1: Can sentiment analysis of Reddit forums be used to predict daily changes of the stock price of selected securities?

2: Which sentiment analysis approach performs best on predefined key performance indicators, based on selected sentiment analysis methods?

3: Which machine learning algorithm delivers the best predictive performance for changes in implied volatility in the options market of selected securities based on the sentiment analysis performed earlier?

4: How can the domain-specific language of the reddit forum wallstreetbets best be incorporated into sentiment analysis?

I recommend writing your related work section in a manner that follows the structure of your research questions. This helps the reader to understand how your research questions are related to the literature and how your thesis contributes to the literature. Furthermore, it gives the thesis an organized feel, which tends to increase the grade for the form and presentation components of your thesis.

Best,

Peter.

Proposal

What should be in the draft:

- Good idea of research

- How will it be approached

- Show that I have enough knowledge

Important points:

Motivate your thesis from both societal and a scientific perspective

Write the background section in a manner that is consistent with your rqs

Describe the exact methods you intend to use (which features, ML methods, etc.)

Working plan: Make it very detailed (best: week by week plan).

Prof. will not give feedback on the actual proposal until we actually submit it. He will only give feedback on the draft.

Think about how you wanna evaluate the algorithms.

Put the rqs in the project definition.

We can write in I form. Some people have issues with that. Prof. will make sure that the second reader doesn’t complain about that.

Outline

Project Definition: Provide a clear description of the problem you plan to address.

Motivation: Explain why this problem is worth addressing (scientific and societal).

Background: What is known in the scientific literature. At least five relevant sources. Sources must be recent, high quality and useful.

Dataset Description: Describe the dataset that is used.

Algorithms and Software: Which algorithms and software will be used? How will ML algorithms be evaluated? What is the baseline? How can ground-truth data labeled data be obtained? -> So I can measure accuracy, precision, or some other metric.

Evaluation Method: How will the results be evaluated?

Milestones and plan: Sketch out some major intermediate milestones.

## Project Definition

The internet has enabled humankind to access information, exchange ideas and become part of a community. Of course, that also applies to message boards. Ever since the internet found mainstream adaption, people joined those message boards to discuss about trading the stock market. Most recently the reddit forum wallstreetbets attracted a lot of interest and now counts over 10 million members. While the wallstreetbets community minted many millionaires, there are also numerous people who lost their life savings.

Even though the reddit subforum was already founded in 2012, it got most of the media attention in 2021 due to a short-squeeze of the GameStop (GME) stock that drove the stock price up several hundred 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)

Interestingly, the story repeated itself when those self-proclaimed “apes” sent other stocks, such as AMC Entertainment and BlackBerry to the moon.

Organizing the mass-coordinated buying of stock, however, requires that enough participants share the same sentiment. Some research shows that social media sentiment has a particularly strong impact on uninformed traders (Danbolt, Siganos, & Vagenas-Nanos, 2015). Interestingly, finance scholar did not consider Reddit as a platform that can have such a big impact on the financial markets and has hence been ignored (Long, Lucey, & Yarovaya).

Hence, this Master thesis will focus on the “meme stock” driven investor sentiment of the wallstreetbets subreddit. By performing sentiment analysis on the aforementioned forum, it is assumed to be able to predict changes in the stock price. Being able to accurately measure the sentiment ensures more efficient, and hence, less volatile markets.

Furthermore, this thesis will analyze how to best incorporate domain-specific language that is commonly used on wallstreetbets into the sentiment analysis. This thesis will try to answer the following research questions:

1: Can sentiment analysis of Reddit forums be used to predict daily changes of the stock price of selected securities?

2: Which sentiment analysis approach performs best on predefined key performance indicators, based on selected sentiment analysis methods?

3: Which machine learning algorithm delivers the best predictive performance for changes in implied volatility in the options market of selected securities based on the sentiment analysis performed earlier?

4: How can the domain-specific language of the reddit forum WallStreetBets best be incorporated into sentiment analysis?

## Motivation

Since the the wallstreetbets subreddit has become very popular just recently, there is very little academic research about the impact of the community on financial markets so far. This thesis not only tries to shine some light on this novel challenge, but also tries to put forward some methods that can be used to perform sentiment analysis. The obtained sentiment will then explore if it is possible to predict daily changes in stock prices. This thesis is intended to contribute some new insights to the currently very limited research with regards to the impact of wallstreetbets on financial markets.

