To the Moon 🚀🌙: Discourse on r/wallstreetbets

Joe Marlo, Nicole Salani, Bilal Waheed

Warning: This paper contains explicit language

Introduction

In early 2021, the stock price of video game retailer GameStop (GME) increased substantially in parallel with the popularity of the internet forum 'wallstreetbets.' It was widely covered in major finance and general news outlets (Bloomberg, Wall Street Journal, New York Times, CNBC) due to the controversial players and their perceived motivation. It is alleged that members of the r/wallstreetbets subreddit were -- and perhaps continue to be -- involved in an unprecedented short squeeze and pump-and-dump scheme that resulted in GME stock price increasing by 20x, a hedge fund reportedly losing more than $4 billion, and a congressional hearing. It spotlighted many current social, political, and economical issues such as online echo-chambers, internet mob-mentality, the obfuscation of financial regulation, the bi-modality of the wealth distribution, and democratization of finance by technology firms.

We conducted a supervised learning text analysis of the subreddit to predict the popularity of comments (via number of upvotes) in the subreddit. The goal of the analysis is to identify text trends (e.g. words, phrases) that drive the popularity of comments. Our intuition is that comments with high upvotes are a strong proxy of the subreddit's latent value system. Examining features that contribute to the prediction of upvotes can potentially give insight into this value system and uncover interesting topical trends specific to the subreddit. We believe this furthers our understanding of online social discord and community cohesion. We will not be directly exploring the legality, effectiveness, or morality of the saga.

We fit statistical and machine learning models on 6.9mm scraped comments made between December 1, 2020 and April 15, 2021. The best model is a boosted tree model; however all fitted models demonstrate weak predictive power measured by RMSE. The boosted model identifies a range of variables as ranking high in importance: sentiment, previous day's GameStop stock price, and hour the comment was made. Of the top ten variables by importance, only one is a key phrase, "retard", which is an endearing remark made to other members of the subreddit. We discuss the results in greater length in the "Results & implications" section.

Background

Wallstreetbets was started in 2012 as a refuge for individuals looking to light-heartedly discuss single-stock trading instead of the usual conservative index funds. The subreddit grew steadily reaching 1.7mm subscribers by December 2020 and by which time it had also developed its own vulgar vernacular. The subreddit self-describes as “like 4chan found on a Bloomberg terminal.” The growth was fueled partly by the increase in retail trading -- made possible by the pandemic and commission-free brokers such as Robinhood. The former left many at home with extra time and some -- including the subreddit’s core demographic -- with extra cash due to stimulus checks and reduced discretionary spending.

In the summer of 2020, Keith Gill -- also known by his handle DeepFuckingValue -- started posting videos to YouTube describing his fundamental analysis of the GameStop stock. He hypothesized the stock was severely undervalued at its then price of ~$4, contrary to the typical institutional wisdom that GameStop is a dying brick-and-mortar business in an increasingly digital world. He highlighted the stock is highly shorted which -- if he was correctly forecasting -- made it susceptible to a short squeeze. A short squeeze occurs when a highly shorted stock price rises and forces shorts to cover their positions by selling the stock, driving the price higher and repeating the process for other shorts.

Gill continued posting videos and stock analysis to YouTube and r/wallstreetbets where his thesis garnered popularity. By late December, the stock price had risen to $20. By mid-January, it had hit critical mass and the short-squeeze was under way. The stock price peaked January 27 at ~$350 and subreddit membership crossed 5mm, reaching 8.5mm only six days later. On the other side of the trade, the hedge fund Melvin Capital reportedly lost $4bn or half of it’s fund and required a bailout from fellow funds including Citadel.

The popular commission-free brokerage trading app Robinhood came under fire on the 28th of January after halting buys of GME (but not sells) due to clearinghouse restrictions on collateral requirements. Many internet denizens allege the halt was due to Robinhood’s relationship with Citadel Securities (a sister firm to Citadel) that purchases order-flow from the brokerage. A congressional hearing was held on February 18 where Gill, Melvin’s Gabriel Plotkin, Citadel’s Ken

Griffin, and Robinhood's Vladimir Tenev spoke. The GME price climbed again in March and, as of May 2021, remains well above the fundamental valuation that Gill set in his original thesis.

