A NLP Based Analysis of Gamestop Mania And WallStreetBets

Luke Palmieri

***Abstract*—Reddit comments from December, 2020 to February, 2021 and a set of newspaper headlines are analyzed using Entropy and Cosine Similarity by day to form a timeseries that is tested using a VAR and Granger Causality. A novel indicator of activity on WallStreetBets is found.**

***Index Terms*—NLP, WallStreetBets, Narrative Economics, Time Series, VAR, Granger Causality, Stock Market**

I. Introduction

THE stock market has experienced rapid swings in prices driven by retail investors and social media hype. The start of these price fluctuations began on a trading community on the popular internet forum Reddit. During these fluctuations, in late January, the stock of Gamestop (GME) saw the largest and fastest increase in it share price ever. This was after years of declining expectations for the company on Wall Street and dismal forecasts for the company's business model. [6]

The most important driver of the events was not any business plan, earnings or other "fundamental" aspects of GME, despite there being a good case for Gamestop being undervalued. Instead it was driven mostly by self full-filling expectations of GME traders. Although research on the events has found it is not necessary for the small group of buyers to be irrational, and in fact there could be an opportunity to make a risk-less profit by orchestrating a "short-squeeze", where because GME was so heavily bet against, it left the short sellers vulnerable to an unpredictable event where a coordinated group could force the short sellers into a positive feedback loop of having to cover their short pushing the price up further.

Regardless, the formation of these expectations among groups determines the price swings. Traditional economic models or valuations would not predict or explain this behavior. Recent research interest has focused on how these self full-filling expectations are created through shared narratives; simplified economic or other stories that aim to explain market dynamics. Millions of retail investors began pilling into Gamestop as a result of these shared narratives going viral. [1]

Internet trends that spread rapidly like this are often completely unexpected black swan events where the emergence of the trend was not predictable, however understanding how the dynamics of forums like WallStreetBets operate could lead to a better understanding of these events and what indicators precede them as early warnings, as well as predict where trends will go and how they change over time.

Using several Natural Language Processing and Statistical tools, this paper examines the case study of WallStreetBets and Gamestop from December 2020 to February 2021

II. Data

The first set of data were collected from the Reddit forum r/wallstreetbets using python pmaw API see [2] 955,967 comments from December 8th, 2020 to January 4th, 2021. 57,963 comments mentioning GME or Gamestop were then selected from the original dataset, along with the date-time and "score" of each comment (Reddit has a up and down voting system, the score is the net result). The only missing data were on January 25th and 26th due to outages in the pmaw API.

A second novel data set was created from the titles of 2,827 English newspapers mentioning Gamestop or GME.

Both data sets were then cleaned by applying tokenization, stripping all non-alpha numeric characters, stop-word removal and Stemming the comments. The comments were then aggregated by day so that each day formed a large document. A few artifacts remain as a result of the formatting of the data from the api.

Frequencies distributions of each day were then calculated using the nltk library.

III. Exploratory Data Analysis

A. Frequency distributions

The frequency distributions in Table 1 contain similar top terms, notice Reddit is mentioned in the newspaper dataset as many articles reported on the emergence of the GME bubble that originated on Reddit.

Reddit	Frequency	Newspapers	Frequency
gamestop	.0384	gamestop	.0447
https	.0100	stock	.0417
stock	.0093	market	.0204
buy	.0092	ET	.0150
like	.0090	In	.0139
gme	.0080	trade	.0127
shares	.0079	gme	.0097
short	.0064	robinhood	.0096
people	.0060	street	.0094
would	.0056	R	.0094
get	.0056	play	.0091
money	.0050	wall	.0090
going	.0048	update	.0086
amp	.0046	hourly	.0084
price	.0046	reddit	.0084
stop	.0046	share	.0082

TABLE I
Comparison of the datasets top frequency terms

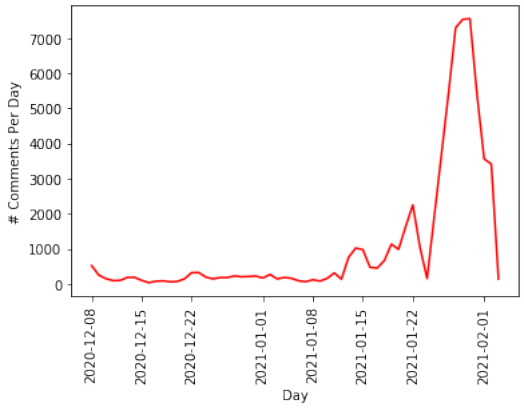

Fig. 1. Comments Per Day

B. Comments Per Day

Figure 1 shows the number of comments made over the time frame of the data set. As with many speculative manias the ”burst” of the bubble is followed by a sharp decline at the day of the peak, January 29th, there were 7,535 comments, yet by Febuary 3rd there were only 135 comments the entire day. This also roughly coincides with the peak of Gamestop Google search trend volume [4]

Next after finding the average comment score by day the relationship between comment score and number of comments was plotted. Score appears to lead the number of comments.

