Predicting stock volatility from descriptive texts

Team:

Julia Ornatowski

Fabian Basler

Evamaria Hammerschmid

1. Introduction

“Measuring and forecasting latent volatility have many important applications in many areas of finance including asset allocation, option pricing, and risk management.“ (Brownlees & Gallo, 2010). The majority of existing analyses are based on a variety of assumptions and information sets, such as historical variances, ranges, or implied volatilities.

Meanwhile, even though the application of natural language processing and machine learning methods has become more accessible and easier to apply on financial data, the existing research in the field of predicting price movements based on qualitative data or descriptive texts has been rather limited. Only relatively young literature focuses on predicting risk with non-quantifiable data from news or social media (Ding etal., 2015; Qin & Yang, 2019), even though Atkins et. al. (2018) state, that "the behaviour of time series data from financial markets is influenced by a rich mixture of quantitative information from the dynamics of the system, captured in its past behaviour, and qualitative information about the underlying fundamentals arriving via various forms of news feeds“ (p. 120).

This project adds to those efforts by examining the additional information content descriptive texts can yield when predicting stock volatility.

2. Objective & Methodology

2.1 Overall objective

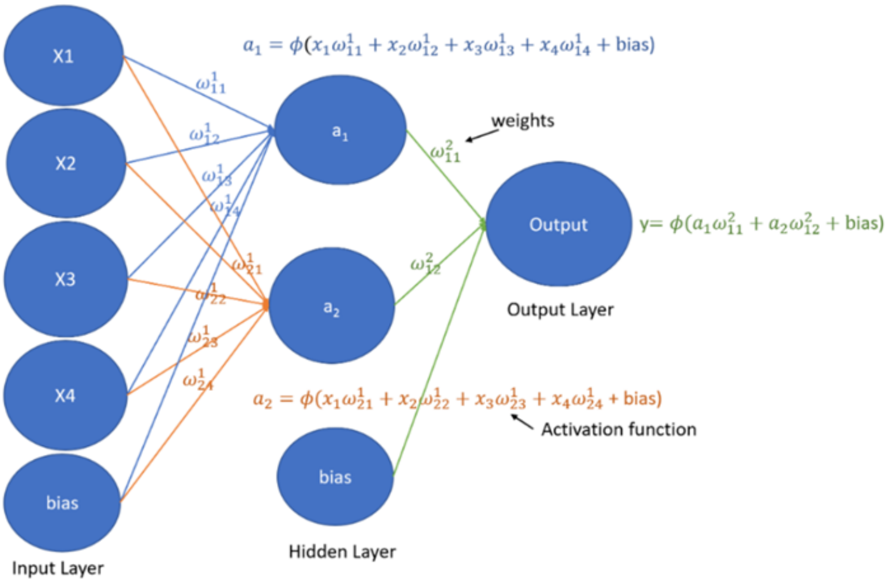
The overall objective is to predict stock volatility using text data as an input. Different Machine Learning models which are listed below were trained based on the company descriptions provided on the Reuters website. The texts give an overview of the company’s market position and current objectives. The idea is that based on the assumption that industries (e.g Tech vs. Automotive) and other yet undetermined distinctive factors, are observable as patterns in the stock’s volatilities. The target for the models is the average daily volatilities of the past trading 252 days (one year), extracted from yahoo finance. Finally, all models are evaluated and compared regarding performance to identify the most successful method.

2. 2 Machine Learning Models:

Different common practice machine learning models were applied. For the Linear Regression, Gradient Boosting and Regression Tree the GridSearchCV package was used to help optimizing the hyperparameters.

2.2.1 Neural Network:

Neural networks are a combination of several Perceptrons. They have an input layer, an output layer, and hidden layers in between. Our neural network consists of 4 hidden layers.


