Like 4chan found the Shannon Entropy: An information theoretic analysis of online discourse

Colton Cunov
(Dated: March 9, 2021)

I. INTRODUCTION

Reddit's r/WallStreetBets is an online forum dedicated to discussing stock and options trading and noted for its role in the meteoric rise of GameStop Corp. stock price in early 2021, which resulted in short position losses to US firms of $70 billion [1]. The forum is associated with a unique culture of colorful language, self-deprecating humor, and pride in extreme capital gains and losses. At the start of the year it consisted of about two million users, exploding to over nine million due to media coverage of the so-called short squeeze.

To examine the effects of this drastic change in user base from an information theoretic perspective, I calculated the correlation information based on post comments for various correlation lengths before, during, and after the rise in user base. My hypothesis was that a drastic increase in the number of participants fueled by a single event would result in less information at various correlation lengths. Additionally, I generated sample comments with remarkable integrity.

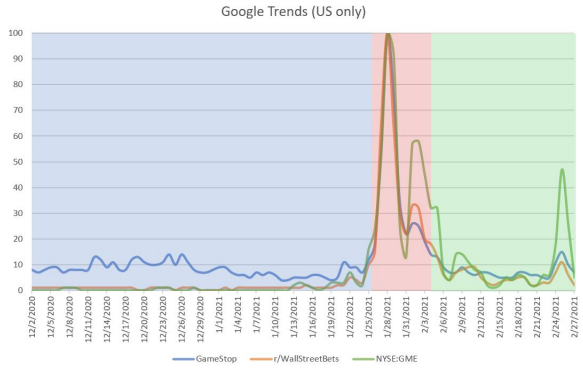

FIG. 1. Google Trends results for the selected keywords, separated into intervals defined by January 25, 2021 and February 5, 2021.

II. DATA COLLECTION AND PRE-PROCESSING

I began by examining Google Trends for the keywords: GameStop, r/WallStreetBets, and NYSE:GME. I chose to bucket the data into intervals separated by January 25, 2021 and February 4, 2021 (Fig. 1). In what follows, *Before* refers to data prior to January 25, *During* refers to data between January 25 and February 4, and *After* refers to data after February 4. To gather a representative sample, I chose to collect highly upvoted (a measure of favorability) data. I used Reddit's API to gather the top posts from the past year and selected the most upvoted comments. Because the focus of comments are more likely to adhere to the topic of the post if they are higher in the comment tree, I only collected top-level comments. This resulted in 1.35 ± 0.05 million words in each range from over 1,300 posts.

Reddit uses Markdown to format text, so strong consideration was given to pre-processing the data. I removed all URLs and Markdown syntax, as well as corrected for missing punctuation and first letter capitalization. While emojis and special characters can add important context on the forum, they are often used to excess and were removed for simplicity. Resource and time limitations restricted the removal of some potentially impactful items (e.g. ASCII art).

III. METHODS

Words were used as symbols to capture syntactical information, which resulted in $46,500\pm3,000$ distinct symbols in each dataset. Additionally, punctuation (e.g. *.*, *?*, *(*, *"*) were considered symbols.

To ensure an appropriate sample size was collected, resampling was performed on all datasets by separating them into 25 smaller samples. The average and median block entropies reflected the entropy for the unrestricted datasets.

Block Entropy of correlation length m was calculated as $S_m = \sum_{x_1,...,x_m} p(x_1,...,x_m) \log_2 \frac{1}{p(x_1,...,x_m)}$. This was used in the calculation of correlation information $k_m = -S_m + 2S_{m-1} - S_{m-2}$. I set $S_0 = 0$ and $k_1 = \log_2 \nu - S_1$, where ν is the number of symbols.

To generate novel text, I combined the datasets and used a Markov Chain. The probability of choosing the next symbol x_{m+1} given the previous m symbols $x_1...x_m$ is $p(x_{m+1}|x_1,...,x_m)$, with x_{m+1} being chosen randomly with probability α. I experimented with different lengths m, starting seeds, and randomness parameters α.

IV. RESULTS

The correlation information k_m decreases significantly when $m > 3$ (Fig. 3). For $m = 2$ *Before* and *During* the buildup were nearly the same, while *After* was lower. For $m = 3$ *During* and *After* were similar while *Before* was higher.

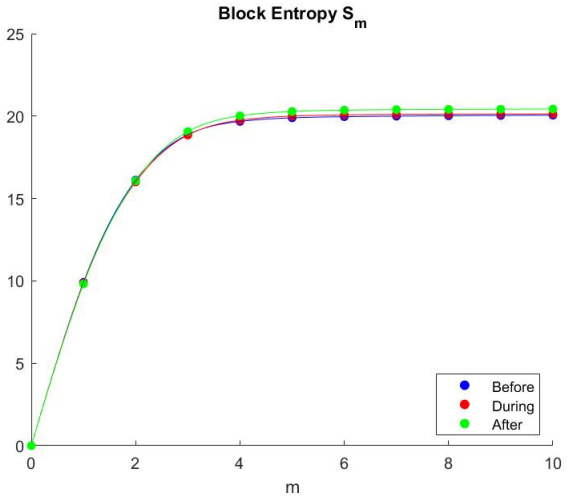

FIG. 2. Block entropy for correlation lengths up to 10.

