



Vienna University of Economics and Business

## **Stockmarket Prediction with Headlines**

Data Processing 2

Submitted to

LV-number: 5451

Course directors:

Dr. Sabrina Kirrane

Michael Sebastian Feurstein MSc

Semester: S23

## Table of Contents

Project Overview and Goal.....	3
Project Data Sources .....	3
Data Source Description .....	3
Dataset Excerpts.....	4
Stock Market Data .....	4
Daily Financial News for 6000+ Stocks .....	5
Analysis Steps and Results.....	5
Analysis Steps .....	5
Results and Interpretation.....	5
Legal and Ethical Issues .....	6
Legal and Ethical Guidelines.....	6
Legal Guidelines .....	6
Ethical Guidelines .....	7
Legal and Ethical Challenges .....	7
Legal Challenges .....	7
Ethical Challenges.....	7
Experience Gained .....	8
Discussion on the challenges encountered .....	8
Summary of experienced gained by each team member .....	8
Alex's Experience.....	8
Gabriel's Experience .....	8
Max's Experience.....	8
Recommendations for future work .....	8
Table of Figures .....	10
References .....	10

## Project Overview and Goal

The stock market is a complex system that is influenced by a variety of factors, including economic news, company earnings and investor sentiment. Financial headlines can provide insights into these factors and can be used to predict future stock price movements.

In our project we created a model that can predict the stock price reaction of a company to a financial headline. Firstly, our model receives a headline as an input. Based on the headline the model tries to predict, which company the headline is about and returns the ticker symbol of the predicted company. Secondly, based on the headline, the model returns a prediction in which direction it expects the stock price of the company to move. The return values are 0 (if it expects the stock price to decline) or 1 (if it expects the stock price to increase) and gives us a probability on how confident the model is about it. We used multiple machine learning techniques and measured them based on their accuracy. Our goal on the project was achieving at least over 50% accuracy, so that our project can be developed further into having real life use cases.

The model can be used to learn more about the relationship between financial headlines and stock prices. By understanding the relationship between financial headlines and stock prices, we can learn more about how the stock market works. This knowledge can be used to improve the accuracy of the model and to develop new investment strategies.

We believe that this project has the potential to be a valuable tool for investors and businesses. By predicting the stock price of a company based on financial headlines, investors can make more informed investment decisions. Businesses can use the model to track the sentiment of the market towards their products or services.

Finally, we will use our project as a future reference for future employers. We believe that our passion for data science to solve economic challenges is a very attractive skillset. A project of this size might not only be interesting for investment firms. Rather it is exciting for every company, that works with big data, because we acquired new experiences and skills in processing and transforming big data with the goal of training machine learning models. Therefore, we are really looking forward to applying these skills in our career and create value based on data.

## Project Data Sources

### Data Source Description

Our data source is **Kaggle**<sup>1</sup>, which is an online platform that hosts data science competitions, provides datasets and kernels for practice, and fosters a collaborative community of data scientists and machine learning enthusiasts. [1]

To achieve our projects goals, we used two datasets from Kaggle:

**“Stock Market Data”**<sup>2</sup>: This is a collection of the shares contained in a given index. For each share there is a separate dataset holding the open-, close-, etc. prices for the share for each day the stock exchange was open, till the 1980s. This data was used for machine learning, to determine whether a stock price had fallen or risen that day. The datasets for each stock are of different lengths, but the most of them have at least ~3000 observations. [2]

---

<sup>1</sup> <https://www.kaggle.com/> (retrieved June 28, 2023)

<sup>2</sup> <https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset?select=stocks> (retrieved June 28, 2023)

**“Daily Financial News for 6000+ Stocks”<sup>3</sup>:** This dataset of financial news, contains headlines and URLs to the articles. For this dataset there was much processing needed to be done. Through NLP we extracted important data that would indicate a rise or fall of a share price, and to determine the stock that is associated with this headline. The dataset has 1.4 million headline entries. [3]

Stock Market Data	Daily Financial News for 6000+ Stocks
<ul style="list-style-type: none"> <li>• <b>Date</b> (text) - specifies trading date</li> <li>• <b>Open</b> (number) - opening price</li> <li>• <b>High</b> (number) - maximum price during the day</li> <li>• <b>Low</b> (number) - minimum price during the day</li> <li>• <b>Close</b> (number) - close price adjusted for splits</li> <li>• <b>Adj Close</b> (number) - adjusted close price adjusted for both dividends and splits.</li> <li>• <b>Volume</b> (number) - the number of shares that changed hands during a given day</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Index</b> (number) – id of observation</li> <li>• <b>Headline</b> (text) – headline of the published article</li> <li>• <b>URL</b> (text) – URL to the published article</li> <li>• <b>article author</b> (text) – author of the article</li> <li>• <b>publication timestamp</b> (text) – publication date and time</li> <li>• <b>stock ticker symbol</b> (text) – abbreviation of the stock name</li> </ul>