$a_1 = \phi(x_1\omega_{11}^1 + x_2\omega_{12}^1 + x_3\omega_{13}^1 + x_4\omega_{14}^1 + \text{bias})$
X1
X2
X3
X4
bias
Input Layer
ω_{11}^1
ω_{12}^1
ω_{13}^1
ω_{14}^1
ω_{21}^1
ω_{22}^1
ω_{23}^1
ω_{24}^1
a1
a2
bias
Hidden Layer
weights
ω_{11}^2
ω_{12}^2
Output
Output Layer
$y = \phi(a_1\omega_{11}^2 + a_2\omega_{12}^2 + \text{bias})$
$a_2 = \phi(x_1\omega_{21}^1 + x_2\omega_{22}^1 + x_3\omega_{23}^1 + x_4\omega_{24}^1 + \text{bias})$
Activation function

1 Basic Neuronal Network (https://towardsdatascience.com/how-to-solve-randomness-in-an-artificial-neural-network-3befc4f27d45)

The layer gives an output of 0 or 1 to the next layer. The activation function is used to predict the 1 or 0. The Neural Network gets trained by gradient descent. The algorithm tries to minimize a loss function. In this case it tries to minimize the mean squared error. With every iteration the neural network becomes more accurate. (IBM, 2021)

2.2.3 Gradient Boosting:

The Gradient Boosting Models is based on a regression tree and in every step generates a model focusing on previous errors. The Gradient Boosting Model was fitted with a learning rate of 0.15 and a number of estimators=120, based on the GridSearchCV results.

2.2.4 Linear Regression:

The Linear Regression model performed best for setting normalize to “True” and fit_intercept to “False”.

2.2. Regression Tree:

For the Regression Tree we tried several combinations for potential optimal hyperparameters with GridSearchCV. The best result was found with max_depth of 2 and min_sample_split of 321.

2.2.5 Mean (as Benchmark):

Lastly, we used the mean of the sample volatilities as predicted values. Then we use the performance of the mean as benchmark to check whether our trained models can outperform this simplistic approach.

