- Eg. “sus” to “suspicious and suspect”

Exploratory Data Analysis

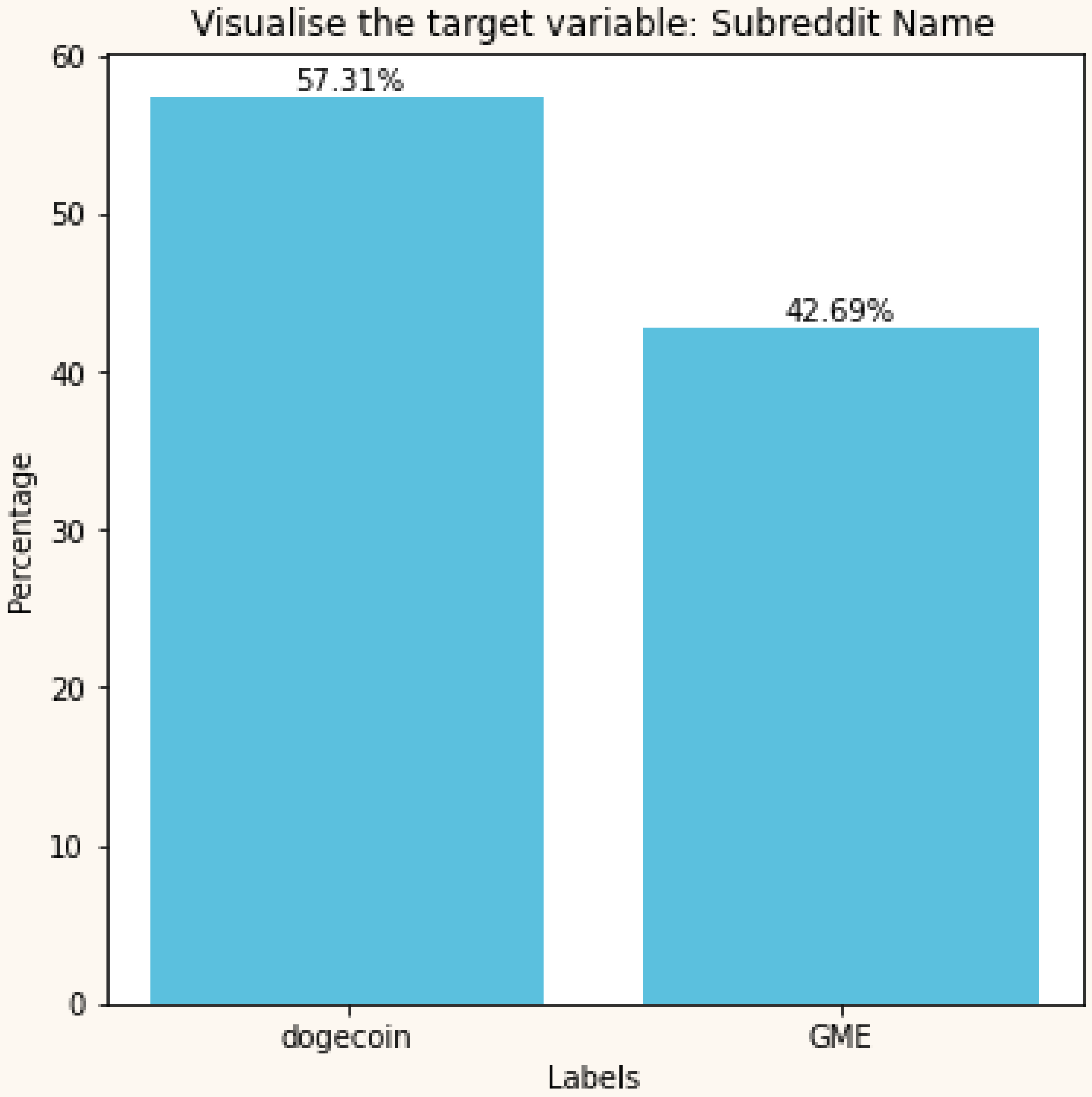
Top 20 Words individual subreddit





Target Variables

Goal: Is there class imbalance in our data?