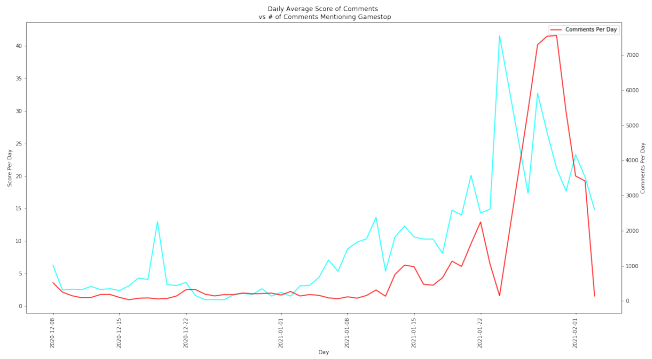

Fig. 2. Score and Comments Per Day

C. TF IDF

See IV-B for more on TF IDF. After applying the method the weighed terms that appear most often before and after the peak of posting on WallStreetBets Occurs. There is not much discernible difference but sell becomes more common relative to buy after the peak crashes. Figure 3.

This relationship was further explored by calculating the ratio of the word/buy to sell for each day.

Figure **??** does seem to confirm after the peak (day 53) that selling was mentioned more often but the data is very noisy. It also does not confirm that sell is a word that indicates negative or positive sentiment just that it was found more often after the peak.

Before

	TF-IDF
gamestop	0.715065
stock	0.172712
buy	0.171867
like	0.168548
gme	0.142468
shares	0.140543
short	0.112694
people	0.111035
would	0.104898
get	0.104241
money	0.092845
going	0.088039
stop	0.087021
price	0.082247
market	0.082137
go	0.082090
sell	0.078271
company	0.076862
amp	0.076533
make	0.073762
one	0.072291
even	0.067939
think	0.067704

After

	TF-IDF
gamestop	0.670644
ðÿ	0.387580
gme	0.104516
stock	0.104516
amp	0.095806
amazon	0.091451
company	0.091451
sell	0.087097
hold	0.087097
people	0.082742
html	0.079567
even	0.078387
amc	0.074032
buy	0.074032
fucking	0.074032
game	0.074032
going	0.069677
also	0.069677
10	0.067326
long	0.065322
cohen	0.065322
one	0.065322
make	0.065322
100	0.061206

Fig. 3. TF IDF Score Words Before And After The Peak of GME Mania

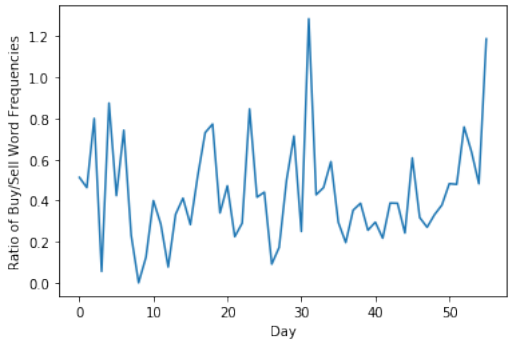

Fig. 4. Caption

IV. Methods

A. Measuring Entropy

The concept of entropy has been applied to NLP papers in economics as a measure of disorder and as a proxy for ”narrative fragmentation” [1].

Entropy is maximized when the probability distribution is uniform. So if there is more narrative fragmentation and divergence of shared sentiment entropy would be expected to be higher. Each day’s frequency distribution of words was used to calculate the daily entropy using equation 1 which shows the formula used to calculate the entropy.

$$-\Sigma_v p_i * \log_2 p_i \tag{1}$$

B. Measuring Similarity

In order to measure how similar the text of each day is a separate measure besides entropy is needed. In order to

calculate the change daily, cosine similarity is used after applying the TF-IDF method.

Where TF is the term frequency adjusted by Inverse Document Frequency (IDF) so that the frequency of a term in a document is then multiplied times the IDF or the log of how many documents contain the term.

Each document or in this case day's worth of comments are then converted in vectors where each word has a TF IDF score. Then the cosine value of an angle between two vectors of each document weight (WD) and the weight of the keyword (WK) is found. The result is that each day can be compared to every other day and roughly compared, because TF IDF is not sensitive to the magnitude of the vectors this approach works well with days that vary highly by number of comments. 10 Shows the pairwise distance of each day in the dataset.