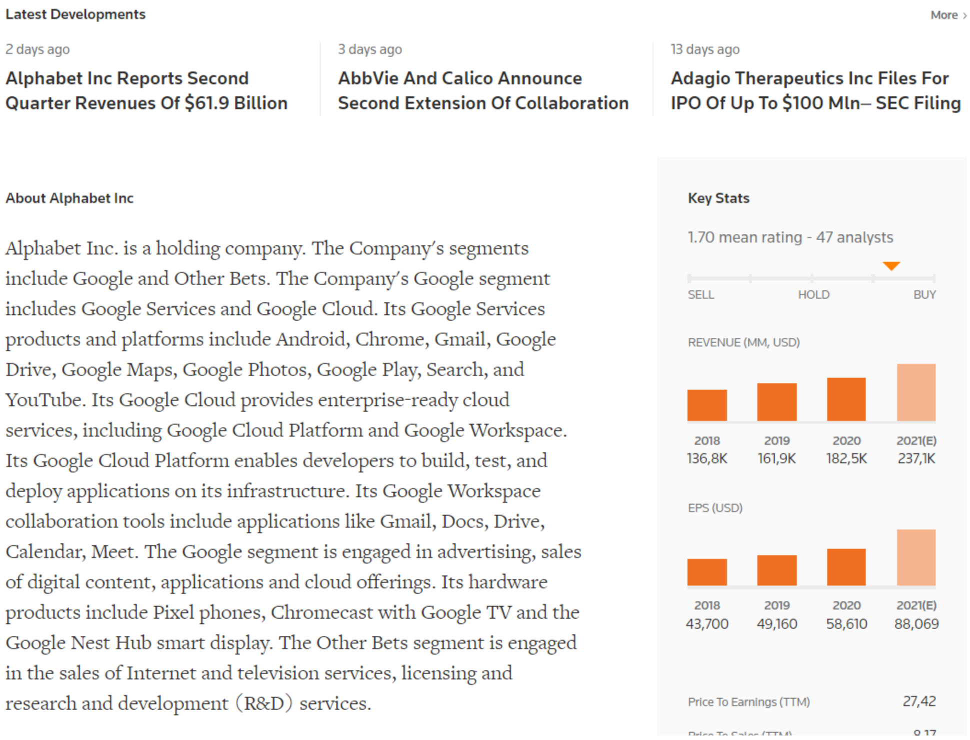
2.3 Data Sources

In order to have a broad portfolio, all stocks from the S&P 500 have been examined, allowing an appropriate separation of the training and test data. In addition to the sample size, the S&P 500 is often considered as a benchmark for the overall U.S economy, including its 500 companies which are broadly spread across different industries. This heterogeneity is additionally supposed to make training the machine learning models more efficient.

2.3.1 Input (X-Variable): Reuters Company descriptions

Reuters provides a 150-200 word long description of every company behind the stocks of the S&P 500. Even though those text do not fully reflect recent events concerning the company, they are regularly updated to sum up key facts of its current market position and objectives.

In order to gather the descriptions of stock , the corresponding tickers were first scraped from a Wikipedia entry with a table of all companies in the S&P 500. Those tickers were then inserted into an URL directed to the matching Reuters page. The sample consists of 505 stocks, since some Since the descriptions were equal, duplicates were simply dropped from the data set.



Latest Developments

More ›

2 days ago	3 days ago	13 days ago
Alphabet Inc Reports Second Quarter Revenues Of $61.9 Billion	AbbVie And Calico Announce Second Extension Of Collaboration	Adagio Therapeutics Inc Files For IPO Of Up To $100 Mln– SEC Filing

About Alphabet Inc

Alphabet Inc. is a holding company. The Company's segments include Google and Other Bets. The Company's Google segment includes Google Services and Google Cloud. Its Google Services products and platforms include Android, Chrome, Gmail, Google Drive, Google Maps, Google Photos, Google Play, Search, and YouTube. Its Google Cloud provides enterprise-ready cloud services, including Google Cloud Platform and Google Workspace. Its Google Cloud Platform enables developers to build, test, and deploy applications on its infrastructure. Its Google Workspace collaboration tools include applications like Gmail, Docs, Drive, Calendar, Meet. The Google segment is engaged in advertising, sales of digital content, applications and cloud offerings. Its hardware products include Pixel phones, Chromecast with Google TV and the Google Nest Hub smart display. The Other Bets segment is engaged in the sales of Internet and television services, licensing and research and development (R&D) services.

Key Stats

1.70 mean rating - 47 analysts

REVENUE (MM, USD)

2018	2019	2020	2021(E)
136,8K	161,9K	182,5K	237,1K

EPS (USD)

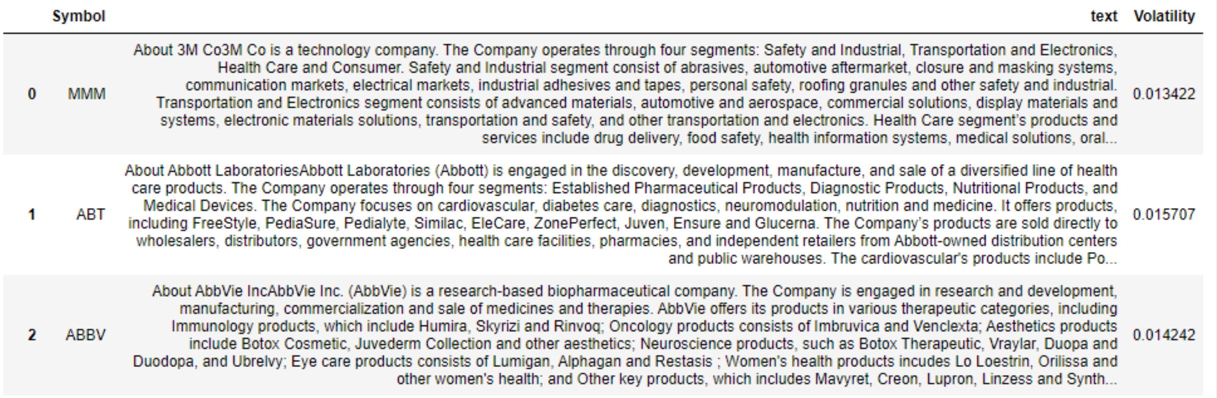
2018	2019	2020	2021(E)
43,700	49,160	58,610	88,069