Figure 1: Dataset structures

## Dataset Excerpts

### Stock Market Data

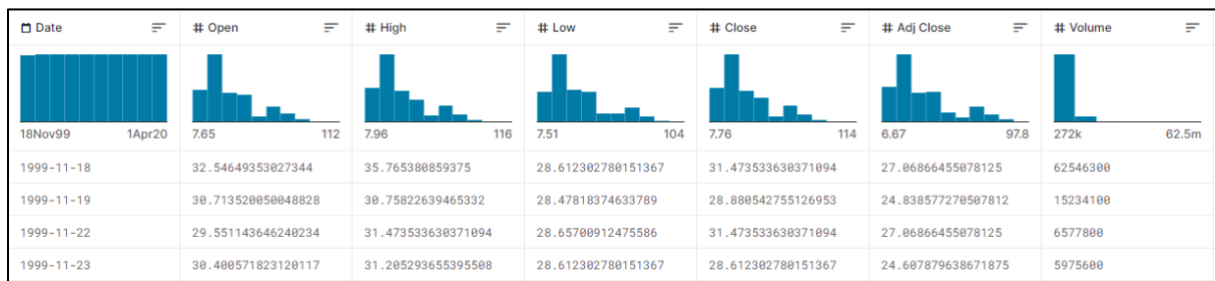


Figure 2: excerpt stock data (stock "A" - Agilent Technologies, Inc.)<sup>4</sup>

[4]

<sup>3</sup> <https://www.kaggle.com/datasets/miguelaelnle/massive-stock-news-analysis-db-for-nlpbacktests> (retrieved June 28, 2023)

<sup>4</sup> <https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset> (retrieved June 28, 2023)

## Daily Financial News for 6000+ Stocks


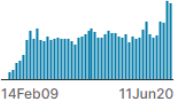
#	headline	url	publisher	date	stock
Index column	Article release headline	URL of article	Author/creator of article	Date of article's release. Note that ~80-90% of the dates don't contain exact timestamps (go to the processed version for	Stock ticker (NYSE/NASDAQ/AMEX only)
	<b>845770</b> unique values	<b>883429</b> unique values	Paul Quintaro 16% Lisa Levin 13% Other (991976) 70%		<b>6204</b> unique values
0	Stocks That Hit 52-Week Highs On Friday	<a href="https://www.benzinga.com/news/20/06/16190091/stocks-that-hit-52-week-highs-on-friday">https://www.benzinga.com/news/20/06/16190091/stocks-that-hit-52-week-highs-on-friday</a>	Benzinga Insights	2020-06-05 10:30:54-04:00	A
1	Stocks That Hit 52-Week Highs On Wednesday	<a href="https://www.benzinga.com/news/20/06/16170189/stocks-that-hit-52-week-highs-on-wednesday">https://www.benzinga.com/news/20/06/16170189/stocks-that-hit-52-week-highs-on-wednesday</a>	Benzinga Insights	2020-06-03 10:45:20-04:00	A
2	71 Biggest Movers From Friday	<a href="https://www.benzinga.com/news/20/05/16103463/71-biggest-movers-from-friday">https://www.benzinga.com/news/20/05/16103463/71-biggest-movers-from-friday</a>	Lisa Levin	2020-05-26 04:30:07-04:00	A
3	46 Stocks Moving In Friday's Mid-Day Session	<a href="https://www.benzinga.com/news/20/05/16095921/46-stocks-moving-in-fridays-mid-day-session">https://www.benzinga.com/news/20/05/16095921/46-stocks-moving-in-fridays-mid-day-session</a>	Lisa Levin	2020-05-22 12:45:06-04:00	A

Figure 3: excerpt headline data<sup>5</sup>

[5]

## Analysis Steps and Results

### Analysis Steps

Our analysis steps for both datasets included following parts. First, we had to do some data processing, including dropping unnecessary columns. To continue to the machine learning part, we split the data into training and test sets. Then by defining and applying a pipeline, the data was brought into a processable structure, using tokenizer, hashingTF, StringIndexer (for the label: stock abbreviation), and for the learning part Logistic Regression. We fitted the pipeline on the training set and made predictions on the test set.

