Using News Sentiment for Better Stock Price Prediction

Tomasz Starakiewicz
Faculty of Economic Sciences
University of Warsaw
tomasz.starakiewicz@student.uw.edu.pl

Dustin Pacholleck
Faculty of Economic Sciences
University of Warsaw
d.pacholleck@student.uw.edu.pl

Abstract **This project aims to find out if Sentiment Analysis improves the accuracy of stock price prediction. It utilizes the pre-trained language model FinBert to predict sentiment of self-scraped financial news articles from 2011 to 2022 of the stock Apple (AAPL). After having assigned a sentiment score to each news article of a particular stock, all news articles of a day are grouped, and an average sentiment of the day is calculated. These average sentiments serve as input for a prediction model, in this case Random Forest, to predict the stock price. The model uses the closing price of the previous day and the sentiment of the day to predict the opening price of the following day.**

The project was not able to show an increase of predictive power using sentiment regarding MAPE compared to the benchmark model. Therefore, it can be concluded that a stronger form of the Efficient Market Hypothesis holds, and the effect of news are already reflected in the stock price.

I. Introduction

Stock exchanges are financial institutions that facilitate the transfer of various goods (monetary values, actions, and precious metals). With a trading turnover of thousands of billions of dollars, this piques people's interest in generating a profit. Items are traded on the market, and their subsequent value is used to judge whether the transaction was profitable.

This market participation can result in gains or losses for a trader, depending on the ability to predict future values. Therefore, stock price prediction gained a lot of interest and different approaches have been developed (Iacomin 2015).

According to the Efficient Market Hypothesis (Fama 1970) commodity prices reflect all available information. Nevertheless, behavioral finance literature has found that investor sentiment has predictive ability for equity returns (McGurk, Nowak, and Hall 2020) and investors sentiment yields predictive power (Zhou 2018).

This is why this project tries to investigate if financial news sentiment can be used for stock price prediction.

II. Data

After having introduced the topic and stated its relevance, this project introduces the used dataset. Within this, there will be a focus on a descriptive part, how the data is cleaned for sentiment prediction and topic modelling to further understand the used data.

A. Description

The dataset used for this analysis is self-scraped by the author sourced by the website investing.com. According to the platform it "is a financial markets platform providing real-time data, quotes, charts, financial tools, breaking news and analysis across 250 exchanges around the world in 44 language editions. With more than 46 million monthly users, and over 400 million sessions, Investing.com is one of the top three global financial websites according to both SimilarWeb and Alexa." (‘About Investing.Com’ n.d.).

The dataset covers worldwide financial stock market news from 2011 till end of 2022 from multiple financial news providers like Reuters, Bloomberg, and investing.com itself. It consists of over 176.224 news articles where one article can be about several stocks.

For the scope of the project the dataset is limited to the news about Apple which yields 9.366 unique articles for prediction.

The news coverage over the years increases but remains from 2017 within a range of 1100-1700 articles per year.

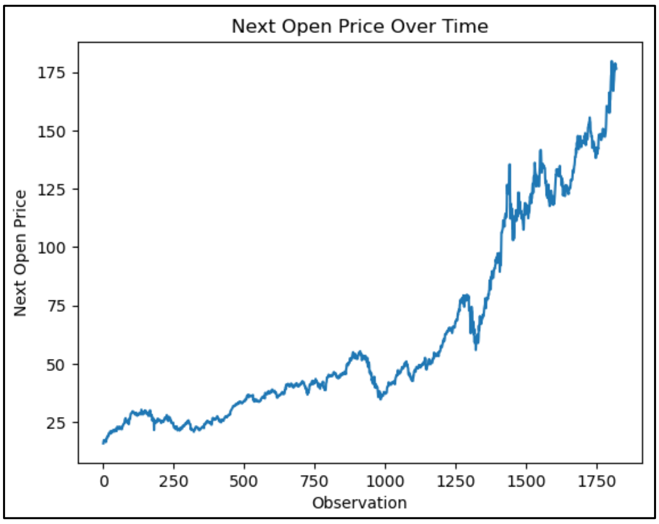
The dataset contains the following variables:

Table I: Dataset Variables

Variable	**Description**
date	Date and time when the news article was published.
headline	Headline of the article.
instrument	Ticker Symbol of the commodity the article is about.
source_url = link	Link to the article.
text	Text of the article.
source	News Publisher of the article.