Model Evaluation

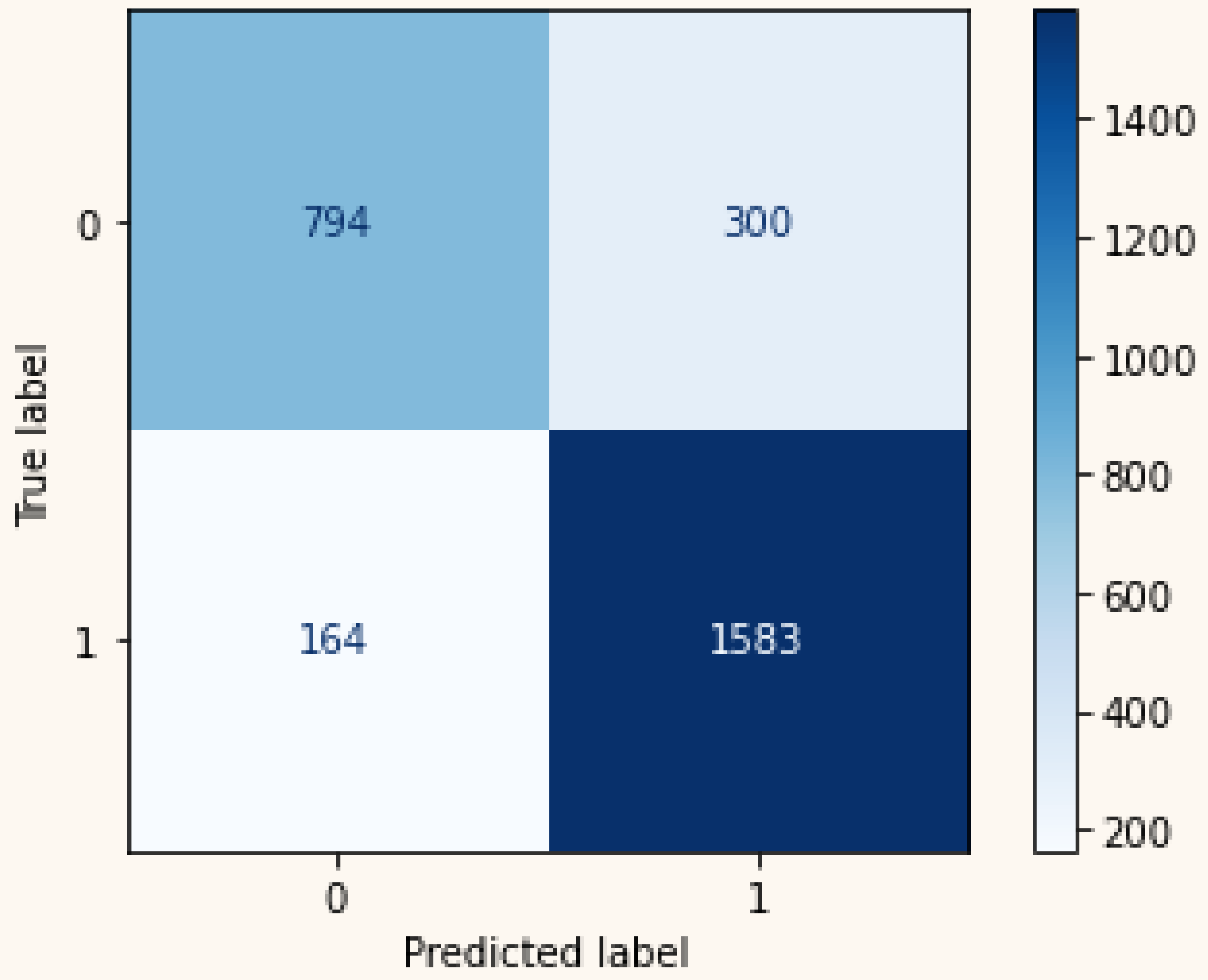
Goal: To get F1-score as close to 1 as possible