For the stock price dataset, we had to additionally calculate whether the stock had risen or fallen that day, which also was our label to make predictions.

Also we tried other classifiers like Random Forest, Gradient Boosting and Support Vector Machines.

### Results and Interpretation

For predicting which stock a headline is referencing to, we achieved an **accuracy of 78.05%**, which is very good considering that we had to use a sample of just 1000 rows out of our data (which had 1.4 million rows). The accuracy would be higher if we had enough computational power to train the model with all 1.4 million rows.

For the prediction, whether a share price rises or falls depending on a headline, out of all the applied classifiers (Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine) the Support Vector Machine seems to perform best. But we don't choose that, because it is computationally expensive, especially for large datasets. We already had to step

<sup>5</sup> [https://www.kaggle.com/datasets/miguelaelnle/massive-stock-news-analysis-db-for-nlpbacktests?select=raw\\_analyst\\_ratings.csv](https://www.kaggle.com/datasets/miguelaelnle/massive-stock-news-analysis-db-for-nlpbacktests?select=raw_analyst_ratings.csv) (retrieved June 28, 2023)

back from a few techniques, because they are too computationally expensive for this project (the Jupyter Server). But applying the Logistic Regression with an additional feature, if positive or negative words appear in the headline, we raised the accuracy by ~1.2% to an **accuracy of 69.27%**.

This value is way more than we ever expected that we would achieve.

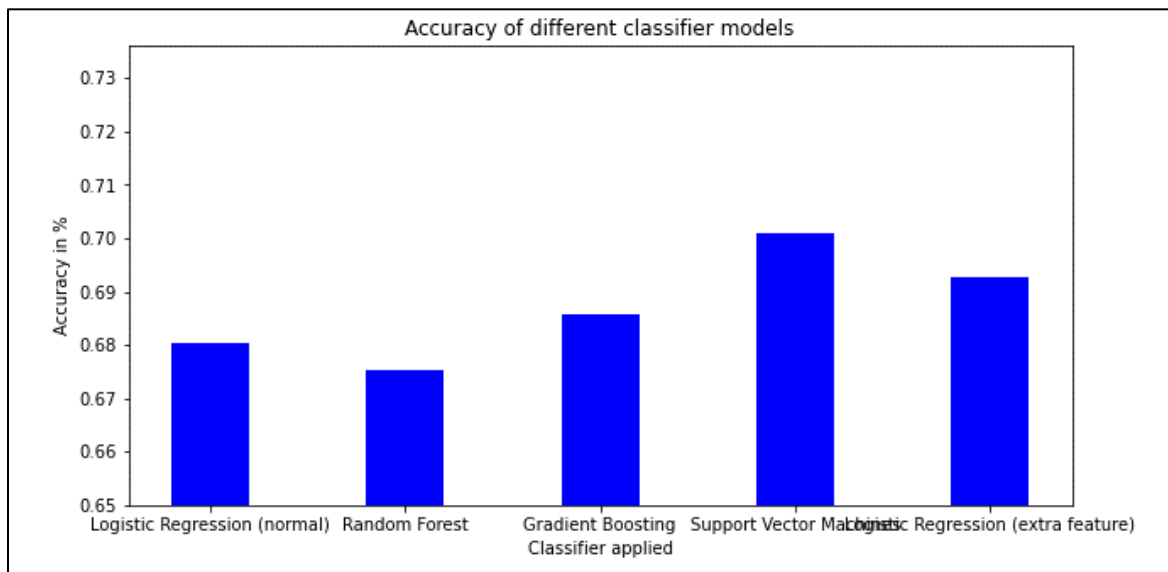


Figure 4: Accuracy of different classifier models

Finally, we chose the Logistic Regression for our project, because it was accurate enough, while still being very fast.

## Legal and Ethical Issues

### Legal and Ethical Guidelines

#### Legal Guidelines

##### *Intellectual Property Rights*

In our project, we ensure that we respect intellectual property rights by using the Kaggle datasets with proper authorization and adherence to the terms and conditions set by the dataset creators. We acknowledge the ownership of the dataset creators and do not infringe upon any copyrights or trademarks.

##### *Data Privacy and Protection*

We prioritize data privacy and protection by implementing appropriate measures to safeguard the personal information of individuals involved in our project. We comply with relevant data protection laws and regulations, and we obtain informed consent from users whose data is utilized. We also take steps to anonymize or aggregate the data whenever possible to protect user privacy. This will especially be relevant for our future work, the dataset from Kaggle has no concrete personal information.

