UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Sentiment analysis on financial data by Large Language Models

Author:
Roman KYPYBIDA

Supervisor:
Nikolay KLIMENKO

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences and Information Technologies
Faculty of Applied Sciences

Lviv 2024

Declaration of Authorship

I, Roman KYPYBIDA, declare that this thesis titled, "Sentiment analysis on financial data by Large Language Models" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Every battle is won or lost before it's ever fought."

Sun Tzu

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Sentiment analysis on financial data by Large Language Models

by Roman KYPYBIDA

Abstract

Lots of people want to make a fortune and one can try himself in trading assets like Bitcoin, but how to determine when to sell or when to buy?
One could try his chances by analyzing the environment events and news about Bitcoin and try to predict if they have positive or negative effect on the price of this asset. But how can you make an analysis of so much information and make it both fast and quality? Human mind can do that, but it would take too much time and it lies on a single subjective opinion. What if there was something more objective? Sentiment analysis can allow you to automate the process of evaluating the sentiments and price movements using automation and computational resources of a cold fast machine. Here, we propose using LLMs as they are new popular approach in NLP and are not yet fully researched. The results of the simulation indicate that increasing number of parameters of the model does not necessarily improve the results and that more advanced LLMs can indeed beat primitive approaches like "holding strategy" and less advanced models like FinBERT. Also, we indeed confirm that sentiments are meaningful and have connection to price movements and can be used for trading and that the best approach is using average raw sentiment as price movement prediction without using ML approaches for the prediction task. These findings suggest one can reach positive return by predicting correctly only 40% of sentiments by using averages of raw sentiments as price predictors.

Acknowledgements

I want to express my gratitude to my scientific advisor, without whom this work would not be possible, Nikolay Klimenko, whose help was pivotal in constructing experiments and structure of this work. His feedback was crucial for making this thesis of higher quality.

Contents

List of Figures

List of Tables

List of Abbreviations

ML	**M**achine **L**earning
DL	**D**eep **L**earning
SA	**S**entiment **A**nalysis
GPT	**G**enerative **P**re-training **T**ransformer
NLP	**N**atural **l**anguage **p**rocessing
LLM	**L**arge **l**anguage **m**odel
BERT	**B**idirectional **E**ncoder **R**epresentations **f**rom **T**ransformers
FinBERT	**Fin**ancial **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
ROI	**R**eturn **o**n **i**nvestment
SM-related	**S**enti**m**ent related
PLM	**P**re-trained **L**anguage **M**odel
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**eccurent **N**eural **N**etwork
LSTM	**L**ong **S**hort-**T**erm **M**emory
MNB	**M**ultibinomial **N**aive **B**ayes
GloVe	**G**lobal **V**ectors
word2vec	**W**ord **t**o **v**ector
FSA	**F**inancial **S**entiment **A**nalysis

Chapter 1

Introduction

1.1 Motivation

Large language models have been gaining popularity for the last few years as a breakthrough technique in Natural Language Processing (NLP). There are numerous applications for this technology and its limitations are not completely known yet. Here, we want to form a conclusion about how well it performs in sentiment analysis.

Sentiment analysis allows us to summarize the emotional aspect of sentences in milliseconds, which the human mind can achieve but lacks the stamina and objectivity of a cold calculated machine. We cannot manually label billions of data points, just because of the time and effort limitations.

In the task of trading we have people already relying on the news about rises and falls of the assets' value and a human can't possibly process and evaluate even all the appearing news headlines. However, we can achieve this by using sentiment analysis. If the model can perceive and process the information and news headlines as good or even better than the human mind and it can process way more data, then it will help make better decisions about managing assets, leading to higher returns and profits. We chose cryptocurrency as our asset. Because of its popularity, there is a lot of data (news and news headlines) about it. Also, it is of a speculative nature, which turns into high dependence on the mood of the market and its fluctuations [11].

Here we test the FinBERT as a baseline against three Llama 2 models from Meta as well as GPT-3.5 for sentiment analysis. Afterward, we use Machine learning techniques - random forest, xgboost, linear regression (see chapter 2) - to predict the price movements based on the sentiments and use the predictions for trading on the historical data and determine the Return On Investment (ROI) of such strategies. Also, we test the performance of the predicting on classification metrics and perform the same test for pure sentiments without ML working on them to evaluate how close the price movements are to the produced markers without ML influence.

The approach is relatively new and there are few papers about it. For example, the trading of companies' stocks has been presented in "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models" [21] and similar research was done in "Sentiment Trading with Large Language Models"

[18]. Both papers have shown positive results, high classification metrics, and high Sharpe ratios[32]. This paper is expected to show similar quality results.

1.2 Goals

The goal of this work is to employ various LLMs for sentiment analysis tasks. We aim to test the feasibility of this approach on sentiment analysis of financial headlines and validate the following hypotheses:

1. Does using ML for price movement prediction outperforms using average raw sentiment as price move prediction;
2. Does advancing ML model improves the results;
3. Does the LLMs beat the baseline holding strategy;
4. Do more advanced models beat the baseline FinBERT model;
5. Will the number of parameters have a positive effect on the performance in terms of f1-score and ROI;
6. Would GPT-3.5, as a model with the highest number of parameters, have the best performance;
7. Are sentiments meaningful - does "traditional" approach outperform "opposite" (see 3.6.4);

We will evaluate the LLMs of various sizes in sentiment analysis using the task of trading and determine if it can be a useful tool for gaining profit through trading.