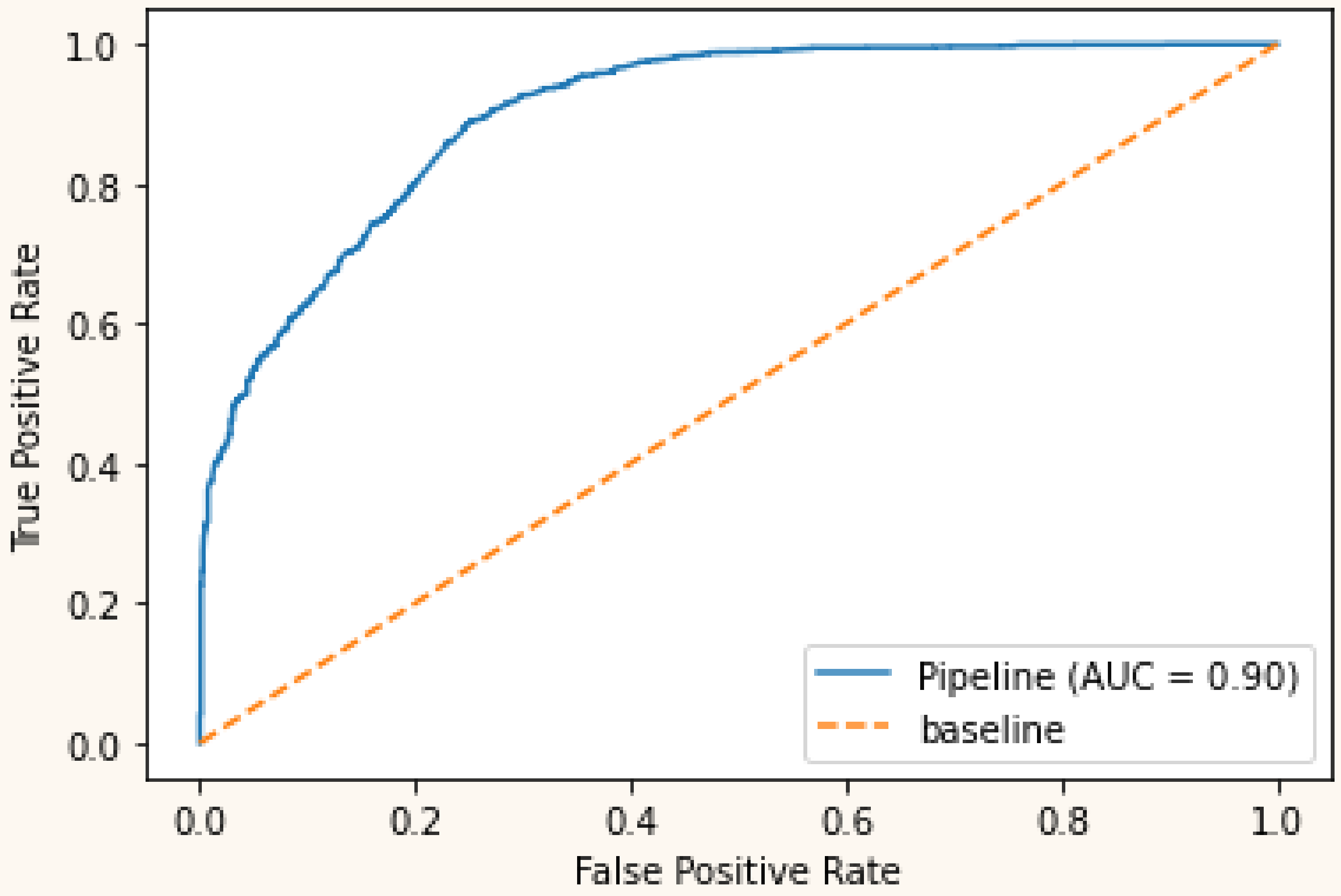
Evaluation metric of the classification models with NLP vectoriser:

Model	Hyperparameters	5 fold CV F1-Score
LogisticRegression	C penalty max_iter	0.8984
MultinomialNB	alpha	0.8999
RandomForestClassifier	n_estimators max_depth	0.8982

