UNIVERSITY OF WATERLOO

Data Science | Machine Learning

Group 13 Final Project

WallStreetBets Community Sentiment Analysis

Amanda Almedia Rocha

Sean Copeman

Lisbeth E. Fernandez

Leena Khote

Colton Mak

Adem Pekediz

Objective

In today's interconnected and digital world, it is clear that social media has become an undeniable force shaping how individuals communicate, consume content, and interact with businesses. As billions of active users across the world engage with each other on various social media platforms, the evolving digital landscape provides businesses with the potential to connect with their audience. The ability to read the general consensus of users on social media is critical for businesses, allowing them to better understand their market audience and plan accordingly. Sentiment analysis is a valuable tool for businesses as it enables them to interpret the public's perception across constantly evolving digital communities with a data-driven perspective.

In our project, we decided to perform sentiment analysis on a niche community known as WallStreetBets, a subforum of Reddit, a popular social media site. WallStreetBets, henceforth abbreviated as WSB, is a community with discussion centred around stocks and option trading. However, unlike other internet communities with a focus on finance-related topics, the WSB community is well-known for its mix of profane and juvenile internet humour along with high-risk and extremely aggressive investment strategies. In 2021, WSB found itself in the headlines of mainstream news as the community performed a short squeeze on GameStop stocks, causing their stock prices to increase dramatically. The unorthodox, internet meme culture of the WSB community along with investment strategies that are antithetical to the traditional norms of Wall Street investing firms makes it a fascinating case study in the intersection of social media and businesses.

Thus, the objective of this project is to identify popular stocks and investment strategies in the WSB community through a sentiment analysis model. We chose to limit the scope of WSB community posts to 2021, as this marked the rising popularity of WSB in mainstream media with their involvement in the short squeeze of GameStop stocks. We hope to identify notable stocks and trade options from the community, identify whether their public sentiment was positive or negative, and correlate the public's opinion with the historical prices of stocks.

Data Preparation

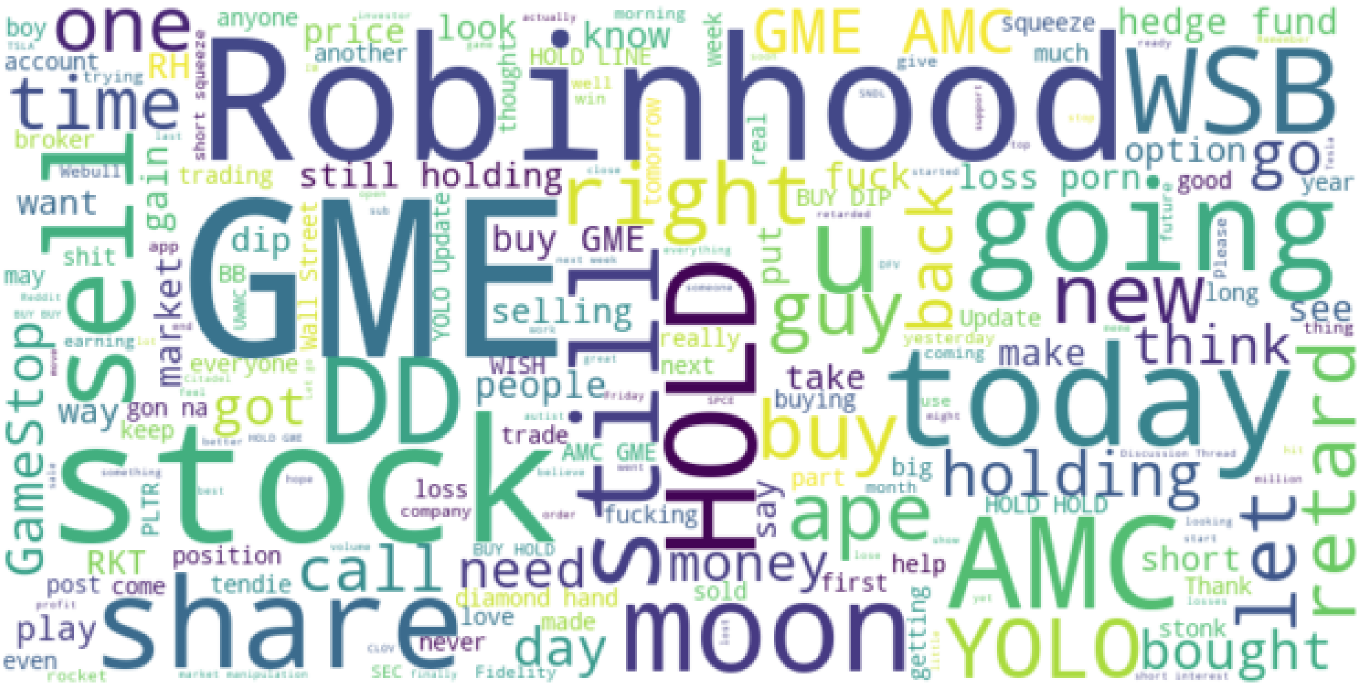
We obtained our data from the following Kaggle dataset:
https://www.kaggle.com/datasets/gpreda/reddit-wallstreetsbets-posts/data