Later, the yahoo finance API is used to retrieve the closing price of the investigated stock on the previous day and the opening price on the following day. Here, it is important to notice that if the previous day or the following day is a banking holiday the closest day's value will be used. Due to the trading hours of NASDAQ (nasdaq.com n.d.), the previous day and next day are used to avoid information leakage. Visualizing the next day's open price over time mirrors the charts of Apple found online ('Apple Inc. (AAPL) Interactive Stock Chart - Yahoo Finance' n.d.).

Figure I: Opening Prices of Apple in the Train Dataset



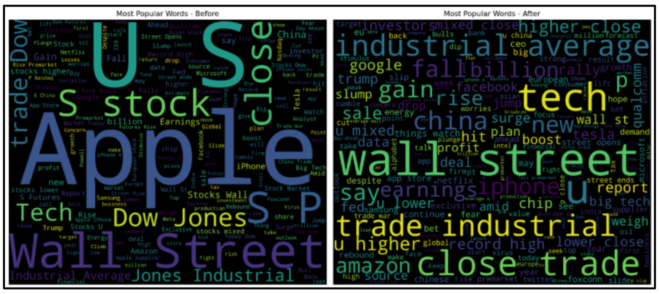
B. Data Cleaning

The data cleaning consists of several parts. One part includes removing special characters and conversion to lowercase the second part includes removing stopwords, the last part stemming with lemmatization This can be achieved by using NLTK ('NLTK : Natural Language Toolkit' n.d.). For the stopwords not only the provided corpus was used but also finance specific stopwords where added. These stopwords were generated by ChatGPT[1] and advanced based on the presented data inspection.

Inspecting the dataset before and after shows the necessity why these steps are necessary. To improve the accuracy of the following topic modelling as also the sentiment prediction (Gharatkar et al. 2017).

[1] Prompt: „Generate domain-specific stopwords for a finance language model."

Figure II: Worldcloud of Raw and Preprocessed Data



Before, it can be observed that Apple itself was a very frequent word which in this case does not provide additional information towards the content.

After the preprocessing different topics become more apparent.

C. *Topic Modelling for Insight Generation*

After having seen the most popular words used in the headlines, this project tries to make the news more comprehensible to understand what they are about. For this, topic modelling can be used due to its capability to automatically organize, understand, search, and summarize large electronic archives (Tong and Zhang 2016).

Topic modelling can be achieved by different methods like Latent Dirichlet Allocation (LDA) which is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar (Ibid.)

Using human interpretability of the topics, three topics seem to be optimal and evaluate as follows.