1.3 Paper structure

In the following chapter 2, we review the related works and studies to the current one and do the research on the results and approaches. Chapter 3 describes the problem setting, the dataset used in the paper, and the methodology of the pipeline from sentiment analysis to trading and testing. Chapter 4 presents the final results. The final Chapter 5 has the conclusions of the paper and proposals for future works in this area.

Chapter 2

Related works

2.1 Sentiment analysis approaches

In the course of the present research we have conducted an analysis of the sentiment analysis domain. There are three main modern approaches we have familiarized ourselves with: machine learning approaches, deep learning approaches and pre-trained language models.

2.1.1 Machine learning

Machine learning solves a classification or regression problem to determine the sentiment. There are three steps: (1) engineering of features, (2) selection of features, and (3) selection of algorithms.

We have four categories of features, as outlined in [14]:

1. Linguistic (e.g., n-grams, RF n-gram, verb, NER, and word cluster);
2. Sentiment lexicon (e.g., the proportion of positive and negative words, maximum, minimum and sum of sentiment score);
3. Domain-specific;
4. Word embeddings

Feature selection s is done by algorithms like Chi-squared, ANOVA (statistical method, which is used to separate observed variance data into different parts to use for additional tests) and others.

ML algorithms used are Support Vector Machines (SVM), Bagging, Random forest boosting trees and others [14].

All three papers [15], [23], [29] confirm that machine learning achieves better results as compared to traditional approaches. However, [30] states that more complex algorithms do not necessarily mean better results.

We can highlight the following advantages of the ML approaches (according to [26]):

- ML makes predictions and improve algorithms without need for human intervention
- ML algorithms able to gain experience improving accuracy

- ML algorithms can handle complex data.
- ML can be applied to a variety of fields

and also these disadvantages (according to [26]):

- ML requires large and high-quality datasets;
- ML needs time and resources to develop algorithms;
- ML is susceptible to errors, especially if the training datasets are biased;

There is existing literature that confirm that sentiments produced with modern techniques like machine learning have correlation with real price movements. For example in the Use of *NLP-Powered Sentiment Analysis in Trading Strategy* [7] showed that *"the sentiment analysis powered trading strategy is able to achieve a similar performance to the benchmark, which is the S&P 500 index, both in periods of bull (growing) markets and bear (falling) markets, with the potential of outperforming under favorable market conditions"*. The correlation is also observable with life events like COVID-19, Machine learning sentiment analysis. *COVID-19 news and stock market reactions* [10] show *"that there is a statistically significant and positive relationship between sentiment scores and S&P 500 market"*.

Also, the positive effect is observed with multiple ML techniques, in *Harvesting social media sentiment analysis to enhance stock market prediction using deep learning* [24] *"experiments were performed using machine-learning and deep-learning methods including Support Vector Machine, MNB classifier, linear regression, Naïve Bayes, and Long Short-Term Memory. The results validate the success of the proposed methodology, which is using Machine learning to provide a more accurate and robust approach to handle SM-related predictions"*.

2.1.2 Deep Learning (DL) methods in sentiment analysis

Deep learning methods utilize approaches like LSTMs, CNNs and RNNs to achieve even better results than machine learning approaches and construct complex representations from textual data with a high level of abstraction. Deep learning methods can be recognized for their capability to capture intricate patterns in financial data.

Deep learning has appeared as a popular method for sentiment analysis due to its capability to incorporate representations of text data. In these methods, text is pre-processed and then encoded using pre-trained embeddings such as GloVe and word2vec. They are fed then into deep learning models like convolutional or recurrent neural networks or long short-term memory or gated recurrent units.[14]

Before the transformers were introduced in [39], it was highlighted that state-of-the-art results across various NLP tasks were dominated by Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)[9]

We can highlight the following advantages of deep learning (according to [26]):

- DL models can process unstructured data;

- DL models can achieve higher accuracy compared to traditional machine learning models;
- DL models automatically learn features, hence avoid manual feature engineering;
- DL models learn to perform a task from input to output, by passing the need of the intermediate steps;
- DL models can be generalized to unseen data;

We can highlight the following disadvantages of deep learning (according to [26]):

- DL models require huge amount of data to train;
- DL models are computationally expensive;
- DL models are often considered black boxes;
- DL models can easily overfit to the training data, resulting in poor performance on new data;
- DL models can be difficult to debug when they fail due to a lack of transparency;

Many works show good performance of deep learning in sentiment analysis. [8] used Bi-LSTM model for sentiment analysis on Twitter data and achieved 81.20% accuracy using embeddings like word2vec and GloVe. [40] used the same model and achieved 90.26 % accuracy in sentiment analysis of Amazon reviews using word2vec embeddings. [19] outperformed traditional LSTM models with the same Bi-LSTM model with specific architecture parameters. In other work [16] used RNNs for sentiment analysis of hotel reviews and achieved 86 % and 84 % accuracy with LSTM and GRU models respectively. [35]

2.1.3 LLMs for sentiment analysis

The recently evolved pre-trained language models can be particularly effective in intricate tasks of NLP. PLMs like BERT were trained on large corpora of text enabling them to capture rich contextual information. After pre-training we can adjust our models for specific purposes to suit the requirements of the target application.
Models like FinBERT and other pre-trained models have shown better performance in FSA tasks [14].

