Classification of news headlines with impact on the stock price change

Maximilian Franke, Sean Mondesire
St. Thomas University
Miami, FL, USA
mfranke@stu.edu, smondesire@stu.edu

***Abstract*—This study examines the relationship of published news and the stock market reaction. Therefore, a deep neural network architecture is proposed to classify news headlines with an impact on the stock price change. For each headline, the top three tags represent an average and normalized stock price change. A tag is a word that has an impact on the stock price change. In this study, the stock price changes will be predicted for the "Apple" stock. Therefore, all news headlines and stock prices will be transformed and given as input layers for a multi-layer feedforward neural network. The optimal architecture of the neural network is 3 hidden layers with 80, 80, and 10 neurons and 10,000 epochs. The corresponding MSE is 0.039 and the accuracy is 0.652.**

I. Introduction

News is part of everyday life. The channels have been extended from traditional news services to social networks. American President Donald Trump, for example, uses Twitter as a news channel. This behavior was heavily discussed in the context of the 2016 presidential election, as statements in the form of tweets about companies affected its stock prices [1]. For example, on the same day as a positive tweet about Ford (Jan 4, 2017), Ford's stock price increased by 4.6%. On the same day of a negative tweet, the stock price decreased by 1.8% for ExxonMobil (Dec 13, 2016) [1]. Even if the stock prices returned to its previous level after some time, a correlation between news and stock price can be deduced.

It is therefore obvious to assume that these correlations do not only exist between Donald Trump's tweets and the stock market but that this correlation can be interpreted in a broader sense. Therefore, this study examines the relationship between published news headlines and the changes in stock prices. This relationship leads to the problem-solving approach of how to classify news headlines to detect a stock price change.

The solution to this problem can clarify the importance of news regarding stock price changes and may increase significantly the accuracy of the stock price change prediction. In specific, the relationship between tags (individual key words) of headlines and stock price changes will be explored. This will lead to design a model of linking a headline to a price change.

The change of stock prices can be ~~controlled~~ with indicators like RSI (Relative Strength Index) or EMA (Exponential Moving Average). The RSI is an indicator that measures from 0 to 100 the extent of the latest price changes to assess overbought or oversold conditions in the price of a stock or other asset. In comparison, an EMA represents the trend of the price change and reacts sensitively to recent price changes depending on the included number of data points from the past. But to predict the stock price change, it is necessary to include daily updated news, which is not reflected in real time in indicators like RSI or EMA, but rather in headlines.

Therefore, the purpose of this study is to underline the high relevance of tags (individual key words) of headlines for stock price changes. In comparison, just analyzing stock price indicators can lead to a wrong trend or even worse, to a wrong decision for buying or selling a position. By implementing tags of headlines and its impact to a stock price change, the decision for buying or selling is based on recent data and includes opinions and thoughts of people.

The overall approach is divided into two parts, Business Intelligence and model building. The first part, Business Intelligence, analyzes the relationship between tags of headlines and stock price changes by answering the following subproblems:

First, it will be analyzed when ~~[min]~~ the stock price is affected, after the news being published. Second, it is critical to identify first the wave of the stock price change and to analyze the development. The aim is to detect the time of the first increase, the maximum, and the time of returning to "normality".

~~Second~~, it will be developed a deep learning neural network that takes in the current stock price, the volume of the stock, and the transformed price change for headline tags. The details of this technique will be presented in the following chapters.

II. Background

~~Since~~ there is a large number of potentially relevant sources of varying quality, the discussion in this paper of the identified problem will be based on an analysis of the literature which follows a standard framework for literature reviewing [2]. First, the scope of review is defined, in a structured and documented form [3]. The focus defines the "type of sources", meaning that research results, research methods, and applications of stock prediction models are

included in the research. The aim is to integrate them to identify research gaps. The procedure in the research is to be regarded as conceptual. The target group is to be defined as the general professional audience, professionals, and practitioners. The framework of the literature analysis is complete and selective, in the sense of complete research in German and English language.

A first structural analysis of the researched literature results shows that the number of articles has increased in recent years. The high proportion of results from recent years in this research area is an indication of the novelty and relevance of the topic. One of the key messages from the analysis is that the literature identifies news sources as an essential component of stock prediction.