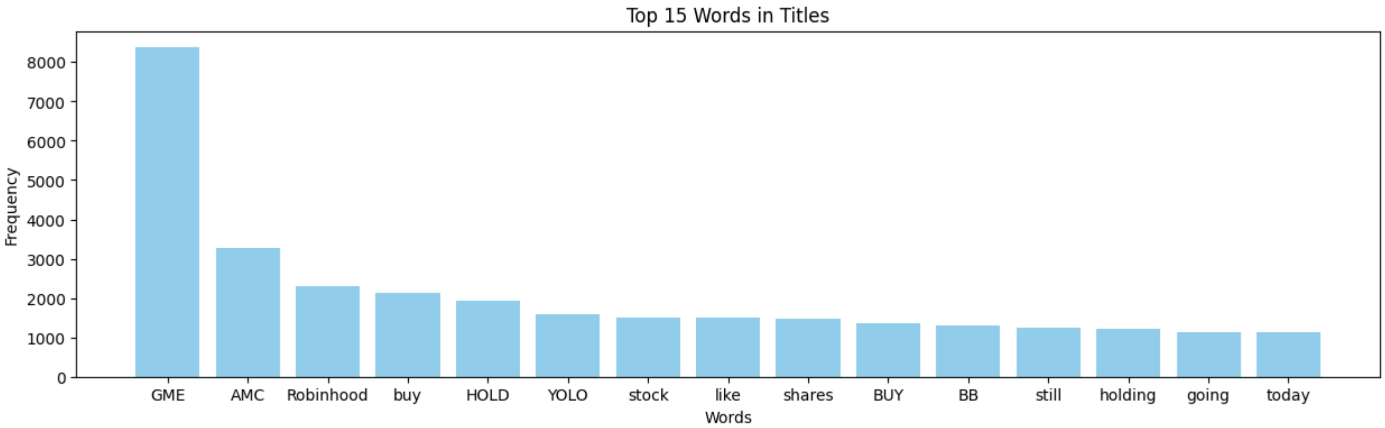
The dataset contained 53187 WSB community posts with the following features:

- Title - Title of the post
- Score - Community scoring of posts, indicated by number of upvotes minus number of downvotes (post’s popularity)
- ID - Unique identifier of post
- url - URL link to post
- comms_num - Number of comments on post (engagement)
- created - Timestamp of post’s creation, expressed in epoch time format

- body - Text content of post
- timestamp - timestamp of post's creation

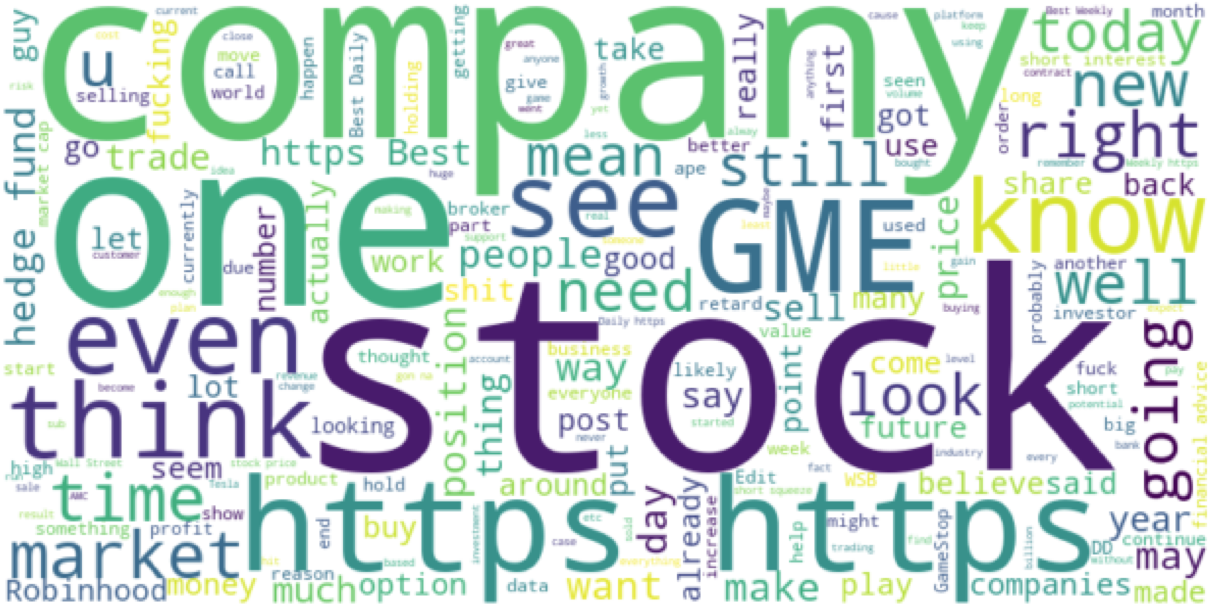
We can visualise the most common words in post titles as such:

