

Stocks: Hype vs Market

Malene Hansen & Sebastian Harvej

Project on Github:

<https://github.com/PBASOFT/Data-Science-Project>



This is an investigation of the possible correlation between hype and stock prices. We want to determine how to detect hyped stocks and, with use of Natural Language Processing techniques, interpret the data we collect on Reddit.com, which will be our source of (possible) hype for this scope.

Null Hypotheses:

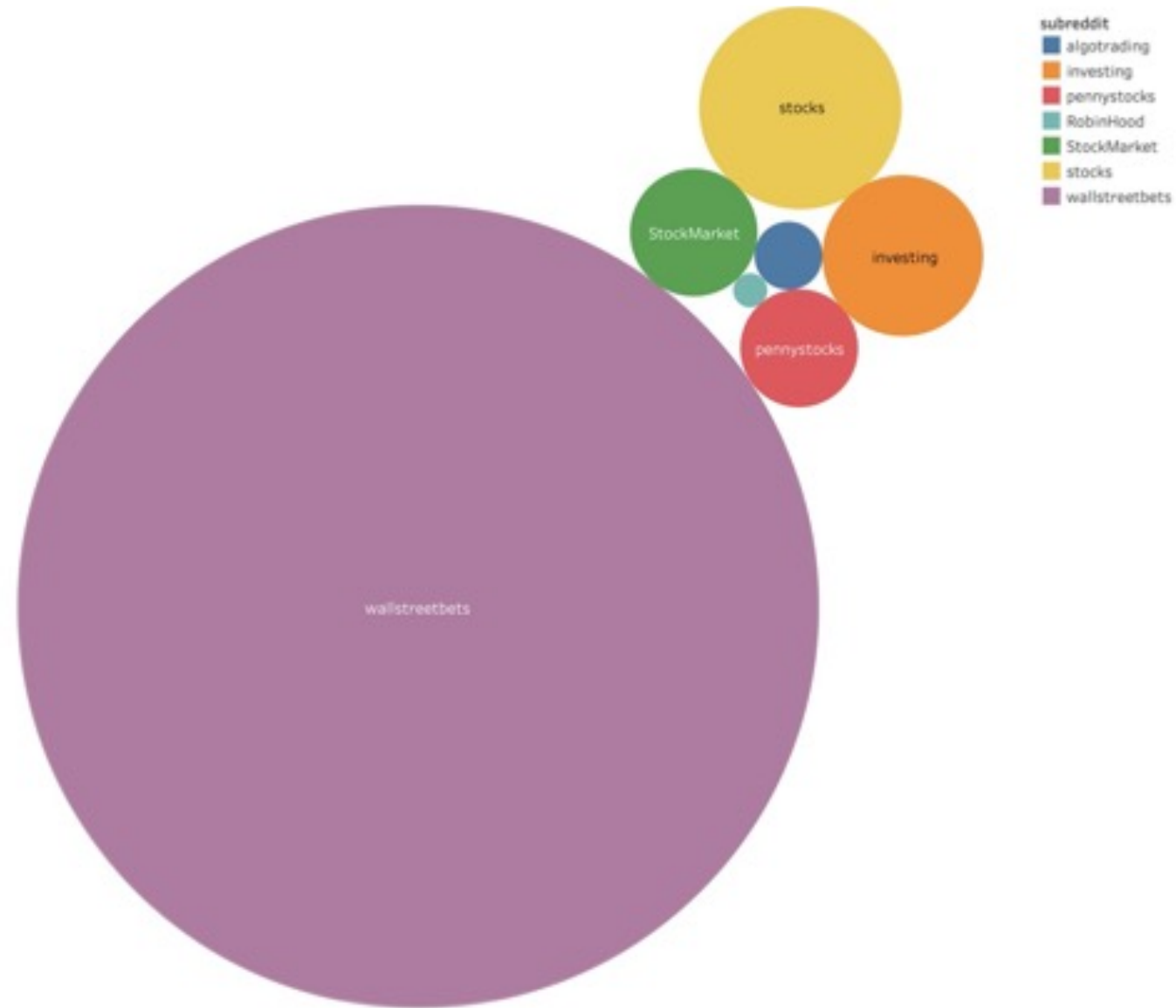
Stock prices can not be affected by hype on social media.