Approaches using indicators as predictors to build classification models are also widely used in this research area. Sezer's study, for example, determines to buy and sell points for stocks by using RSI values. Therefore, a stock trading system was developed, which combines different classification algorithms. The results of their proposed system showed, in comparison to other trading systems, similar results [4]. In comparison, Balaji created in their study 14 different models, which were based on four main different techniques of deep learning and tried to forecast the stock price with accuracies above 0.5 [5]. Furthermore, another study has determined that there is a high correlation between stock market prices and returns [6].

In comparison to the quantitative approaches above, unstructured data as text form has also already been applied in many kinds of research, and predictions were made. For example, Tweets (special tags), were used to predict the spread of diseases [7]. Natural Language Processing is a key element in this process and represents a sub-area of computer science and linguistics, where interaction between computers and natural language takes place. The sentiment analysis is a common use case of text classification. The aim is to classify an attitude (sentiment) in a text as positive or negative.

According to Atkins, the extracted information from news sources is better for predicting the direction of stock volatility movements than the direction of price movements. Using a Latent Dirichlet Allocation (LDA) followed by a Naïve Bayes classification model, the accuracy achieved an increased value for predicting the stock's volatility of 56% [8].

Another approach predicts the daily stock price changes by using a deep neural generative model (DGM) of news articles by creating a market simulation [9].

Besides, in the study of Y. Xie text mining and support vector machine was chosen to forecast the Chinese stock market. First, Chinese online news was analyzed by text mining technology and sentiment analysis as a basis for predicting the stock price by using a support vector machine (SVM). However, this approach was better for predicting a specific stock price than predicting the trend of stock [10].

Another relevant approach was shown by Evans. This study detected cashtags ($) on twitter and, by using classification models, classifies if a tweet relates to a stock exchange-listed company. They used twitter as a data source by claiming that investors share information and discussions about the stock market [11].

Based on the documented literature research and analysis, similar approaches can be identified, but with different thematic focuses. Similar to other studies is that the approach of this study is based on the thesis that news headlines have an impact on the stock price change [8], [9], [11]. However, exploring the relationship between tags (individual key words) of headlines and stock price changes to design a model of linking a headline to a stock price change is not tackled in the literature results presented.

Atkins's approach of using LDA to model the semantic of the news is similar but used Naïve Bayes as a classification model, which shows a simplistic assumption. Therefore, the learnings of this analyzed study are being integrated but beyond that, the relationship between tags (individual key words) of headlines and price changes will be explored so it will lead to design a model, linking a headline to a stock price change. This study explores, therefore, the use of neural networks. Especially in speech or image recognition neural networks are applied since solutions in these fields are difficult to achieve with logical programming. To define a neural network, it can be stated that a network consists of at least one or more parallel working neurons. These neurons send information in directional connections using activation signals while each neuron receives one or more inputs and returns an output [12].

III. METHODOLOGY

The aim of this study is to determine the relationship between published news, the publication time, and the changes in close stock prices. To accomplish this, a deep learning artificial neural network will be produced to automate the classification of tags and close price changes by predicting the price change with a random split of 60% as training set, 20% as validation set, and 20% as a testing set.

A. Data Sources

The data sources differ between news and stock prices. For collecting news data, the open source API of "News" with different client libraries are chosen [13]. This API can be used for receiving articles from many news sources and filtering with different criteria, like a keyword. For this study, the Python client library is used to implement the News API into a Python application [14]. For stock prices, the API of Dukascopy is used, which provides historical stock prices and

more [15]. The result is historical news and stock prices for the year 2019.

The collected data fit the Big Data characteristics by fulfilling the five "Vs" of Big Data [16]. The volume is limited to the time frame but can be extended to live data or a bigger time frame. The velocity can be identified as fast because considering the data via the open source API's can be generated in less than a minute. The variety is limited to JSON data, so the challenges for integration, transforming or processing the data do not cause a serious challenge. The quality of the data is very high despite the sources are publicly accessible. Nevertheless, some invalid and noisy data are removed during the process of data Validation and Cleansing. This study focuses on the close price, therefore the prices for open, high, and low are removed. As a result, the data show high veracity and is analyzed effectively and quickly, so the value of the data, defined by the usefulness of the data, is very high.