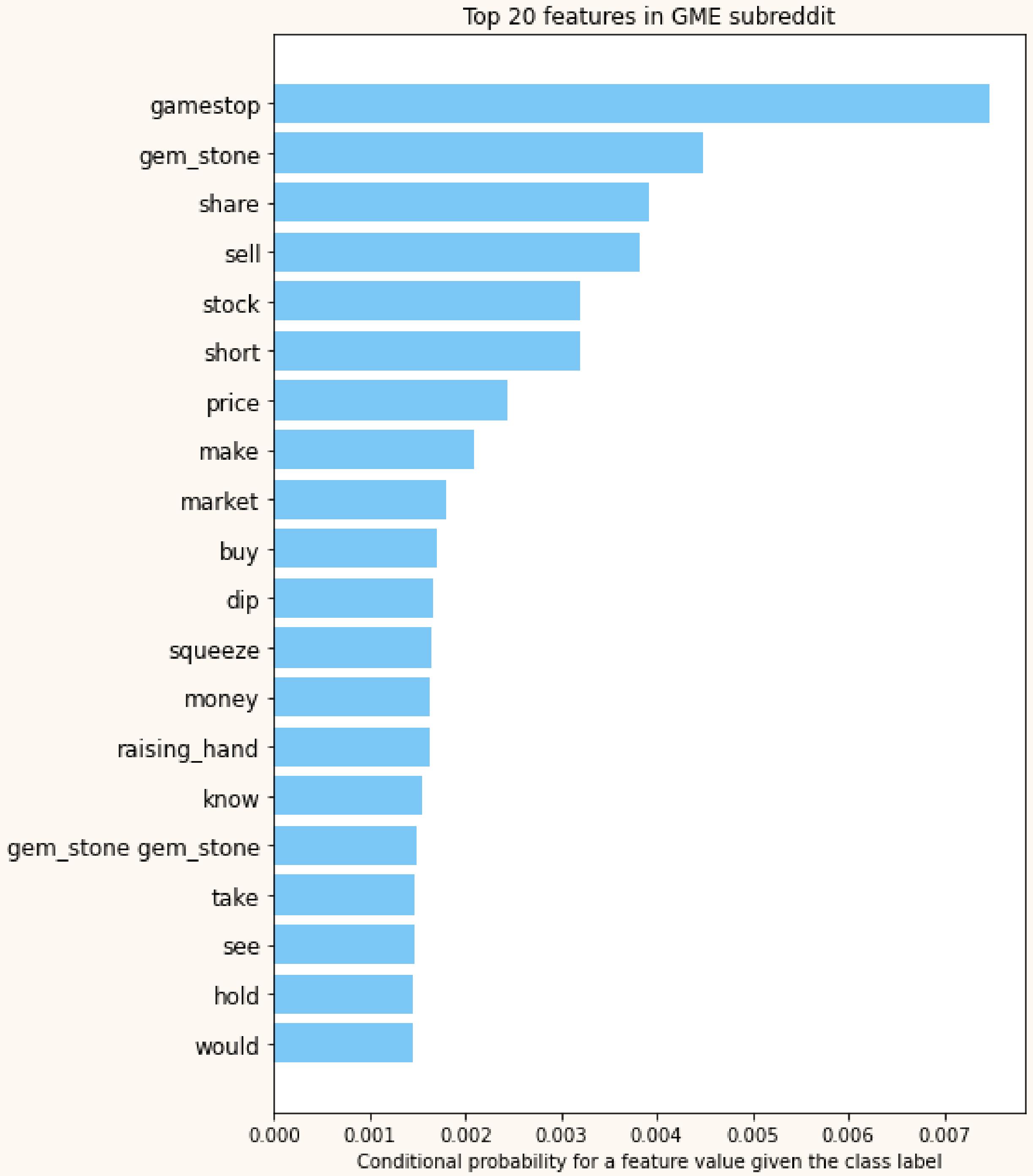
Model Evaluation