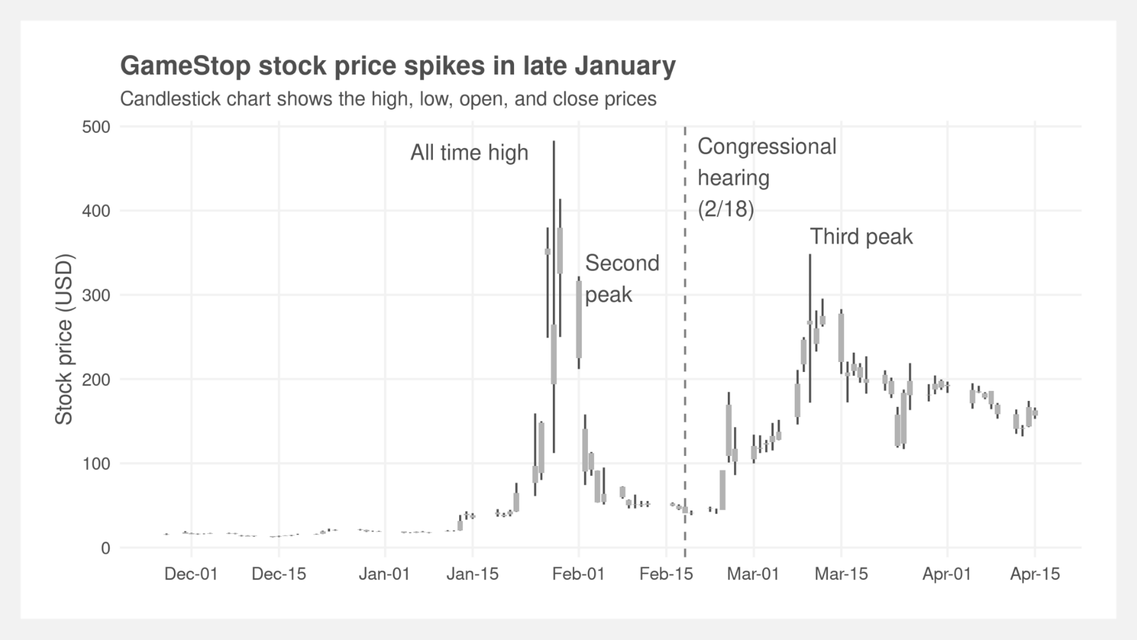


Figure 1: GameStop stock price

Little academic attention has historically been given to specific subreddits however the recent extreme growth in r/wallstreetbets -- 10mm members as of May 2021 -- has garnered interest. Boylston et al. (2021) studied the growth of the subreddit and found that the core humor promoted "in-group cohesion and in providing an unconventional third place for traders." Bradley et al. (2021) claim that WSB members are skilled traders as evidenced by increases in stock prices in the days following "due diligence" posts by members -- a specific subcategory of the subreddit devoted to detailed stock analysis. Eaton et al. (2021) found the opposite: that Robinhood -- the previously popular brokerage for subreddit members -- users' trades are unrelated with future stock prices and that these investors "behave as noise traders."

More attention has been paid to reddit as a whole. Horne et al. (2017) implemented supervised machine learning models to predict the popularity of posts and identify key features across different subreddits. They found it is possible to predict the popularity of comments with some certainty using ridge regularized regression models and carefully crafted features. Their resulting key features are sentiment, fluency based on use of specific 'insider' language, comment-to-post relevancy, newness (e.g. new vs. stale memes), and user flair. They used precision-at-k and Kendall-tau distance to evaluate the models. In some sense, our approach mirrors Horne et al.'s framework but with a focus on one subreddit community, r/wallstreetbets.

Thukral et al. (2018) take a different angle in their analysis of social interactions in online communities by looking at how temporal patterns and group dynamics influence popularity of posts based on both the number of comments and votes. They find that popular posts exhibit early bloomer properties in that they gain the most traction (comments and votes) within the first day of their lifespan and that only a small percentage of popular posts exhibit steady or late bloomer characteristics where engagement activity is constant or increases over the lifespan of a post, respectively. In our analysis, we also investigate whether popularity of comments (based on upvotes) has temporal properties by including the time (hour, day of the week, weekend or market day) when the comment is made. These temporal features are also generally important to consider given that external 'market' events seemed to drive engagement activity as well as underlying themes of some of the discussions' within the r/wallstreetbets subreddit.

Barker and Rhode (2019) leverage topic modelling to explore major themes related to e-cigarettes and vaping on reddit. Their approach is guided by the understanding that examining digital groups dedicated to specific discussion topics provides useful insights for monitoring communities that share similar e-cigarette beliefs. This is because reddit users have the ability to actively curate lists of their favourite subreddits and so the posts and comments that users encounter are likely to be from preferred topic discussion groups. With this framework in mind, we leverage topic modelling to get a sense of the underlying themes or issues that permeate throughout discussions and ultimately some insight into users' shared beliefs or values.

Methods and approach

We center our approach to this analysis on building a predictive model of a user's comment based on the comment's text. We apply five supervised regression algorithms (linear regression, k-nearest-neighbors, decision tree, random forest, and gradient boosted trees) with feature selection defined by a LASSO model. We wish to create an interpretable model to attempt to understand which words and/or phrases drive upvotes and propagate ideas and sentiments. Extensive feature engineering was conducted in conjunction with a bag-of-words approach to the comments.

Approaching this problem as a regression is intuitive because we ultimately want to quantify discourse around the GME saga, and gain insight into the inherent value system within r/wallstreetbets. The goal is to learn about the specific language used in the subreddit that could help understand the potential “echo chamber” effect (i.e. infer which language is ‘valued more’ than others).