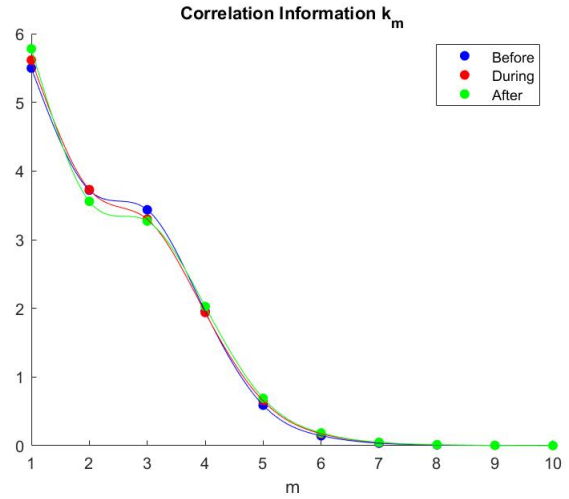

FIG. 3. Correlation Information for correlation lengths up to 10.

Following are some samples of generated text with $\alpha = .001$, and some whitespaces removed for readability. The seed is italicized.

- $m = 3$: *like the stock*. WE LIKE THE STOCK! Troll level achieved. Love it! That's a huge gain. Almost makes me want a Tesla model S for each share of GME in the green. We like the stock. Thats leading the answer. YES!!
- $m = 3$: *the world is* proud. GUH oh sh**. This was mint. This is a long hold company. You guys have to remember we're a small group of people who are smarter than us, then we sell. Hold the line! Absolute cringe. Mods please add a song
- $m = 2$ *i just* wanted to. I ain't budging. Bought at the music at the bass boosted BRRRRRRRRRRRRRRRRRRRRRRRR. If this hits $1,000. The climb has begun. Today and now we're FINALLY getting exposure. Down 80%? **** this couldn't have described my rollercoaster trading career in
- $m = 2$ *is not* financial advice! LOLOLOL!!!!. ****. My strategy is better than you need a list of affected people to buy back in there claiming they are ****ing adamantium! Ride the god**** moon. Well, with a profit shorting. Lets go! Ret***
- $m = 2$ *you can* slide in to selling tomorrow, I also sold my AMC FD's have any questions or concerns. My comrades failed me. I don't understand this post at least be buying anymore AMC. I wish. Why you cashing out tanking the whole time we played our trap
- $m = 2$ you are a couple people here? The best yolo trades. If someone is sticking their **** around, I AINT ***ING SELLING. Gl hf. That was a coup! Just keep your options. The cost to purchase these contracts to buy so much more sense to liquidate

V. DISCUSSION

Compared to *Before*, information correlation for $m = 3$ was lost *During* the buildup in user. Information correlation for $m = 2$ was lost *After*. For $m > 3$ the differences in correlation information were less significant. Because there were $46,500 \pm 3,000$ distinct symbols in the datasets, notably higher than most English corpora (Charles Dickens' *A Tale of Two Cities* has under 11,000 unique words), I hesitate to make conclusions on the grounds of questionable data.

Text in social media, especially those with anonymity and unique culture, are difficult to structure. I had difficulty balancing the removal of unimportant "spam" and preserving the integrity of the data, which may have impacted my findings. For example, three of the top ten most frequent 3-symbol sequences were (- - -), (*! ! !*), and (. . .). Additional factors may include spelling errors, epizeuxis (e.g. "never, never, never sell"), stretchable words (e.g. "GME to the moooooooon!!") [2], and inserting spaces between letters for emphasis.

In my data, 25% of comments were five words or less. This may have been the reason the correlation information begins to decrease precipitously when $m > 4$. An example is the now ubiquitous "I like the stock", which was commented hundreds of times on some posts.

Due to the use of words as symbols α can have a magnified effect if a seldomly occurring word is selected, resulting in low comprehensibility for the rest of the generated text. A very low α is appropriate for this dataset, as is $m < 4$ to prevent reproducing individual comments due to the large decrease in correlation information for $m \geq 4$.

[1] S. Rao and T. Adinarayan, Losses top $70 billion on short positions in u.s. firms - ortex data, Yahoo! Finance (2021).

[2] T. J. Gray, C. M. Danforth, and P. S. Dodds, Hahahahaha, duuuuude, yeeessss!: A two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings, PloS one **15**, e0232938 (2020).