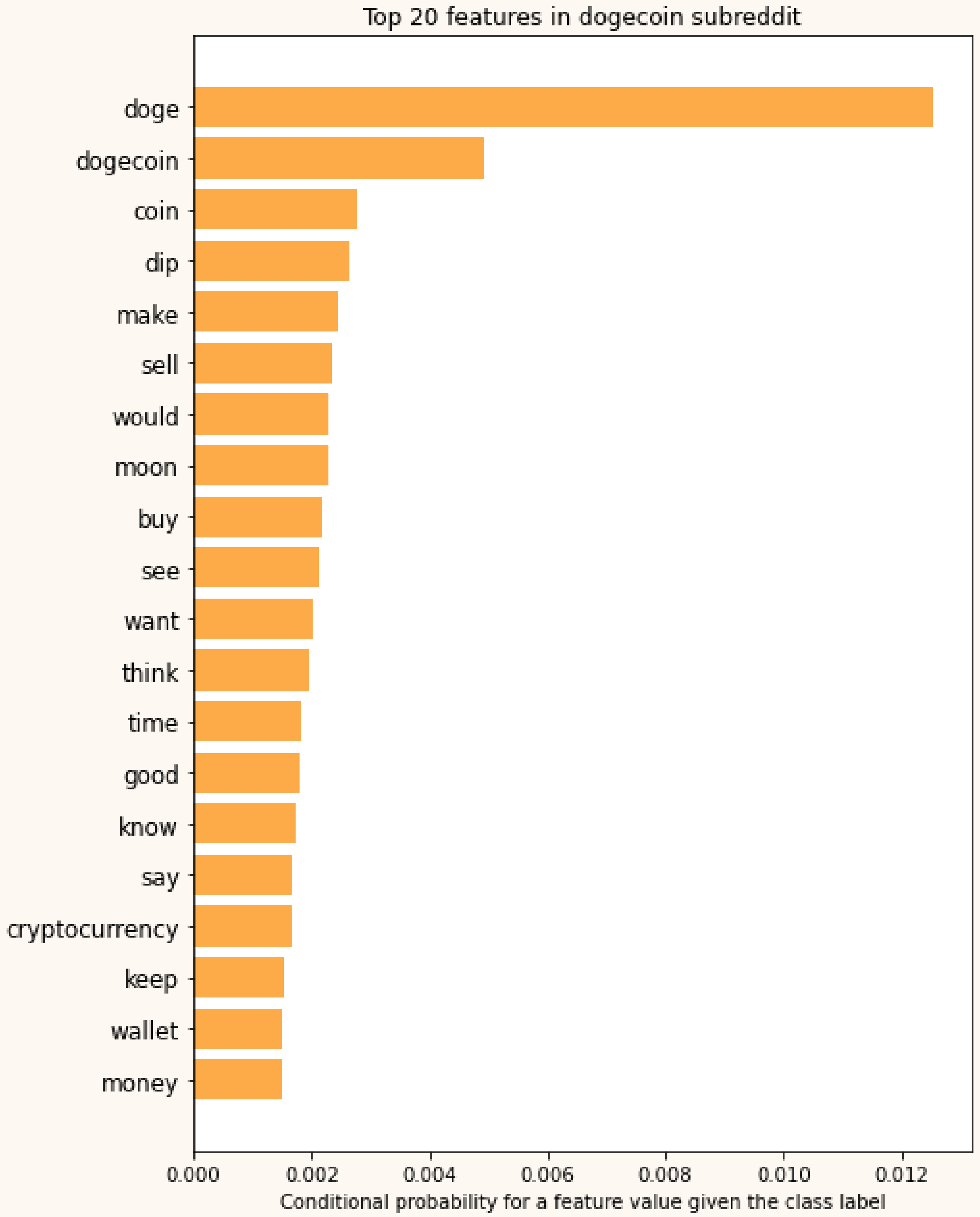
Goal: Is our data able to classify correctly?





