WallStreetBets: The Future of Retail Trading, are Hedge Funds at Risk?

Sebastian Dixon

April 24, 2024

Abstract

The market events of the January 2021 GameStop short squeeze, led by Reddit's WallStreetBets, started discussions by institutional investors to re-evaluate the retail investor effect on the market. Previous to this event, retail investors were thought to have little influence on stock market value. This assumption is justified by Wall Street trading teams with greater accuracy and access to more privileged data. This paper researches the statistical relationship between WallStreetBets and the equity stock market to determine the quality of financial advice. By using a dataset with posts from 2012 to 2021 we can investigate this relationship using both an empirical analysis of the investors' proposed portfolio, and a machine learning random forest classification model. The portfolio is composed of the top 16 most mentioned traded equities and ETFs in the dataset. The features in this model consist of sentiment values for the post's titles, portfolio Boolean quantities, and the post score. The sentiment is composed using VADER with a custom vocabulary. The portfolio in conjunction with our trading strategy generates a 950% return over the 9-years outperforming the S&P 500. Our classification model outputs 78% accuracy in calculating if the stock price will increase or not in a week's time thus confirming a relationship between the posts and stock market value. By both labelling posts as proactive financial advice and finding strong portfolio returns, traditional funds should consider retail investors as a genuine predictive power in the markets.

I certify that all material in this dissertation which is not my own work has been identified.

Contents

1 Introduction

1.1 Background

WallStreetBets is an online financial retail trading community with a reputation for being rebellious in the investment industry. Users on the platform have developed an unconventional investment culture centred around trading equities with a large risk appetite. Typically trading is related to online culture and popularly discussed companies in categories such as gaming, technology, biopharmaceuticals, and defence. These companies fall into the category of "meme stocks", relating to companies that have gained a cult-like following on social media [6]. The community prefers short-term trading and exploiting leverage, exhibiting gambling-like behaviours. This is summarised in the term YOLO investing, coined on the website. Retail traders on the website commonly execute trades through the use of Robinhood, a commission-free stock trading platform in US markets [7]. The integration of Robinhood into WallStreetBets has exemplified the gamification of investing in the stock market with motives by the platform similar to that seen in gambling [8].

In January of 2021, the brick-and-mortar store GameStop (GME) was being heavily shorted. Large hedge funds, such as Melvin Capital, collectively adopted a position of more than 139% of existing shares in GME [9]. WallStreetBets users became aware of this position and collectively made the decision to perform a short squeeze against the hedge funds [10], which led to Melvin Capital losing billions of dollars by the day. By the end of the first quarter of 2021, the hedge fund reported a loss of 42% in capital at the hand of the short squeeze. By May 2022 Bloomberg News reported Melvin Capital planned to close in June of that year [9].

Since the events, WallStreetBets has gained significant media attention both in financial news and being covered in a film premiere. The growth of the community over the events skyrocketed and has continued to grow since reaching 15 million active subscribers at the time of writing [11].

Importantly, recent trends in the wider financial market have been closely linked with behaviours in online financial communities like WallStreetBets. The popularisation of these communities has further compounded these trends. The World Federation of Exchanges 2020 report suggests a positive trend in retail trading participation and index futures and options volumes [12]. These equity derivate markets data were more limited, yet the trends are still visible. The retail trading participation in equity ETF markets also exhibits these trends, to the extent that significant increases in the value of retail trading in 2020 can be seen [12].

1.2 Motivations

My motivation is to determine if the WallStreetBets posts have a statistical relationship with the value of the future stock price. Portfolio backtesting and machine learning models empower informed decisions on investment strategies. Through simulating an investment strategy using historical prices to see how well the strategy would have performed in the past [14]. The fundamentals of investing state that past performance is not a reliable indicator of future results [15]. Therefore, my motivations are not to predict the future but to demonstrate that there are strong statistical relationships existing in the properties used in the portfolio backtests and machine learning models.

The portfolio backtests are to be compared against the return of the S&P 500 index. Should the portfolio generate returns outperforming the index, there is quantifiable evidence to suggest that the portfolio is a strong investment [17]. The portfolio is to be tested using a trading strategy similar to the context in which the users of WallStreetBets would engage themselves. The machine learning model will quantify the predictive power to classify the posts as proactive as opposed to reactive financial advice. Proactive posts will exhibit a positive correlation against the future value of the stock being referenced [16]. The context of the model's accuracy will therefore indicate the probability of a post on the site having this predictive property.

1.3 Objectives

In this paper, our objective is to determine if the macro WallStreetBets posts strategy and portfolio constitute reliable financial advice. This will be achieved through two main methods leveraging a dataset. The first is an empirical analysis of the portfolio in conjunction with a trading strategy representative of the users in the community. Should the portfolio demonstrate returns greater than the S&P 500 index, there is sufficient evidence to suggest that the companies targeted by the community are good choices and that the trading strategies are effective on the equities and ETFs chosen [17].

The second objective is a machine learning analysis of the posts made on the site leveraging sentiment analysis for labelling bullish and bearish sentiment [18]. The classification of posts' price data will enable a supervised machine learning model to be trained and tested on the dataset. Should the accuracy of the machine learning classifier be greater than the probability of a random guess, the post's macro trends exhibit proactive financial advice.

2 Specification and Literature Review

2.1 Financial Preliminaries

In this section, we will inspect the financial background knowledge that is relevant to our task. We will first review the differences between institutional and retail traders in the context of risk. Then go over derivates and one common financial instrument, Options, for hedging risk. These are appropriate for understanding the data behind the retail trader's motivation and strategy for investing in the stock market.

2.1.1 Retail and Institutional Traders

Trading on the stock market is becoming more accessible to independent investors through the mobilisation and democratisation of the markets as we will later discuss. Traditionally those to invest capital choose professional institutional investors, who manage their funds, balancing risk against reward in the client's portfolio. Those who trade independently participate in the market on their behalf, known as retail investors. These retail investors, or retail traders, lack the knowledge, experience, and important investment research of institutional investors, therefore putting them at a disadvantage in finding opportunities in the markets [1]. Professionals are, however, restricted to a specific risk tolerance when designing a portfolio for their clients, to minimize losses. Retail traders can have as little or as large of an appetite for risk as they choose. This risk influences the equities chosen to be invested in by these different groups. Exploring different categories of the market can open opportunities in both cases. Risk can be reduced through the use of multiple different financial instruments or investment methods. The ability to hedge is to help reduce the risk of loss on any existing position [2]. Institutional investors, such as Hedge Funds, are derived from this term [3], therefore investing to intelligently produce portfolios with maximised profit and minimised risk.

2.1.2 Derivates Market and Options

As aforementioned financial instruments create the opportunity for users to hedge their risk on an existing position in the market. One category of financial instruments is derivatives. Derivatives are products whose value is derived and dependent on the value of an underlying asset. Therefore, creating an additional derivatives market purely based on the value of already existing assets. An Option is a derivative that gives the buyer of the option (but not the obligation) to buy or sell the underlying financial asset at a particular price, called the strike [4]. The contract is voided at a date known as the 'expiry' and can be exercised either on the expiry or at any time up to it [5].

An option's intrinsic value is influenced by several factors and market indicators. The most influential include the underlying asset price, strike price, time to expiration, and volatility [27]. Volatility is a measure of risk (uncertainty), or variability of price of an option's underlying security [27]. Option volatility is based on the market's anticipated future volatility. The higher the volatility the more expensive they become as estimates indicate greater expected fluctuations (in either direction) in underlying price levels [27].

2.2 Institutional and Retail Investors

In this section we will demonstrate the differences between institutional investors and retail investors relevant to our research question, allowing us to make better and more informed decisions on understanding the relationship between these online communities and the wider financial market. This is achieved through understanding these investors' differences in the context of risk management, trading behaviour, and market opportunity. Existing research by investment banks is also available to us to gain a high-level perspective of how they benefit and model retail trading behaviours.

2.2.1 Risk Management

Risk tolerances are controlled and assessed differently between retail and institutional traders. In the professional environment, strict measures are employed in researching the market and commodity or equity before investing capital. This often places institutional investors with low levels of risk tolerance to maintain investment reputability. The retail investor will have varying risk tolerance depending on several financial characteristics and objectives with their money. Ultimately, they lack the high level of research and due diligence taken to investigate the fundamentals of a company determining if it's a safe and valuable investment [19]. Institutional investors have access to far greater amounts of capital and, therefore can technically incur greater levels of risk than retail traders.

The best comparison of an institutional investor comes from Hedge Funds as their goal is to pursue absolute returns, similar to that of a retail trader [19]. The trading behaviours shown on WallStreetBets often belong to unhedged, high-risk trades. This shows the general risk tolerance of WallStreetBets to be fundamentally greater than that of Hedge Funds. The risk appetite determines the equity and ETFs invested in.