We can highlight the following advantages of using LLM (refer to [36], [5], [33], [27], [31]):

- Good performance in many languages and tasks
- Improved language understanding and generation
- Improved machine translation with larger training data.
- Continuous Learning - LLMs gain new knowledge and skills in the process of training

- Flexibility - can be used for a variety of tasks

We can highlight the following disadvantages of using LLMs (refer to [36], [5], [33], [27], [31]):

- High computational requirements and potential biases in the training data
- Hallucinations - the generation of content that is irrelevant, made-up, or inconsistent with the input data
- Require a lot of training data

We want to test if these pros will outperform the provided cons and yield a positive return on investment.

While the ML techniques discussed in the previous section have been shown to classify financial sentiment better than traditional approaches, alternative approaches such as the transformer architecture [35], [25], [34] have demonstrated itself as having better capabilities in capturing sentiment. [37].

Transformer architecture showed State Of The Art (SOTA) results in many domains. For example, [28] introduced a text-to-text transfer transformer (T5), which has achieved state-of-the-art performance on the SQuAD question and answering task, [4] trained an autoregressive language model (GPT3) on 175 billion parameters,35 which returned the highest accuracy of 86.4% on the LAMBADA language modeling task and [12] introduce BERT, which showed, that not pre-trained on any specific data only fine tune towards the specific tasks, performed competitively (80.5% accuracy, representing a 7.7% absolute improvement on GLUE) [37].

Positive results of the trading strategies we can see in the papers *Stock Market Sentiment Classification and Backtesting via Fine-tuned BERT*[22] and *Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?* [42]. In these papers, it is demonstrated that *"incorporating sentiment analysis in the trading strategy has positive results on the returns"*.

We have familiarized ourselves with two papers, which use the same methodology as we do to trade on the stock market. In *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models* [21] it is shown that *"incorporating advanced language models into investment decisions can improve prediction accuracy and trading performance"*. Also, it demonstrates that *"only advanced language models can interpret complex news and press releases"*. In our work, we abstain from answering questions about which news models work best and evaluate the sentiment analysis based on trading success. In the aforementioned work 11 LLMs were used, in our work we use different sets of LLMs and on different assets - Bitcoin, whereas this paper works with companies' stocks. The same methodology was run in Sentiment trading with large language models [18], but different LLMs were used on different stocks. But both works show high Sharpe ratios from 1 to over 3 for less and more advanced models correspondingly, which is more than a good result. In both papers, the most advanced models show the best results and we also want to test this hypothesis in our work. All works show a positive connection between the sentiments and price movements and we also expect such results in our work. All papers we examined had observation of this positive connection and not a single

paper of different views had been found, although the results vary from technique to technique and less advanced models can't handle complex and more ambiguous news headlines. In our work, we concentrate on the performance of the LLMs in sentiment analysis relative to the number of parameters of the model and we evaluate that performance based on success in trading.

2.2 Technical background

2.2.1 ML

Machine learning techniques are an important part of the field of AI engineering and an important part of this work. Here we explore three ML techniques: linear regression, random forest, and xgboost.

Linear regression

Regression analysis is an important statistical method, which enables the identification and characterization of linear relationships between variables. It is the most common and comprehensive statistical and machine learning algorithm and we use it as a baseline ML technique here.

Random forest

Random forests are a combination of tree predictors, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. [3]. We use it as a second ML technique in this work.

Xgboost

XGBoost, a scalable tree-boosting system, can solve real-world scale problems using a minimal amount of resources [6]. It is one of the ML techniques with the highest accuracy and we use it as a third in our work.

2.2.2 LLMs

Large Language Models have become very popular due to their performance on a variety of NLP tasks, since the release of ChatGPT in November 2022. The ability to understand and generate text is acquired by training billions of model parameters on massive amounts of data, as predicted by scaling laws [17]. Many different tasks are solved using LLMs and the research evolves rapidly. LLMs refer to transform-based neural language models that contain a large number of parameters, which are pretrained on massive datasets, examples of LLMs are Llama and GPT-4 [43]. These language models are used to get sentiments on news headlines to use in trading afterward.

BERT

BERT stands for Bidirectional Encoder Representations from Transformers and is designed to train deep bidirectional representations from unlabeled text, by conditioning simultaneously on right and left contexts in all layers [13].

FinBERT

FinBERT is an LLM based on BERT, specifically, trained for the financial domain. The results of the research on this model show that it archives improvements in every metric on current SOTA in financial sentiment analysis. After pre-training even a small part of the model, it outperforms SOTA models in financial sentiment analysis. This model is used as a benchmark model for this paper and as opposed to other models in this research is an encoder-only model [1].

Llama 2

Llama 2 is a collection of pre-trained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. The fine-tuned LLMs, called Llama 2-Chat, are optimized for dialogue use cases. These are open-source models and, according to research, they outperform open-source chat models and may substitute the closed-source models [38]. This collection of models is good for comparison, as each model in the collection has a similar architecture, but a varying number of parameters, which makes it great for testing the proposed hypotheses (refer to 1.2).

GPT

GPT - the Generative Pre-trained Transformer - is based on the transformer architecture, a deep neural network designed for NLP tasks. It has gained popularity in the field, due to its high performance and ability to effectively converse [41]. This model was chosen as it is the most popular model in the world and has an impressive 175 billion parameters.

Chapter 3

Approach to Solution

3.1 Problem Setting

We have familiarized ourselves with many methods of sentiment analysis, but the pre-trained models like FinBERT remain the SOTA in the field. Taking into account the success of LLMs in other domains (see Chapter 2), it is therefore an interesting task to try them out in Financial Seniment Analysis (FSA). There are many different LLMs, and one of the key differences between them is the number of parameters.