The Data

Data is collected from investing related subreddits.

wallstreetbets by far contains the largest chunk of data. And also, the largest community; more than 10 million members.

To compare hype to the market, we need to analyze the data in order to identify hype.



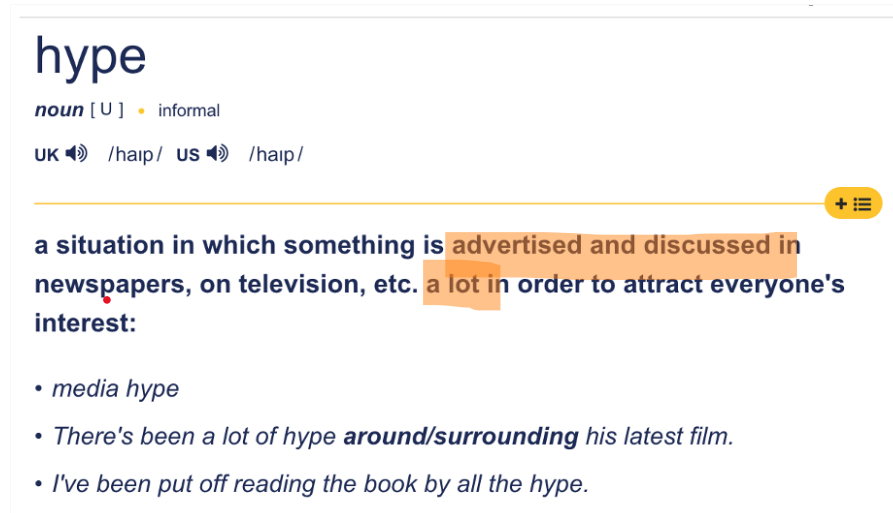
HYPE

To detect hype, we need to define hype.



Based on the meaning provided by the Cambridge Dictionary, we conclude that hype can be measured on two parameters:


Hype measures:

1. Discussed a lot
2. Advertised a lot



The screenshot shows the Cambridge Dictionary entry for the word 'hype'. The word is in a large, bold, blue font. Below it, the part of speech 'noun' is shown in a smaller, grey font, followed by '[U]' and 'informal'. The UK and US pronunciations are listed as '/haɪp/'. A yellow button with a plus sign and a list icon is to the right. The definition is: 'a situation in which something is advertised and discussed in newspapers, on television, etc. a lot in order to attract everyone's interest:'. The words 'advertised and discussed in' and 'a lot in' are highlighted in orange. Below the definition, there are three bullet points: 'media hype', 'There's been a lot of hype around/surrounding his latest film.', and 'I've been put off reading the book by all the hype.'

hype
noun [U] • informal
UK  /haɪp/ US  /haɪp/

+ 

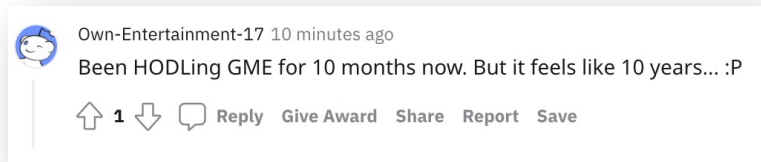
a situation in which something is advertised and discussed in newspapers, on television, etc. a lot in order to attract everyone's interest:

- *media hype*
- *There's been a lot of hype **around/surrounding** his latest film.*
- *I've been put off reading the book by all the hype.*