Furthermore, as mentioned earlier, the “to the moon” movement had a tremendous impact on the lives of individuals, both to the positive and negative. Besides that, however, many investment funds have been negatively impacted by the recent short-squeezes. While it might seem noble to root for individuals who try to force large funds out of their positions at big losses, it is easy to forget that many of those funds manage money for charitable endowments, pensions and others. Furthermore, such disruptions to the financial markets can harm its stability, thus causing spillover effects which can also negatively impact the lives of many people (Lyócsa, Baumöhl, & Vyrost, 2021).

## Background

Gauging sentiment of online forums to predict movements in stock prices has been a research subject for many years now. (Das & Chen) 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.

However, the impact of the wallstreetbets subreddit set an unforeseeable precedent. It wasn’t expected that the message boards can have such a big influence on certain stocks. Since the wallstreetbets meme-stock movement is a relatively recent phenomena, there is very little research on the impact of wallstreetbets on individual stocks. (Long, Lucey, & Yarovaya) try to establish a foundation for future research of sentiment analysis derived from Reddit on the stock market in, what they believe, to be the first paper on that topic. They try to uncover if specific emotions, such as *“Angry, Fear, Happy, Sad and Surprise”*, of comments on reddit posts impacts intraday returns of a specific stock. They conclude that the impact of tone, as well as the number of comments do have an impact on returns. However, they show that the number of comments are not related to sentiment. Instead, 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 one 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. Any asset that is targeted by the wallstreetbets forum.

The research by (Jemai, Hayouni, & Baccar) proposes a system, according to which a sentiment analysis project should be structured. The first phase is the *data collection* phase. In that phase, data is to be obtained from a source. In the second phase, the *preprocessing* phase, the text is cleaned up. As a result, it will be easier to feed the text into a machine learning algorithm. In this phase, several steps are taken. One of the steps is *data tokenization*. This is a popular technique, in which a body of text is broken down into several sentences and each sentence into a list of words. Another step in the preprocessing phase is to delete stop words, such as *is, the, a* and other common words. Furthermore, special characters such as *@* or deleting urls. Furthermore, it is proposed to change the text to lowercase. As a final step, they propose *lemmantization*. By doing that, the structure of a word is analyzed and then converted into its normalized form.

The next step is *data preparation*. In that step the data is prepared for sentiment analysis by converting the tokens into a dictionary. The dictionary is then split up into train and test sets.

In the final *classification* phase machine learning algorithms can be used to learn from the training data.

In addition to the proposed steps, the paper also touches on related work by peers. For example, they briefly explain the work carried out by (Parveen & Pandey, 2016). In that work they showed that preprocessing data with emoticons, leads to more accurate results than preprocessing data without emoticons.

Since the wallstreetbets community uses many domain-specific words, those words also need to be accounted for. Since those words with strong positive or negative sentiments are oftentimes used, it is very important to identify those word’s polarity for determining the semantic orientation. One way to handle domain-specific words, is by having a dictionary that is customized for those words. This dictionary can then be searched for finding and scoring the sentiment of the word (Asghar, 2014). Other research deviates from the aforementioned dictionary based approach. Instead, they examine how deep learning methods can be used to automatically detect and identify domain-specific words from sentences. By doing so it is assumed that the algorithm can not only detect whether domain-specific words are used (sentence-level detection), but also to identify the exact position of the term in the sentence (token-level identification). By doing so it is possible to detect new meanings of words in an already existing dictionary. In addition, this approach also allows to classify newly created words, that do not yet exist in a dictionary. This can be achieved, by having models that formulate domain-specific word detection as a sequence-labelling task. It is shown in experiments, that the flexibility of a part of speech feature is best to detect domain-specific words. That is because domain-specific words often entail a structured part of speech transformation of existing syntactic uses of words. Novel domain-specific tokens can be learnt, by understanding the contextual structure within a sentence. Those out-of-vocabulary tokens can be learnt in the hidden layers of LSTMs (Hochreiter & Schmidhuber, 1997). The model can be improved, by applying a character-based convolutional neural network to encode the spelling of words (Pei, Sun, & Xu, 2019).