B. Procedure

First, the news must be transformed to serve as an input for the deep learning network. A possible solution is to manually annotate the sentence components of the news. Words can be represented as vectors of real numbers for the calculations in deep neural networks. This type of representation is also called *Word Embeddings*. This is, from a mathematical point of view, an embedding, a mapping, in which an object can be seen as a part of another object. Word Embeddings make it possible to transfer semantic meanings into a geometric space [17]. These semantic relationships are shown in the following illustrated example by assuming that v(x) is the corresponding vector for the word x [18]:

$$v(Germany) - v(Berlin) \approx v(France) - v(Paris) \quad (1)$$

Another approach is to automatically tag news with corresponding stock symbols. A tag refers to the marking of content. During this process, additional information is assigned to the news. Therefore, tagging gives information on a specific keyword, which makes the users search simple. Both approaches result in input data for the deep neural network, which explores a close price change. This study used the approach of tagging as followed:

First, each headline was transformed by replacing special characters, the text was cleaned, and common words like "and" were removed. Afterwards, a term-document matrix was built with the tag and the frequency (the tag "update" was found 162 times for headlines with reference to "Apple"). Then, the average price differences from minute 1 to 15 were calculated for each tag and with the z-Score function *Z(x)* normalized, so outlies are processed:

$$Z(x) = \frac{x - \mu(x)}{\sigma(x)} \quad (2)$$

Also, the counts are normalized by using the z-Score function. Then, the normalized counts and normalized price differences for each tag were summarized by using a weighted average. The following table shows the process of the transformation for the tag "update" for headlines with reference to Apple:

Tag	Default		z-Score normalization		Weighted average
	Count	*Avg. price diff.*	*Count*	*Avg. price diff.*	
update	162	0.0675	22.312	0.00132	0.0029

Figure 1: Process of input transformation

Lastly, the top three tags by the weighted average for each headline are extracted. These three values, the current close price and the volume at the pub date of the news are the input for the deep neural network. The dependent variable is the actual price change which is concerning the correct comparison also transformed with the z-Score function *Z(x)*. The following table summarizes the input for the deep neural network for one unique headline example:

Input					Output
Close price	*Volume*	*Tag 1*	*Tag 2*	*Tag 3*	*Price difference*
150.163	822,766	-0.1407	0.1759	-0.0137	-0.0678

Figure 2: Input variables for the deep neural network

This first row of this table represents a headline with the current close price, the volume and the top three tags with the weighted and normalized price difference. The last column represents the output variable as the normalized actual price difference.

Since this study focuses on stock price changes at the minute-level, all published news between 4 pm and 9:30 am are excluded because there is no market reaction when the market is closed. Also, all stocks are limited to the symbol "AAPL" and the news are filtered for Apple. This restriction provides that the selected approach can be evaluated in this study and will be developed and tested in a pilot project in a further research design with more stock symbols.

As a deep neural network, a multi-layer feedforward neural network is proposed to implement the defined inputs and to predict the close price change. The information in this network moves forward in one direction from the input nodes, through the hidden layers to the output nodes. This deep learning neural network has five input layers and one output layer (compare figure 2).

Concerning the deep neural network architecture and strategy, this study presents two approaches. The first approach is looking at the error between the predicted close price change and the actual close price change and is, therefore, about error minimizing. To detect the best architecture of the deep neural network in this approach, the

permutation of the hidden layers and the number of epochs are tested. In detail, up to 3 hidden layers with a permutation of 5, 10, 20, 40, and 80 neurons and with different number of epochs (10, 100, 1000 up to 10,000) are created. As a result, this study will compare the *Mean Squared Error (MSE)* of $n(x) = 500$ different network architectures:

$$n(x) = \Sigma\, Neurons^{\Sigma\, Layers} * \Sigma\, epochs \qquad (3)$$

The second approach is about a deep neural network classification model. The model predicts a close price change for the incoming headline. But in this approach, it is explored to distinguish if the predicted close price change is positive (1) or negative (0). Therefore, the prediction and the actual close price changes are transformed into categories and the accuracy is the metric which is used as a binary classification model, by dividing the errors in relation to the test data set into four categories: True positive (TP); True negative (TN); False positive (FP); and False negative (FN). This metric will be compared through the different architectures of the neural network.