We decided to test if there is a connection between the number of parameters and their performance. In other words, does size matter?

3.2 Data

For the set objectives, we need the following data: (1) open and close prices of bitcoin for some time with corresponding dates of when the prices were relevant to compute the price movements and the success of the trading strategy, (2) sets of news headlines for each date to turn them into the sentiments.
We have found this data on the Kaggle platform [2].

The dataset contains both English and Spanish headlines, but for this work, we filter out the headlines in Spanish. Since the headlines are presented as an array in one column - "articles" we have to parse this column and create a new file with headlines. The close and open prices are in the "close_price" and "open_price" columns correspondingly.The "begins_at" column represents the according dates for the prices and news.

The dataset had been collected using web scraping from the first page of Google News by date for several years, pricing data was collected using Robinhood API. While headlines for many cryptocurrencies are included, the best results were obtained by limiting models to data from 2021, when cryptocurrencies gained their popularity,

begins_at	open_price	close_price	high_price	low_price	symbol
2/25/2018	9680.2	9584.45	9864.64	9300.02	BTC
2/26/2018	9592.495	10318.785	10452.52	9378.91	BTC
2/27/2018	10318.785	10548.385	10861.01	10140.93	BTC

TABLE 3.1: An example of the first six columns of the dataset

articles
['Original Pizza Day Purchaser Does It Again With Bitcoin Lightning Network', 'This 11-year-old just wrote a book on bitcoin that hopefully a kid can \nunderstand','Without Mentioning Blockchain, Putin Says That Russia Must Stay Ahead In \nTechnology','El comprador original del Pizza Day lo hace de nuevo con la Lightning \nNetwork de Bitcoin','Meet the strippers tattooed with BARCODES so sneaky punters can tip in \nBitcoin without their partner...', 'Biostar TB250-BTCPRO - test i opinia','Elon Musk, Ne Kadar Bitcoinâ€™i OlduÄŸunu AÃ§Ä±kladÄ±!','Comprar bitcoin aumenta tu estatus social: el factor que ...','Polis, BNM kerjasama tangani jenayah babit Bitcoin', "Les cryptomonnaies Ã l'assaut des produits structurÃ©s"]

TABLE 3.2: An example of the last seventh column of the dataset

3.3 Methodology

The plan for the work is the following:

- the sentiments must be computed from the news headlines dataset
- the predictions of the prices must be computed by ML techniques
- classification metrics computed against the price movements
- trading simulation run using the predictions and sentiments

In this section, we describe each part of this methodology.

3.3.1 Sentiment score

The first step in our pipeline is retrieving sentiments and sentiment scores from news headlines.

Prompt

To get sentiments using generative models like Llama 2 and GPT we need a prompt, that would retrieve the sentiments from the headlines. Here we use a similar prompt, proposed in [21] with minor changes: in our prompt, we asked for sentiment only about Bitcoin's price and also we changed "YES", "NO" and "UNKNOWN", to "Positive", "Negative" and "Neutral". Also, we got rid of the request for a short sentence and asked to be concise and give a one-word answer.

Large Language Models Role

We retrieve the sentiments from pre-trained LLMs: we feed a headline into a model and the model returns a sentiment score or a sentiment, which can be turned into a sentiment score.

Example 1 - FinBERT: headline → model → scores → max score → sentiment → sentiment score (See Figure 3.1).

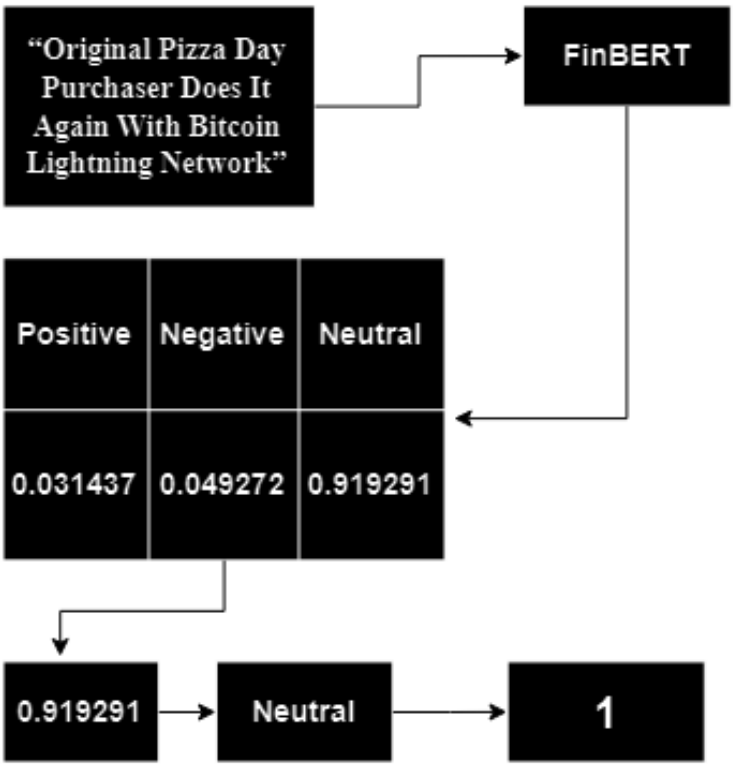

FIGURE 3.1: Example 1. Sentiment retrieval using FinBERT model

Models

We decided to choose FinBERT as our baseline as it is promising in FSA and is currently a SOTA in the field. Also, it is easy to deploy and use and it does not take a lot of computational resources. This model is the only encoder-only model in our set of models, which means it does not use any prompt and just gives away scores of the sentiments of input headlines.