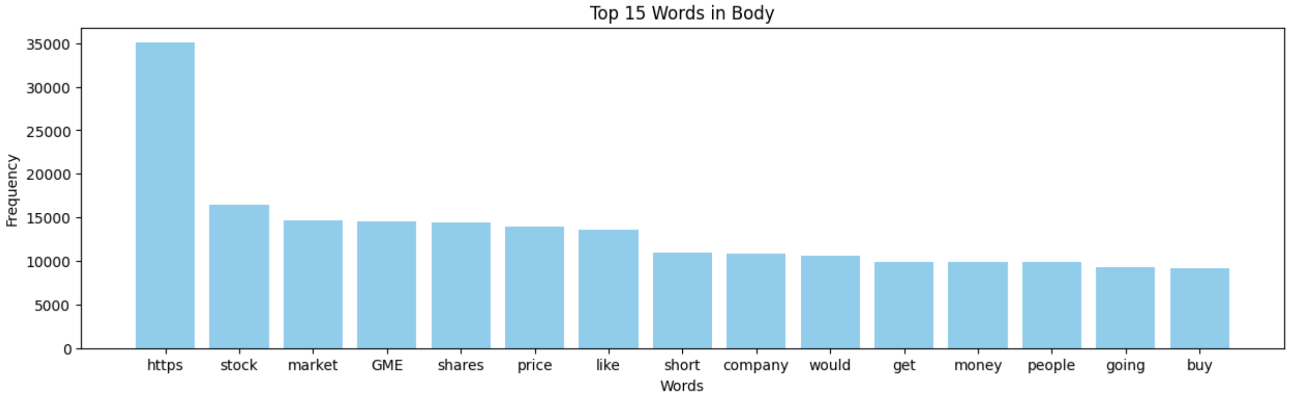




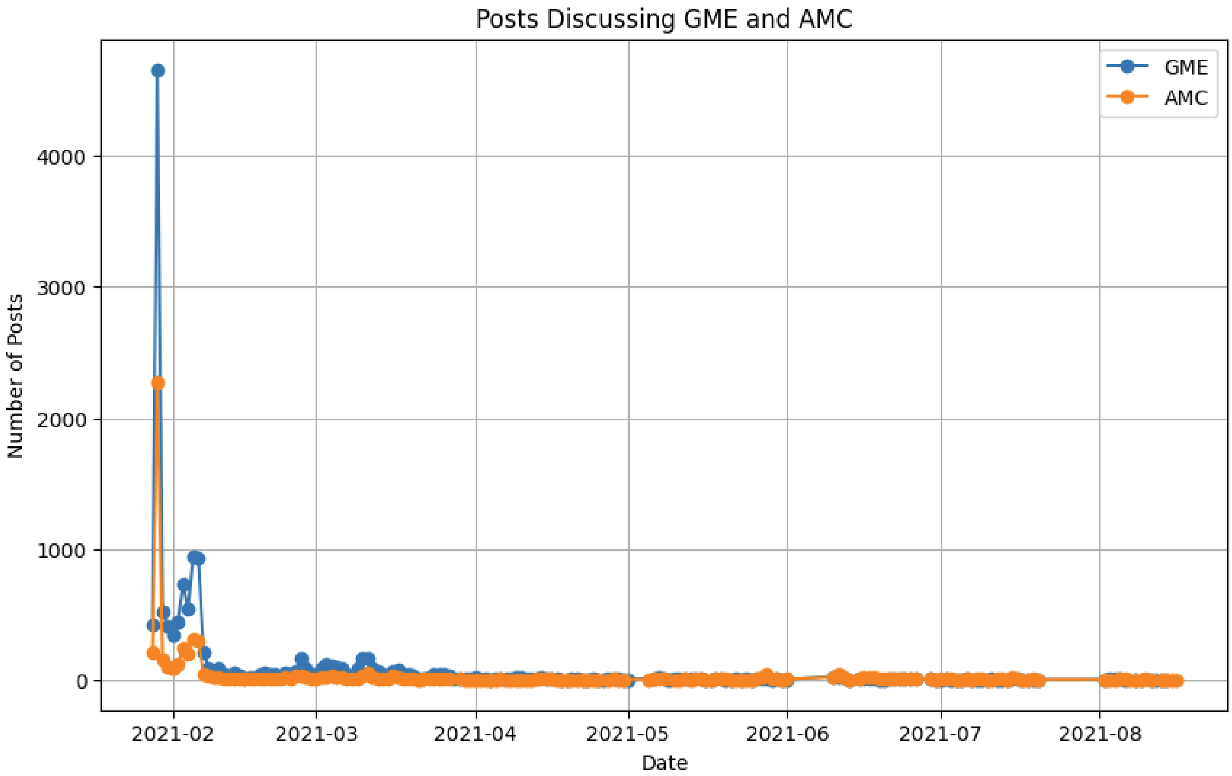
From these graphs, we can identify some of the most popular stocks mentioned in WSB. These include GME (GameStop), AMC (AMC Entertainment), BB (BlackBerry). From the word graph, we can also identify specific language vocabulary and slang that are popular within the community, such as tendie, ape, stonk and loss porn; indicative of the community's reputation for juvenile humour.