Figure III: Results Topic Modelling Part I

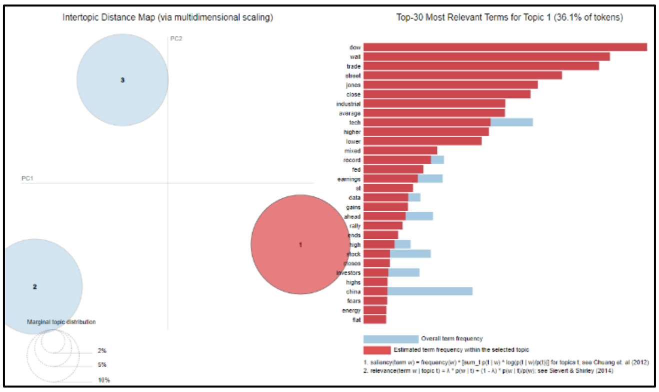


Figure IV: Results Topic Modelling Part II

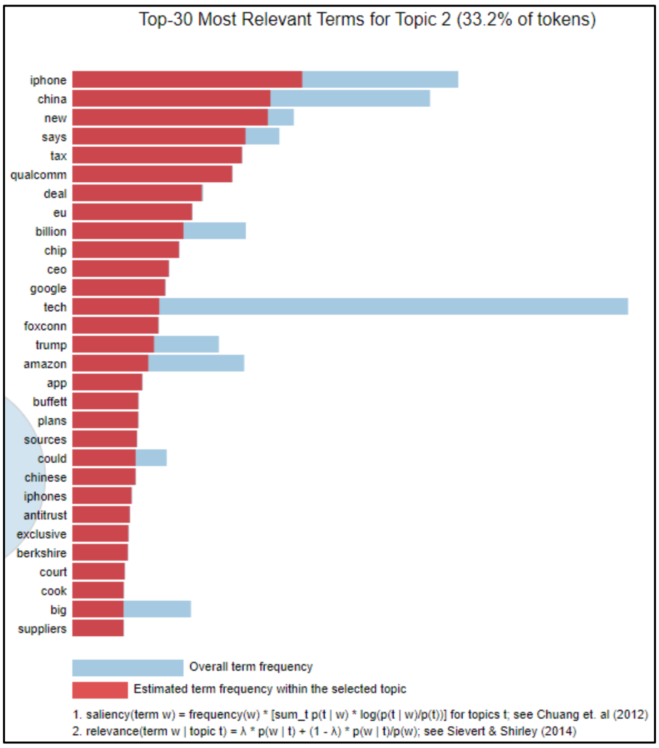
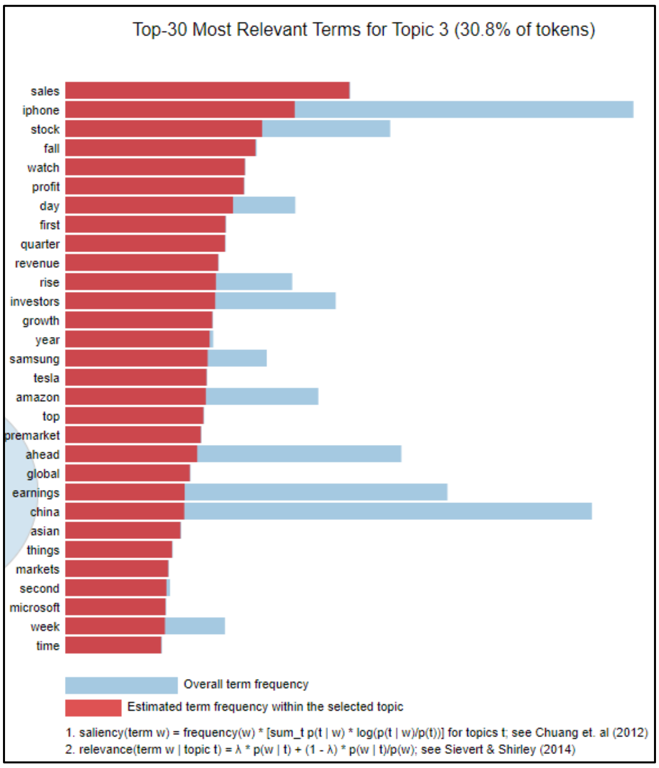


Figure V: Results Topic Modelling Part III



From the above figures a naming of the topics can be derived as follows:

1. Stock Market News
2. Suppliers News
3. Operational Performance News

After having a good understanding of the data and having it preprocessed, modelling can be performed.

III. Modeling

The preprocessing and topic modelling set the fundamentals for further data modelling. In this project the modelling will consist of sentiment predicting using FinBert and afterwards using the predicted sentiment for stock price prediction.

A. *Sentiment Modelling Using FinBert*

Sentiment Analysis is used to systematically identify, extract, quantify, and study affective states and subjective information ('Sentiment Analysis' 2023). This can be done either by having prelabeled data to train a language model or to rely on pretrained models.

Here, the FinBert model is used. It is based on the general-purpose BERT model but is fine-tuned on a domain-specific corpora for the financial context. It is said to outperform NLP task in the financial domain in comparison to other state-of-the-art methods (Araci 2019).

The sentiment analysis has two components. The first and for the stock price prediction relevant one is to determine the average positive, neutral, and negative sentiment on a given day since this could be used as a proxy for market sentiment of the given stock Apple. For this, the scores for each headline are determined and then averaged over a day.