Data

Our data were scraped from r/wallstreetbets using the Pushshift API, a third party service that allows more extensive historical data than the official reddit API. Custom functions were required to construct queries with our required parameters and convert the resulting JSON object to a dataframe. It took approximately 10 hours to download the posts and comments. The API is rate-limited to 20,000 comments per request and 200 requests per minute. This artificially reduces the number of comments captured during the first peak at the end of January.

These data are cleaned to separate comments from posts, and remove all comments and posts that are primarily image-based (memes). Upvotes are extracted and matched to each comment as this will serve as the primary outcome variable. The subreddit has a unique vernacular -- that can be extremely vulgar at times -- so special consideration was made to account for the vocabulary through a lexicon based approach to sentiment analysis as well as identification and interpretation of key phrases as detailed out in the feature engineering section.

Discussions in the r/wallstreetbets subreddit span a wide variety of Wall Street related topics and content, so we subsetted the scraped data to posts and comments that included the keywords “GME”, “gamestop”, or “gamestonk”. We also wanted to capture authentic conversations so we filtered out any obvious bots-generated posts and comments. Following that, we removed all original posts and focused on only comments in our final analysis. Comments generally contain more substantive discussion than the original posts which are often boilerplate text created daily by moderators. The final data consisted of 6.9 mm comments that were posted between December 1st 2020 to April 15th 2021 (as shown below).

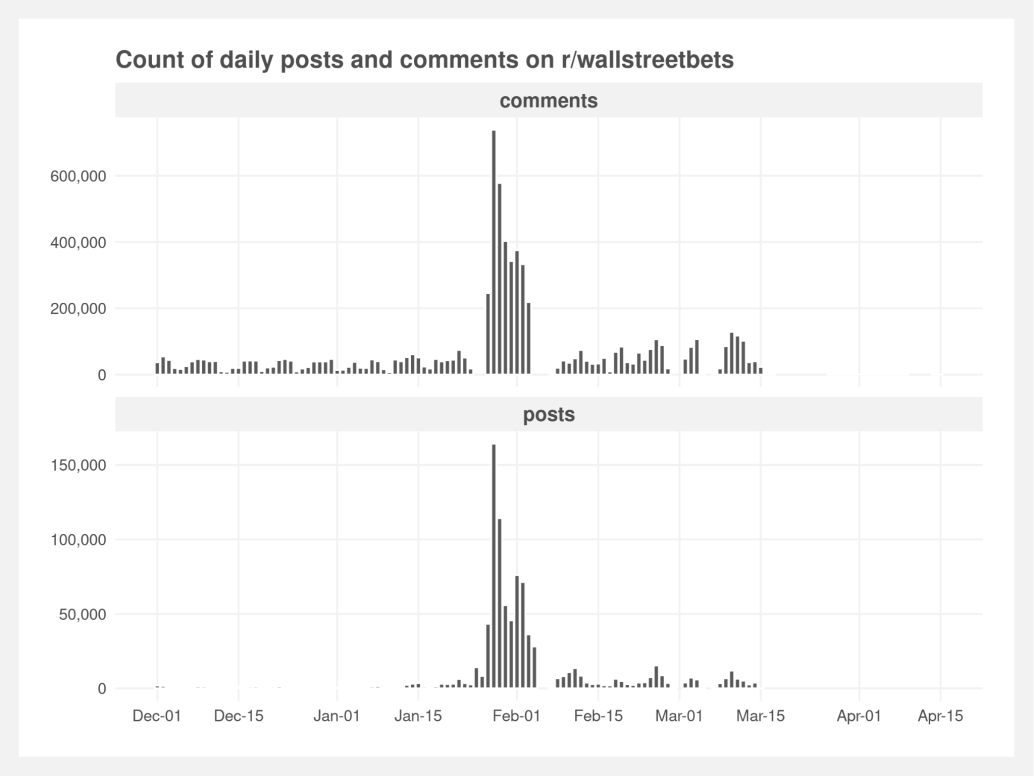


Figure 2: Frequency of collected comments and posts

Data imbalance

The data are highly unbalanced with approximately 80% of comments having less than 10 upvotes (shown below). We balanced the data by binning the comments into six bins with breaks at [-∞, 0, 5, 10, 100, 1000, 1e6] and randomly sampling 10,000 comments from each.