Institutional investors often invest in an equity which has proven to show strong fundamentals, therefore investing in solid companies with strong financials, and show to be undervalued. This strategy is present in one of the world's most successful hedge funds Berkshire Hathaway [45]. The majority of retail investors on WallStreetBets don't follow this principle and instead invest based on the community's current viral obsession [20]. Therefore, as these two trading groups show polarising risk tolerances, they will invest in different parts of the market. This is analogous to the WallStreetBets behaviours which avoid investment in large ETFs which have lower risk. This strategy is a part of the wider WSB investment behaviour whereby they are anti-financial institutions, therefore staying away from the safer trades made [21].

The WallStreetBets trading strategy most popular in the community is known as the YOLO. The YOLO involves making one single unhedged trade with a user's entire portfolio worth into one stock [20]. About the phrase, You Only Live Once, this strategy disregards its inherent extreme risk with the opportunity for making a huge profit. Common strategy extensions include using leverage and multiplying the returns [24]. As well as purchasing a single contract on the derivatives market as an Option, reducing the capital requirements. The most popular time until a contract expires is within the same day, known as day trading [22]. This creates a 0DTE option, a zero-day till expiry option, the closer to expiry, the greater the risk and reward. Following the growth of WSB, the daily notional value of trading in 0DTE options has grown to about $1 trillion, according to J.P. Morgan [23].

2.2.2 Barclays US Equity Derivatives Strategy

In September 2020, institutional investment research on retail investors conducted by the Barclays equity team developed two trading strategies. These trading strategies outperform the market by capitalising on new retail options trading volume [26]. Simply put, a blueprint for exploiting less experienced new retail traders.

The first sentence summarises the retail trader community trading behaviours, "We show that retail investors have been driving a significant increase in option (mostly short-dated calls) volumes for large-cap tech stocks" [26]. Highlighting retail investor trends on the market, the use of options, and an interest in large-cap tech stocks. This confirms that the analysis is targeted at those similar to WallStreetBets.

Barclays analysts recognise that the new retail traders are correlated against the US government's stimulus checks [26]. Stimulus checks are not that large, therefore more profit can be found using options as they require less capital to be exposed to greater return [31]. With high-risk tolerance, investing in 0DTE being popularised in the WSB group increases the volume in these derivatives markets [26]. Often the underlying asset is a large-cap tech stock, simply a viral technology company with a market capitalisation of $10 billion or more [30].

The strategy involves the option value, priced on volatility. The Barclays analysts are focusing on this property. By determining opportunity in the market where there is a convergence or divergence in properties of this volatility [26] the analysts developed two investment strategies to outperform the index.

The understanding is that institutional investors are intelligent enough to harness the growing retail traders and outperform the index by leveraging the changing market. The caveat to this applying to all investment

strategies is that Barclays is one of the world's largest financial institutions and the second-largest bank in the United Kingdom [29]. This position is accompanied by a strong research team able to identify and leverage these market movements. This capability might not extend beyond top research firms, benefitting retail traders. It is also clear that there is little risk for the Barclays group releasing this research publicly. Both as there is a conflict of interest affecting the report's objectivity, and that the details of the strategy are undisclosed [26].

2.2.3 Democratization of the Market

The percentage of retail investors as a share of overall US trading volumes has been rising steadily over the recent decade but has increased dramatically since the beginning of the pandemic [32]. The crucial factor behind this change is the rise of the trading app. Applications such as Robinhood have spearheaded this mobile trading generation, targeting at the younger demographic [33]. Done through approaches such as gamification, hype, education content, and referral systems. All lowering the barrier to entry for new investors in the stock market [33]. The online brokerage states in their About Us page, that it's their mission to democratize finance for all [34].

The consequences of this growth in retail traders using online brokerages are they often incur hidden financial incentives. Robinhood attracts retail traders to the platform with the allure of commission-free trading [35]. The saying "if you aren't paying for the product, you are the product" applies in this case. Market makers such as Citadel Securities pay Robinhood for the right to execute their customer's trades [35]. Additionally, the company works with J.P. Morgan, the biggest US bank by assets, to process their transactions [36]. By feeding the retail traders' data to financial institutions, the institutional investor wins from learning about the trends in trading behaviour and executing their trades. This contradicts the company's mission of democratising the market.

2.3 Portfolio Backtests

A portfolio analysis of WallStreetBets comprises finding the most popularly invested and mentioned stocks in the community over the period being analysed. This develops a macro-style impression of the users' opinions of the market. Our method is to allocate shares of individual assets in the portfolio according to the community's semantic opinion of the asset at that time. Should the opinion of an asset be positive during a specified time frame, a larger share can be allocated, and a smaller share for negative opinions accordingly.

A comparison of the WSB portfolio against the wider market performance can be summarised by the S&P 500 index, referred to as the index. The index tracks the 500 largest US companies listed on the stock market [38]. Benchmarking investment portfolios against the index is important for understanding both performance and diversification metrics [39] [40]. The objective is to outperform the index, meaning investment in your portfolio has been successful [40]. A similar line of work investigating the performance of WSB compared to investment banks proposes the use of a reference portfolio and a similar method of comparing the WSB portfolio to the reference portfolio. The paper also suggests the use of the S&P 500 as the reference portfolio [37].

2.4 Machine Learning Classification

Classification of posts as valid financial advice can be constituted through first finding if the post is proactive or reactive to market events, and then if the source of the advice is reliable. A proactive post proves that the information occurs before the market event, and reactive posts are in response to the market [16]. This has a predictive quality to it, where the information is related to future market events. This can be quantified as a statistical correlation between the sentiment of the financial advice and the value of the asset being referenced.

The reliability of financial advice will come as a result of the accuracy of performance metrics connected to the statistical relationship between the information and the price information. If a strong statistical relationship exists, the reliability is better, and vice versa.

The machine learning model requires a dataset to be trained and tested on. From this dataset, different features can be developed through feature engineering methods. The objective is to produce a series of valuable features with correlation against the target variable [41]. In our case, the target variable is a binary classifier to

determine if the price increased or decreased associated with a post exhibiting positive or negative sentiment about an asset in the portfolio.

The features used for the model belong to different types of data. The price data on the stock market belongs to time series data. The post-text data will produce categorical features for sentiment, stock ticker information, as well as the raw text data itself [41]. Converting the stock ticker information into a representation able to be trained and tested through encoding measures such as one-hot encoding, is appropriate for our portfolio of equally prioritised data values where no ordinal relationships exist [42].

The different metrics for assessing the quality of the fit of the model to the data each measure independent qualities and for different purposes. to balance the generalisability and specification of the model to the problem we avoid underfitting and overfitting respectively. The accuracy of the model will determine the probability of making correct predictions, which is particularly useful when false positives and negatives have equal importance [43]. Precision and recall quantify the trade-off between false positives and false negatives. Often in financial applications, it's especially important to reduce false positives as they can cause the greatest difficulties, where false positives can cost merchants up to 75% times more than an act of fraud [44]. This makes precision and recall important to assessing the quality of the classification model we are developing.

2.5 Research Approach

The research approach of this paper combines and builds on research from similarly inspired works trying to determine the significance of WallStreetBets in the financial markets. The WallStreetBets topic has become increasingly more popular on account of the media coverage most recently concerning the GameStop short squeeze in January 2021. The importance of a topic such as quantifying the financial trends set to determine the future of our financial system is increasingly important now entering the age of artificial intelligence.

Different research approaches to the data surrounding the hype of 2021 choose to label the data as anomalous concerning the previous posts made on the site on account of the huge change in the number of users, therefore changing the overall concentration of real financial advice. It’s our opinion that the data should be regarded in the same respect as more historical data. The data in this time frame is the very reason the community has been so interesting. Further comment from capital.com reporter suggests that the phenomenon in early 2021 could be repeated as this type of trading behaviour and volume only becomes more commonplace and accessible to those interested in the markets. The reporter had this to say “Trading volumes linked to WSB meme stocks did not decrease over the course of 2021 and remained strong in the second half of the year” [32].

By combining the empirical analysis of the WallStreetBets portfolio against the index with the machine learning classification of the quality of the financial advice, we can comprehensively quantify the extent to which the market correlates with these users. This analysis of WallStreetBets is unique. Should the market have a significant statistical relationship, then the behaviours with which the users of the community portray on the market may have an effect on institutional investors with the same objective of generating sheer return on investment such as Hedge Funds.

3 Design

3.1 Dataset Design

The higher the quality of the dataset used in an analysis and machine learning model, the greater the potential for strong statistical relationships to be calculated and identified [47]. This therefore prioritises developing a usable dataset before processing, cleaning, and normalisation. Important consideration will be given to the timeframe of the dataset which in this case will need to to capture the earliest and most recent market events in the community. This design will lead to a better understanding of the evolution of retail traders over time.