Also documented are the number of articles on that given day as additional input variable the provide a weight to the sentiment.

The second component is assigning each article a dominant sentiment to use it for further analysis of the found sentiment groups.

B. *Predictive Modelling Using Random Forest Regressor*

After having determined the sentiment mix of a given day by using Finbert, Random Forest is used for stock price prediction.

The Random Forest model is arbitrarily chosen because this project focuses on text mining. Random Forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time ('Random Forest' 2023).

For stock price prediction the data is split into training and testing, where the data from 2022 serves as test sample.

For this project, first a base model is created which in a second step is optimized by Randomized Search ('Sklearn.Model_selection.RandomizedSearchCV' n.d.).

At the end the optimized model is used to make predictions.

This procedure is used on the data with the sentiment information and on the data without such information.

The model without sentiment is used as reference model to evaluate if adding sentiment improves forecasting accuracy in form of Mean Absolute Percentage Error (MAPE) ('Mean Absolute Percentage Error' 2023).

IV. Evaluation

Having modelled the different sentiments and predicted the stock prices with and without using the sentiment, this project evaluates the achieved results.

A. *Sentiment Evaluation*

Since no true labels are available, the quality of sentiment prediction cannot be evaluated. Nevertheless, the identified sentiment groups positive, neutral, and negative can be further investigated.

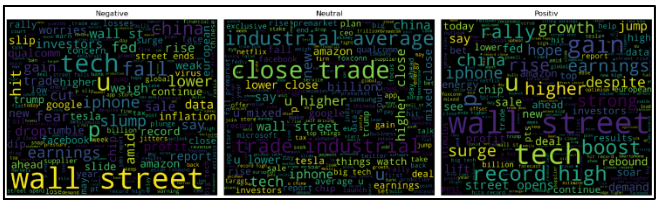
From the table it becomes apparent that most articles have a dominantly neutral sentiment which is in line with the expectation. Nevertheless, we can see more negative news articles then positive which could be explained by a negativity bias (‘Negativity Bias’ 2023).

Table II: Sentiment Distribution

Sentiment	**Number of Articles**
Negative	2,891 (30.87%)
Neutral	4,837 (51.64%)
Positive	1,638 (17.49%)

To investigate the different sentiments, a wordcloud gives more insights into the data.

Figure VI: Wordcloud of the Sentiments



As it can be seen, negative sentiment often includes words like fall or lower, neutral sentiment often close or trade, and positive sentiment often higher or record. Such words are in line with the expectations.

B. *Predictive Modelling*

After gaining an understanding about the different sentiment groups, the predictive Random Forest models will be evaluated.

As it can be seen in the table below, the difference in predictive accuracy between the sentiment and the reference model is small and the reference model even performs slightly better.

Table III: Model Evaluation

Model	**MAE**	**MSE**	**RMSE**	**MAPE**
Reference Base	3.230221	16.692650	4.085664	0.021351
Reference Tuned	3.171608	16.144776	4.018056	**0.020978**
Sentiment Base	3.332613	17.267890	4.155465	0.021906
Sentiment Tuned	3.203805	16.212730	4.026503	**0.021099**

This might lead to the conclusion that in the used test setting using sentiment does not improve predictive accuracy. The data itself already seems to reflect the price movements sufficiently for the Random Forest model to predict market movements. This can be an indicator that the Efficient Market Hypothesis might hold true in a stronger form.

V. SUMMARY

Having evaluated the sentiment and the predictive models, this chapter summarizes the found results.

First, further findings of this project will be summarized and afterwards the stated aimed to analyze if Sentiment Analysis improves the accuracy of stock price prediction on the example of Apple will be answered. Last, this project gives an outlook for further research.

1. The news about Apple can be split into three topics: Stock Market News, Suppliers and Operational Performance.
2. News tend to be neutral but have a negativity bias.
3. Random Forest can be used for predicting stock prices.
4. Finally, we cannot conclude that using sentiment based on this project with its specific assumptions yielded a better predictive accuracy than the benchmark

model. This gives reasons to believe that in these settings a stronger form of Efficient Market Hypothesis seems to hold.