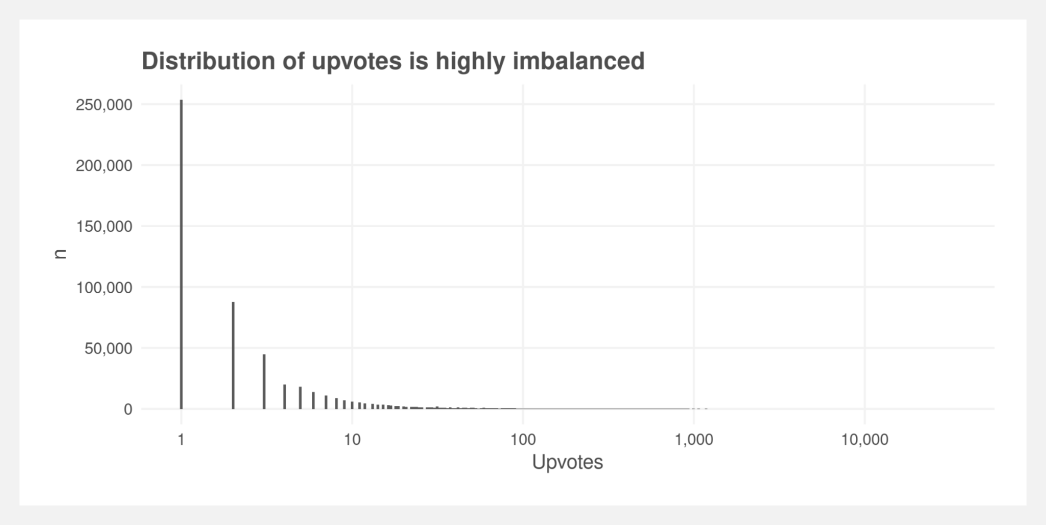


Figure 3: Imbalance distribution of upvotes

Feature engineering

To build our models, we created distinct sets of features that could be most predictive of a comment's number of upvotes. These are broadly divided into text, sentiment, and topic features. The text features served as the main features of the data and were created using a bag-of-words approach.

We converted each comment into its constituent words, after stemming and removing stop words (i.e. 'the', 'and', etc.), and allowed for each word to be a unique binary feature of comment. In addition, we created a unique lexicon of words, phrases, and emoji combinations specific to r/wallstreetbets that we felt were indicative of a comment's popularity (e.g. "to the moon", "diamond hands", "short squeeze", etc.). We included these as unique binary features. The full lexicon is provided in the attached appendix.

For sentiment, we applied the lexicon and rule-based Vader algorithm via Python to measure the overall sentiment of the comment ("positive" or "negative", inclusive of emoji's), as this algorithm is well-suited for social media posts. We modified the underlying sentiment dictionary for financial words and r/wallstreetbets key phrases and words. Four final sentiment measures are included -- positive, neutral, negative, and a "composite" sentiment score reduces the prior three to one dimension to capture the overall sentiment of a specific post.

For the topic feature, we apply Latent Dirichlet Allocation (LDA) topic modeling to cluster comments into five distinct topics. These topics uniquely distinguished comments from each other based on shared sentiments and themes (e.g. topic 1 included phrases such as "short", "stock", "squeeze", whereas topic 4 included "buy", "time", "hold", "share"). The full plot of each topics' key words and phrases is provided in the appendix. The goal is to extract inherent themes in users' comments that help predict a comment's popularity.

Apart from these three main categories of features, we also created several other features related to general interest in the r/wallstreetbets subreddit and GME stock prices. We expanded our text features to include bi-grams with greater than 100 total mentions (these represent potentially popular phrases). We previously tested uni-grams but performance as measured by RMSE was better for bi-grams. In addition we included temporal features (month, day, time), an indicator for whether the comment includes a number, if the comment was posted on a market holiday, and an indicator of whether the comment was a direct comment to the original post, or a comment on another user's comment. This is of interest because it is reasonable to assume that a direct post comment could be more visible to users and spark a larger discussion.

Another key feature about the r/wallstreetbets community is its collective response to external stock market events. To factor in how such events shaped conversation, we included previous day GME stock price data as a feature to measure the impact of real time GME stock price fluctuations on the subreddit's popularity.

Feature selection

Given the size and inherent variation of the dataset, constructing our features resulted in a feature set of over 150 features. We used LASSO regression to parse out the most important features in an effort to build a flexible and parsimonious model for adequate test set performance. Additionally, LASSO aids in avoiding overfitting and in interpretability for addressing our main research question.

The cross-validated LASSO model retained about 50 key features ranging from the actual comment text (i.e. specific bi-grams and key phrases that we deemed to be significant to a comment's popularity), sentiment scores, to temporal features including the hour of day, and the comment's topic.

Results & implications

Five models -- linear regression, KNN, decision tree, random forest, and XGBoost -- were evaluated using root mean squared error (RMSE) and precision-at-k. The former is a commonly used metric to measure the difference between the predictions from the actuals. The latter measures the overlap in the rankings of predictions and rankings of the actuals. Horne et al. (2017) used this method to assess their models and described it as the "percentage of the posts ranked among the top k as predicted by the learned model that are also among the top k posts by true scores, averaged over all posts."

Predictive performance

Based on these metrics, the XGBoost model performs the best. It is only slightly worse on RMSE than the random forest but has greater precision across many k values. We believe the precision metric is more in line with our goals -- it prioritizes the ranking of the predictions whereas RMSE prioritizes the magnitude of the residuals. In other words, we are emphasizing that highly upvoted comments are predicted above low upvoted comments, but the difference in predicted upvotes and actual upvotes is less important.