Next, we use the set of three Llama 2 models from 7 to 70 billion parameters. These models are good choices for testing our main hypothesis as they possess the same architecture, but a different number of parameters.

The last model in our set is GPT, which is the most popular model in the world and also possesses the highest number of parameters in the set, so expected to yield the best results.

All models were used with parameter of randomness equal to 0.5.

3.4 Price movement prediction

After we produce the sentiments, we need to form the predictions of price movements using the ML techniques described next. Also, we use the raw version of

sentiments for the testing, which means using an average of sentiments for a particular date as a prediction, to see if the usage of ML is an improvement or not.

We use the following formula for tuning these ML techniques:

$$Prediction = avgsent + n_{neg} + n_{pos} + n_{neu}$$

where avgsent is the average sentiment for the date of interest, n_neu is the number of negative sentiments for the date of interest, and analogically n_pos and n_neu are numbers of positive and neutral sentiments for the date of interest.

We used the period of one year (last year in the dataset) for testing and the rest of the dataset for training.

We use regression as our baseline as it is the most popular and the simplest of the ML models, the next is random forest as one of the most advanced models in the field, and xgboost is the last model since boosted trees yield very high accuracy.

3.5 Trading simulation

Next, we run the "simulation of trading" (see code at [20]):

1. We define variables for the balance of dollars we have and the balance of bitcoins we own, each equal to zero in the beginning;
2. Then we give a particular value - 100 - to the balance of dollars we own in the beginning;
3. We iterate over the predictions of price movements;
4. If the prediction is negative then, we convert the balance of bitcoin in dollars, using the current open price, and add the retrieved value to the balance of dollars. Then we set the value of the balance of bitcoins to zero.
5. If the prediction is positive, we convert the number in the balance of dollars to the number of bitcoins, using the current open price, and set it to the balance of bitcoins variable. Then we set the balance of dollars to zero.
6. If the prediction is 0, we do nothing.
7. We convert the bitcoins to dollars if they were not converted before then to dollars and update the balance of dollars with it.
8. We finish the simulation and collect the results: the returns, the final balance of dollars the starting balance, and some other variables needed for testing.

We run the same simulation of trading for each set of sentiments received from the models and each set of predictions from each ML technique.

3.6 Evaluation

3.6.1 Classification metrics

After we have received the predictions from running our ML models, we evaluate the performance of our predictions with the f1-score.

We compute the price movements by subtracting each row of the array of prices number k -1 from the row number k, the row before the one, for all k from 1 to the number n, where n is the number of rows.

After we retrieved the price differences, we convert them to percentages and price movements: we divide each number k by the number k - 1 and multiply by 100; then we convert all numbers higher than 0.5 to 1, lower than -0.5 to -1 and in between -0.5 and 0.5 to 0. This way we get the price movements.

3.6.2 ROI

To evaluate the success of the trading we compute ROI on the results of each simulation of trading. Then we collect all metrics and compare them in terms of the ML techniques used to produce them, the LLM used to produce them, and the number of parameters the LLM has. Also since we compute the same simulation for raw sentiments we take results from those simulations to compare with other pipeline procedures that were run before.

3.6.3 Baseline holding strategy

We start with evaluating the baseline strategy, which is holding Bitcoin for a year (last year since we predicted the last year) and than selling it.

It returns 64.25 dollars after investing 100, since bitcoin dropped since the time of 2/24/2022 to 2/24/2023 which is the period of evaluation. This means we can lose at most 35.75 % with our trading strategy to beat the baseline.

3.6.4 Opposite approach

To decide if our sentiments are meaningful we test our pipeline as follows: we run the trading simulation with opposite interpretations of sentiments - positive, then sell, and negative means buy. If the results indicate higher results on average than from the traditional pipeline, then we can state that our sentiments cannot be used for price predictions since their nature is random. If our results are the opposite of the original ones or worse than the original ones, then we can conclude that they are meaningful in terms of price predictions and can be used for the task of trading. We compute ROI for both scenarios and compare the results.

Chapter 4

Results

4.1 Selecting Score Interpretation Method

Model/Metric	f1-score (%)	ROI (%)
Regression	40	9.89
RandomForest	40	11.49
XGBoost	39	13.33
Raw	41	13.79

TABLE 4.1: Average metrics across models for ML techniques and raw sentiments.

We start our selection of score interpretation methods. Here we explore the average metrics across models for different ML models versus each other and raw sentiments. We observe the growth of the ROI and stable f1-score, while advancing the ML model and we end the table with the highest ROI for raw sentiments. Here we can infer that it is not particularly important how much price movements are predicted correctly, but rather when we predict them correctly, since the f1-score stays quite the same around 40%. It is enough to be correct at least a third of the time to get positive results. However, the raw sentiments show the highest f1-score of 41 %, which could indicate the crucial importance of correctly classifying false negatives and false positives.

For the rest of the paper, we will deal with the raw sentiment experiments, since they yielded both the highest f1-score and ROI. We conclude that adding ML techniques in the pipeline on average does not outperform using average raw sentiment as price movement prediction.