Two main datasets will be required for this analysis, the first is a historical stock market dataset. This must include daily valuations of listed equities and ETFs tradable on the world's largest exchanges, accounting for all types of investment opportunities that users of WallStreetBets could reference. The daily valuations of each of these companies will allow further analysis of this time series data for investigating the relationship between stock price movements and the posts on the community. Reliable and accurate pricing information is also necessary for producing a reliable and accurate representation of the market.

The second main dataset is a WallStreetBets posts information dataset. This must include every post made in the financial group over the reference period. The post includes several components: the title, score, author, time of creation, number of comments, and more. To implement the research approach defined previously, only the title, score, and time of creation are necessary to provide insight into these users. Reddit hosts the community therefore popular methods of web scraping must be employed to extract this information. Popular Reddit scraping API PushShift along with other similar tools have recently been revoked [48]. This prevents data collection on the website. Therefore, the best method of still accessing the dataset required is to use publicly available datasets from hosts such as Kaggle [46]. These datasets are licensed under the Creative Common's public domain, allowing us to work with this data for our research [49].

A large and varied dataset will fundamentally prevent measures of over-specification of the machine learning model. If there is enough variance in the dataset, it becomes more difficult to overfit to a training set. Greater size allows for further preventative overfitting measures such as cross-validation, allowing the algorithm to better generalize [50].

3.2 WallStreetBets Data

As previously mentioned, the best source for this data is from dataset host, Kaggle. Here users uploaded datasets having previously scraped from the website Reddit. The dataset chosen, uploaded into the public domain by user Raphael Fontes, features the required elements for analysis along with the appropriate size for adequate performance [52] [46]. A Reddit post made on WallStreetBets primarily contains two components, the title, and the body. Posts on WSB commonly contain the majority of the information and sentiment available for analysis in both the title and body. The body of the post can also contain image data, which is beyond the scope of this analysis, requiring image processing steps.

In determining if the title alone is adequate for sentiment analysis of these posts, we conducted an exploratory data analysis on a separate dataset containing both the title and the body to compare the two. The results show that the frequency of positive, neutral, and negative sentiment in the two groups of data is largely the same, as shown in Figure 1. Therefore, continuing our analysis with the title text alone provides us with the same sentiment for a post and greater efficiency from processing less text.

The title data also provides us with the opportunity to determine which component of the portfolio the author is mentioning. Stocks listed on the exchange are represented by a unique shorthand representation called a ticker. An example of this would be to take the stock of Apple Inc. listed on the Nasdaq exchange as AAPL. The text can contain none or multiple tickers. While determining in the case of multiple tickers which one is being referenced to exactly by the overall sentiment of the post is a challenge, we can assume that it is likely to be the first mentioned ticker in the post.

The chosen datasets column called score is an integer value default set to 1. This can be incremented or decremented by other subscribers to the WallStreetBets community. This can be thought of as a method of showing appreciation, respect, or kindness to the author's post. This scoring system also influences the number of users on the platform who view the post, as posts with a higher score are made more visible. A post with a higher score can be interpreted as more valuable, especially in the context of financial advice and hence can prove to be a useful feature in predicting the reliability of the sentiment in the title text.

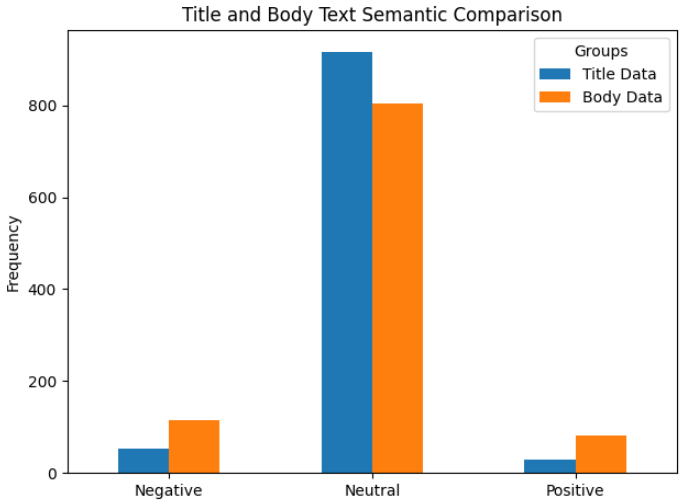

Figure 1: A histogram comparing the frequency of sentiment categories in title and body data.

3.3 Finance Data

Yahoo Finance provides accurate real-time and historical financial market data. The website produces detailed information on the individual equities and ETFs market data, presenting itself as a suitable and reliable source for our valuation calculations. This data can be accessed through the Python programming language library yfinance. Developing a database for the companies and ETFs present in the WallStreetBets portfolio will be made using this library. By finding the range of the post's timelines in combination with the exchanges listed ticker for the company, we can create a list of open valuations for the company in the specified range. Therefore, we can efficiently query this database for the value of the company at any date.

Relevant to the research we are conducting, the open value of an equity is the market value at the beginning of that day's regular trading hours. The open price will form the representation of the market value for that stock on that day. Given that WSB posts are made outside of trading hours on the weekend and bank holidays, accounting for valuations of companies outside of tradable hours will have to be estimated. This estimation will be that of the most recent market open price for the stock.

Another smaller dataset necessary for calculating the price information of posts ticker data is a list of all SEC-recognised tickers still tradable on the exchange. This simple dataset of stock tickers and their associated company name will be used in pairing mentions of the company via the company name with the ticker representation for the previously mentioned portfolio stock open value database.

3.4 WallStreetBets Portfolio

The WallStreetBets portfolio is to be comprised of the top 16 most mentioned stocks in the community over the length of the dataset. This choice will provide sufficient variance in the dataset. Upon initial inspection into the top 16 companies in the WallStreetBets dataset found the top 16 companies and ETFs to be the following shown in Table 1, the following data provided by Yahoo Finance [54].

The ranking of the most popularly mentioned tickers mentioned from 21st May 2012 to 12th February 2021 shows a bias towards the technology sector with consumer discretionary ranking second to that. These two sectors comprising 75% of the total portfolio make a majority share, therefore there might be a risk of lacking diversification. The portfolio also includes the SPY ETF, which tracks the S&P 500 index, this is important to consider when making a comparison of the portfolio against the same index. Plotting the value of each ticker from their most recently available price history from 21st May 2012 up to 12th February 2021. Figure 2 shows us that three particular equities have undergone consistent volatile price movements in comparison with a relatively steadily increasing SPY valuation. The three include AMC, RH, and TSLA, all of which have been the result of online hype becoming the target of these retail traders with a large risk appetite.

Rank	Ticker	Name	Sector
1	SPY	SPDR S&P 500 ETF Trust	Index Tracker
2	GME	GameStop Corp.	Consumer Discretionary
3	AMC	AMC Theatres	Consumer Discretionary
4	TSLA	Tesla Inc.	Consumer Discretionary
5	PLTR	Palantir Technologies Inc.	Commercial and Government
6	AAPL	Apple Inc.	Technology
7	AMD	Advanced Micro Devices	Technology
8	BB	BlackBerry Limited	Technology
9	AMZN	Amazon.com Inc.	Consumer Discretionary
10	NIO	NIO Inc.	Consumer Cyclical
11	NVDA	Nvidia Corp.	Technology
12	MU	Micron Technology Inc.	Technology
13	RH	RH	Consumer Cyclical
14	SNAP	Snap Inc.	Technology
15	NOK	Nokia Corp.	Technology
16	SPCE	Virgin Galactic Holding Inc.	Consumer Discretionary

Table 1: A table of Portfolio asset rank, ticker, name, and market sector

3.5 Trading Strategy

The trading strategy we will employ to test the WallStreetBets portfolio against the market during the backtest portion of development is important to represent the average WallStreetBets trader mentality. This trading strategy is best described as a confidence measure in buy signals. Financial advice is often categorised as bullish or bearish. Bullish investor's financial advice is the belief that the price will rise, and bearish investors believe that the price will fall [53].

We can therefore categorise the posts on WallStreetBets into one of these two categories. This can be achieved through the analysis of sentiment. Should the post exhibit a high probability of positive sentiment we label it as bullish, and a post with a high probability of negative sentiment is bearish. Accordingly, the more bullish posts made give us greater confidence to buy the asset being mentioned, and more bearish posts on the asset gives us less confidence to buy.

Several assumptions are being made in this strategy. The first is that a WallStreetBets user is never in a position to not consider buying the stock, there is simply a greater or lower percentage chance of them buying the stock. In the context of a portfolio where you already have exposure to the stock, this will indicate the percentage of the portfolio in which you own that stock. Therefore, this instead translates to buying or selling proportional quantities of the stock.