We can similarly repeat the same visualisations to see the most commonly used words in the post's text body:

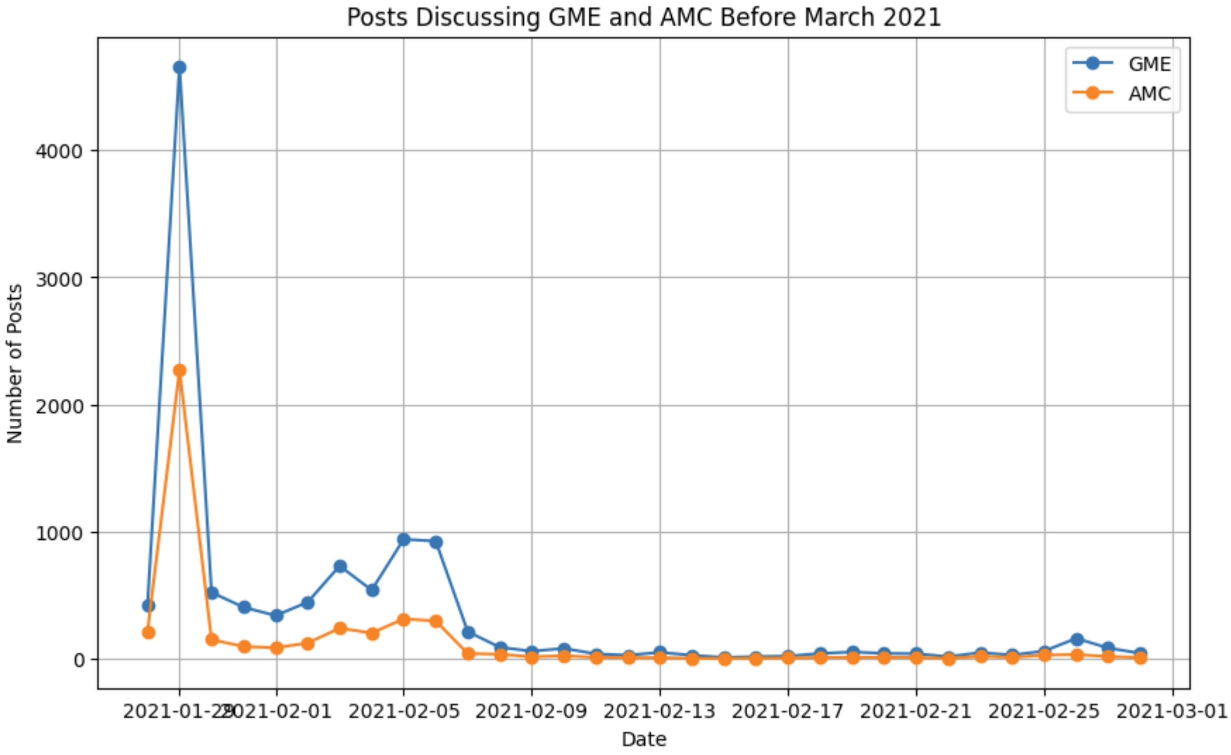


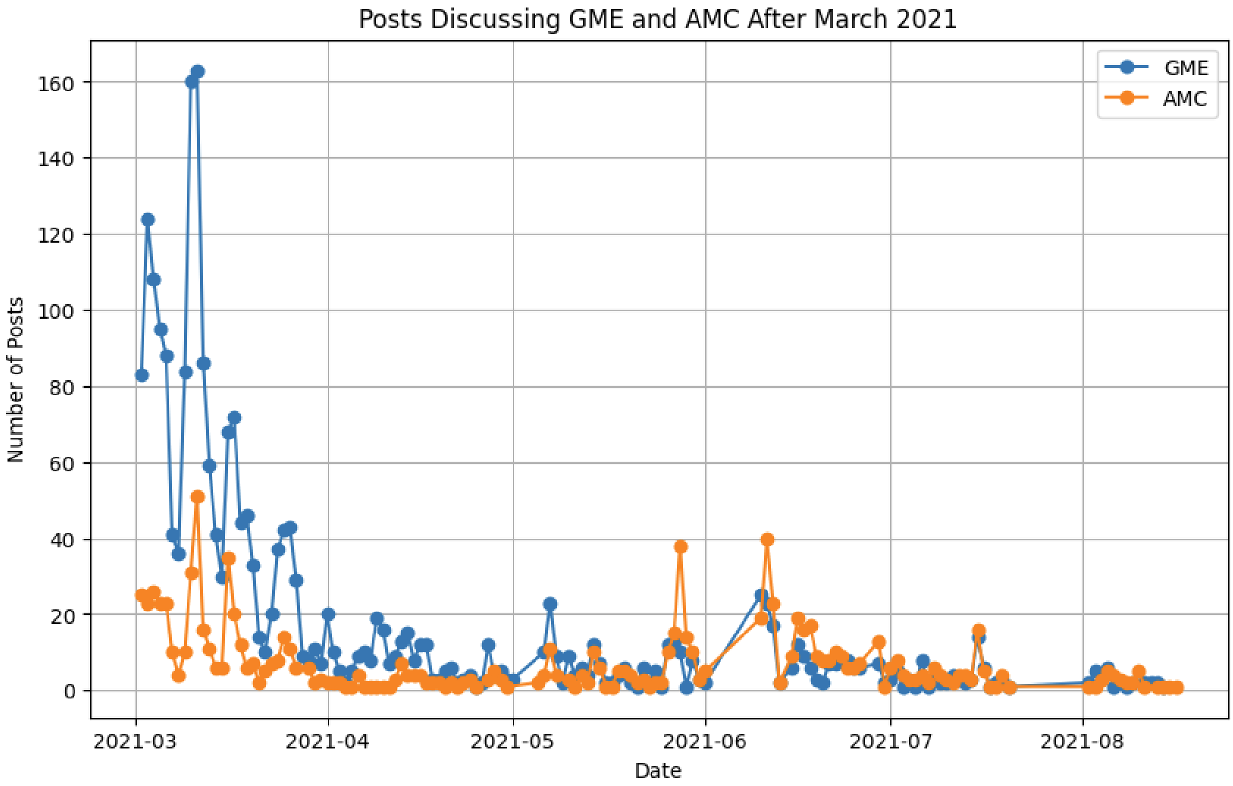


Once again, we can see that the most commonly mentioned stock is GME (GameStop). Thus, we have chosen to further narrow the scope of our project to performing sentiment analysis on two primary stock options: GME (Gamestop) and AMC. Despite being commonly discussed, RobinHood the investment trading platform commonly used in the United States has been eliminated from our analysis because it had not yet gone public, the discussion was around RobinHood as a product, not as an investment.



From the above graph, we can see that the vast majority of discussion surrounding GME and AMC stocks took place early in 2021. Hence, we create two separate visualisations to show the number of posts discussing GME and AMC before and after March 2021.