Price To Earnings (TTM)	27,42
Price To Sales (TTM)	8,17

2 Example of the text data (GOOGL.O - Alphabet Inc Profile | Reuters)

2.3.2 Target (Y-Variable): Volatility

To match the description of S&P 500 stocks with an appropriate time horizon, volatility was measured as the average over the past 252 days. To download the required data, the earlier derived tickers were applied to fetch the Closing Prices of all stocks between the 13.07.2020 and the 13.07.2021. Next the percentage change was calculated over the whole period and finally the standard deviation applied.



	Symbol	text	Volatility
0	MMM	About 3M Co3M Co is a technology company. The Company operates through four segments: Safety and Industrial, Transportation and Electronics, Health Care and Consumer. Safety and Industrial segment consist of abrasives, automotive aftermarket, closure and masking systems, communication markets, electrical markets, industrial adhesives and tapes, personal safety, roofing granules and other safety and industrial. Transportation and Electronics segment consists of advanced materials, automotive and aerospace, commercial solutions, display materials and systems, electronic materials solutions, transportation and safety, and other transportation and electronics. Health Care segment's products and services include drug delivery, food safety, health information systems, medical solutions, oral...	0.013422
1	ABT	About Abbott LaboratoriesAbbott Laboratories (Abbott) is engaged in the discovery, development, manufacture, and sale of a diversified line of health care products. The Company operates through four segments: Established Pharmaceutical Products, Diagnostic Products, Nutritional Products, and Medical Devices. The Company focuses on cardiovascular, diabetes care, diagnostics, neuromodulation, nutrition and medicine. It offers products, including FreeStyle, PediaSure, Pedialyte, Similac, EleCare, ZonePerfect, Juven, Ensure and Glucerna. The Company's products are sold directly to wholesalers, distributors, government agencies, health care facilities, pharmacies, and independent retailers from Abbott-owned distribution centers and public warehouses. The cardiovascular's products include Po...	0.015707
2	ABBV	About AbbVie IncAbbVie Inc. (AbbVie) is a research-based biopharmaceutical company. The Company is engaged in research and development, manufacturing, commercialization and sale of medicines and therapies. AbbVie offers its products in various therapeutic categories, including Immunology products, which include Humira, Skyrizi and Rinvoq; Oncology products consists of Imbruvica and Venclexta; Aesthetics products include Botox Cosmetic, Juvederm Collection and other aesthetics; Neuroscience products, such as Botox Therapeutic, Vraylar, Duopa and Duodopa, and Ubrelvy; Eye care products consists of Lumigan, Alphagan and Restasis ; Women's health products incudes Lo Loestrin, Orilissa and other women's health; and Other key products, which includes Mavyret, Creon, Lupron, Linzess and Synth...	0.014242

3 Final Data Set including Tickers, Corresponding Texts (X-Variable) and the Volatility (Y-Variable)

2.4 Preprocessing and Preparing for ML methods:

For preprocessing the data, a CountVectorizer was applied to create a bag of words model. Additionally a Lemmatizer, not part of the CountVectorizer package, was added.

Next the TfidfVectorizer used to further clean the data and create a feature matrix. Uninformative stopwords were included by setting the parameter stop_words to “english”, both uni- and bigrams were included for the feature matrix through setting “ngram_range” to “1,2”. The previously created function “my_preprocessor” was applied through setting the preprocessor-parameter to “my_preprocessor”.