Source : <https://dictionary.cambridge.org/dictionary/english/hype>

Detecting stocks

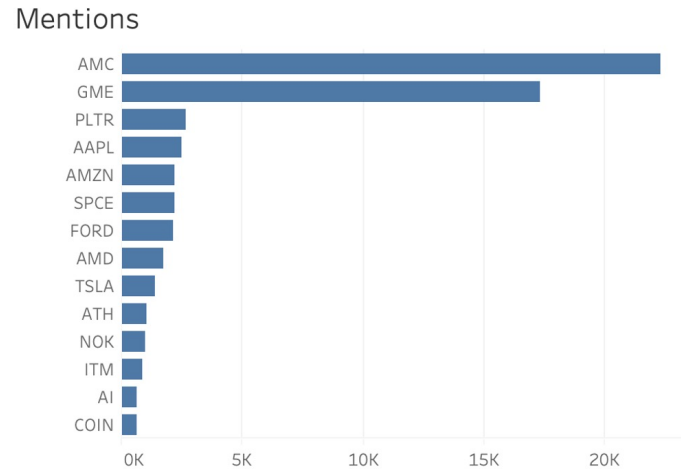
Named Entity Recognition is applied to extract stock mentions from our data.



Been HODLing **GME** **ORG** for **10 months** **DATE** now. But it feels like **10 years** **DATE** ... :P

Capable of recognizing "real world concepts", it is a powerful tool to identify stock mentions.

Stocks that are being discussed a lot



AMC and GME are without doubt being discussed ***a lot*** in comparison to the rest of the field.

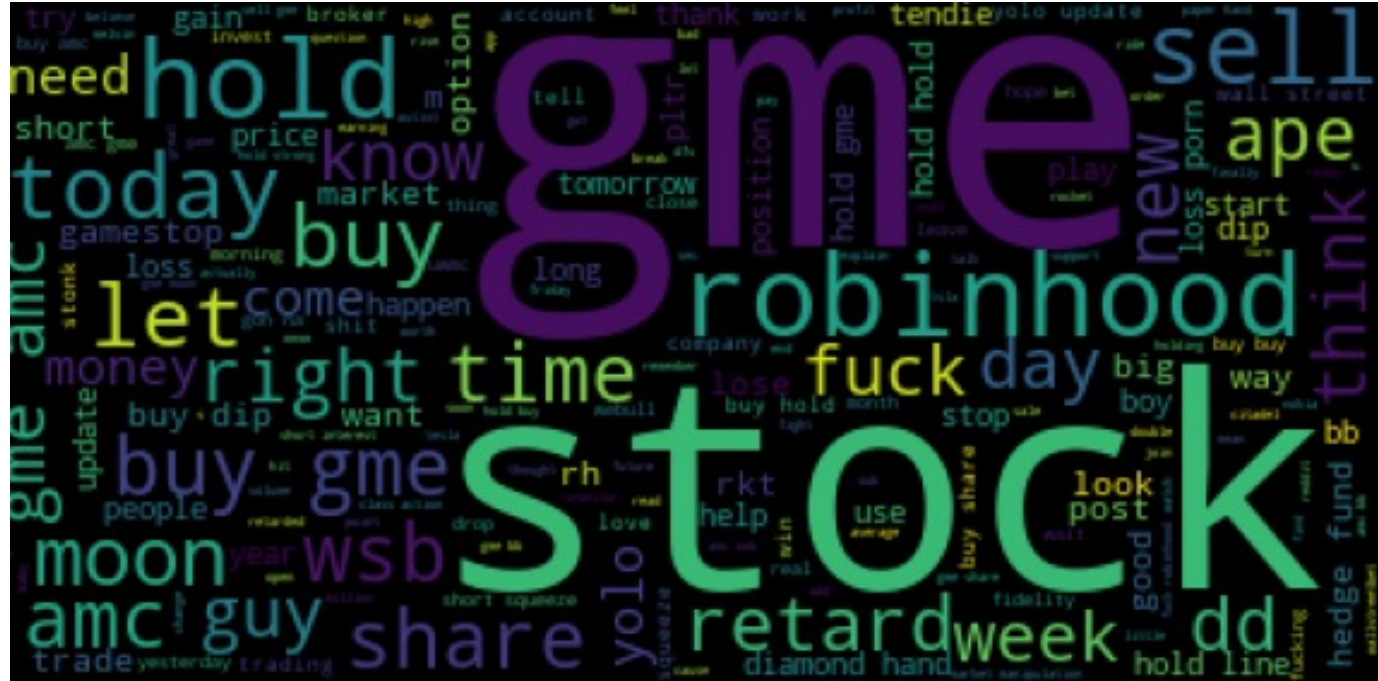
Stocks that are being advertised a lot

To identify a stock as hyped we also need to know the sentiment of the context, as a hyped stock is also being ***advertised*** a lot.