The second assumption is that there is no position to short a stock. Technically should the bearish sentiment be great enough there would be a threshold value at which you would choose to short the asset in question as the probability of the price falling is significant. In this strategy, the integration of shorting a stock can’t be accurately implemented given we are not simulating a buying and selling market separately from the trading strategy. The trades made are purely theoretical as they do not change the underlying value of the asset being exchanged.

The third assumption is that every trade being made is being executed within the same time window. This window is fixed and has been determined via the overwhelming majority of WallStreetBets traders operating on short time frames [26]. The longest time frame commonly traded is contracts with less than two weeks to maturity [26]. Therefore, the assumption states that all contracts are made at the start of the trading week and finished at the end.

The trading therefore occurs week-to-week and operates based on the confidence to buy a stock in the portfolio dependent on the bullish and bearish sentiment surrounding the stock. The quantification of buying confidence is calculated by the ratio of the number of bullish to bearish mentions of the stock. This ratio summarises the WSB community's collective opinion on the stock that week and therefore can be used to calculate the proportionality of that specific stock in the portfolio each week. At the end of the week the total profit can be calculated, representing selling the position. Repeating this process for each tradable week will result in a plot of the value of the portfolio over time. allowing comparison against the S&P 500.

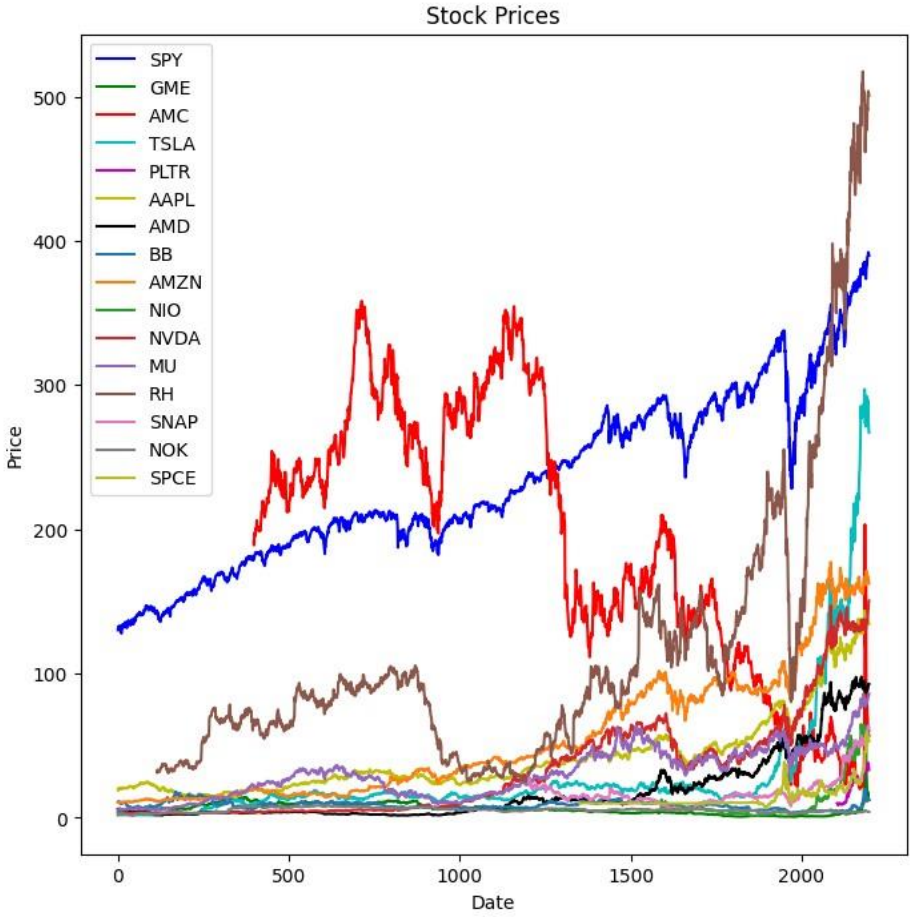

Figure 2: Plot of WallStreetBets portfolio equity and ETF market values.

4 Development

The development of quantifying the statistical relationship required a coding language with appropriate tools and libraries for data analysis and machine learning. The Python programming language provides this environment with a multitude of libraries to import and call functions from. Development inside a Python Jupyter Notebook is the industry standard for this task of data analysis. We will explain each of the components of the development, comparing different approaches we considered and producing code examples to support the explanation. The development of the system can be separated into five main steps following the direction of data in the system. The first is the definition and justification of libraries, followed by data and post-semantic processing. Having prepared the data, the empirical analysis containing portfolio testing can be conducted. Lastly, the data is segmented into training and testing sets for machine learning classification. These steps are illustrated in Figure 3.

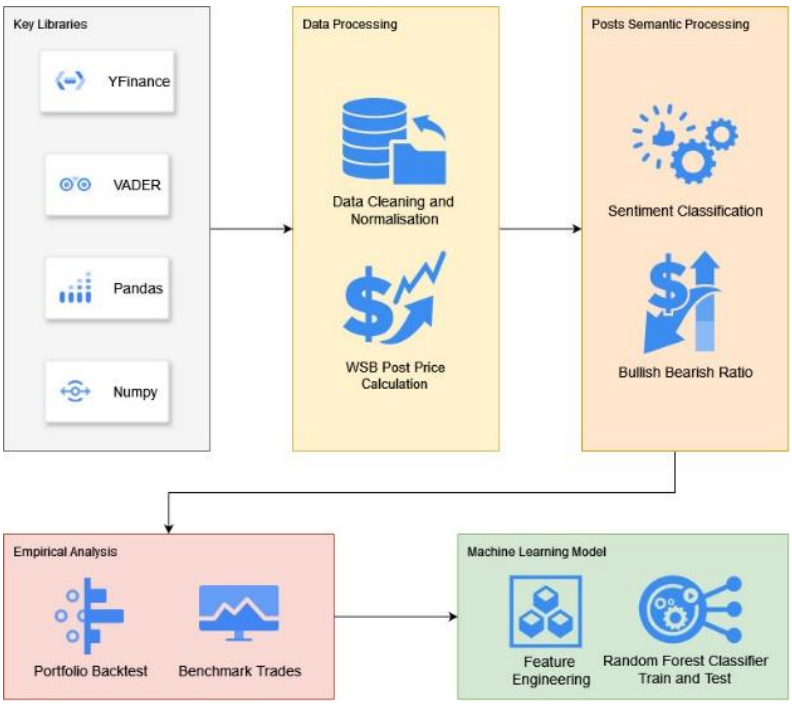

Figure 3: A diagram showing the data flow in system development.

4.1 Libraries

The first two libraries are essential to data analysis in Python, NumPy and Pandas. NumPy is for multi-dimensional arrays and matrices mathematical operations. Pandas is a powerful data analysis tool for transforming and manipulating data structures quickly. These two are integrated into each stage of the analysis and further libraries utilise these libraries' functions. Pandas' data frames form the data structure for column and row-based database views used throughout the project in both financial market and WSB datasets.

The natural language toolkit, or NLTK, provides a suite of text-processing functions for classification, tokenization, and semantic reasoning. Built on top of this library is VADER (Valence Aware Dictionary and sEntiment Reasoner), which develops stronger rule-based sentiment analysis tools specifically for social media text. Therefore, this library is essential to analyse Reddit post data. The VADER lexicon can also be customised, in our case to include a list of words specific to the WallStreetBets vocabulary and to account for common uses of sarcasm. Each new word added includes a sentiment score of either negative or positive float values. A competing sentiment analysis library to VADER is TextBlob. TextBlob features no social media score specificity and therefore the sentiment values are less accurate than VADER. Accordingly, VADER as well as NLTK were chosen for the sentiment analysis library in this research.

Data visualisation libraries for producing graphs similar to that of Figure 1 and Figure 2 are produced using the Matplotlib and Seaborn libraries. Both can generate information and statistical graphics are essential for understanding relationships in the data. Seaborn can more quickly represent data stored in Pandas' data frames, which are used in this research application.

Finance market data library yfinance is used for both real-time and historical daily market value information. As previously mentioned in the development of a WallStreetBets portfolio, the library is used to capture market open data of the equities and ETFs across different trading exchanges.

4.2 Important Functions

The function of calculating the sentiment using the VADER sentiment analysis processes each row of the WSB database and returns an integer value of 1 for positive sentiment, -1 for negative sentiment, and 0 for neutral sentiment. These values are calculated on a threshold hyperparameter, based on the compound score generated

in the following function. This is done in combination with the predict sentiment function which also takes the text of the title from the WSB database rows, but instead of returning a categorical variable, returns a float value called the compound score. The compound score is a float variable from -1 to 1 and provides a greater granularity of quantifying the sentiment. These two representations of sentiment are useful as features for the machine learning model as there is more variance in the training set.

