## Abstract

## Introduction

Be specific and clearly structured. Answering the sub-rqs leads to a clear solution of the main RQ.

1. Talk about recent developments.

2. Those developments lead to my RQs (show how they are connected).

3. Try to answer the sub-questions to come up with a convincing solution to the main RQ.

Modern society has been able to access information, communicate ideas, and become part of a community due to the advent of the internet. Of course, online discussion boards have been playing a critical role to provide a platform where people can do so. Those are also used by a variety of people to talk about the stock market and discuss trading strategies. Recently, the Reddit forum wallstreetbets has become one of the most well-known and influential investing online-forums.

Even though the Reddit subforum was already founded in 2012, it received the majority of its media exposure in 2021 as a result of a short-squeeze of the GameStop (GME) stock, which drove the stock price hundreds of percent. However, it was not the rapid price appreciation that amazed market participants. Instead, it was the unprecedented decentralized and coordinated buying of Gamestop shares by members of the wallstreetbets community that attracted attention. (Anand & Pathak, 2021)

Organizing the mass-coordinated buying of stock, however, requires that enough participants share the same sentiment. According to several studies, social media sentiment has a particularly strong impact on uninformed traders (Danbolt, Siganos, & Vagenas-Nanos, 2015). Write a bit more about retail investors.

Interestingly, finance scholars did not consider Reddit as a platform capable of having such a significant impact on the financial markets. As a result, the site has been neglected in their research (Long, Lucey, & Yarovaya, 2021).

Hence, this thesis will attempt to contribute to the scientific community by answering the following Research Question:

*Can sentiment analysis of the WallStreetBets Reddit-forum be used to predict daily changes in the stock price of Gamestop?*

To answer this research question several fields in the domains of machine learning and finance need to be explored. To begin, it must be determined how the discussions about the Gamestop stock on WallStreetBets should be handled in order to serve as suggestive input features for sentiment analysis. One of the challenges, is the heavy use of peculiar terminology and domain-specific phrases on the WallStreetBets forum, as well as many novel words (Anand & Pathak, WallStreetBets Against Wall Street: The Role of Reddit in the GameStop Short Squeeze, 2021). According to recent research, sentiment lexicons with a focus on a certain domain produce superior sentiment analysis results than a general-purpose sentiment lexicon (Park, Lee, & Moon, 2015). Furthermore, the text data needs to be cleaned and pre-processed in order to be accurately processed by a machine learning algorithm (Jemai, Hayouni, & Baccar, 2021). As a result, the following sub-research question was formed:

*How can the domain-specific language of the Reddit forum WallStreetBets best be incorporated into sentiment analysis?*

Subsequently, the machine learning models can be trained to perform sentiment analysis. However, each machine learning algorithm has its own idiosyncrasies and assumptions, and no single classifier works optimally in all possible scenarios. Hence, it is a good idea to evaluate the results and performance of different machine learning algorithms. As a result, the best model with a given set of hyperparameters can be selected to solve a particular problem (Raschka & Mirjalili, 2019). This thesis will explore traditional machine learning methods such as Naïve Bayes (NB) and Support Vector Machines (SVMs), as well as deep learning methods like Long Short Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). Due to the high dimensionality of textual data, deep learning methods have shown to outperform traditional machine learning techniques in recent research. That can be explained by the ability of deep learning methods to automatically learn the most important features, whereas traditional methods may suffer from the curse of dimensionality (Fu, Yang, Li, Fang, & Wang, 2018).

As was mentioned earlier, however, no classifier works best on all scenarios which is why the next research question needs to be answered:

*Which sentiment analysis approach performs best on predefined key performance indicators?*

The impact of sentiment on stock prices has gained attention by researchers in recent years. For example, it is shown that social media sentiment can have a direct effect of how market participants perceive a company, which can lead to changes in the stock price of companies. This is especially true for smaller firms with low analyst coverage (Feng & Johansson, 2019). Other researchers show that sentiment obtained from Twitter can be used to predict returns of a broader stock market index (Gu & Kurov, 2020). In other research the emotions of discussions on WallStreetBets are studied by performing sentiment analysis. The research suggests that only some emotions demonstrate a significant impact on one-minute returns of the Gamestop shares (Long, Lucey, & Yarovaya, 2021).

There are several models that have shown strong results with regards to forecasting time-series. The most prevalent in the financial industry is Auto Regressive Integrated Moving Average (ARIMA) which captures temporal structures in time-series data. However, it is not designed to include other features, such as sentiment. This is why this thesis will also compare other models such as LSTMS and XGBoost, which have also demonstrated strong predictive capabilities with regards to time-series data. As a result, the final sub-research question will explore the final part of the main research question.