4.2 Models performance

Hypothesis 1 states that LLMs beat the baseline holding strategy. We can observe it in the table and also confirm statistically: the P-value of the t-test, 4.907e-7, is less than 0.05 for results of LLMs against the baseline holding strategy, which statistically confirms that our models beat the baseline holding strategy

Our second hypothesis is that our models would beat the baseline FinBERT model. We again observe it in the table and the P-value of the second t-test, 0.00001716, is less than 0.05 for values of ROIs of LLMs against the result of FinBERT, which confirms that our models beat the baseline statistically.

Model	# of parameters	Price move prediction F1 score (%)	Trading simulation ROI (%)
Baseline #1:Hold	0	-	-35.75
Baseline #2:FinBERT	110M	39	0.89
Llama7B	7B	43	17.67
Llama13B	13B	38	16.60
Llama70B	70B	43	16.08
GPT-3.5	175B	41	17.70

TABLE 4.2: Results for pipelines with raw sentiments for all models

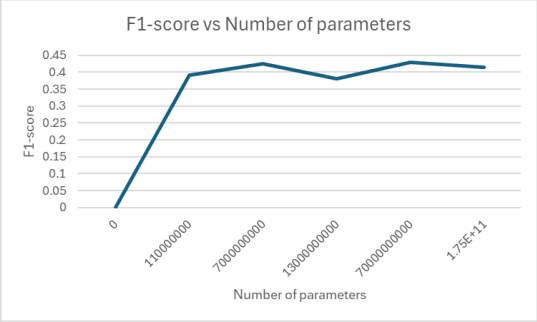

FIGURE 4.1: F1-score vs Number of parameters

Our third hypothesis is that the ROI and f1-score will grow with growth in the number of parameters of our models. We can partially confirm this hypothesis since we observe that the worst model outperforms only the baseline and loses to all other models. The most advanced model - GPT 3.5 - outperforms all models. The results for Llama 2 models are inconclusive. The first model with 7 billion parameters loses to GPT 3.5 but wins over the next two models with 13 billion and 70 billion parameters. Also, we don't have a stable answer if f1-score would grow with an increased number of parameters (see Figures 4.1, 4.2).

What does statistics tell us about this hypothesis? The p-value of .40618 of ROIs against the number of parameters is not significant at $p < 0.05$, hence, we can not confirm statistically that the results grow with a number of parameters.

In the same, we can tell about the connection between the f1-score and the number of parameters.A P-value of 0.685284 is not significant at any chosen level of significance, hence, we cannot conclude that the f1-score grows with a number of parameters of the model.

The fourth hypothesis states that GPT 3.5 as a model with the largest number of parameters will yield the best results and we confirm this hypothesis (See Table 4.2)

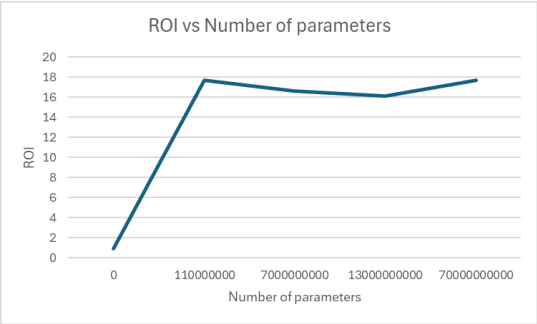

FIGURE 4.2: ROI vs Number of parameters

4.2.1 Opposite trading

	# of parameters	ROI (%)	ROI Opposite approach(%)
FinBERT	110M	0.89032	14.80
Llama7B	7B	17.6718	-3.74
Llama13B	13B	16.6025	-2.85
Llama70B	70B	16.0841	-2.42
GPT 3.5	175B	17.7077	-3.76

TABLE 4.3: ROIs of opposite approach versus traditional

The results of the "opposite approach" are, as expected, opposite to the original approach. All models beat the baseline holding strategy (refer to 3.6.3), but only one model - FinBERT - yields positive results, but other models return negative ROIs, and the best model originally - GPT 3.5 - gets the worst results and loses almost 4%. In fact, four out of five models lose more than 2% of the starting capital. We can't state that ROI gets worse with growth in parameters, which is proved statistically: p-value - 0.49299 - is not significant at level 0.05 nor at level 0.1.

The p-value for the Pearson coefficient between two sets of ROIs is .001199, which is less than the significance level of 0.05, which means it is significant. So we can conclude that there is strong evidence of a negative correlation between two sets, traditional and opposite. Hence, we can conclude that the sentiments are meaningful and can be interpreted as price movement predictions, it means it is important to match our sentiments to price movements to get high ROI results.

4.3 Other meaningful results

4.3.1 FinBERT paradox

It is interesting that with regression it has yielded the second highest result, but has yielded the third worst result with raw sentiments. It could be explained by a high number of "neutral" labels among the raw sentiments of FinBERT and a high number of predicted "positive" and "negative" with regression (see Figure 4.5), which means a low number of transactions of bitcoin and high number holding decisions. Since bitcoin has fallen in the period of interest, it has an adverse effect on the ROI of the strategy. On the other hand, a high number of right transactions can lead us to a profitable way. We have zero "neutral" labels predicted by regression markers,

so with the highest ROI result in hand it means the regression has yielded the right movements of bitcoin price (see Figure 4.6). So here it is less about how much we trade and more about when we trade.