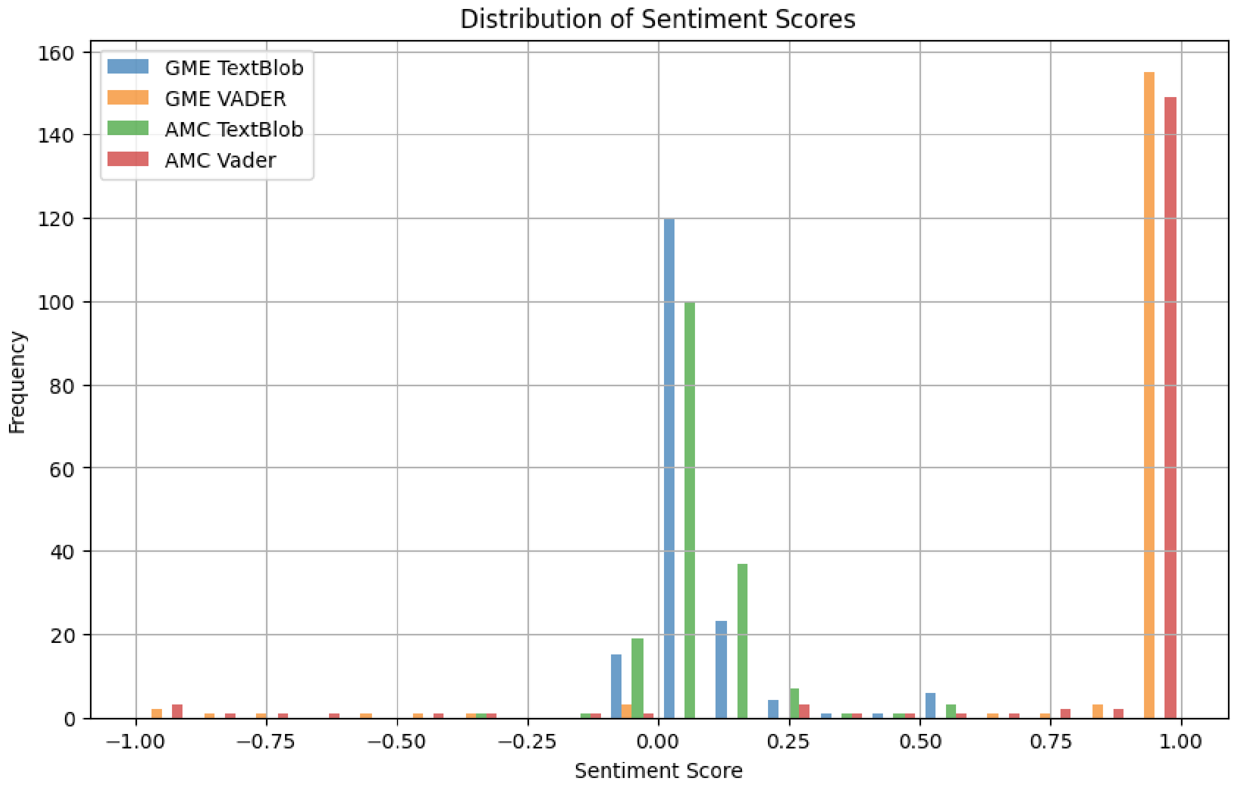




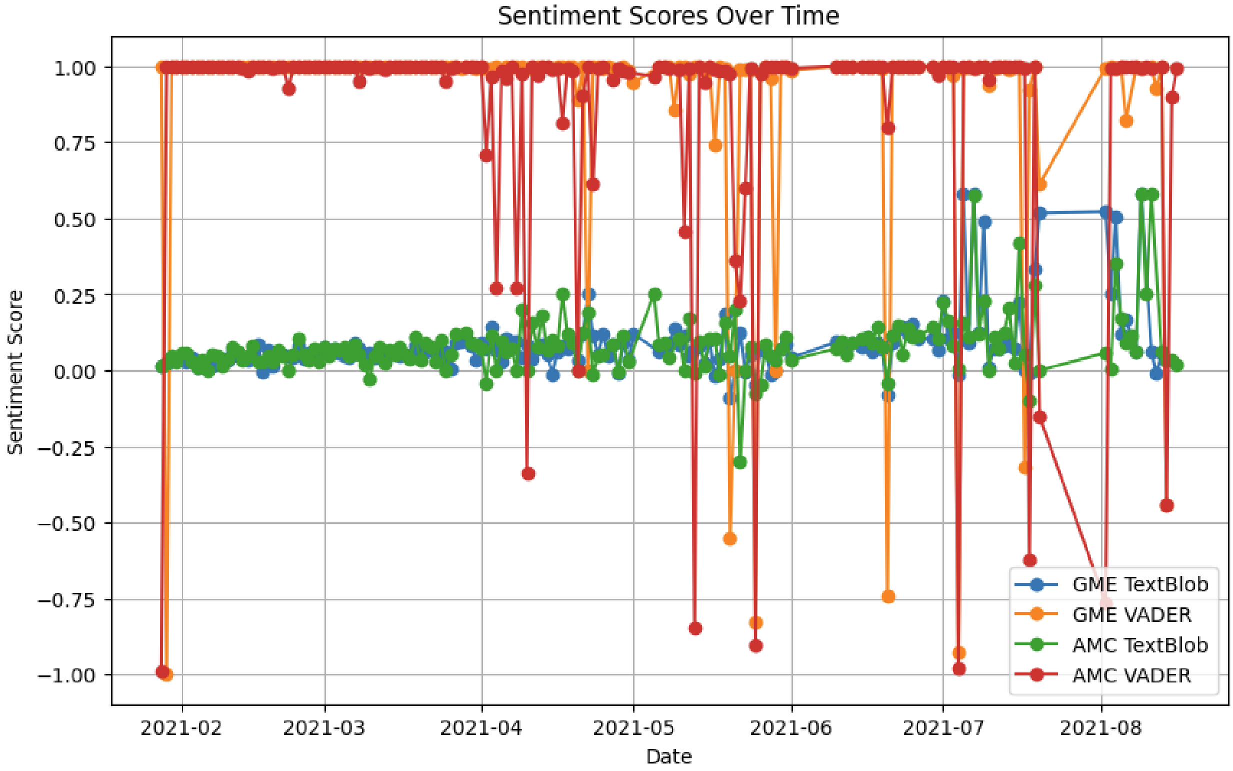
From these visualisations, we observed that popular discussion surrounding GME and AMC stocks spiked in late January with over 6000 combined posts. However by April, discussion had slowed down to below 40 combined posts.

Model Design and Evaluation

We chose to perform sentiment analysis on two different models: TextBlob and VADER. TextBlob is a popular Python library used for various natural language processing (NLP) tasks such as classification, noun phrase extraction and sentiment analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is another Python library that performs sentiment analysis tasks, but is also specifically trained on language used in social media.