(Gupta, et al., 2019) introduce SLANGZY, an algorithm that uses a mathematical “slang” factor to better judge social-media word definitions found in the Urban Dictionary, which is the largest crowd-sources slang dictionary available on the internet. The research shows that SLANGZY succeeds in normalizing the unstructured meanings of internet jargon in the Urban Dictionary. Hence, the algorithm can be provide more accurate meanings of non-standard words.

## Dataset

This Master thesis relies on two data sources. To perform sentiment analysis, posts from wallstreetbets need to be mined. In order to predict the stock prices of selected securities, the stock prices need to be accessed. Both datasets will be explained in the following section.

While reddit does offer an official API, the API is most useful for streaming data. There are some strict limitations of loading historical data. However, pushshift.io provides a solution for that. Pushshift is maintained by the /r/datasets mod team. The FAQ on the pushshift subreddit states, that pushshift data is best used to:

* Analyze large quantities of reddit data
* Grab data for a specific date range in the past
* Search for comments
* Aggregate data

Pushshift copies data from reddit at the time it is posted. Since Pushshift uses the document based database Elastic, it is extremely fast to query data. However, currently Pushshift does not regularly update certain metadata, such as scores, edits to a submission’s text or comments. Hence, there might be some minor inconsistencies of what is shown on reddit and what is in the database. To access the Pushshift API, I used an API wrapper called PMAW. Since requests are I/O-bound, PMAW is multithreaded. Hence requests can be run asynchronously which allows the data to be loaded much faster.

When making the API request, the most important parameters are the following:

* subreddit: Name of the subreddit
* q: The search term based on which the subreddit is queried
* before: The starting date of the query
* after: The end date of the query

The query returns 89 columns. Most of which, however, can be dropped since they either aren’t useful or contain no data. The most important columns are the number of columns, the title of the post and the content of the post. Emoticons are also included in the text.

The stock market data is obtained from…

## Algorithms

Lemmantization, stopword removal, tf-idf.

Naïve Bayes (NB): NB is a probabilistic supervised machine learning algorithm. The relatively simple algorithm works probabilistic, meaning that it assigns the probability of belonging to a class (Jemai, Hayouni, & Baccar, 2021).

Since text naturally has many dimensionalities, which can be handled very well by NB, this algorithm established itself as one of the standards for sentiment analysis. In this thesis Multinomial Naïve Bayes will be used for text classification. That is due to the strength of the model to handle larger vocabulary sizes. Furthermore, the algorithm is relatively easy to implement, can be used for real-time applications and is highly scalable. The downside, however, is that the prediction accuracy of the algorithm oftentimes is lower than other sentiment analysis techniques (Song, Kim, Lee, Kim, & Youn, 2017).

Support Vector Machines (SVM): SVMs can be used for 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 fit the hyper-plane by using spatial transformations, also known as kernel functions. 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).

DeepForest (gcForest): Tree ensemble method

LSTM

## Evaluation Methods

## Milestones and Plan

The thesis can be completed in time if the following schedule is adhered to. Note that a lot of time is allocated to cleaning the reddit text data. This is because I also have exams during that time interval.

*Note: The scaling of the bars is off. I will fix it asap.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Category | Days From Now | Days To Complete Milestone | Start Date | End Date |
| Finish Proposal Draft | Proposal | 0 | 4 | 27.09.21 | 01.10.21 |
| Finish Final Proposal | Proposal | 11 | 18 | 08.10.21 | 15.10.21 |
| Prepare Thesis Proposal Presentation | Proposal | 24 | 32 | 21.10.21 | 29.10.21 |
| Clean Reddit Text Data | Data | 18 | 39 | 15.10.21 | 05.11.21 |
| Load Stock Data for Securities | Data | 34 | 39 | 31.10.21 | 05.11.21 |
| Sentiment Analysis | Algorithm & Evaluation | 39 | 53 | 05.11.21 | 19.11.21 |
| Compare Sentiment Analysis Algorithms | Algorithm & Evaluation | 53 | 56 | 19.11.21 | 22.11.21 |
| Predict Changes in Stock Prices | Algorithm & Evaluation | 57 | 60 | 23.11.21 | 26.11.21 |
| Compare Prediction Algorithms | Algorithm & Evaluation | 60 | 63 | 26.11.21 | 29.11.21 |
| Write First Draft of Thesis | Writing | 63 | 87 | 29.11.21 | 23.12.21 |
| Finalize Thesis | Writing | 92 | 116 | 28.12.21 | 21.01.22 |

## References