TF IDF worked better as a measure than Doc2Vec see [1]

Two outliers, day 7 and 8 (visible as the darkest line in 10 were removed from dataset before calculating the daily change in cosine similarity because their low similarity score was due to a lack of comments on those days.

C. Vector Auto Regression

A Vector Auto Regression can be used to model a stochastic process. It is useful in modelling variables over time and determining the relationship with a minimum of required information. In the case of this data set a linear model would be useless and the VAR overcomes this.

Economic work has used VAR models of different time series as the approach lends itself to multivariate time-series. Previous work has modelled many macro economic variables such as the simultaneous interactions between the stock and bond markets. [5] Macro variables have also been used in VAR models together with stock market data. [5].

The VAR model can be written as:

$$y_t = c + A_1 y_{t-1} + ... + A_{y-p} + e_t \tag{2}$$

Where t is the time, c is the constant, and p is the number of "lags" in the equation. If p is 3 lags then the equation for a variable will include all variables including auto-regression of the earlier lags as terms in the equation.

One of the key assumptions in a VAR is that the data is a stationary stochastic process with the same first moment and second moment at each time. None of the variables used from the data in this paper were stationary. A traditional method used in this case is if possible to difference the data. Taking the VAR in difference of the data and applying the Augmented Dickey-Fuller Test confirmed the first difference was then stationary.

The differenced variables used were:

- Number of Reddit Comments Per Day
- Number of News Publications Per Day
- Entropy of The Comment Frequency Distribution Per Day
- Average Score of The Reddit Comments Per Day
- Cosine Similarity of The Reddit Comments Per Day

Next, using the Akaike Information Criterion to select the p "lags", 6 was the optimal number. The VAR model has

[1] $https://cs.stanford.edu/\ quocle/paragraph_vector.pdf$

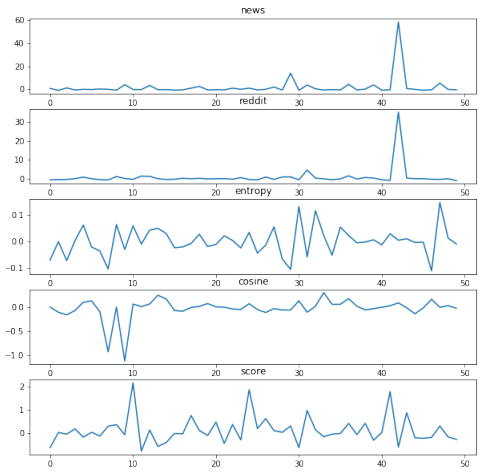

Fig. 5. Variables used in VAR.

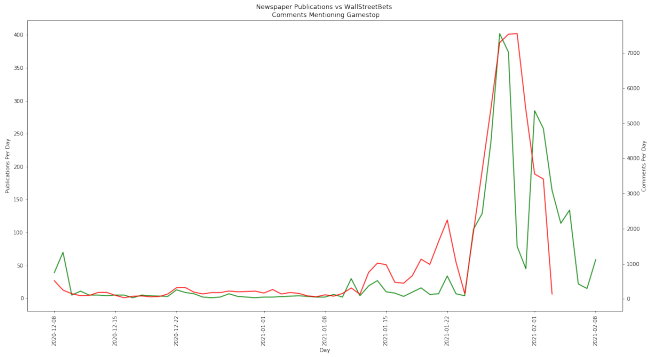

Fig. 6. Newspaper Publications in Green

five eqautions, one for each variable. Since the endogenous variable of interest is the number of Reddit comments per day the terms for that equation were observed. Of the terms based on the six lags none were significant at the .05 level.

The correlation matrix of residuals figure 7 show the residuals of the number of Reddit comments per day vs the number of news publications per day as being .9846. Considering the residuals are highly correlated and their terms are not significant at any lag > 0 , Granger Causality can be ruled out. A test was performed to double check and there was no convincing evidence to reject the null hypothesis there was no Granger Causal relationship at the .05 significance level. Without differencing the data beforehand earlier results did show a Granger Causal relationship that ran one way from the Reddit variable to the news variable.

That result should not be taken seriously before differencing the data for numerous reasons. Considering the time series was not stationary and most data points seem to follow a similar pattern until the last few days when variance and mean are much higher than earlier, it is likely that the terms are a spurious correlation, especially given that the lags where the Granger Test was significant was 4. As the non-differenced Figure 6 shows, in light of this plot the lag of 4 can be seen to not make sense visually.