We attempt to predict the sentiment of the stock mentions' context by applying *sentiment analysis*.

Sentiment analysis

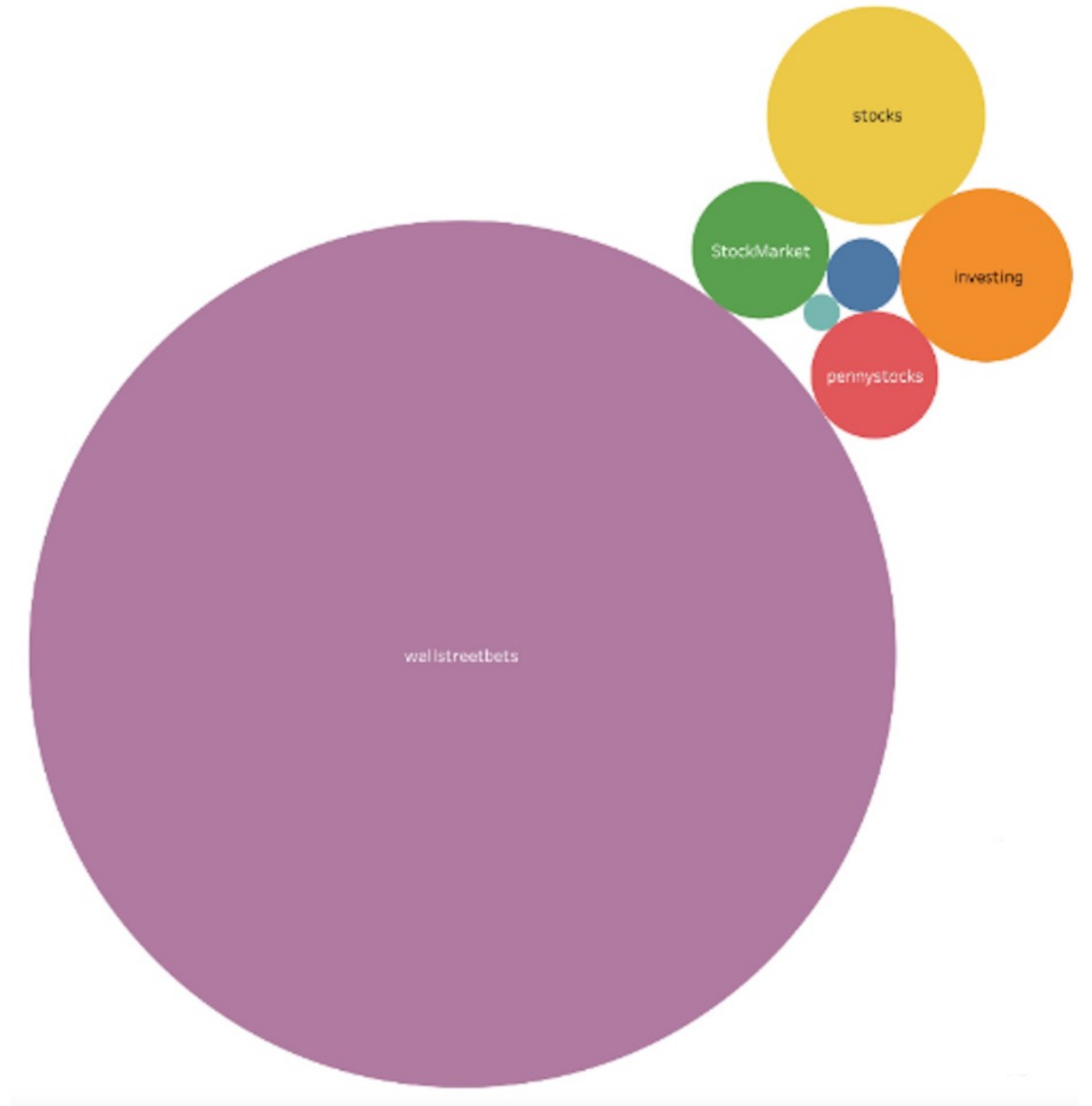
While researching stock themed subreddits we discovered that the language was very different from what is used other places. Especially on the subreddit wallstreetbets.