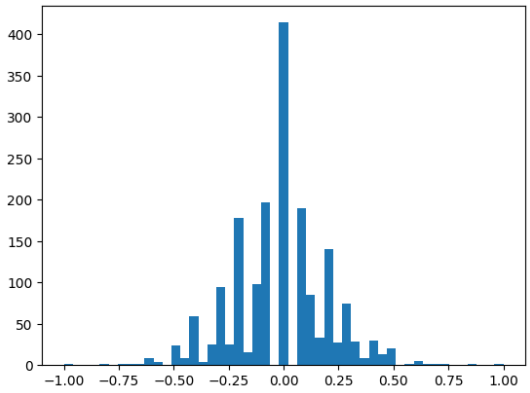

FIGURE 4.3: Distribution of raw sentiments of the FinBERT model.

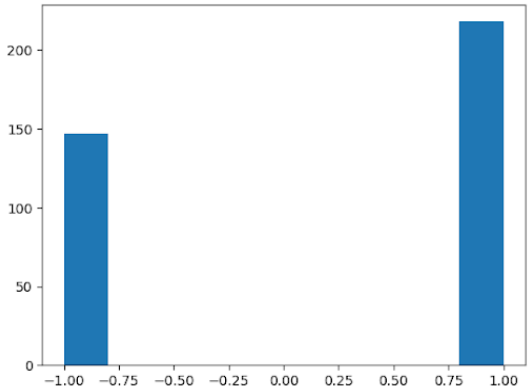

FIGURE 4.4: Distribution of predicted by regression labels for FinBERT

Chapter 5

Conclusions

Here we summarize the findings and results of the research.

We begin by summarizing the results of testing different ML models against raw senitments. Firstly, we find that on average using average raw sentiment as price movements yields higher ROI than using ML approaches to construct predictions of the price movements. Secondly, we find that in fact advancing the ML model improves the results - F1-score and ROI - of the trading iteration, but still does not outperform using average raw sentiment.

Next, we summarize results of using different LLMs against each other in terms of ROI and f1-score. Firstly, we conclude that using LLMs makes sense and does yield higher ROI than just holding your fortune and expecting the growth of Bitcoin. Secondly, we find that more advanced models - with higher numbers of parameters - do perform better than the baseline FinBERT model.

Now we investigate the influence of number of parameters on performance. We conclude that advancing a model by increasing its number of parameters does not necessarily mean better results in terms of ROI or f1-score. And we indeed find out that GPT 3.5 performs the best as a model with highest number of parameters.

Finally, we tested the approach of "opposite" trading and confirmed that the sentiments are meaningful and must be used in a traditional way, when positive sentiments indicate growth in price and negative indicate fall of the price. But in the same time, we find that it is not necessary to guess right the sentiment many times - just about a third times is enough for a positive return.

Here, we tested 5 LLM models against each other and we did not mention other pretrained for FSA models, besides FinBERT. But one could test many other models and find out different results. Here we were constrained by computational resources, which is very important factor. Also, the data is very important: the source of the news should give more descriptive labels of events, rather than creating news by itself. Architecture of LLMs is an important factor too: it makes more sense to compare similar LLMs, rather than completely different, like we did with Llamma 2 models.

For future works we suggest using the more advanced ML models like RNNs and LSTMS for predicting the price movements and using more advanced equation of regression by adding extra terms not mentioned in this work. Also, one could

test different financially pretrained models to detect the best approach for detecting the sentiment. The third suggestion is testing sentiment analysis by large language models on different languages and testing the results one against the other. Also, one could test different periods of producing a sentiment - in this work we used a period of one day, one could test if longer periods would present in better returns.

Here, we've made a new research on Bitcoin sentiment analysis and tested different ML techniques one against another to find that using average raw sentiment to predict price movements of Bitcoin is the best approach and we found out that among the tested models, GPT 3.5 performs the best and it's raw sentiments should be used for predicting movements of price of Bitcoin.

.1 Appendix A

In this table we present results of all experiments with traditional and opposite approach.

Model	Regression	RandomForest	XGBoost	f1-score	ROI (%)	"Opposite" ROI (%)
FinBERT	o	x	x	34	25.56	-9.80
FinBERT	x	o	x	36	6.71	8.54
FinBERT	x	x	o	34	4.35	8.54
FinBERT	x	x	x	39	0.89	14.81
Llama7B	o	x	x	41	-1.67	15.19
Llama7B	x	o	x	38	4.78	8.09
Llama7B	x	x	o	39	17.13	-3.30
Llama7B	x	x	x	43	17.67	-3.74
Llama13B	o	x	x	43	4.33	8.56
Llama13B	x	o	x	43	15.53	-1.96
Llama13B	x	x	o	44	16.65	-2.90
Llama13B	x	x	x	38	16.60	-2.86
Llama70B	o	x	x	43	12.02	1.11
Llama70B	x	o	x	42	30.15	-12.97
Llama70B	x	x	o	43	22.98	-7.90
Llama70B	x	x	x	43	16.08	-2.42
GPT 3.5	o	x	x	41	9.25	3.68
GPT 3.5	x	o	x	35	0.29	12.93
GPT 3.5	x	x	o	36	5.56	7.29
GPT 3.5	x	x	x	41	17.71	-3.77

TABLE 1: Results of all experiments

Bibliography

[1] Dogu Araci. "Finbert: Financial sentiment analysis with pre-trained language models". In: *arXiv preprint arXiv:1908.10063* (2019).

[2] Aaron Bastian. Link to dataset: https://www.kaggle.com/datasets/aaroncbastian/crypto-news-headlines-and-market-prices-by-date?select=BTC.csv.