The function used for calculating the bullish-to-bearish ratio takes its argument the starting date of the week of trading. From this start date, the range of the dates for the trading week is assigned. Iterating over the days of the trading week, two Pandas' series are initialised, the first is the list of posts made in the week, and the second is a set of the tickers in those posts. Then iterating over the unique list of tickers the total number of positive and negative mentions of the ticker is calculated. After iterating over each ticker and day of the week, a dictionary data structure is produced, with keys of the ticker, and values of the frequency count that week for positive and negative sentiment posts including that ticker. The function returns a tuple object, with the first value containing the date of the start of the week, and the second object being the dictionary data structure containing the ticker's bullish-to-bearish ratio.

4.3 Data Processing

In preprocessing the data from WallStreetBets, we decided to only focus on a subset of the data with a score greater than 1. This would remove posts with the default score value of 1, as these posts although potentially providing financial contributions to the community, received no recognition and therefore are considered to be less important by the millions of users. By focusing on a smaller group of the community we should expect higher quality posts, therefore having more statistically significant properties. An alternative subset of the dataset was tested in those posts with a score greater than the mean score, however in later testing this proved to be too small a subset of the dataset and limited the size of training and testing sets for the machine learning model.

Further steps of preprocessing the WSB dataset include producing a new column for the Date the post was made. The dataset included UTC timestamps of when a post was made, however, this is a less comparable representation of time, therefore conversion of this to the format YYYY—MM—DD is more useful, particularly in comparison to our posts with time series stock market data which uses this format.

The stock market dataset of tickers includes every recognised ticker listed by the Securities Exchange Commission, SEC [59]. This dataset of tickers is therefore comprehensive in accounting for every actively tradable equity in different exchanges. This benefits from not having to merge and concatenate different exchanges ticker listings, as we found there to be duplicates in this practice. The dataset benefitted from converting all ticker values to uppercase.

Producing an additional column in the WSB dataset to represent the tickers found in the title text was achieved using the ticker database. The post ticker and date columns are used to reference against the market value database for determining the market open value for the post that day. Then further calculate the market open value in one week past the post. This spread of price values allows for a simple comparison to represent the price direction of the equity in that trading week. We label the price movement week as a Boolean variable to show either the stock value increased as 1 or decreased as 0, one week relative to the post being made.

4.4 Posts Semantic Processing

Understanding the opinions of the posts can be achieved by quantifying the sentiment of the text. Natural language processing (NLP) is used to convert this human language into numerical representation for computers to understand. An extension of this processing method is semantic processing and analysis. Each fundamental word is a part of a lexicon where it is given a sentiment rating either positive or negative. The challenges of working with social media data are catered towards in the VADER semantic analysis library of functions we are using in this research. Reddit and WallStreetBets are home to a very unique dialect of social media language, where now the sentiment surrounding individual words and emojis (text pictograms and ideograms) is different [55]. Understanding this language in our sentiment analysis is key to the community's financial opinions on the market for the most accurate possible statistical analysis.

We start developing an understanding of the language by first reviewing posts and developing a table of contents for words and emojis with unique semantic meanings. Additional resources on semantic value scores effective in research in this same line with a similar semantic approach have been implemented [56] [57]. This table will be converted into a dictionary to be added to the VADER custom lexicon. The table generated is shown below, Table 2.

Table 2: New words dictionary and semantic value scores [57].

Word	Score	Word	Score	Word	Score	Word	Score
available	0.8	diamond_hand	3	cash	0.6	advice	1.3
awesome	3.7	dip	−0.4	concern	−1.3	alternative	0.9
baby	1.2	dumb	−1.9	crash	−3.2	amazing	3.2
bad	−2.7	earning	1.8	crazy	0.7	ass	−1.9
ball	0.4	easy	1.6	crypto	0.5	attack	−1.9
bull	2.8	end	−0.8	damn	−1.7	capital	1
bullshit	−2.4	enough	0.1	diamond	2.9	fact	0.3
buy	1.9	hype	1.2	hard	−1.1	fake	−2.3
call	0.9	idiot	−2.6	hedge	0.5	fight	−1.2
future	1.1	illegal	−3.2	hell	−2.5	fine	1.3
gain	2.2	interest	1.1	high	2.4	flair	1.4
gamma	0	issue	−1.1	hodl	2.8	fuck	−2.8
gang	−0.3	joke	−0.5	hold	1.5	fucking	−2.7
gold	2	jump	1.4	holding	1.6	fun	1.9
good	2.5	least	−0.4	hope	1.5	funny	1.9
great	3.1	legal	1.9	limit	−0.4	problem	−2.3
green	2	manipulation	−2.3	lmao	2.6	profit	2.5
hand	0.1	margin	−0.1	lol	1	proud	2.1
party	0.8	moment	0.7	long	1.8	pump	−0.5
penny	−0.2	moon	2.1	loss	−2.5	purchase	1.3
poor	−1.9	movement	0.9	love	2.3	push	0.5
possible	0.8	naked	−1.1	low	−1.7	quick	0.8
potential	1.4	nice	2	luck	2.1	retard	−2.2
power	2.2	order	0.4	revolution	2	share	0.8
pretty	2.3	panic	−3	rich	2.5	shit	−2.6
probably	0.4	straight	1	ride	1	short	−1.8
top	2.4	strong	2.1	rocket	2.8	silver	−0.2
trade	0.6	stupid	−2.1	sale	−0.7	small	−0.3
value	1.3	support	2.2	scare	−2.3	squeeze	−1.6
win	2.7	target	1.3	scared	−2.6	star	2.4
worth	1.9	tendie	1.7	sell	−1.8	stonk	1.5
wrong	−1.8	to_the_moon	3.5	seller	−1.3	stop	−0.8
yolo	2.4			selling	−1.9		

4.5 Bullish to Bearish Ratio

The bullish to bearish ratio (Bull: Bear) is to be calculated using the WallStreetBets and stock value databases and the list of equities and ETFs in the portfolio. From the WSB database, the posts are collected by week in the date column. Posts in each week are separated into those referencing individual tickers in the portfolio. Then the frequency count of each ticker's positive sentiment and negative sentiment posts is found. From this the Bull: Bear ratio can be calculated by dividing the sum of all positive posts by the total sum of positive and negative posts. If a ticker in the portfolio has not been referenced that week, a neutral sentiment score of 0.5 is output. this represents that no change in the exposure to the stock should be altered that week against the baseline. Therefore, a ratio value between 0.5 and 1 represents a strong positive sentiment for that asset, and a value between 0 and 0.5 represents a stronger negative sentiment.

Date	SPY	GME	AMC	TSLA	PLTR	AAPL	AMD	BB	AMZN	NIO	NVDA	MU	RH	SNAP	NOK	SPCE
2021-02-16	0.5	0.50	0.50	0.50	0.5	0.5	0.50	0.50	0.50	0.5	0.5	0.5	0.5	0.5	0.50	0.5
2021-02-08	1.0	0.00	0.00	0.57	1.0	1.0	1.00	0.74	0.25	0.5	0.5	0.5	1.0	0.5	0.00	0.5
2021-01-30	1.0	0.13	0.05	0.90	0.3	0.0	0.33	0.00	1.00	0.0	0.5	0.5	1.0	0.5	0.03	1.0
2021-01-23	0.5	1.00	0.50	0.33	1.0	0.5	1.00	1.00	1.00	0.5	1.0	0.5	0.0	1.0	1.00	1.0
2021-01-16	0.5	1.00	1.00	0.50	0.6	0.5	0.50	1.00	0.50	0.0	0.5	0.5	0.0	0.5	0.50	1.0

Table 3: Selection of portfolio assets Bull to Bear ratio by the date of the week.

The results for each asset in the portfolio are compiled into a Pandas data frame. The data is indexed by the date of the start of the week. Columns for each asset in the portfolio show the Bull: Bear ratio calculated in the iterative method. This data can be seen in Table 3.

4.6 Empirical Analysis

The empirical analysis constitutes the quantitative assessment of the WallStreetBets portfolio. Through back-testing trading strategies and benchmarking against the S&P 500 index, we can conclude if the most popularly invested stocks on the platform are good choices.

The trading strategy utilises the portfolio bullish to bearish ratio as well as the financial market value database. First, the ratio is scaled by a factor of 2. This scaling operation is required so that the ratio can act as a multiplier for the value of the asset's price that week. Therefore, a neutral sentiment previously of 0.5 becomes a multiplier of 1 time. The greatest factor we can generate is therefore a 2 times multiplier on the price of the asset. Having calculated the value of the asset in the portfolio we can sum the total value of each asset to generate the total portfolio value that week. Plotting the total portfolio value over the 204 weeks in the data we can output the backtest.