IV. Results

This chapter summarizes the results of this study. Firstly, it will show the results of the data processing part and then the performances of the models. After the news were filtered for Apple content, the top three most frequent tags are "update" (162), "iphone" (140), and "china" (108). All in all, 2,927 tags were found for 1,221 headlines regarding Apple.

In the next step, the relationship between these tags and the corresponding average close price difference are analyzed. Concerning the timeframe, all close price differences from minute 1 to minute 15 were calculated for each tag. This is equivalent to 52,686 calculations for all tags and all minutes. The tag "update" shows an average close price difference for minute 1 of $ 0.0132 and for minute 15 an average close price difference of $ 0.0623. This means that the average close price change of the tag "update" is increasing from minute 1 to minute 15 by $ 0.0491.

The next process step is to analyze the relationship between the time frames, the price differences, and the tags. First, the question, which time frame shows the highest average price difference, can be answered with 3 minutes after the news headlines are published. Also, there is a decreasing trend after 3 minutes concerning the average price difference. But there is no statistically difference between the different minutes from 1 to 15. The following graph summarizes these results and shows the decreasing trend after 3 minutes.

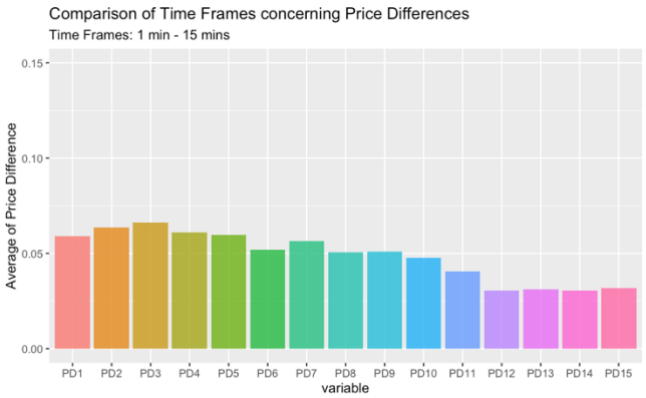

Figure 3: Time frames and average close price difference

Although, there is no statistically difference between the time frame, a time frame must be chosen concerning the input data for the deep neural network. So, this study will focus on the average close price differences after 3 minutes because at that minute the average close price difference is in comparison to the other time frames the highest.

The relationship between the unique tags and average close price difference shows that the tag "get" with an count of 12 and an average close price difference after 3 minutes of $ 9.2961 is the highest in positive terms. The tag with the highest negative value of the average close price difference is "quarterly" with 14 counts and $ -2.8677 after 3 minutes. So, news headlines which have the word "quarterly" in its headline can decrease the close price for Apple. The following graph shows the top 15 tags concerning the average close price change after 3 minutes:

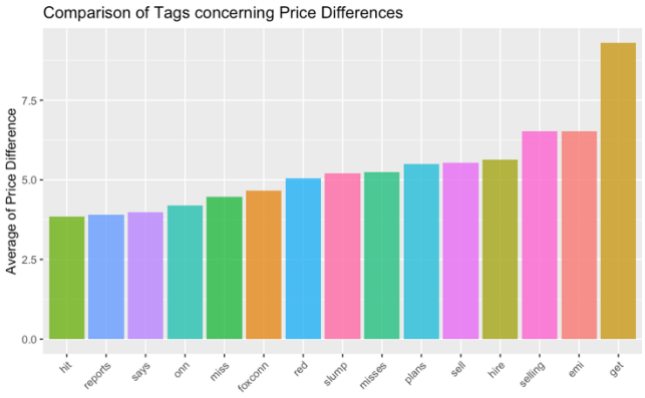

Figure 4: Top 15 tags by the average close price change after 3 minutes

After the counts and the average price differences for each tag were normalized (z-Score), the total range is between 19.2338 and -5.9115. The first tag for each headline shows in comparison to the second and third tag a wider range (compare the following boxplot graph):