[3] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[4] B. Mann N. Ryder M. Subbiah J. D. Kaplan P. Dhariwal A. Neelakantan P. Shyam G. Sastry Brown T. B. and A. Askell. "Language models are few-shot learners. Advances in Neural Information Processing Systems". In: (2020).

[5] Ivan Ivanovich Bulyko. *Modular approach to building large language models*. US Patent 7,774,197. 2010.

[6] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[7] Zanyang Neil Chen. "Use of NLP-Powered Sentiment Analysis in Trading Strategy". In: *Proceedings of the 2nd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2023)*. Atlantis Press. 2023, pp. 109–115.

[8] Sakib Chowdhury et al. "A RNN based parallel deep learning framework for detecting sentiment polarity from Twitter derived textual data". In: *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE. 2020, pp. 9–12.

[9] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[10] Michele Costola et al. "Machine learning sentiment analysis, COVID-19 news and stock market reactions". In: *Research in international business and finance* 64 (2023), p. 101881.

[11] Kelly Ann Coulter. "The impact of news media on Bitcoin prices: modelling data driven discourses in the crypto-economy with natural language processing". In: *Royal Society Open Science* 9.4 (2022), p. 220276.

[12] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[13] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[14] Kelvin Du et al. "Financial Sentiment Analysis: Techniques and Applications". In: *ACM Computing Surveys* (2024).

[15] Li Guo, Feng Shi, and Jun Tu. "Textual analysis and machine leaning: Crack unstructured data in finance and accounting". In: *The Journal of Finance and Data Science* 2.3 (2016), pp. 153–170.

[16] Md Sagar Hossen et al. “Hotel review analysis for the prediction of business using deep learning approach”. In: *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE. 2021, pp. 1489–1494.

[17] Jared Kaplan et al. “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361* (2020).

[18] Kemal Kirtac and Guido Germano. “Sentiment Trading with Large Language Models”. In: *Available at SSRN 4706629* (2024).

[19] D Ashok Kumar and Anandan Chinnalagu. “Sentiment and emotion in social media COVID-19 conversations: SAB-LSTM approach”. In: *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE. 2020, pp. 463–467.

[20] Roman Kypybida. Link to github: https://github.com/normangalt/Diploma-2023—2024.

[21] Alejandro Lopez-Lira and Yuehua Tang. “Can chatgpt forecast stock price movements? return predictability and large language models”. In: (2023). DOI: http://dx.doi.org/10.2139/ssrn.4412788.

[22] Jiashu Lou. “Stock Market Sentiment Classification and Backtesting via Fine-tuned BERT”. In: *arXiv preprint arXiv:2309.11979* (2023).

[23] Zachary McGurk, Adam Nowak, and Joshua C Hall. “Stock returns and investor sentiment: textual analysis and social media”. In: *Journal of Economics and Finance* 44 (2020), pp. 458–485.

[24] Pooja Mehta, Sharnil Pandya, and Ketan Kotecha. “Harvesting social media sentiment analysis to enhance stock market prediction using deep learning”. In: *PeerJ Computer Science* 7 (2021), e476.

[25] Manish Munikar, Sushil Shakya, and Aakash Shrestha. “Fine-grained sentiment classification using BERT”. In: *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. Vol. 1. IEEE. 2019, pp. 1–5.

[26] Emanuele Neri et al. “Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology”. In: *La radiologia medica* 128.6 (2023), pp. 755–764.

[27] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. “Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages”. In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. 2021, pp. 116–126.

[28] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[29] Thomas Renault. “Intraday online investor sentiment and return patterns in the US stock market”. In: *Journal of Banking & Finance* 84 (2017), pp. 25–40.

[30] Thomas Renault. “Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages”. In: *Digital Finance* 2.1 (2020), pp. 1–13.

[31] Todd Richard Watson. “Large-Sized Language Classes”. In: (2012). DOI: doi:10.1002/9781405198431.WBEAL0666.

[32] William F Sharpe. “The sharpe ratio”. In: *Streetwise–the Best of the Journal of Portfolio Management* 3 (1998), pp. 169–185.

[33] Irene Solaiman et al. "Release strategies and the social impacts of language models". In: *arXiv preprint arXiv:1908.09203* (2019).

[34] Chi Sun et al. "How to Fine-Tune BERT for Text Classification?" In: *Chinese Computational Linguistics*. Ed. by Maosong Sun et al. Cham: Springer International Publishing, 2019, pp. 194–206. ISBN: 978-3-030-32381-3.

[35] Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. "A survey of sentiment analysis: Approaches, datasets, and future research". In: *Applied Sciences* 13.7 (2023), p. 4550.

[36] Peng Xu. Thorsten Brants. "Distributed Language Models." In: (2013).

[37] Andrew Todd, James Bowden, and Yashar Moshfeghi. "Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions". In: *Intelligent Systems in Accounting, Finance and Management* 31.1 (2024), e1549.

[38] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[39] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[40] JS Vimali and S Murugan. "A text based sentiment analysis model using bi-directional lstm networks". In: *2021 6th International conference on communication and electronics systems (ICCES)*. IEEE. 2021, pp. 1652–1658.

[41] Gokul Yenduri et al. "Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions". In: *arXiv preprint arXiv:2305.10435* (2023).

[42] Haohan Zhang et al. "Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?" In: *arXiv preprint arXiv:2306.14222* (2023).

[43] Wayne Xin Zhao et al. "A survey of large language models". In: *arXiv preprint arXiv:2303.18223* (2023).