4.7 Classification Model

Machine learning models for a binary classification task each have individual strengths and weaknesses. The choice of determining the model is dependent on the data's properties. Four common classifiers we tested in determining the model of choice include Logistics Regression, Random Forest, Support Vector Machines, and Gradient Boosting Machines. An additional training tool we used to better optimise the data to the model is cross-validation. This tool provides more reliable estimates of the model by reducing the variance in the training and testing data split. By creating different training and testing splits on each iteration of training, the model is exposed to different training and testing samples from the same dataset. The number of different subsets we chose in training was five, called 5-fold cross-validation. This hyperparameter is tuned relative to the size of our dataset. The implementation of cross-validation and machine learning models was done using the scikit-learn library set of tools and functions.

Logistic regression provides a strong binary classification of the data when the features and the target variable are approximately linear. This is done by calculating the binary outcome using a logistic function. This model is basic and lacks any complicated statistical relationship or data structuring, therefore is likely to perform well considering our dataset has a low polynomial complexity and the features exhibit linear correlation against the target variable.

Random Forest uses decision trees for classification and performs well on different varieties of datasets, therefore making it a more general-purpose classification mode. We do not expect this model to generate the best performance metrics but still do well to generalise to the data given its robustness to overfitting.

Support Vector Machines (SVM) are a more powerful classification model attempting to calculate the plane in a lower dimension which accounts for the greatest percentage of variance in the data. This will likely be an accurate model in predicting the data's labels are our dataset is well-regularised. The benefit of this model is the ability to work well with higher dimensions of data, which would accommodate a larger more diversified portfolio if we chose to test this research against a new dataset.

Gradient Boosting builds on top of the decision trees used in Random Forest and sequentially creates new models correcting the errors made by the previous tree. We expect this method to yield high performance on our dataset, at the cost of computational complexity. However, the model may return lower performance in comparison to some simpler models as our data is simpler than that required for this model.

Comparison of each of these models will be done using model performance metrics: accuracy, precision, recall, and F1 score. We are looking to prioritise precision over recall to reduce false positives to the greatest extent possible.

5 Testing and Evaluation

In testing the data with the statistical measures, we defined for financial advice quality we must evaluate two components of this paper. First, the performance of the WallStreetBets Portfolio, and second the quality of our machine learning classification model. Should both these components generate a statistically significant result, the research question can be quickly concluded as per our means of analysis.

First, the S&P 500 index generates a 299.52% return from 21st May 2012 to 12th February 2021. Interestingly the fundamental WSB portfolio generates a 329.81% return over the same period. This demonstrates that even without a trading strategy, the portfolio already outperforms the index. The trading strategy applied to the WSB portfolio generates a return of 957.67% over the 9 years. This backtest performance is independently very strong, and when combined with the result of more than tripling the index return this portfolio strategy satisfies strong investment quality. The comparison of the portfolio and index is shown in Figure 4.

Second, the machine learning classification models are each trained on the WSB posts dataset. Each model produced an accuracy of over 70%. This accuracy is contextualised as a 0.7 probability of a post being correct in predicting the direction of the asset in a week. this is a resounding result for accuracy but doesn't account for other performance metrics. Each model produced a uniform result for the precision, outputting a value of 0.78, which translates to 78% per cent of all instances predicted as being positive. This value indicates we have a low false positive rate, beneficial in our scenario for minimising costly false positives. Each classifier output a recall of close to 1 with the exception being the Random Forest precision of 0.87. This indicates that the classifier has a low false negative rate and captures the majority of positive instances. The F1 score is calculated as the harmonic mean of precision and recall, therefore providing a balance between these two previous metrics.

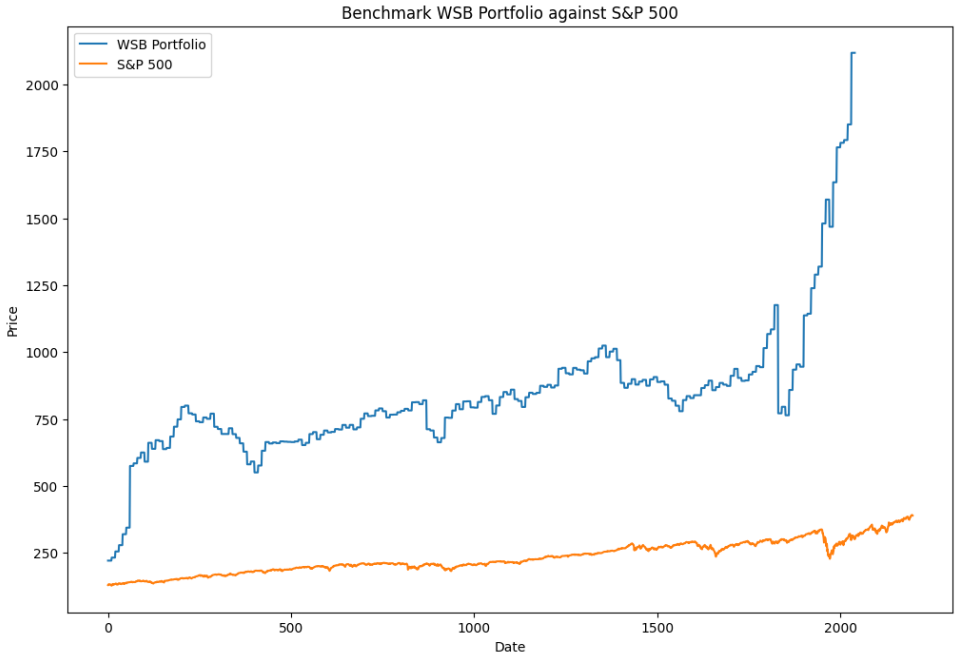

Figure 4: Portfolio backtest benchmarked against the S&P 500

From these results, we have identified the Support Vector Machine classifier to generate the highest accuracy of 0.78 in combination with high precision and recall scores of 0.78 and 0.99 respectively. The SVM model therefore satisfies our objectives for accuracy and generalisability of the target feature. The results in comparison of these classifiers can be seen below in Figure 5.

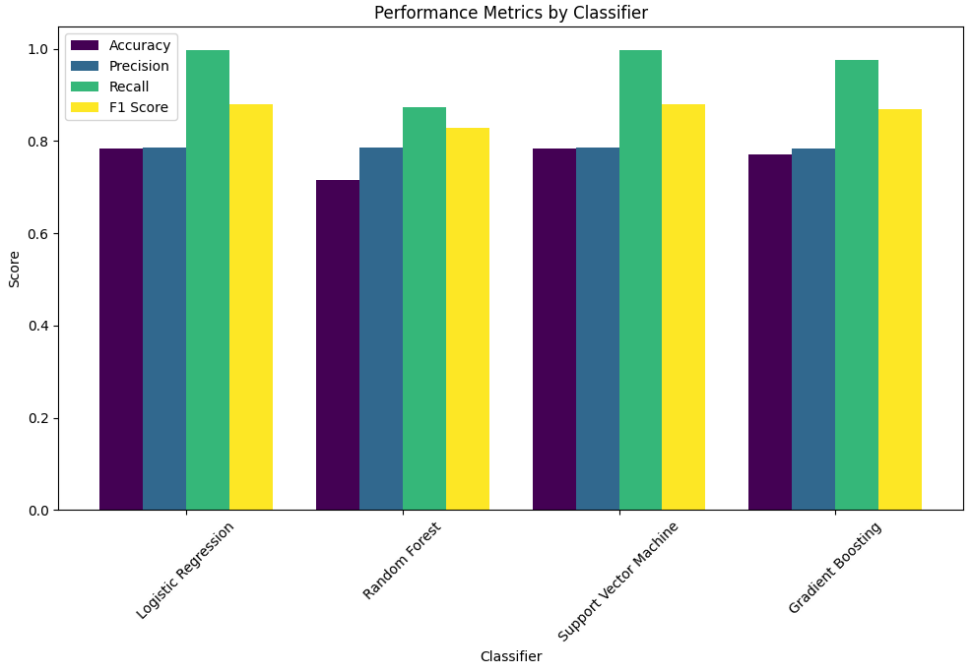

Figure 5: Machine learning classifier performance metrics comparison.

6 Critical Research Assessment

Observations and trends made during the research of this topic and others closely related show there is a general trend in the title's sentiment over time. The trend shows titles are slowly becoming more positive, hence there could well be an inherent bias in the data in more recent data which may account for the high recall scores across the machine learning models [57].

Potential extensions for further research may include consideration in the dataset that retail investor communities outside of WallStreetBets also influence the financial markets, although they might not exhibit the same phenomenon taking on Hedge Funds. This inclusion of the broader communities would provide a more generalised comment on the competition hedge funds face in the investment market.