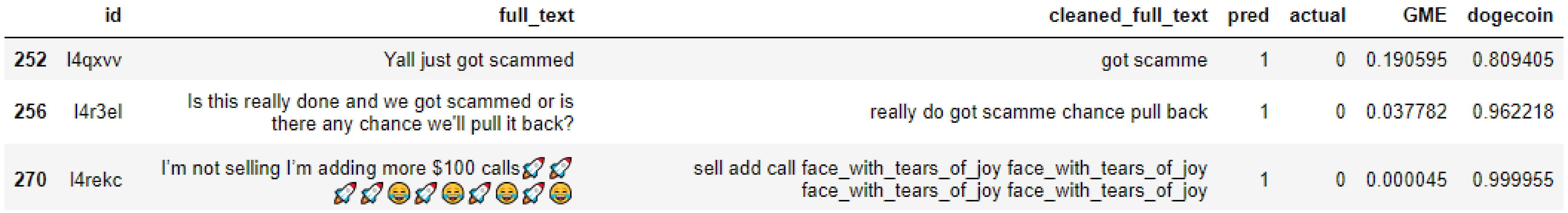
Inferential Findings & Analysis





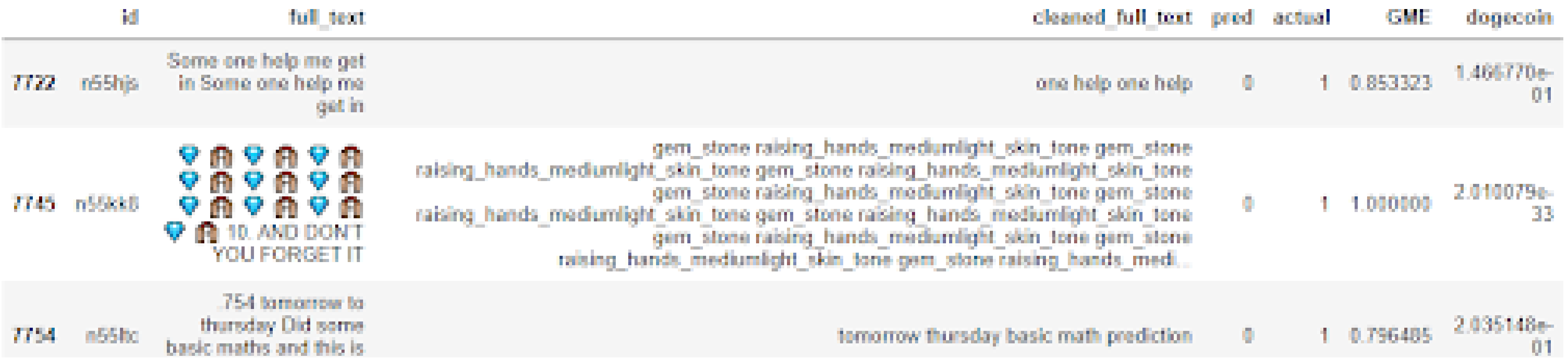
Limitations

- Example of false-positive posts: Predicted as Dogecoin but it actually comes from GME subreddit.



	id	full_text	cleaned_full_text	pred	actual	GME	dogecoin
252	l4qxvv	Yall just got scammed	got scamme	1	0	0.190595	0.809405
256	l4r3el	Is this really done and we got scammed or is there any chance we'll pull it back?	really do got scamme chance pull back	1	0	0.037782	0.962218
270	l4rekc	I'm not selling I'm adding more $100 calls🚀🚀🚀🚀😂🚀😂🚀😂🚀😂	sell add call face_with_tears_of_joy face_with_tears_of_joy face_with_tears_of_joy face_with_tears_of_joy	1	0	0.000045	0.999955

- Example of false-negative posts: Predicted as GME but it actually comes from Dogecoin subreddit.



	id	full_text	cleaned_full_text	pred	actual	GME	dogecoin
7722	n55hjs	Some one help me get in Some one help me get in	one help one help	0	1	0.853323	1.466770e-01
7745	n55kk8	💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼💎🙌🏼 10. AND DON'T YOU FORGET IT	gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_mediumlight_skin_tone gem_stone raising_hands_medi...	0	1	1.000000	2.010079e-33
7754	n55ltc	754 tomorrow to thursday Did some basic maths and this is	tomorrow thursday basic math prediction	0	1	0.796485	2.035148e-01

Conclusion & Recommendation

Exploratory Data Analysis

>

Modeling & Evaluation

Types of NLP vectorizers:

- CountVectorizer
- TFIDVectorizer

Types of Classification Model:

- Logistic Regression
- MultinomialNB
- Random Forest

=

Recommended

Best Model:
Multinomial (Naive Bayes) with CountVectorizer

F1 Score: 90%

Predicted classified correctly:

- GME: 83%
- Dogecoin: 91%

Conclusion & Recommendation

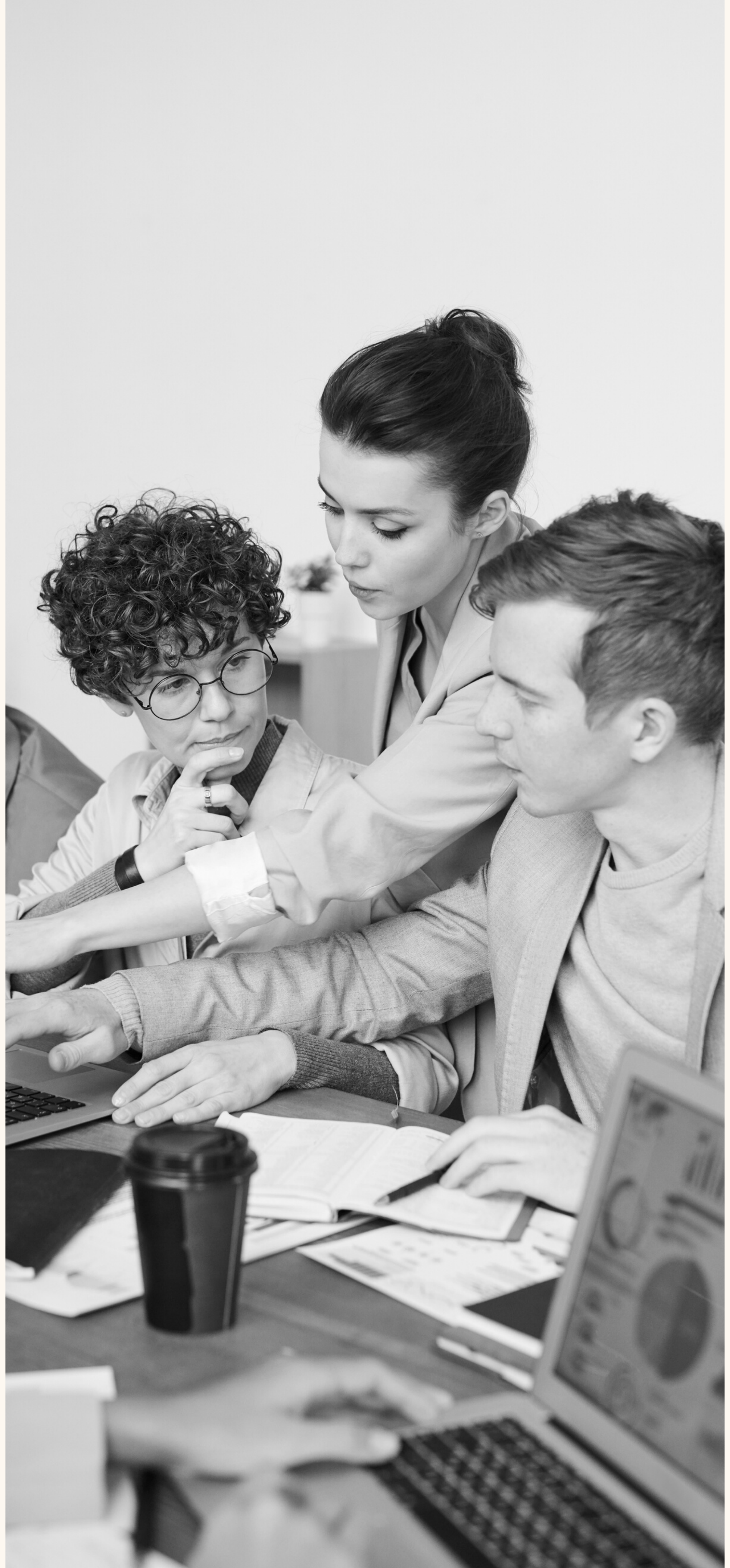
- The current data is based on a recent event. If the model is trained with data from other similar events, we may be able to improve the score further and differentiate 2 different data clearly.

- As unstructured data is now explored in many real-time businesses and depending on the cause of fluctuation of individual price, other media can be explored, for example, financial blogs and news sites

Hi!

If you have any questions at all

Don't hesitate to ask.



Thank You