Classification of news headlines with impact on the probability of stock prices changes

Maximilian Franke, Sean Mondesire

St. Thomas University

Miami, FL, USA

mfranke@stu.edu, smondesire@stu.edu

*Abstract*—Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# Introduction

News is part of everyday life. The channels have been extended from traditional news services to social networks. The 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 of 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 the probability of stock price changes?

The solution of 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 price changes will be explored. This will lead to design a probabilistic 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 in order to predict the probability of a 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 their probability of a stock price change, the decision for buying or selling a position 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 three subproblems:

First, it will be analyzed when [min / h / d] the stock price is affected, after the news being published. Second, it is critical to identify first the wave of the stock price change, after or even before news being published, and then to analyze the development. The aim is to detect the time of the first increase, the maximum, the time of returning to “normality” and if there is a price change before the general public is informed.

Second, it will be developed a classification model that classifies the impact on the probability of a stock price change by using deep learning techniques.

# 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, developed by vom Brocke [2]. The procedure, which comprises 5 phases, is transformed into 3 major phases in terms of result orientation, focusing on identifying the research gap. The transformed procedure of literature analysis used in this thesis is summarized in the following figure.

Fig. 1: Transformed framework for literature reviewing

1. Phase 1

First, the scope of review is defined, in a structured and documented form. The requirement of a structured definition of review scope is fulfilled by the taxonomy according to Cooper[3]. This is divided into five properties: (1) focus; (2) goal; (3) organization; (4) target group; (5) framework. 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 in order 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 a complete research in German and English language.

Second, the conceptualization of the topic is performed via a concept map [2] with the topic’s stocks and model, in order to identify key terms and understand the subject matter. As a result, a list of keywords for the research is created, focusing on stocks and prediction which is summarized in the following search word list:

|  | Search word list | |
| --- | --- | --- |
| Related Stocks | Related method |
| 1 | Headlines / news | Classification |
| 2 | stocks | prediction |

1. Phase 2

The literature search is carried out in the publication databases Karlsruher Virtueller Katalog, ScienceDirect, and Web of Science in the stages defined by the search word list[4]. The results are verified or falsified in a first iteration by restricting the abstract, title and keyword list with respect to research relevance. In a second iteration, in the form of a full text analysis, the selected results are examined more closely and relevant articles for the classification of news headlines are included in the analysis so that a total of 9 relevant articles are identified. 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 stocks prediction. According to Atkins the extracted information from news sources is better for predicting the direction of stocks volatility movements than the direction of price movements. Using a Latent Dirichlet Allocation (LDA) followed by naïve Bayes classification model, the accuracy achieved an increased value for predicting the stocks volatility of 56% [5].

Another approach predicts the daily stock price changes by using a deep neural generative model (DGM) of news articles. By creating a market simulation, the result showed that the proposed model achieved the highest profit [6].

In addition, 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 support vector machine (SVM). However, this approach was better for predicting a specific stock price, but less accurate for predicting the trend of a stock [7].

Approaches using indicators as predictors to build classification models are also widely used in this research area. Sezer’s study, for example, determines 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 [8]. 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 [9]. In addition, another study has determined that there is a high correlation between stock market prices and returns [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 an data source by claiming that investors share information and discussions about the stock market [11].

1. Phase 3

In this section, the results of the literature analysis are combined with the problem selection to generate a discussion of a research agenda, which is the starting point for further action.

Firstly, it should be noted that a wide range of approaches and solutions have been identified with the main topics according to the search word list *(see Appendix)*. 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 is based on the thesis that news headlines have an impact on the stock price change [5], [6], [11]. However, exploring the relationship between tags (individual key words) of headlines and price changes in order to design a probabilistic model of linking a headline to a price change is not tackled in the literature results presented.

Databases

KVK, ScienceDirect, ScienceDirect, Web of Science

N = 134

N = 505

N = 60

1st iteration step (abstract, title, keyword list)

N = 9

2nd iteration step (full text analysis)

N = 184

N = 187

# Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# Results and discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

##### Acknowledgment

##### References

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

[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] C. Leyh, T. Schäffer, and S. Forstenhäusler, “SIMMI 4.0-Vorschlag eines Reifegradmodells zur Klassifikation der unternehmensweiten Anwendungssystemlandschaft mit Fokus Industrie 4.0,” *Multikonferenz Wirtschaftsinformatik, MKWI 2016*, vol. 2, no. March, pp. 981–992, 2016.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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