## 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 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 specific 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 the machine (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. This is why it is a good idea to evaluate the results and performance of different 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 wavelet coherence framework (???), which have also demonstrated strong predictive capabilities with regards to time-series data. As a result, the final sub-research question will hopefully be able to answer 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?*

## Related Work

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

## Methodology

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

### Cost of Annotation

### Active Learner

Since the data obtained from Reddit is unlabeled, it cannot be fed into the Machine Learning Algorithms. However, supervised machine learning methods need to utilize labeled data (Sazzed & Jayarathna, 2021). Talk about why Active Learner was chosen instead of Unsupervised Learner, etc.

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 desirable. As a result, the total cost of annotation can be reduced drastically.

The Active Learner implementation for this thesis is as follows:

First, an Active Learner Object is created.

Results

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

## Discussion

## Conclusion