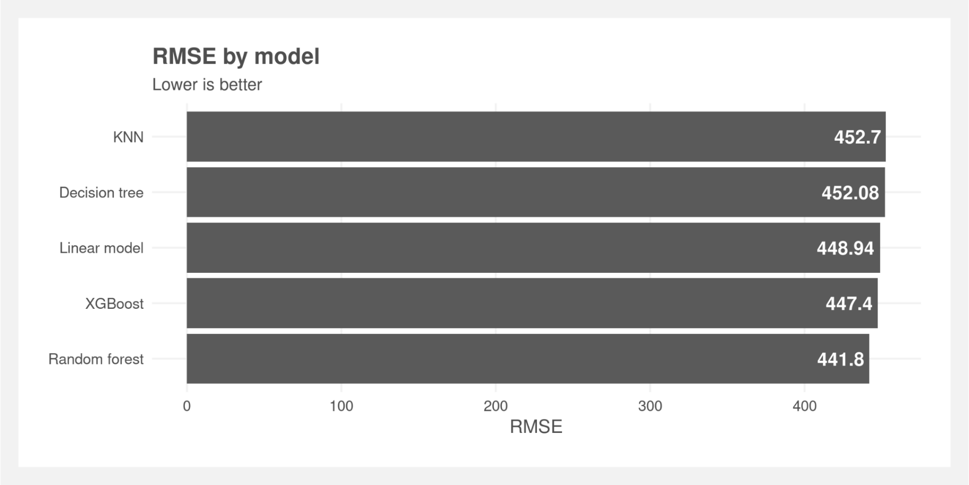
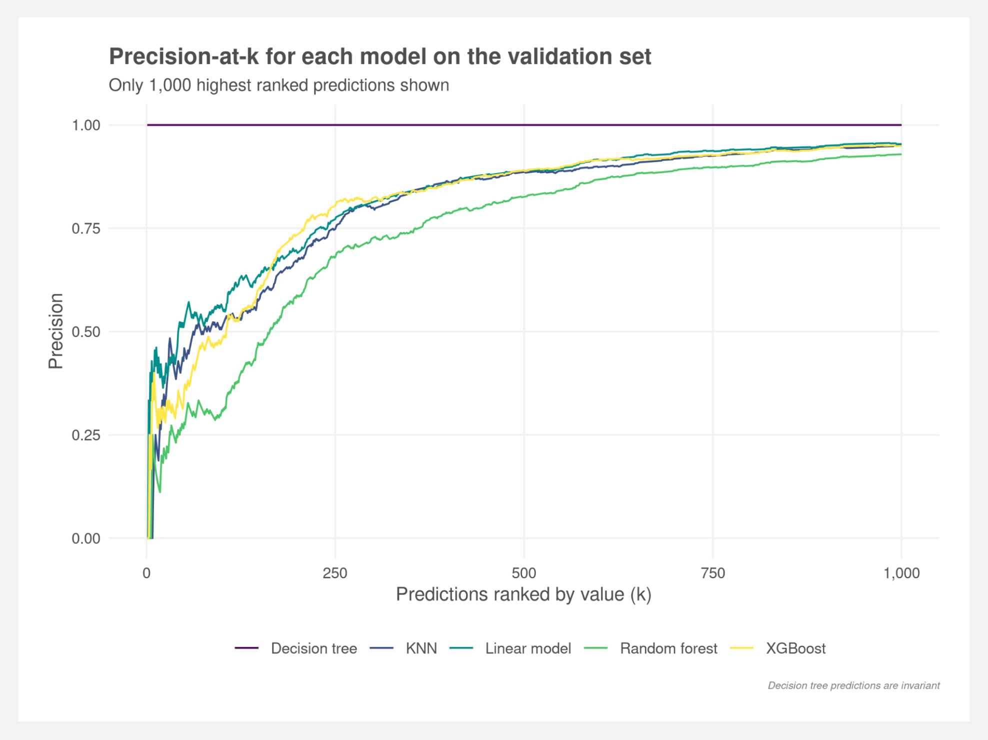


Figure 4: Predictive performance of models by RMSE



Decision tree predictions are invariant

Figure 5: Precision-at-k of models

Note that the decision tree is severely underfit and predicts the same value for every observation.

Overall, the predictive performance of all the models is poor as demonstrated by the high RMSE. This may be driven by the imbalanced dataset, poor feature engineering and selection, or because the bag-of-words approach does not capture enough information. There could also be other latent variables that are driving the popularity of comments such as timing or relevance. See the limitations for further discussion.

Quantifying discourse

Pure predictive capability does not address our original research question. Our goal is to identify trends in vernacular used by the subreddit community members associated with popular comments. We believe these trends serve as a strong proxy for community members' values. And because the comments are collectively voted on via the upvote system, looking at predictors of upvotes provides deeper insights as to which topics or issues they reward. The advantage to this supervised approach is insight into both understanding popular language and which language strongly resonate with those in the discussion.