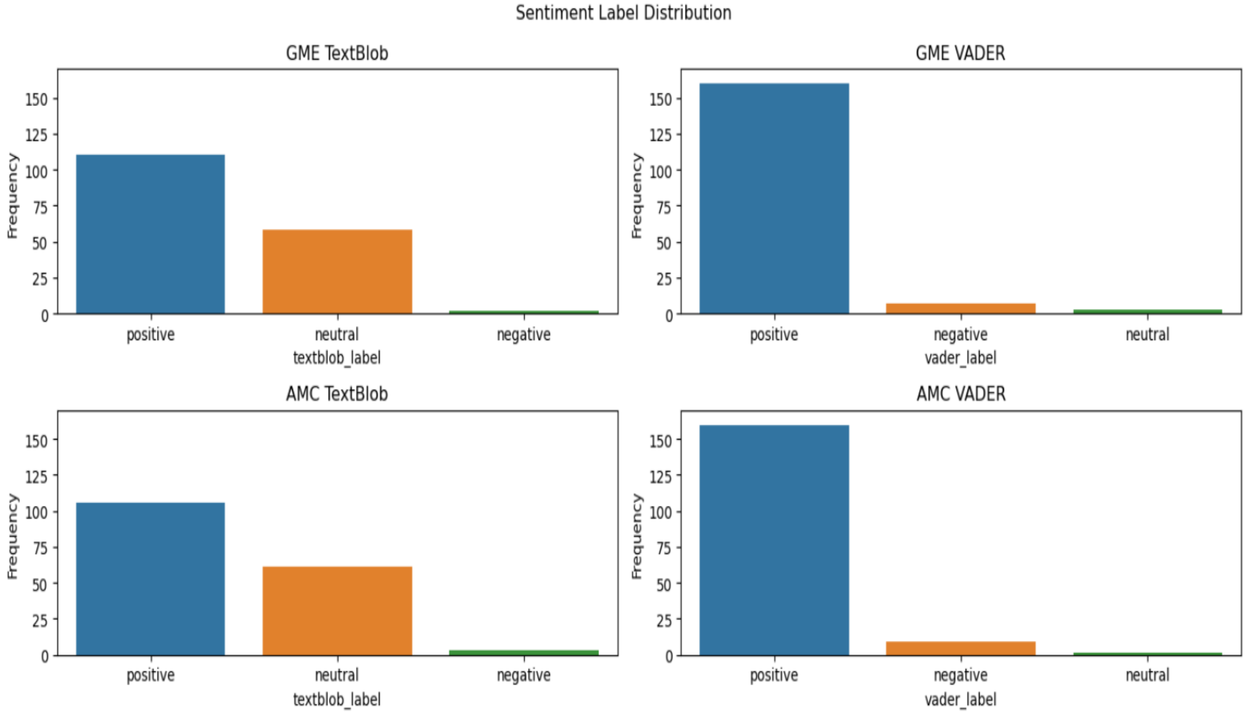


Distribution of Sentiment Scores



Sentiment Scores Over Time

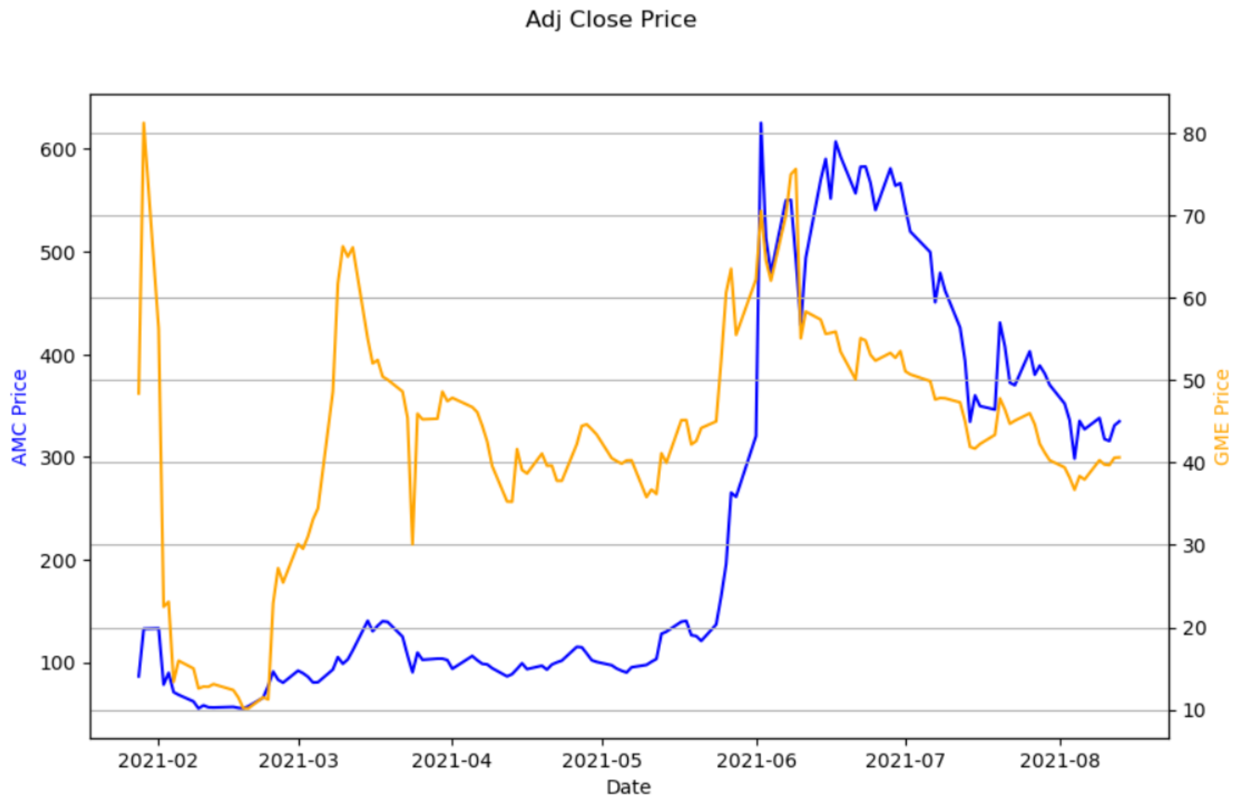
The above graph visualises results of training both sentiment analysis models. Both models score the sentiment of each WSB post on a scale from -1 to 1, with -1 indicating a negative sentiment, 0 indicating a neutral sentiment, and 1 indicating a positive sentiment. The graphs also show that VADER tends to score with much more variability than TextBlob, heavily leaning towards the positive extremes. In comparison, TextBlob scoring starts off with close to neutral scores before trending towards a positive sentiment after July.