Sentiment analysis

To accommodate this issue we trained our machine learning model with data from wallstreetbets.

This is very important as it is the subreddit with the most comments by far.



Sentiment analysis

To get the best accuracy on our prediction we tested two different machine learning models 1 using k-nearest neighbors as the classifier and one using Linear Support Vector Classification.

The results of the second was very better than the first one so we decided to go for Linear Support Vector Classification.

	precision	recall	f1-score	support
-1	0.60	0.26	0.36	1042
0	0.54	0.93	0.68	2304
1	0.77	0.22	0.34	1480
accuracy			0.57	4826
macro avg	0.64	0.47	0.46	4826
weighted avg	0.62	0.57	0.51	4826

	precision	recall	f1-score	support
-1	0.80	0.73	0.76	1042
0	0.87	0.94	0.90	2304
1	0.87	0.81	0.84	1480
accuracy			0.85	4826
macro avg	0.85	0.83	0.83	4826
weighted avg	0.85	0.85	0.85	4826

Sentiment analysis

This is an example of what the model predicted right, but in most cases, it wasn't accurate enough, so we decided not to use this as a factor when looking at the correlation of mention and price of a stock.

Text	Sentiment
ok I put it all in!!!	1
should buy SNDL now ??	0
Can apple not freaking suck for just one day	-1

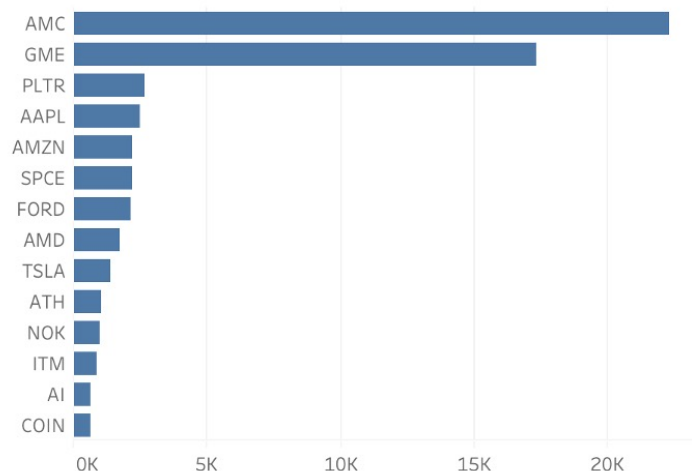
HYPE VS MARKET

We haven't yet implemented sentiment analysis to our model to fulfil the second leg of the hype definition "*advertised a lot*".

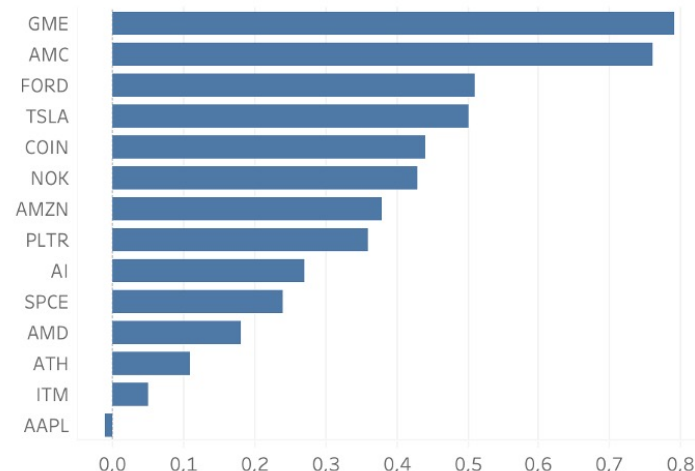
But we were able point out two stocks that fulfil the first leg of the hype definition "*discussed a lot*": **AMC** and **GME**.