Despite the low predictive value of the models, we believe the models may provide some insight into the subreddit. We pull the feature importance of the XGBoost model and the coefficients from the LASSO feature selection model and cross-reference them with the features that refer to the content of the post.

The XGBoost models do not lend themselves well to interpretability like linear regression. We extract variable importance -- roughly explained as the average reduction in variance after each split for a given feature. A larger value means a more important predictive feature. Of the top ten features, only three refer to actual content of the post -- the positive and negative sentiment and the keyphrase "retard," an endearing remark made to signal community inclusion.

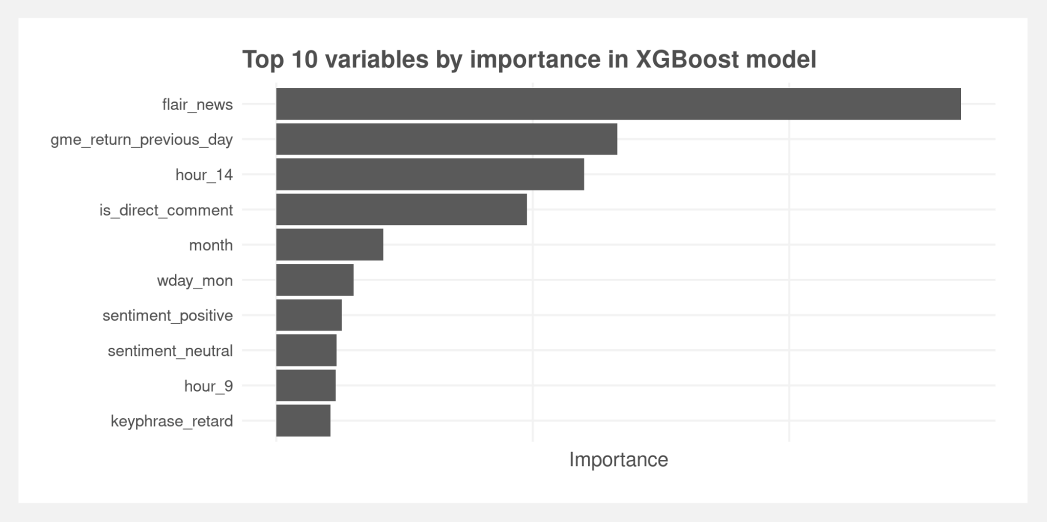


Figure 6: XGBoost variable importance

There are 13 content features within the XGBoost model -- ranked by importance in the plot below. The bigram “short position” and “GME shares” are self-evident. “Lambo” refers to Lamborghini and used akin to “longshot” where users are showing or referencing large gains as the eventual windfall will allow them to buy a Lamborghini. “Retail investors” is often a self-reference derived from institutions or the media disparaging retail investors as “dumb money.” “Hedge funds” is usually discussed in the context of the funds on the other side of the GME trade. And “gem stone raising hands” is the emoji reference “diamond hands,” a compliment indicating the user is great at trading.

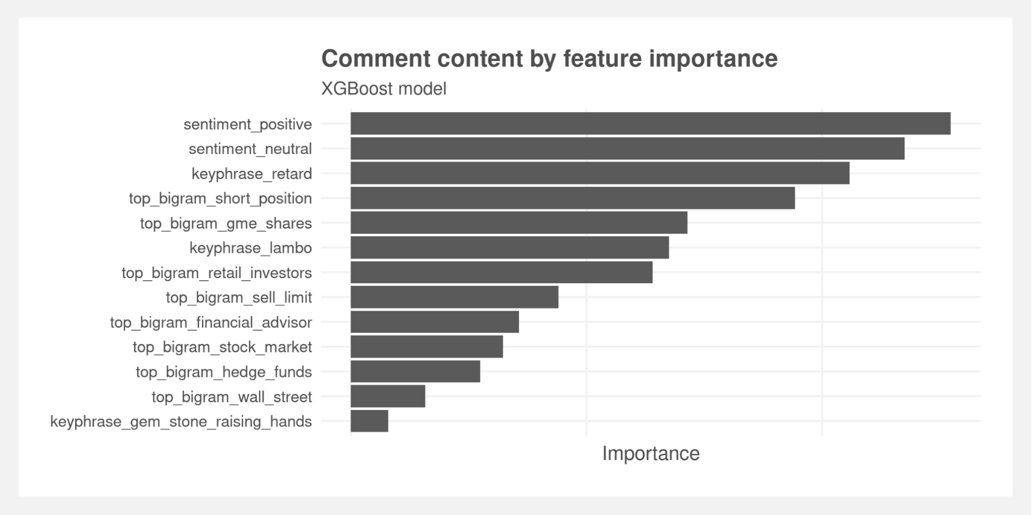


Figure 7: XGBoost variable importance of features containing content

The LASSO model contains all of the above features (since those are derived from the LASSO). The most notable ones are “Melvin Capital,” the hedge fund that lost $4bn; “autist” a remark used similar to “retard;” and “paper handed” which is the opposite of “diamond hands.” The features with the largest negative coefficient values are “hold hold”, “tendies”, and two variations of “buy GME.” All three represent positive sentiment but they are often used as filler or throwaway comments.