##### *Future Consumer Protection and Model doublecheck*

We take consumer protection seriously and ensure that our project does not engage in misleading or deceptive practices. We provide clear and transparent information about the limitations, accuracy, and potential risks associated with the predictions generated by our models. We also avoid making any false or exaggerated claims regarding the stock market performance or the reliability of our predictions.

## **Ethical Guidelines**

### *Respect for User Privacy*

We prioritize user privacy by implementing measures to protect the personal information of individuals involved in our project. We handle data responsibly, ensuring it is securely stored and accessed only by authorized personnel. We also adhere to privacy principles such as minimizing data collection, obtaining user consent, and providing transparency about how data is used.

### *Responsible Data Usage*

We are committed to responsible data usage by adhering to ethical standards and best practices. We use data for the intended purpose of our project and take measures to ensure its integrity, accuracy, and security. We also respect the rights of data providers and comply with any restrictions or limitations set by the dataset creators regarding data usage and sharing.

## **Legal and Ethical Challenges**

### **Legal Challenges**

#### *Intellectual Property Infringement*

To avoid potential intellectual property issues, it is important to ensure that the datasets used in our project are properly licensed and obtained from authorized sources. We overcame this challenge by thoroughly reviewing the terms and conditions of the datasets, obtaining necessary permissions, and providing proper attribution to the dataset creators. This is also described on our proposal.

#### *Data Privacy Compliance*

Given the sensitive nature of financial and personal data, it is crucial to comply with data privacy laws and regulations. To overcome this challenge, implement strong data protection measures were implemented, we obtained informed consent from users, securely stored and handled data, and considered anonymization techniques whenever possible.

#### *Compliance with Financial Regulations*

Financial markets are heavily regulated, and using stock market data for predictive purposes may involve compliance challenges. We ensured that we are aware of and comply with relevant financial regulations to avoid any legal issues. We consulted with legal professionals or experts in the field to ensure compliance with regulations specific to our project.

### **Ethical Challenges**

#### *Bias*

Addressing bias and discrimination in machine learning models is crucial. In our project, we have taken steps to identify and mitigate biases by employing fairness metrics such as Disparate Impact, Equalized Odds, and Statistical Parity. These metrics allow us to assess and address potential biases in our models and datasets. By regularly evaluating and monitoring our models for fairness, we aim to reduce discriminatory outcomes and ensure inclusivity.

We understand the importance of evaluating our models for fairness and inclusivity. Therefore, we have incorporated techniques like data augmentation and diverse training data to minimize the potential for bias. Additionally, we have implemented bias-aware algorithms that consider factors such as protected attributes or sensitive features to reduce discriminatory outcomes.

By adopting these measures, we strive to create machine learning models that are more equitable and less prone to bias. However, it is crucial to acknowledge that bias detection and

mitigation are ongoing processes. We will continue to evaluate our models, monitor fairness metrics, and refine our approaches to address any identified biases.

The most important way to avoid the biases was that we used a lot of newspaper sources in the financial sector. It would also be possible to extend the input with the publisher to train the model a bias on the publisher.

### *Accountability and Responsibility*

For us, it is essential to take responsibility for the outcomes and impacts of our project. We developed a framework for accountability, established guidelines for responsible data usage, and considered conducting regular audits or third-party reviews to ensure compliance with ethical standards. In addition, we engaged with relevant stakeholders and addressed any concerns or feedback that arise.

## **Experience Gained**

### **Discussion on the challenges encountered**

During the project, we faced several challenges. One of the main challenges was understanding and working with natural language processing (NLP) techniques, particularly in the context of predicting short forms from headlines and the positive effect of headlines on stock differences. NLP can be complex and requires a deep understanding of language processing and machine learning algorithms. Additionally, the preprocessing of data proved to be challenging at times, and merging datasets resulted in some data loss due to a lower intersection. Overcoming these challenges required collaborative problem-solving and continuous learning.

### **Summary of experienced gained by each team member**

#### **Alex's Experience**

Alex focused on the middle part of the NLP pipeline, analyzing and processing headlines. He applied techniques like sentiment analysis and keyword scanning to extract insights and contributed to embedding the keyword scanner for positive effect prediction.

#### **Gabriel's Experience**

Gabriel gained experience in NLP, focusing on the beginning and ending stages of the machine learning process. He worked on preprocessing textual data, feature extraction from headlines, and developing models for predicting short forms.