Further extension into this research related to this topic would be to investigate more recent phenomena identified by retail investors to determine if there is yet another statistical relationship between these users and the market. One example of this most recently is the meteoric rise in the valuation of the company Nvidia, taking the interest of retail and institutional investors alike [60]. The analysis we provided was limited to the events before this company made more relevant headlines in the markets on account of the limited data-scrapping abilities. Should future research be able to alleviate this data collection restriction this would satisfy an area for improvement in our research.

Limitations surrounding data collection extend to the dataset not including the body text of posts. Having previously justified that the distribution of sentiment across the title and body text is very similar, it would still provide a greater quantity of text to derive sentiment from further bolstering the results developed by statistical models and backtests.

A comparison of our results and research against existing research highlights differences in departments of research. We found the majority of related research questions developed solutions that are confined to one domain of research. Examples of these domains include economics, psychology, language studies, and computer science [22] [37] [61] [62]. These approaches develop further details in their respective interests in

comparison to similar topics we touch on. However, our approach blends domains of research from finance, economics, and data science therefore producing a unique cross-disciplinary perspective to the research topic on WallStreetBets.

This unique contribution to the research topic utilises our own provable effective trading strategy representative of WallStreetBets traders. This approach has not been identified in other works, and we would like to attribute this insight to a greater contextual understanding of the fundamentals of the community and retail traders alike.

A comprehensive assessment of the results against the research question might conclude the research hasn't necessarily quantified the competition between hedge funds and WSB. Outside of describing both types of investors compete in a fair market and have the same objectives. In defence, our research has mentioned that WSB is against hedge funds beyond the example of Melvin Capital's bankruptcy [9]. Therefore, WSB is inclined to make market moves similar to that of the GME short, putting their capital and exposure to the market at risk. However, this risk has not been quantified and developing more evidence would satisfy an extension into this research topic.

7 Conclusion

Our results show that the community of WSB produces an effective portfolio when paired with the macro trading strategy capable of outperforming the S&P 500 index. In addition, WSB features posts which are proactive and predictive of the future asset price direction. This combination presents the community's collective efforts as strong and provably effective financial advice. The WSB community can identify assets with high investment potential and provide accurate advice for investment similar to that of institutional investors seeking shear profit such as hedge funds. With the growth in this community and those alike, there is the risk that hedge funds may lose business from investors now approaching these online forums for financial advice instead.

The growing population of retail investors online in communities such as WallStreetBets suggests that the golden age of retail investors is yet to come with emerging technologies in blockchain-based finance such as cryptocurrencies, NFTs, and decentralized finance all largely driven by retail investors. This further suggests that the new cohort of retail traders will play a key role in shaping the financial markets of the future. This mobility of market adoption is something that more traditional institutional investors lack. The tech boom of the early 2000s and now the artificial intelligence boom of this decade are reasons to believe that the change is not slowing. New and emerging companies in the wake of change are the target of retail investors such as WallStreetBets, presenting themselves as the future of financial markets. Hedge funds and related institutional investors are at risk of falling behind should they not pay attention to the trading behaviours of these online financial communities.

8 Bibliography

[1]"Retail Investor: Definition, What They Do, and Market Impact," *Investopedia*, Apr. 24, 2024. https://www.investopedia.com/terms/r/retailinvestor.asp#:~:text=Retail%20investors%20are%20non%2Dprofessional (accessed Apr. 24, 2024).

[2]"What is hedging? | Advanced trading strategies & risk management | Fidelity," *www.fidelity.com*, Apr. 24, 2024. https://www.fidelity.com/learning-center/trading-investing/hedging

[3]"What Is a Hedge Fund? | Preqin," *www.preqin.com*, Apr. 24, 2024. https://www.preqin.com/academy/lesson-3-hedge-funds/what-is-a-hedge-fund

[4]G. D. Gay and J. Hull, "Options, Futures, and Other Derivative Securities.," *The Journal of Finance*, vol. 45, no. 1, p. 312, Mar. 1990, doi: https://doi.org/10.2307/2328826.

[5]R. Stok and P. Bilokon, "From Deep Filtering to Deep Econometrics," *Social Science Research Network*, Jan. 2023, doi: https://doi.org/10.2139/ssrn.4571299.

[6]A. Hayes, "What are meme stocks?," *Investopedia*, Aug. 31, 2023. https://www.investopedia.com/meme-stock-5206762

[7]"Robinhood UK - Commission-free US Stock Trading & Investing App," *Robinhood*, Apr. 24, 2024. https://robinhood.com/gb/en/

[8]"Game Over: Robinhood Pays $7.5 Million to Resolve 'Gamification' Securities Violations | Insights | Vinson & Elkins LLP," *www.velaw.com*, Apr. 24, 2024. https://www.velaw.com/insights/game-over-robinhood-pays-7-5-million-to-resolve-gamification-securities-violations/

[9]"Melvin Capital," *Wikipedia*, Jan. 29, 2022. https://en.wikipedia.org/wiki/Melvin_Capital

[10]R. Davies, "GameStop: how Reddit amateurs took aim at Wall Street's short-sellers," *the Guardian*, Jan. 28, 2021. https://www.theguardian.com/business/2021/jan/28/gamestop-how-reddits-amateurs-tripped-wall-streets-short-sellers

[11]"Subreddit Stats - statistics for every subreddit," *subredditstats.com*, Apr. 24, 2024. https://subredditstats.com/r/wallstreetbets

[12]P. Gurrola-Perez, K. Lin, and B. Speth, "Retail trading: an analysis of global trends and drivers," 2022. Available: https://www.world-exchanges.org/storage/app/media/WFE-Retail-Investment%20Sep%2020%202022.pdf

[13]P. Gurrola-Perez, K. Lin, and B. Speth, "Retail trading: an analysis of global trends and drivers," 2022. Available: https://www.world-exchanges.org/storage/app/media/WFE-Retail-Investment%20Sep%2020%202022.pdf

[14]"Backtest your portfolio performance," *www.lseg.com*, Apr. 24, 2024. https://www.lseg.com/en/data-analytics/asset-management-solutions/portfolio-management/backtest-your-portfolio-performance#what-is (accessed Apr. 24, 2024).

[15]J. Brown, "Past Performance Is Not Indicative Of Future Results," *Forbes*, Apr. 24, 2024. https://www.forbes.com/sites/johnbrown/2016/09/29/past-performance-is-not-indicative-of-future-results/?sh=c7da0b73bf5b (accessed Apr. 24, 2024).

[16]W. J. CFP®, "Thoughts From Willis | What it Means to Have a Proactive Versus a Reactive Advisor," *insights.wjohnsonassociates.com*, Apr. 24, 2024. https://insights.wjohnsonassociates.com/blog/thoughts-

from-willis-what-it-means-to-have-a-proactive-versus-a-reactive-advisor (accessed Apr. 24, 2024).

[17]"Am I Outperforming the S&P 500?," *Magnifi.com*, 2023. https://magnifi.com/learn/am-i-outperforming-the-s-and-p-500 (accessed Apr. 24, 2024).

[18]"Bullish and Bearish," *Corporate Finance Institute*, Apr. 24, 2024. https://corporatefinanceinstitute.com/resources/equities/bullish-and-bearish/

[19]"Institutional Traders vs Retail Traders: What's the Difference? | FXOpen," *Market Pulse*, Feb. 06, 2024. https://fxopen.com/blog/en/institutional-and-retail-traders-where-the-difference-lies/ (accessed Apr. 24, 2024).

[20]Editor, "What Is YOLO Trading?," *Financhill*, Jun. 23, 2021. https://financhill.com/blog/investing/what-is-yolo-trading (accessed Apr. 24, 2024).

[21]"A Comprehensive WallStreetBets-vs-the-Establishment Explainer," *kogod.american.edu*, Jan. 29, 2021. https://kogod.american.edu/news/wallstreetbets-vs-estrablishment (accessed Apr. 24, 2024).

[22]C. Boylston, "(PDF) WallStreetBets: Positions or Ban," *ResearchGate*, Jan. 01, 2021. https://www.researchgate.net/publication/348860912_WallStreetBets_Positions_or_Ban (accessed Apr. 24, 2024).

[23]"r/WallStreetBets," *Wikipedia*, Jan. 27, 2021. https://en.wikipedia.org/wiki/R/WallStreetBets (accessed Apr. 24, 2024).

[24]"wsb yolo - YouTube," *web.archive.org*, Nov. 06, 2019. https://web.archive.org/web/20191106045411/https://www.youtube.com/watch?&v=A-tNkuYV4_Q (accessed Apr. 24, 2024).

[25]"Explainer: The rise of 0DTE stock options and how they could be a risk to markets," *Reuters*, Feb. 22, 2023. https://www.reuters.com/markets/us/rise-0dte-stock-options-how-they-could-be-risk-markets-2023-02-22/ (accessed Apr. 24, 2024).