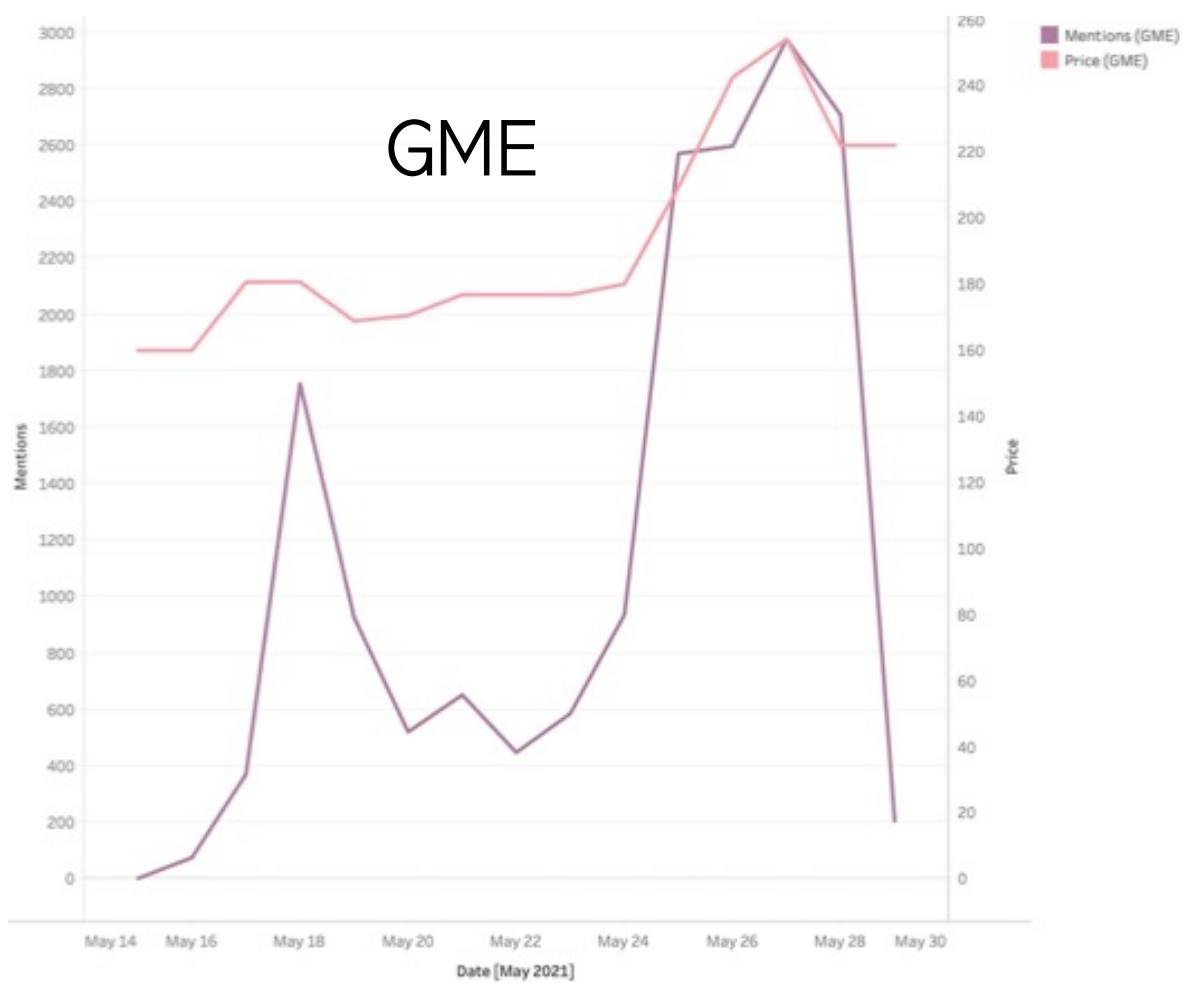
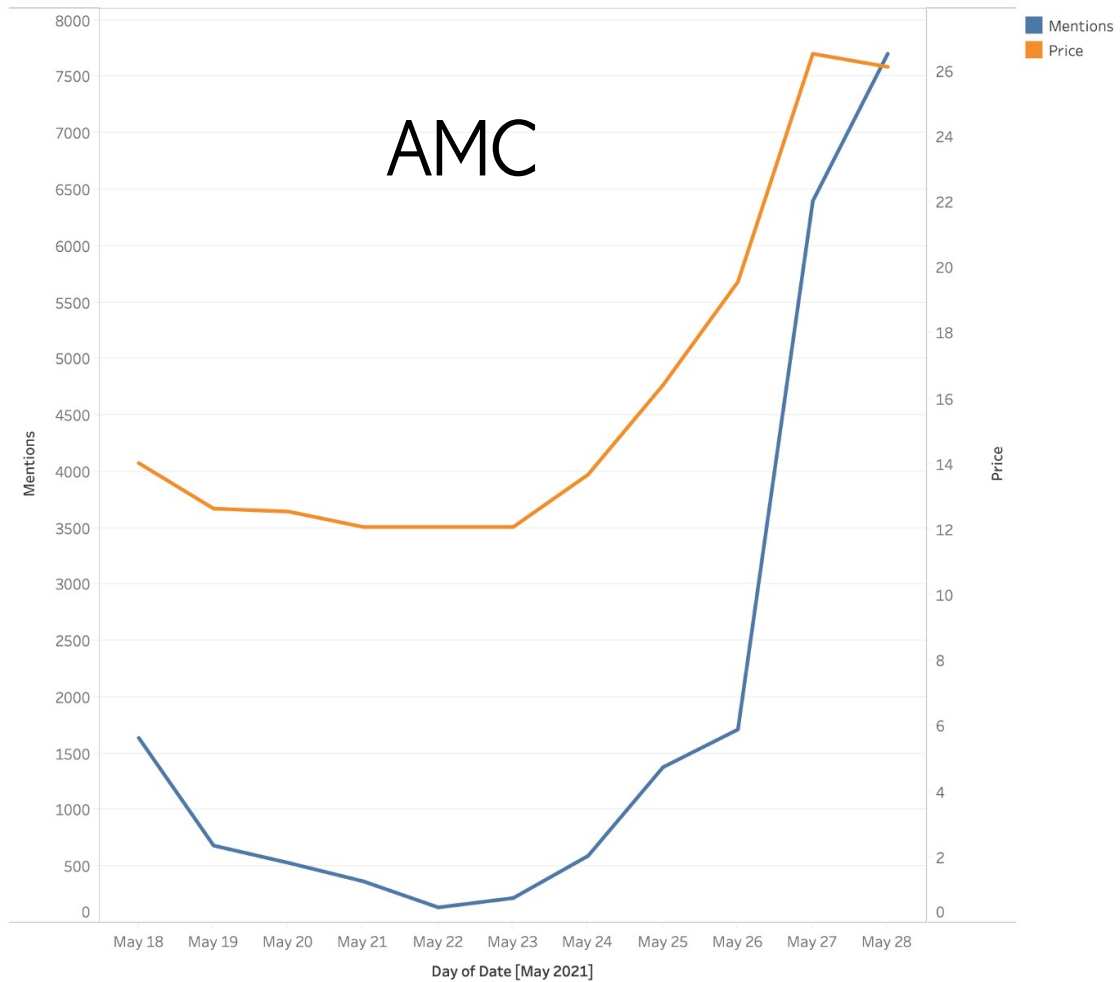
They are not only clearly discussed more than the average. They are also clearly the most correlated with the stock price:

Mentions



Correlation Coefficient





Conclusion

With such a clear correlation between the two most discussed stocks and their corresponding prices, we can not claim that our results support the null hypothesis at this point.

We therefore **reject the null hypothesis.**

Our hope for the future is that by improving the sentiment analysis we can learn more about the mentions. And possibly determine whether the hype is causing the price, or the price is causing the stock. And hopefully keep rejecting the null hypothesis.