*Which machine learning algorithm delivers the best predictive performance for changes in daily stock prices of Gamestop based on the sentiment analysis performed earlier?*

## Literature

Talk about literature. Relevant, use a lot, processed with new insights provided. Clearly connected to RQ.

Gauging sentiment of online forums to predict movements in stock prices has been a research subject for many years now. Das & Chen (2007) did a study on the Yahoo! message board, which was amongst the first ones on the internet for investors to exchange ideas. In their paper, they show that the relationship of stock price to sentiment is significant and that market activity is related to activity of the message boards. Other researchers, such as (Lyócsa, Baumöhl, & Vŷrost, 2021) also showed that as the discussion volume on WallStreetBets increased, the volatility of certain stocks got amplified. (Umar, Gubareva, Yousaf, & Ali, 2021) also found that sentiment of investors on WallStreetBets affected the returns of the Gamestop stock. However, they also show that other features such as the put-call ratio and the short-sale volume had a strong impact on the stock price.

Long, Lucey, & Yarovaya (2021) tried to uncover the impact of specific emotions such as *“Angry, Fear, Happy, Sad and Surprise”* from the comments on WallStreetBets discussions on intraday changes of the stock price of the affected stock. While they conclude that the tone as well as the number of comments has an impact on the stock price, they show that the number of comments is not related to sentiment. Additionally, they argue it is the number of comments that is posted within an hour that has the biggest effect on one minute changes in the stock price. Furthermore, the paper shows that the emotions *Sad, Anger* and *Surprise* have a significant impact on the gamestop 1-minute stock price. The *Happy* sentiment does not show a significant impact on 1-minute price changes, however, a causality test showed a link between the *Happy* sentiment and intraday returns of the GME stock. In addition, the paper shows, that sentiment only impacts intraday returns if a thread has more than 2000 comments. Hence, the authors confirm that Reddit sentiment has an impact on the stock market. They also argue that any asset that is targeted by a large crowd from wallstreetbets can become a subject of excessive volatility, without being driven by any fundamental reasons.

Since the WallStreetBets meme-stock movement is a relatively recent phenomenon, there is very little research on the impact of wallstreetbets on individual stocks, especially with regards to sentiment analysis. However, of all the published research, none account for the domain-specific language used on the forum. This also applies to research in other fields, who usually also use a general-purpose sentiment lexicon, because of the cost associated with building a domain-specific one. However, it has been demonstrated that using a domain-specific knowledge base results in more accurate sentiment analysis (Park, Lee, Moon – 1-s2.0-S0….).

It is argued that there is no general-purpose sentiment lexicon that can be optimally applied on all domains. In different domains, some terms can have different meanings. A good example is the word “unpredictable”, which would have negative sentiment for electronics but can be a positive label for movies. It has been demonstrated that by adapting sentiment lexicons to a certain domain performance for sentiment classification can be enhanced (Lu, Castellanos, Dayal, & Zhai, 2011). This adapted lexicon can then be searched to find and score the sentiment of a specific word (Asghar, 2014).

While lexicon-based methods have found widespread adoption, mainly due to their simplicity, other machine learning methods have also shown strong performance (tfidf/Wang2020\_Article\_...pdf).

With the support of an Active Learner, a complete domain-specific corpus with its respective labels can be created using only partial annotations (Park, Lee, Moon).

## Methodology

Clear explanation of methods with connections drawn to other methods, appropriate robustness checks of assumptions, consistent, transparent, and correct.

### Data Preprocessing

The research by Jemai, Hazouni, and Baccar (2021) presents a system for structuring a sentiment analysis project. The data collection phase is the first step, where textual data is obtained from a source. The data is then cleaned in the second step, the data pre-processing phase. To do so, several actions need to be performed. Data tokenization is one of the actions. This is a common technique in which a large body of text is broken down into multiple sentences, each of which is then broken down into a list of words. Stop words such as *is, the, a* and other common words are also removed during the pre-processing phase. In addition, special characters such as @ and urls should also be removed. It is also suggested that the text is converted to lowercase. As the final step, the research proposes lemmatization. By doing so, the structure of a word is analyzed and converted to its normalized form.

### Data Preparation

The literature writes about various methods to prepare text data for machine learning.