#### **Max's Experience**

Max specialized in the positive effect prediction part, utilizing logistic regression and mathematical techniques. He gained expertise in fine-tuning the logistic regression model, handling class imbalance, and enhancing statistical rigor for accurate predictions.

### **Recommendations for future work**

Moving forward, we have identified some recommendations for future work. Firstly, given the interest in real-time prediction, we would like to embed the system in a data stream. This involves incorporating XML RSS feeds from multiple newspapers to capture the influence of headlines on stock differences. Additionally, we would integrate an API from a trading portal to generate offers and closing tags with weights, allowing the system to prioritize the most accurate predictions. This integration would require further development and testing to ensure seamless data flow and accurate predictions. Continuously refining the NLP models, exploring advanced techniques, would also be important areas to focus on in future iterations. Especially the linear regression part for the exact prediction to be able to tell the exact amount



of change in the stock market. Also, a weight checker and self-learning part will be introduced in the following flowchart:

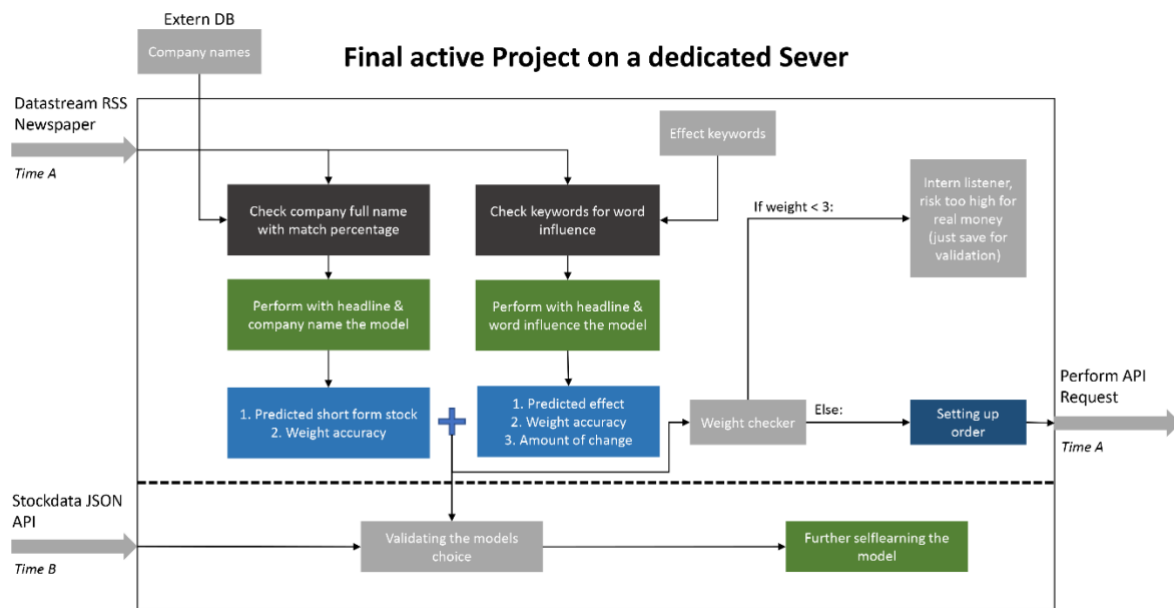


Figure 5: Final project flow

## Table of Figures

Figure 1: Dataset structures .....	4
Figure 2: excerpt stock data (stock "A" - Agilent Technologies, Inc.) .....	4
Figure 3: excerpt headline data .....	5
Figure 4: Accuracy of different classifier models .....	6
Figure 5: Final project flow .....	9

## References

- [1] “Start with more than a blinking cursor”. Kaggle Inc. <https://www.kaggle.com/> (accessed July 7, 2023)
- [2] O. Onyshchak. “Stock Market Dataset Historical daily prices of Nasdaq-traded stocks and ETFs”. Kaggle Inc. <https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset?select=stocks> (accessed July 7, 2023)
- [3] BOT\_Developer. “Daily Financial News for 6000+ Stocks ~4m articles for 6000 stocks from 2009-2020”. Kaggle Inc. <https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests> (accessed July 7, 2023)
- [4] O. Onyshchak. “Stock Market Dataset Historical daily prices of Nasdaq-traded stocks and ETFs”. Kaggle Inc. <https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset?select=stocks> (accessed July 7, 2023)
- [5] BOT\_Developer. “Daily Financial News for 6000+ Stocks ~4m articles for 6000 stocks from 2009-2020”. Kaggle Inc. <https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests> (accessed July 7, 2023)