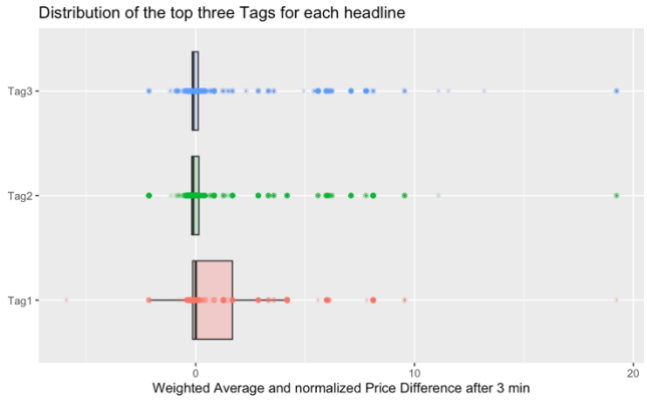

Figure 5: Distribution of the top three tags for each headline

The first model approach is about to minimize the error by predicting the close price change after 3 minutes. Therefore, the model has 5 input layers (current stock price, current Volume, Weighted average and normalized price difference after 3 minutes for the top 3 tags of each headline). The output is the actual normalized stock price change after 3 minutes. Concerning this approach different numbers of epochs are tested but there is no statistically difference between the different numbers of epochs concerning the MSE. However, after 100 epochs, the average MSE is the smallest:

epochs	**10**	**100**	**1,000**	**10,000**
Avg. MSE	0.8274	0.8105	0.8305	0.8295

Figure 6: Average MSE for different epochs

But by just looking at the epochs, the minimal MSE is not represented. So, the minimum MSE value is 0.0394 for a deep neural network architecture with 80 neurons for hidden layer 1, 80 neurons for hidden layer 2, 10 neurons for hidden layer 3, and 10,000 epochs. The following figure shows the top 5 architectures concerning the smallest MSE:

Hidden Layer			**epochs**	**MSE**
1	**2**	**3**		
80	80	10	10,000	0.03937635
20	80	5	10,000	0.24899562
80	40	80	10,000	0.26936258
40	80	5	10,000	0.27109301
40	10	40	10,000	0.27588953

Figure 7: Top 5 neural network architectures for MSE

The first architecture shows an MSE value which is 84.2% smaller than MSE value in the second architecture. The average percentage change of MSE from the second to the fifth place is 3.3% and, therefore, 80.9% smaller than the change rate from the second to the first. Also, the epochs are still 10,000 which could lead to the statement that for a small MSE the epochs should be set to 10,000 even there is no statistically difference between the average MSE of the different number of epochs.

The second explored approach in this study is about accuracy. Therefore, the predicted price changes and the actual price changes are transformed in the categories positive price change (1) and negative price change (2). In this approach, different numbers of epochs are also tested but, again, no statistically difference between the different numbers of epochs concerning the accuracy. However, after 10,000 epochs, the accuracy (ACC) is the highest:

epochs	**10**	**100**	**1,000**	**10,000**
Avg. ACC	0.5056	0.5170	0.5098	0.5176

Figure 8: Average accuracy for different epochs

But, just by looking at the epochs, the maximal accuracy is not represented. So, the maximum accuracy value is 0.6522 for a deep neural network architecture with 5 neurons for hidden layer 1, 80 neurons for hidden layer 2, 5 neurons for hidden layer 3, and 10,000 epochs. The following figure shows the top 5 architectures concerning the highest accuracy:

Hidden Layer			**epochs**	**Accuracy**
1	**2**	**3**		
5	80	5	10,000	0.65217391
40	80	5	10	0.62318841
80	80	80	10	0.62318841
40	80	5	100	0.62318841
40	20	5	10	0.60869565

Figure 9: Top 5 neural network architectures for accuracy

By analyzing the two tables for MSE and accuracy, it is obvious that the same number of epochs (10,000) can be detected for the MSE and accuracy strategy. Also, the deep neural network architecture shows for the minimal MSE and the highest accuracy similarities (same number of neurons for hidden layer 2). That can lead to the hypothesis that the architecture with 3 layers has 80 neurons for the second layer and can lead to the optimal architecture for a deep neural network by implementing the limitations of this study.