Visualization with t-sne:

### Chart, bubble chart Description automatically generated

### Cost of Annotation

### The Case for a Supervised Method over an Unsupervised Method

Since the data obtained from Reddit is unlabeled, it cannot be fed into the Machine Learning Algorithms. However, many promising sentiment analysis methods rely on labeled data (Sazzed & Jayarathna, 2021). One approach to label data is using unsupervised machine learning models. Unsupervised models are commonly applied in Natural Language Processing and text classification (Unsupervised.pdf, Jung Lee). However, unsupervised models are a better choice for uncovering hidden patterns in a dataset, especially without any a priori knowledge of the structure of the data. As a result, unsupervised models excel at summarizing or exploration a large text corpus. However, if there are predefined categories, supervised learning methods can be more applicable than an unsupervised method. One of the major disadvantages of supervised models, however, is the cost associated with manually labelling the data (active-learning-approaches-for-labelling-text, Miller, Linder, Mebane).

### Active Learners to Reduce the Cost of Annotation

One of the key concepts of Active Learners is that if a machine learning algorithm is allowed to choose the data from which it learns, it will achieve higher accuracy with less training data. If a considerable amount of the data is unlabeled, this is especially desirable. As a result, the total cost of annotation can be reduced drastically. Research shows that the total number of manual annotations can be reduced by 80% when using an Active Learner instead of randomly selecting data to label (Active Learning and Cost of annotation – Jason Baldridge and Miles Osborne).

If data is manually annotated at random, the annotator will invest a lot of time into labeling irrelevant instances. This may incur costs which could be avoided with an Active Learner. It is argued that Passive Learning, or randomly selecting instances to be labeled by an annotator, is especially costly if the class distribution of the data is imbalanced or if there are many very similar documents. For example, if a specific feature set appears on only 1% of instances, the annotator would have to label 1000 documents to cover the feature set on 10 relevant documents. When it comes to document similarity, large clusters of very similar documents might be identifiable. Because features may be barely distinctable, the annotator might spend a lot of effort labeling uninformative instances when selecting them random. An Active Learner, on the other hand, suggests which instances the annotator should label. Those instances can be determined on various quantitative metrics (active-learning-approaches-for-labelling-text, Miller, Linder, Mebane).

### Using Active Learning to Capture the Domain-Specific Language of WallStreetBets

I kinda repeat this from the literature section. However, I still will want to have the explanation of an AL in relation to a domain specific lexicon!

~~Many industry applications and researchers use general-purpose sentiment lexicons, because of the cost associated with building a domain-specific one. However, it has been demonstrated that using a domain-specific knowledge base results in more accurate sentiment analysis (Park, Lee, Moon – 1-s2.0-S0….).~~

~~It is argued that there is no general-purpose sentiment lexicon that can be optimally applied on all domains. In different domains, some terms can have different meanings. A good example is the word “unpredictable”, which would have negative sentiment for electronics but can be a positive label for movies. It has been demonstrated that by adapting sentiment lexicons to a certain domain performance for sentiment classification can be enhanced (Lu, Castellanos, Dayal, & Zhai, 2011). While lexicon-based methods have found widespread adoption, mainly due to their simplicity, other machine learning methods have also shown strong performance (tfidf/Wang2020\_Article\_...pdf).~~ Since this thesis also tries to compare the performance of different machine learning algorithms for sentiment analysis, the focus is on creating a domain-specific corpus which can be used to train the models.

With the support of an Active Learner, a complete domain-specific corpus with its respective labels can be created using only partial annotations (Park, Lee, Moon).

### Active Learner Workflow

The illustrated workflow provides an overview of how an Active Learner works. To begin with, cleaned and pre-processed data needs to be available that can be used by the Active Learner. Furthermore, the Active Learner can also be trained with some initial training data, which is also referred to as the seed. By using clustering algorithms, the *seed* data can be selected methodologically, which allows the Active Learner to achieve higher accuracy faster when compared to randomly picking the initial seed data (Kang, Ryu, & Kwon, 2004). All the unlabeled instances will become the *pool* data, which need to be labeled. The labeled data is fed into the Active Learner. In addition, a query strategy needs to be defined, based on which the Active Learner queries new instances from the aforementioned pool. A query strategy evaluates the informativeness of unlabeled samples. Common strategies are *uncertainty sampling, query-by-committee, expected model change, expected error reduction and variance reduction*.