This plot also seems to explain the extremely strong corre-

lation between the residuals of both the news and Reddit data. Any equation fit should have about the same change per day (differenced) for both news and Reddit.

	news	reddit	entropy	cosine	score
news	1.000000	0.984556	-0.134629	-0.098212	-0.272258
reddit	0.984556	1.000000	-0.085325	-0.023594	-0.240308
entropy	-0.134629	-0.085325	1.000000	0.664366	0.286895
cosine	-0.098212	-0.023594	0.664366	1.000000	0.111998
score	-0.272258	-0.240308	0.286895	0.111998	1.000000

Fig. 7. Correlation Matrix of Residuals

However, the term with the smallest p value with six lags in the Reddit equation was Score with lag 1 =

$$Lag1Score * 6.45$$

Based on this finding Figure 2, the plot of the differenced variables, and Figure 8 which shows score leading Reddit posts by about a day, I ran another VAR test with just those two variables and an optimal lag of 1 instead of six based on the AIK for just these two variables. (it is important to note with 6 lags and 5 variables odds of a spurious inference increase)

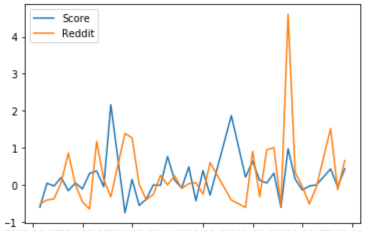

Fig. 8. Differenced Average Score Vs Reddit Comments Per Day

Results For VAR			
	coefficient	standard error	prob
constant	0.348891	0.719795	0.628
L1.score	3.379216	1.173025	0.004
L1.reddit	0.019379	0.137521	0.888

TABLE II
VAR TABLE

The results of the VAR in Table II show it is worth testing the score variable for Granger Causality. Also important is that the L1.reddit term, or in otherwords the term for auto-correlation when the lag is 1 is not significant and also has a small coefficient. Which means likely there is no auto-correlation with score which soldifies the robustness of the finding.

V. RESULTS

A. Granger's Causality

H_o : Score does not Granger Cause Reddit Comments
H_a : Score does Granger Cause Reddit Comments

Results For Granger Causality	
Test	p-value
F	.0060
Chi^2	.0029

TABLE III

Based on the results for Granger Causality in Table III, we have convincing evidence that score does Granger Cause Reddit Comments and we can reject the null hypothesis at the significance level of .05.

The causality only runs one direction, when testing the variables reversed the Chi p value is = .4. The result seems to be then that Score is a leading indicator of Reddit by a Day.

This result demonstrates a phenomenon that could be important to understanding the GME mania on Reddit. Original highly "upvoted" posts like the one by $u/deepfuckingvalue$ that lead to the influx of millions to the WallStreetBets subreddit were followed by millions of users. Clearly there are fluctuations day by day and it is possible that high quality or high enthusiasm bring a surge of comments that either decrease quality or lead to boredom. there could be other variables that explain both score and comments as well.

This finding is novel to the best of our knowledge, and could be an important indicator for predicting the behavior of WallStreetBets as well. In other work that looked at google search trends and volume traded in GME in order to predict the price for example, search trends exhibit a similar pattern as the number of comments, which means score leads both and could be a better indicator used to predict the behavior of WallStreetBets.

B. Entropy

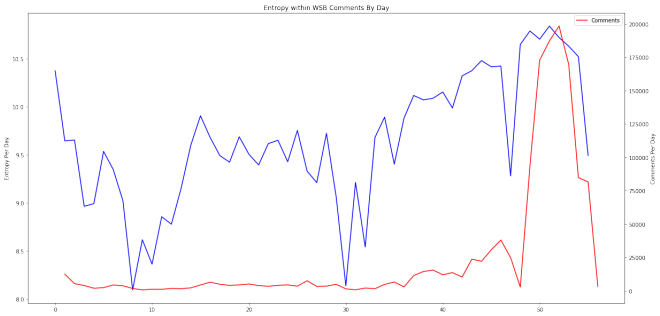

Fig. 9. Within Day Entropy (Blue)

The calculated entropy did not Granger Cause Reddit comments or have a significant term in any of the VAR models. Previous work on business cycles found that entropy decreased during expansions and increased during recessions. [4] At least in this case it is clear stock price bubbles do not have a decrease in entropy and the highest entropy was found during days with the highest volume of posts. It is is possible that shocks or highly unpredictable socially driven behavior like the GME run-up in price are related to entropy but more work would need to be done, and more data used instead of just this specific case study, as the previous work was done over a longer time period.