V. Discussion

This study explores two approaches for measuring the deep neural network architecture concerning the MSE strategy and the accuracy strategy. But this study is limited to the "AAPL" stock symbol which means that just headlines and stock prices regarding Apple are included in this study. As mentioned, this restriction provides that the selected approaches used in this study will be developed and tested in a pilot project in further research designs with more stock symbols.

Also, the number of layers is limited to 3 which was made for the purpose that this study is a starting point for deeper analysis which can implement more hidden layers to detect the optimal architecture of the deep neural network in order to minimize the MSE or maximize the accuracy. The models

used in this study can be used on normal machines with appropriate computational power.

The approach used in this study to link tags to close price changes is rarely new and shows promising results for wider use of stock symbols. This used approach needs to filter all headlines for different stock symbols and creates for each stock symbol input data because the tag “update” will have another average price change for different stock symbols. Also, the time frame can be different, meaning that for Apple 3 minutes were selected as the optimal time frame concerning the average price difference. But other stock symbols may have another optimal time frame concerning the average close price change.

REFERENCES

[1] N. Popper, "A little birdie told me: Playing the market on trump tweets.," The New York Times, 2017.

[2] J. M. vom Brocke, A. Simons, B. Niehaves, K. Riemer, R. Plattfaut, and A. Cleven, "Reconstructing the Giant : on the Importance of Rigour in Documenting," *17th Eur. Conf. Inf. Syst.*, pp. 1–13, 2013.

[3] H. M. Cooper, "Organizing knowledge syntheses: A taxonomy of literature reviews," *Knowl. Soc.*, 1988.

[4] O. B. Sezer, M. Ozbayoglu, and E. Dogdu, "A Deep Neural-Network Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters," *Procedia Comput. Sci.*, vol. 114, no. 2016, pp. 473–480, 2017.

[5] A. Jayanth Balaji, D. S. Harish Ram, and B. B. Nair, "Applicability of deep learning models for stock price forecasting an empirical study on bankex data," *Procedia Comput. Sci.*, vol. 143, pp. 947–953, 2018.

[6] A. Joseph, M. Larrain, and C. Turner, "Daily Stock Returns Characteristics and Forecastability," *Procedia Comput. Sci.*, vol. 114, pp. 481–490, 2017.

[7] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, 2010.

[8] A. Atkins, M. Niranjan, and E. Gerding, "Financial news predicts stock market volatility better than close price," *J. Financ. Data Sci.*, vol. 4, no. 2, pp. 120–137, 2018.

[9] T. Matsubara, R. Akita, and K. Uehara, "Stock price prediction by deep neural generative model of news articles," *IEICE Trans. Inf. Syst.*, vol. E101D, no. 4, pp. 901–908, 2018.

[10] Y. Xie, "Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method," *J. Comput.*, vol. 12, no. 6, pp. 500–510, 2017.

[11] L. Evans, M. Owda, K. Crockett, and A. F. Vilas, "A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach," *Expert Syst. Appl.*, vol. 127, pp. 353–369, 2019.

[12] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*. 2015.

[13] "News API Documentation," 2020. [Online]. Available: https://newsapi.org/docs. [Accessed: 09-Apr-2020].

[14] M. Lisivick, "News API - python," 2019. [Online]. Available: https://github.com/mattlisiv/newsapi-python. [Accessed: 09-Apr-2020].

[15] Dukascopy, "Dukascopy," 2020. [Online]. Available: https://www.dukascopy.com/trading-tools/widgets/quotes/historical_data_feed. [Accessed: 09-Apr-2020].

[16] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals Concepts, Drivers & Techniques*. 2016.

[17] F. Chollet, "Using pre-trained word embeddings in a keras model.," 2016. [Online]. Available: https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html. [Accessed: 09-Apr-2020].

[18] D. Selivanov, "Linguistic regularities," 2020. [Online]. Available: https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html. [Accessed: 09-Apr-2020].