After the query function selected instances from the pool, an oracle needs to label those. An oracle is normally at least one human with training on how to annotate the data at hand (Settles, 2009). However, human annotators are oftentimes inconsistent and the result may vary from person to person (Ali & Gayar, 2019). Once the new instances are labeled, those instances need to be removed from the pool, since they are now part of the labeled data. The Active Learner then needs to be taught the new instances, which he can use to adjust the model. After each iteration, the results can be evaluated. A common performance measure for Active Learners is accuracy. *To see how Accuracy is calculated go to page xy in section Evaluation*. If a predefined stopping criterion is not yet met, the query strategy selects more instances from the pool and repeats the process. If the stopping criterion is met, the process ends (Lu, Henchion, & Namee, 2019).

### Active Learner Implementation

To implement the Active Learner the modAL package was used. modAL was designed with modularity, flexibility and extensibility as high priorities (Danka & Horvath, 2018). The estimator defined in the Active Learner object is a Support Vector Machine. A SVM was chosen because of its strong generalization performance (Alves, Baptista, Firmino, de Oliveira, & de Paiva, 2014). Uncertainty sampling was chosen as the query strategy because it has been demonstrated to be a strong baseline strategy. This query strategy assumes, that instances that are far from the decision boundary are adequately explainable and instances close to the decision boundary are uncertain. As a result, the Active Learner queries the samples about which it is most uncertain about (Osborne & Baldridge, 2004).

## Sentiment Analysis Models

The next section explores the machine learning models that will be used to perform sentiment analysis on the domain-specific corpus created by the Active Learner. The classification of the best performing model will then be used as an input to predict changes in stock prices.

Naïve Bayes (NB): NB is a probabilistic supervised machine learning model. By working probabilistically, the classifier assigns the probability of belonging to a given class based on certain features (Jemai, Hayouni, & Baccar, 2021). Because of the high dimensionality of text data, which can be handled very well by NB, this algorithm has established itself as one of the standards for sentiment analysis. This thesis will use Multinomial Naïve Bayes to classify the sentiment of the text. This is due to the model’s ability to handle larger vocabulary sizes. In addition, the algorithm is simple to implement, suitable for real-time applications, and highly scalable. However, the algorithm’s prediction accuracy is frequently lower than that of other sentiment analysis techniques (Song, Kim, Lee, Kim, & Youn, 2017).

Support Vector Machines (SVM): SVMs can be used to solve both regression and classification problems. Classification is done by finding a hyper/plane with the biggest margin, meaning it looks for the greatest distance to the nearest sample points (Jemai, Hayouni, & Baccar, 2021). SVMs use spatial transformations, commonly known as kernel functions, to fit the hyperplane. Kernels can be linear, RBF or others. The radial basis function (rbf) kernel is best used for non-linear problems and is a general-purpose kernel that is often used in pattern recognition problems. The linear kernel, on the other hand, is typically used when there are only two classes present. A good example for that might be positive and negative sentiment (Alves, Baptista, Firmino, de Oliveira, & de Paiva, 2014).

Long Short Term Memory (LSTM): LSTMs are becoming increasingly popular for sentiment classification. LSTMs are built on a recurrent neural network architecture (RNN). In an RNN the neurons are connected to themselves through time. As a result, the input from a time instance ti will also be used as an input for the next time instance ti+1.That leads to the problem of vanishing gradients. LSTMS are designed to overcome that problem.

The LSTM architecture does so via its four constituents: A memory cell which can remember a lot of information from previous states, an input gate which controls the inputs into the neurons, an output gate with an activation function and lastly a forget gate which resets the neuron (Priyantina & Sarno, 2019).

BERT: BERT is a relatively new machine learning algorithm developed by Google in 2018 and mainly designed for natural language processing. BERT is pretrained on the English Wikipedia and BooksCorpus. Because of the pretraining users won’t need as much computing power to achieve good results, even if the dataset is relatively small (Devlin, Chang, Lee, & Toutanova, 2019). The BERT github page even states that “Most NLP researchers will never need to pre-train their own model from scratch” (google-research, 2020).

### Hyperparameter Tuning

In order to find the best performing models, some hyperparameter tuning steps were taken. For the implementation of the Naïve Bayes and Support Vector Machine models, five-fold grid search cross-validation was used to find the best parameters in a pre-defined parameter grid.

For the deep learning models, LSTM and BERT, a loop was created that iterates over a set of parameters. Within each iteration, the model is fit on the training data while setting aside 20% of the data for validation.

Results

Clear, transparent, and original presentation of results. Detailed visualizations, insightful multi-level evaluation of model performance.

## Discussion

## Conclusion