The measure could be adjusted so that it is not calculated over all lemmatized words, and instead a threshold is set to reduce the amount of noise. Alternative measures of change in a text distribution could also be used along side entropy and used a feature in future work. Other papers have used sentiment analysis of Reddit, if a classification model were

used along with entropy for each class of sentiment then emerging trends could be identified.

C. Cosine Similarity

Figure 10 shows the pairwise distance for each day in the data set. The two outliers removed can be visualized by the dark line in the upper right corner of the correlation matrix. An issue with this measure may be that the higher number of post days are more similar to most days, even days that are much more temporally distanced. This could make the measure less informative on a day by day basis to compare two days and overemphasize the size of total comments in a day. It still is better than euclidean distance, but better alternative such as using pre-trained embeddings on each day could be better.

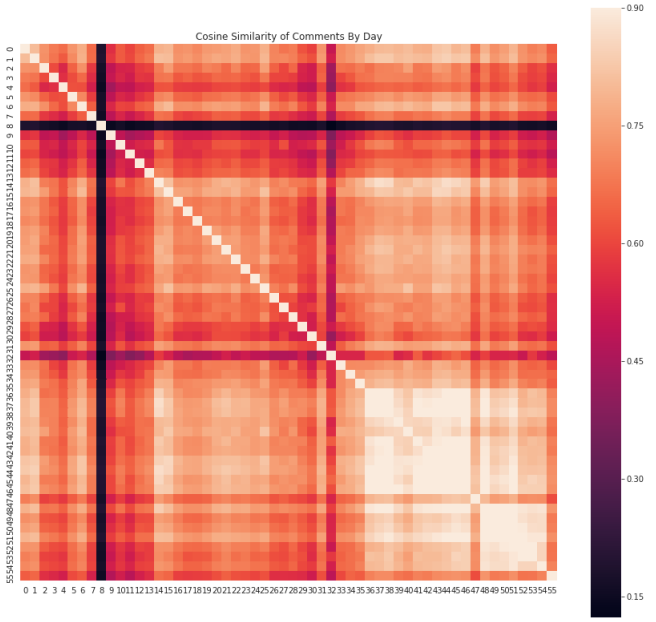

Fig. 10. Shows Cosine Similarity Matrix of Days

VI. Conclusion

In conclusion, while the approach used based on entropy and cosine similarity in this paper did not yield useful results yet, statistical testing based on calculating daily score of "upvotes" was found to be, to the best of our knowledge, a novel indicator of the trends of WallStreetBets.

As explained in the last section this builds on previous work using google search trends.

The test indicating newspapers and Reddit had roughly the same frequency of mentions of Gamestop despite the activity resulting from the coordinated behavior of investors on WallStreetBets. However, this study only looked at a slice of time, it is unlikely news picked up coverage of GME until long after the hype began on Reddit. Another important limitation of this study is that all tests were conducting on daily variables. In markets a day is a long time when trades are conducted on tiny time scales. For instance the paper on Google search data found a relationship that disappeared after an hour. In this context finding an indicator that is lagged a day behind another is useful.

Previous work on entropy and macro-economic fluctuations seems to demand further work that could look at the relationship between entropy of various social media platforms and self-fullfilling expectations.

There seems to be a lot of potential applications of temporal text datasets and difference metrics that was not explored in this paper, which has some very fundamental limitations because of the small number of days analyzed whether or not the methods employed are useless or not.

Finally "*Apes Together Strong*" -WallStreetBets aphorism found in bigrams

References

[1] C. Bertsch, I. Hull, and X. Zhang, Narrative fragmentation and the business cycle, Economics Letters, vol. 201, p. 109783, 2021, doi: https://doi.org/10.1016/j.econlet.2021.109783.

[2] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The Pushshift Reddit Dataset.Proceedings of the International AAAI Conference on Web andSocial Media, 14:830–839.

[3] P.Bafna, D. Pramod, A Vaidya, "Document clustering: TF-IDF approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 61-66, doi: 10.1109/ICEEOT.2016.7754750.

[4] Vasileiou, Evangelos and Bartzou, Eleftheria and Tzanakis, Polydoros, "Explaining Gamestop Short Squeeze using Intraday Data and Google Searches", March 16, 2021. http://dx.doi.org/10.2139/ssrn.3805630

[5] A. Nasseh, J. Strauss, "Stock prices and domestic and international macroeconomic activity: a cointegration approach", The Quarterly Review of Economics and Finance 40 (2000) 229 –245.

[6] Richmond Federal Reserve Bank, https://www.richmondfed.org/publications/research/eco 13