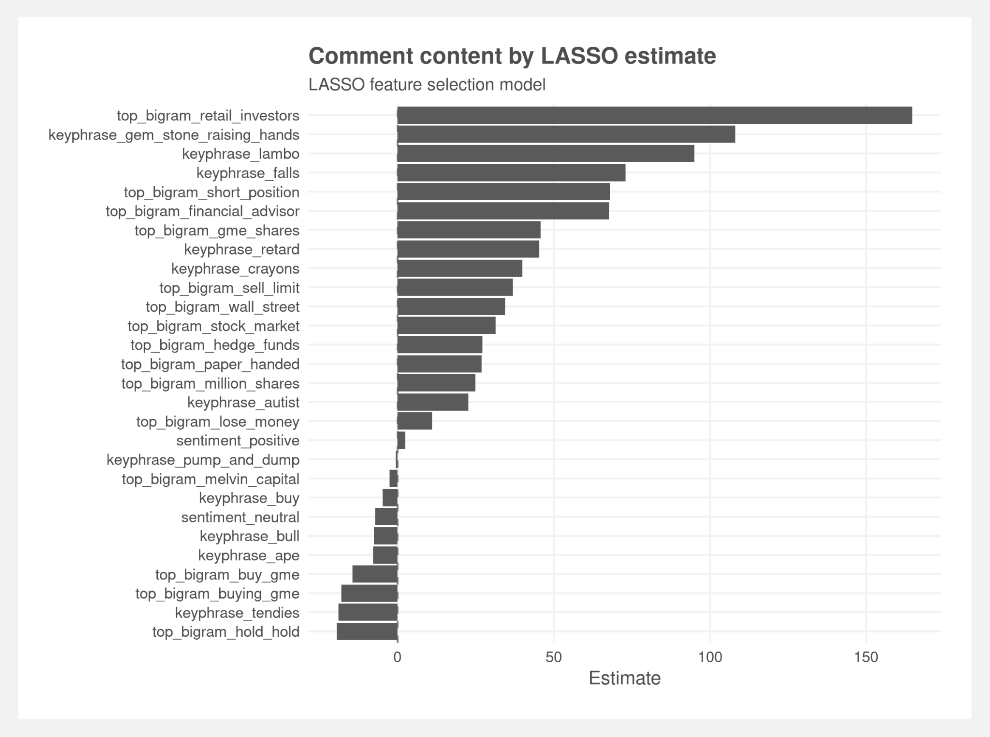


Figure 8: LASSO estimates

Limitations and further research

Our results show that predicting upvotes is a challenging task, requiring creative text analysis and extensive domain knowledge to capture key information relevant to this specific event. Despite challenges, we attempted to gain a deeper look into an interesting social media phenomenon. We explored the landscape of the r/wallstreetbets language and quantified discourse, but ultimately discourse did not prove useful in predicting upvotes. Therefore, we could not definitively identify the unique language associated with the hype of r/wallstreetbets.

We focused on the time period December 2020 to mid-April 2021 to capture the beginning stages to the end of the first bubble. There is a possibility that the subreddit's value system changed during the intense subscriber growth period of January 2021. Including data from before, during, and after this period may be conflating value system regimes. It may be important to look at the temporal evolution of discourse 'themes' and to also separately look at predictors of engagement activity based on 'upvotes.' Such an approach probably provides a more comprehensive view into collectively shared ideas or values within a 'topical' digital community like r/wallstreetbets where external market events are a big part of the discussion.

This modeling process does not account for the natural nesting of comments within other comments and within posts. We excluded posts and we included a flag if a comment was a "top-level" comment -- meaning it was in response to the original post and not another comment -- but this is not capturing the potential importance of the full hierarchy. Horne et al. (2017) included a relevance metric of the comment to the post based on cosine similarity of the texts. We may also be missing a timing element; timing of comment on another comment and the timing of the comment's content compared to popular events.

Other models may be better suited to these data. The linear regression -- which has similar RMSE and precision-at-k -- may provide more insight into the language. Models that incorporate the sequencing of text -- in lieu of the bag-of-words approach -- like a recurrent neural network (RNN) may pick up more predictive patterns than the simple n-gram approach.