Lastly a fit_transform function was applied to create the final X-values.

2.5 Comparison of different models

In order to compare the different models three variations of the Mean Error were compared. Firstly, the mean absolute error values were compared. The MAE takes the average of all absolute deviations of the predicted value to the actual value.

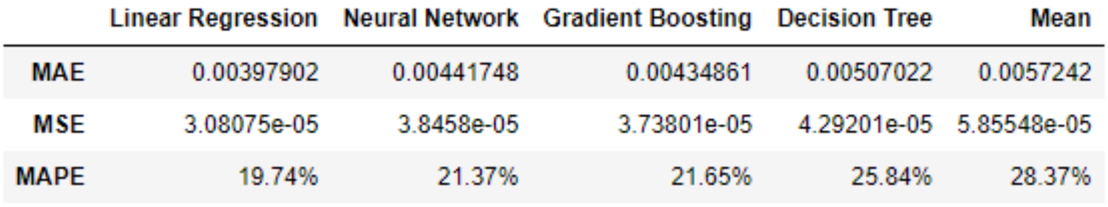
$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Next the Mean Squared Error was compared, which is the average of all deviations of the predicted value to the actual value squared. Lastly, for easier comparability we calculated the Mean Absolute Percentage Error.

3. Results:

The results show that every model outperforms simply taking the mean as predicted value, therefore some predictive information can be attributed to the texts.

Surprisingly the standard Linear Regression model performed best, closely followed by the Neural Network and the Gradient Boosting Method. As a less optimized version of the Gradient Boosting Model the Decision Tree performed worse and generated the worst results of all Machine Learning Models, only three percentage points better than simply looking at the mean.



	Linear Regression	Neural Network	Gradient Boosting	Decision Tree	Mean
MAE	0.00397902	0.00441748	0.00434861	0.00507022	0.0057242
MSE	3.08075e-05	3.8458e-05	3.73801e-05	4.29201e-05	5.85548e-05
MAPE	19.74%	21.37%	21.65%	25.84%	28.37%

4 Performance comparison of different applied Machine Learning Models

4. Limitations and Future work:

One major difficulty was comparing the different models, since the volatility measure is not an easily classifiable category, but rather a continuous value. It should also be pointed out, that even though it was attempted to define ideal hyperparameters for every model, some constraints were imposed through the available processing power.

There also remains room for improvement regarding the analysed data. Though the company descriptions are regularly updated, they do not reflect or include recent news regarding the company or larger macroeconomic influences.

Future work could therefore include the "Latest News" and "Key development" section from Reuters to predict volatility. To match this with the volatility measure of the last 250 days, we suggest to also gather textual data based on this timeframe.

Another improvement possibility would be to use a more dynamic model, checking how the predicted volatility changes if the company description text is updated or renewed, which would require monitoring the descriptions over a certain period to detect changes.

Another possibility of strengthening the results would be to broaden the amount of "static" data used to predict the volatility. It should also be noted that only a limited amount of around 150 words was used for these predictions. In addition, it only provided qualitative descriptions and no metrics. A suggestion would be to include "Financials" and "Key Metrics" to future models.

Sources

Ding, X., Zhang, Y., Liu, T., and Duan J. (2015). Deep learning for event-driven stock prediction. In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15). AAAI Press. 2327–2333.

Qin, Y. & Yang, Y. (2019). What You Say and How You Say It Matters: Predicting Financial Risk Using Verbal and Vocal Cues. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).

Atkins, A., Mahesan, N., and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science, 120-137.

Brownlees, C., & Gallo, G. (2010). Comparison of Volatility Measures: a Risk Management

https://www.ibm.com/cloud/learn/neuralnetworks#:~:text=Neural%20networks%2C%20als 20known%20as,neurons%20signal%20to%20one%20another