Nevertheless, further research is needed. It should be investigated if the effect remains similar with other trading intervals, with other predictive regression models, and with different stocks. Furthermore, other pre-trained language models should be evaluated for sentiment prediction, social media data should be considered for prediction, and using the full news body instead of the headline should be considered.

References

‘About Investing.Com’. n.d. Investing.Com. Accessed 26 February 2023. https://www.investing.com/about-us/.

‘Apple Inc. (AAPL) Interactive Stock Chart - Yahoo Finance’. n.d. Accessed 27 February 2023. https://finance.yahoo.com/quote/AAPL/.

Araci, Dogu. 2019. ‘FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models’. arXiv. http://arxiv.org/abs/1908.10063.

Bukovina, Jaroslav. 2016. ‘Social Media Big Data and Capital Markets—An Overview’. Journal of Behavioral and Experimental Finance 11: 18–26.

Dhuriya, Ankur. 2021. ‘What Is Topic Modeling?’ Analytics Vidhya (blog). 2 February 2021. https://medium.com/analytics-vidhya/what-is-topic-modeling-161a76143cae.

Fama, Eugene F. 1970. ‘Efficient Capital Markets: A Review of Theory and Empirical Work’. The Journal of Finance 25 (2): 383–417.

‘FF Project FinBERT’. n.d. Accessed 27 February 2023. https://kaggle.com/code/hathalye7/ff-project-finbert.

‘FinBERT: Financial Sentiment Analysis with BERT’. (2019) 2023. Jupyter Notebook. Prosus AI. https://github.com/ProsusAI/finBERT.

Gharatkar, Sandesh, Aakash Ingle, Tanmay Naik, and Ashwini Save. 2017. ‘Review Preprocessing Using Data Cleaning and Stemming Technique’. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 1–4. https://doi.org/10.1109/ICIIECS.2017.8276011.

Iacomin, Radu. 2015. ‘Stock Market Prediction’. In 2015 19th International Conference on System Theory, Control and Computing (ICSTCC), 200–205. https://doi.org/10.1109/ICSTCC.2015.7321293.

Kapadia, Shashank. 2022. ‘Evaluate Topic Models: Latent Dirichlet Allocation (LDA)’. Medium. 24 December 2022. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

McGurk, Zachary, Adam Nowak, and Joshua C. Hall. 2020. ‘Stock Returns and Investor Sentiment: Textual Analysis and Social Media’. Journal of Economics and Finance 44 (3): 458–85. https://doi.org/10.1007/s12197-019-09494-4.

‘Mean Absolute Percentage Error’. 2023. In Wikipedia. https://en.wikipedia.org/w/index.php?title=Mean_absolute_percentage_error&oldid=1136183305.

nasdaq.com. n.d. ‘Stock Market Trading Hours for Nasdaq’. Accessed 26 February 2023. https://www.nasdaq.com/stock-market-trading-hours-for-nasdaq.

‘Negativity Bias’. 2023. In Wikipedia. https://en.wikipedia.org/w/index.php?title=Negativity_bias&oldid=1136133821.

‘NLTK :: Natural Language Toolkit’. n.d. Accessed 26 February 2023. https://www.nltk.org/.

‘ProsusAI/Finbert · Hugging Face’. n.d. Accessed 26 February 2023. https://huggingface.co/ProsusAI/finbert.

‘Random Forest’. 2023. In Wikipedia. https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1134454815.

‘Sentiment Analysis’. 2023. In Wikipedia. https://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=1134856605.

‘Sklearn.Model_selection.RandomizedSearchCV’. n.d. Scikit-Learn. Accessed 27 February 2023. https://scikit-learn/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.

Tong, Zhou, and Haiyi Zhang. 2016. ‘A Text Mining Research Based on LDA Topic Modelling’. In , 201–10.

Zhou, Guofu. 2018. ‘Measuring Investor Sentiment’. Annual Review of Financial Economics 10 (1): 239–59. https://doi.org/10.1146/annurev-financial-110217-022725.