Appendix

Custom lexicon

Phrase	Sentiment		Phrase	Sentiment
💎🙌	100		silverback	25
🚀	100		this is the way	25
bull	100		we like the stock	25
buy	100		elon	10
deep fucking value	100		musk	10
DFV	100		crayons	0
mars	100		degen	0
ROARING KITTY	100		JPOW	0
stronger together	100		my wife’s boyfriend	0
tendies	100		pump and dump	0
to the moon	100		retard	0
YOLO	100		retards	0
💪	50		stonk	0
🦍	50		🌈🐻	-50
Ape	50		buy high sell low	-50
autism	50		dip	-50
autist	50		drill	-50
autistic	50		gay bears	-50
degenerate	50		bear	-100
hold the line	50		falls	-100
lambo	50		paper hands	-100

moon	50		put	-100
printer	50		puts	-100
i like the stock	25		smooth brain	-100
money printer go brrr	25		weaklies	-100

Table 1: Custom lexicon and associated sentiment

Topic Modeling Analysis

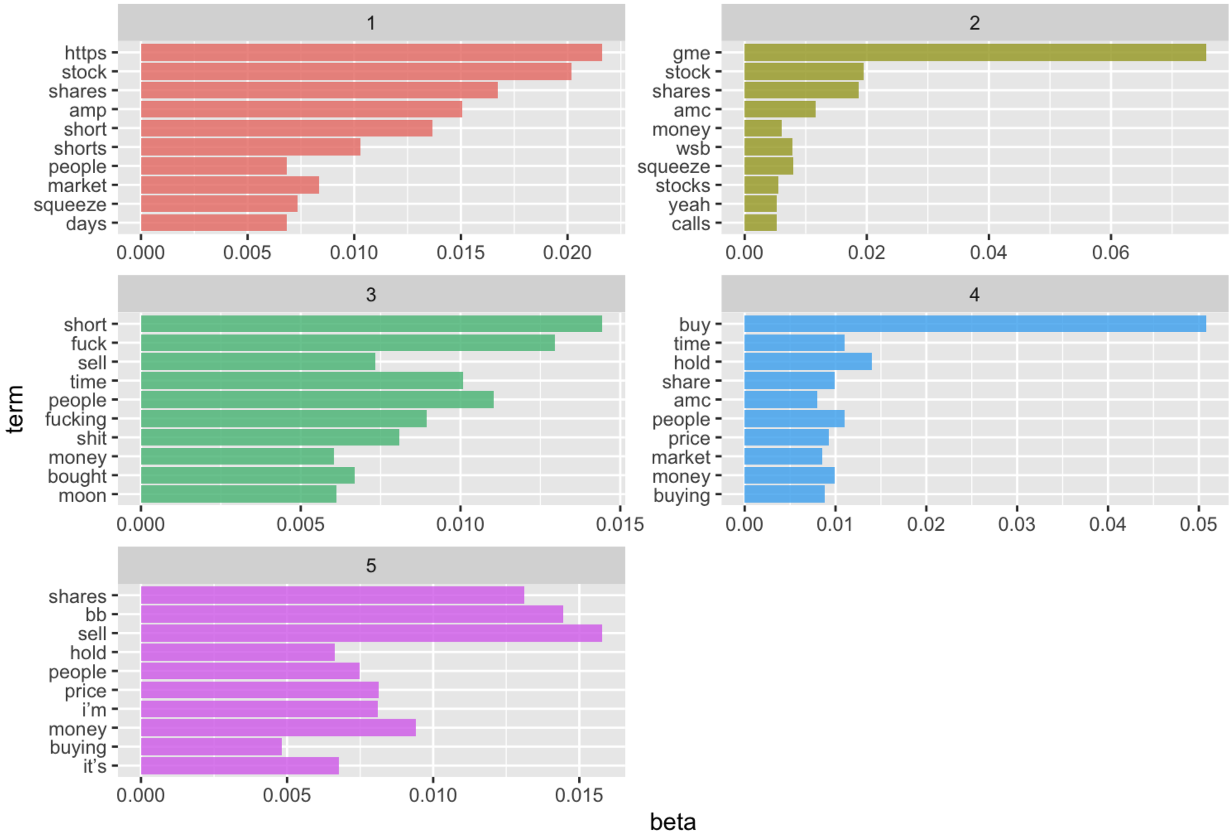


Figure 9: Latent Dirichlet Allocation topics

References

- Joshua O. Barker, & Jacob A. Rohde (2019). Topic Clustering of E-Cigarette Submissions Among Reddit Communities: A Network Perspective. *Health Education & Behavior, 46(2_suppl), 59S-68S*.

- Christian Boylston, Beatriz Palacios, Plamen Tassev, & Amy Bruckman. (2021). WallStreetBets: Positions or Ban.

- Bradley, Daniel and Hanousek Jr., Jan and Jame, Russell and Xiao, Zicheng, Place Your Bets? The Market Consequences of Investment Advice on Reddit's Wallstreetbets (2021).

- Eaton, Gregory W. and Green, T. Clifton and Roseman, Brian and Wu, Yanbin, Zero-Commission Individual Investors, High Frequency Traders, and Stock Market Quality (2021).

- Benjamin D. Horne, Sibel Adali, & Sujoy Sikdar. (2017). Identifying the social signals that drive online discussions: A case study of Reddit communities.

- Sachin Thukral, Hardik Meisheri, Tushar Kataria, Aman Agarwal, Ishan Verma, Arnab Chatterje & Lipika Dey (2018). Analyzing behavioral trends in community driven discussion platforms like Reddit