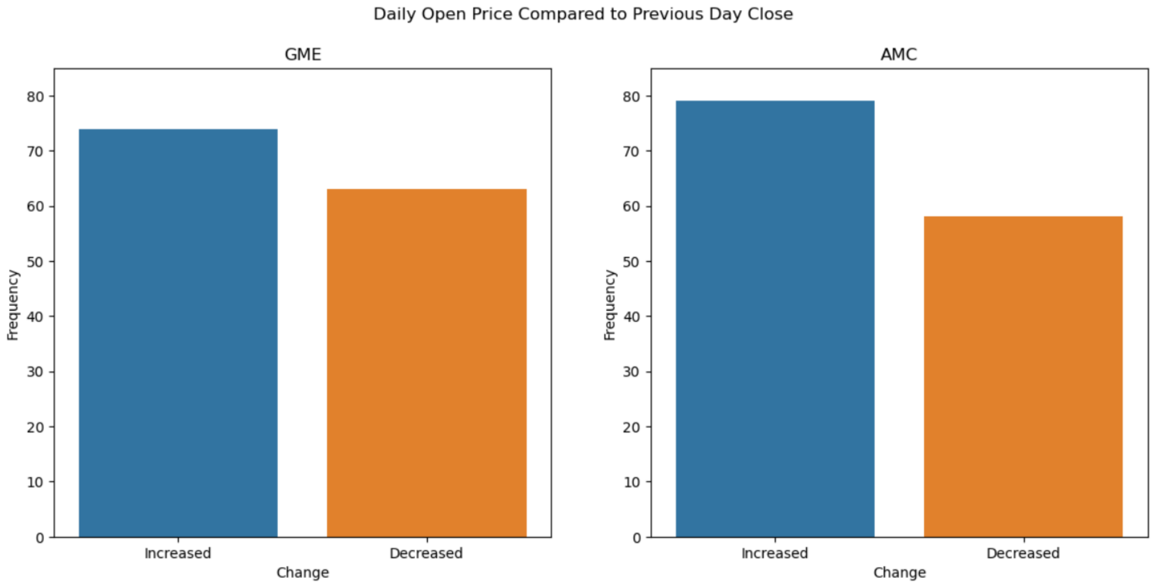
From the distribution of sentiment labels, we can see the community sentiment towards both GME and AMC stocks are extremely positive, with very few negative posts. However, the TextBlob model classifies more posts as a neutral sentiment compared to VADER.

Financial Analysis

Next, we chose to visualise historical stock prices for GME and AMC to determine whether they correspond with the the results of our sentimental analysis models. We utilised the yfinance Python library to download and visualise stock prices during 2021.



GME saw high price levels in the first half of the chosen period, while AMC saw a large surge in price in June which brought them to similarly high price levels before beginning a decline together.



Most days both stocks opened higher than their previous day’s close, Though AMC had a higher ratio of increases, mostly due to its run up in June.

We measured possible correlation between stock price movements and the results of our sentiment analysis model by comparing whether a positive community sentiment corresponds to a stock price increase or if a negative sentiment corresponds to a stock price decrease on the next day. Thus we reached the following accuracy scores:

Sentiment Analysis Model	Accuracy
AMC TextBlob	25.88%
AMC VADER	40.00%
GME TextBlob	26.47%
GME VADER	35.29%

Conclusion

In conclusion, our sentiment analysis model for the WSB community shows that the community's sentiment and public opinion fails to correlate with stock price movements. Accuracy scores of TextBlob and VADER models for both AMC and GME stocks are all below 50% accuracy.

This can be attributed to several reasons that can help provide a greater insight to the WSB community. WSB is known for their extremely aggressive and highly speculative trading strategies, which can fail to yield consistent results. Although this high-risk, high-reward strategy worked for the GME short squeeze event at the start of 2021, these successful results will not persist for a long-term strategy and may not be applicable to other stocks such as AMC.

From our word cloud visualisation, we can also see that the language in many WSB posts are an indication of the community's highly speculative behaviour along with unfounded levels of hype regarding certain stocks. For example, both the rocket and moon are common emojis used to predict that a stock's price will be rapidly rising upwards, similar to how a rocket rapidly rises as it travels to the moon. Other examples include the specifically all-capitalised HODL (an intentional misspelling of HOLD) showing a strong belief in holding onto a stock despite its falling price in hopes that it will rise in the future.

Although our sentiment analysis model fails to show that the WSB community sentiment is able to predict stock price movements, it can serve as a practical stepping stone for future applications. Sentiment analysis models can serve as a valuable decision-making tool for businesses to assess the intricacies of public opinion. As the reach of social media continues to grow on a global scale, it continues to remain as an invaluable medium for businesses to leverage their influence over customer behaviour and perception. Sentimentl analysis models provide businesses with the opportunity to adapt market strategies and tailor products to underlying trends and emerging patterns. Thus, a better understanding and application of sentiment analysis models stands as a cornerstone for businesses to continue to thrive in the ever changing landscape of consumer sentiments.