[26]"Barclays US Equity Derivatives Strategy Impact of Retail Options Trading | PDF | Option (Finance) | Stocks," *Scribd*, Apr. 24, 2024. https://www.scribd.com/document/521690968/Barclays-US-Equity-Derivatives-Strategy-Impact-of-Retail-Options-Trading

[27]MERRILL, "Understanding Option Pricing: Intrinsic & Time Value," *Merrill Edge*, Apr. 24, 2024. https://www.merrilledge.com/investment-products/options/options-pricing-valuation

[28]"Difference between Implied, Realized and Historical Volatility - Macroption," *www.macroption.com*, Apr. 24, 2024. https://www.macroption.com/implied-vs-realized-vs-historical-volatility/

[29]T. text provides general information S. assumes no liability for the information given being complete or correct D. to varying update cycles and S. C. D. M. up-to-Date D. T. R. in the Text, "Topic: Barclays Bank," *Statista*. https://www.statista.com/topics/3913/barclays-bank/#topicOverview (accessed Apr. 24, 2024).

[30]J. Fernando, "Market Capitalization: How Is It Calculated and What Does It Tell Investors?," *Investopedia*, Mar. 16, 2023. https://www.investopedia.com/terms/m/marketcapitalization.asp (accessed Apr. 24, 2024).

[31]"Update: Three rounds of stimulus checks. See how many went out and for how much. | Pandemic Oversight," *www.pandemicoversight.gov*, Feb. 17, 2022. https://www.pandemicoversight.gov/data-interactive-tools/data-stories/update-three-rounds-stimulus-checks-see-how-many-went-out-and

[32]“WallStreetBets: Yesterday’s Meme or Here to Stay?,” *Bloomberg*, Apr. 24, 2024. https://sponsored.bloomberg.com/article/capital/wallstreetbets-yesterday-s-meme-or-here-to-stay (accessed Apr. 24, 2024).

[33]E. Boyd, “Case Study: Robinhood Removing Barriers to Entry,” *Financial Marketer*, Aug. 19, 2021. https://financial-marketer.com/case-study-robinhood-removing-barriers-to-entry/

[34]Robinhood, “About Us,” *Robinhood*, 2022. https://robinhood.com/us/en/about-us/

[35]K. R. Fitzgerald, “Here’s how Robinhood is raking in record cash on customer trades — despite making it free,” *CNBC*, Aug. 13, 2020. https://www.cnbc.com/2020/08/13/how-robinhood-makes-money-on-customer-trades-despite-making-it-free.html

[36]“Robinhood Taps JPMorgan to Process Transactions,” *BrainStation®*, May 03, 2021. https://brainstation.io/magazine/robinhood-taps-jpmorgan-to-process-transactions (accessed Apr. 24, 2024).

[37]T. Buz, “DEMOCRATIZATION OF RETAIL TRADING: CAN REDDIT’S WALLSTREETBETS OUTPERFORM INVESTMENT BANK ANALYSTS?,” Dec. 2022. Accessed: Apr. 24, 2024. [Online]. Available: https://arxiv.org/pdf/2301.00170.pdf

[38]W. Kenton, “S&P 500 Index: What It’s for and Why It’s Important in Investing,” *Investopedia*, Mar. 23, 2021. https://www.investopedia.com/terms/s/sp500.asp

[39]K. Fisher, “Why you should benchmark your portfolio against a global stock index,” *The National*, Oct. 03, 2023. https://www.thenationalnews.com/business/money/2023/10/03/why-you-should-benchmark-your-portfolio-against-a-global-stock-index/ (accessed Apr. 24, 2024).

[40]“Understanding Benchmarks | PIMCO,” *Pacific Investment Management Company LLC*, Apr. 24, 2024. https://www.pimco.co.uk/en-gb/resources/education/understanding-benchmarks

[41]“8 Feature Engineering Techniques for Machine Learning,” *ProjectPro*. https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423 (accessed Apr. 24, 2024).

[42]J. Brownlee, “Why One-Hot Encode Data in Machine Learning?,” *Machine Learning Mastery*, Jul. 27, 2017. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ (accessed Apr. 24, 2024).

[43]“Top Performance Metrics in Machine Learning: A Comprehensive Guide,” *www.v7labs.com*, Apr. 24, 2024. https://www.v7labs.com/blog/performance-metrics-in-machine-learning

[44]“How to reduce false positives in fraud prevention,” *Ravelin*. https://www.ravelin.com/blog/reduce-false-positives-fraud (accessed Apr. 24, 2024).
[45]J. Reece, “Warren Buffett Detailed Fundamental Analysis - HD.” https://www.nasdaq.com/articles/warren-buffett-detailed-fundamental-analysis-hd-8 (accessed Apr. 24, 2024).

[46]“Reddit - r/wallstreetbets,” *www.kaggle.com*, Apr. 24, 2024. https://www.kaggle.com/datasets/unanimad/reddit-rwallstreetbets (accessed Apr. 24, 2024).

[47]“Why is Data Quality Important in ML?,” *censius.ai*. https://censius.ai/blogs/importance-of-data-quality (accessed Apr. 24, 2024).

[48]lift_ticket83, “Reddit Data API Update: Changes to Pushshift Access,” May 01, 2023. https://old.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/

(accessed Apr. 24, 2024).

[49]Creative Commons, “Creative Commons — CC0 1.0 Universal,” *Creativecommons.org*, 2019. https://creativecommons.org/publicdomain/zero/1.0/

[50]Jason Brownlee, “Impact of Dataset Size on Deep Learning Model Skill And Performance Estimates,” *Machine Learning Mastery*, Aug. 06, 2019. https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/

[51]IBM, “What Is Exploratory Data Analysis? | IBM,” *www.ibm.com*, 2020. https://www.ibm.com/topics/exploratory-data-analysis (accessed Apr. 24, 2024).

[52]“Raphael Fontes | Master,” *www.kaggle.com*, Apr. 24, 2024. https://www.kaggle.com/unanimad (accessed Apr. 24, 2024).

[53]“Bullish vs. Bearish Definition,” *NerdWallet*, Sep. 20, 2023. https://www.nerdwallet.com/article/investing/bullish-vs-bearish#:~:text=The%20main%20difference%20between%20bullish (accessed Apr. 24, 2024).

[54]“Virgin Galactic Holdings, Inc. (SPCE) Stock Price, News, Quote & History - Yahoo Finance,” *finance.yahoo.com*, Apr. 24, 2024. https://finance.yahoo.com/quote/SPCE/

[55]“Honorary WSB Autist award goes to Chamath Palihapitiya! He is out here defending us retail investors and calling out hedge funds for their manipulation and bullshit practices that have left retail holding the bag for YEARS. GME going 🚀 🚀 🚀 🚀 🚀 🚀 ,” *Reddit*, Jan. 27, 2021. https://www.reddit.com/r/wallstreetbets/comments/l69jz5/honorary_wsb_autist_award_goes_to_chamath/ (accessed Apr. 24, 2024).

[56]C. Smythe, “The r/WallStreetBets Glossary: A field guide to apes, stonks, tendies, and more,” *www.businessofbusiness.com*, Jun. 17, 2021. https://www.businessofbusiness.com/articles/the-wallstreetbets-glossary-a-field-guide-to-apes-stonks-tendies-AMC-GME-reddit/

[57]S. (Cheng) Long, B. Lucey, Y. Xie, and L. Yarovaya, “‘I just like the stock’: The role of Reddit sentiment in the GameStop share rally,” *Financial Review*, vol. 58, no. 1, pp. 19–37, Oct. 2022, doi: https://doi.org/10.1111/fire.12328.

[58]“Reddit WallStreetBets Posts Sentiment Analysis,” *kaggle.com*, Apr. 24, 2024. https://www.kaggle.com/code/thomaskonstantin/reddit-wallstreetbets-posts-sentiment-analysis (accessed Apr. 24, 2024).

[59]SEC, “SEC.gov | Home,” Sec.gov, 2024. https://www.sec.gov/ (accessed Apr. 24, 2024).

[60]D. Clark, “Nvidia Says Growth Will Continue as A.I. Hits ‘Tipping Point,’” The New York Times, Feb. 21, 2024. Accessed: Apr. 24, 2024. [Online]. Available: https://www.nytimes.com/2024/02/21/technology/nvidia-earnings.html

[61]“The Mobilisation of r/wallstreetbets,” 2020. https://www.diva-portal.org/smash/get/diva2:1606894/FULLTEXT01.pdf (accessed Apr. 24, 2024).

[62]G. Kinch and C. Tjernberg, “r/wallstreetbets Influence on the Stock Market Sentiment Analysis on r/wallstreetbets during one of the loudest and most noticeable periods of financial debate on social media,” 2022. Accessed: Apr. 24, 2024. [Online]. Available: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9082927&fileOId=9082966