UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Stock return predictability based on world news sentiment

Author:
Ostap KHARYSH

Supervisor:
Dr. Yarema OKHRIN

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2019

Declaration of Authorship

I, Ostap KHARYSH, declare that this thesis titled, "Stock return predictability based on world news sentiment" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"The stock market is filled with individuals who know the price of everything, but the value of nothing."

Phillip Fisher

"Every once in a while, the market does something so stupid it takes your breath away"

Jim Cramer

"In the 20th century, the United States endured two world wars and other traumatic and expensive military conflicts; the Depression; a dozen or so recessions and financial panics; oil shocks; a flu epidemic; and the resignation of a disgraced president. Yet the Dow rose from 66 to 11,497."

Warren Buffett

"The investor's chief problem - and his worst enemy - is likely to be himself. In the end, how your investments behave is much less important than how you behave."

Benjamin Graham

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Stock return predictability based on world news sentiment

by Ostap KHARYSH

Abstract

Modern stock trading leans on various algorithms, attitudes and data resources in order to win on arbitrage. One of such is news. There is a compelling amount of investigations based on the news provided by specific agencies, but none of them cover the world news. In our research, we investigated how stock prediction could benefit from the world news. Based on the GDELT Project data we created the news preprocessing algorithms to capture the stock related news and made several investigations to explore the benefits which trader could receive in case of including this source to his/her trading strategy. We proved the efficiency of world news for stock return volatility predictions.

Acknowledgements

First of all, I would like to thank my academic supervisor Prof. Yarema Okhrin (Head of the Department of Statistics at University of Augsburg, Germany) who guided me through this research, provided me with valuable recommendations and advice which let this research come true.

Also, I would like to thank Oles Dobosevych (Ukrainian Catholic University) for his consultation on data mining and storage, Prof. Yaroslav Prytula (Ukrainian Catholic University) for the introduction to the topic of stocks, which helped me to determine the final topic of my research.

Additionally, I would like to thank the University of Augsburg that provided me with a cozy accommodation for the period of research on the university campus.

Finally, I am grateful to Ukrainian Catholic University and, particularly, to the Faculty of Applied Sciences for the great atmosphere and considerable academic schedule which made a worthwhile impact on my life.

Contents

List of Figures

List of Tables

List of Abbreviations

APARCH	Asymmetric Power ARCH
AR	Autoregression
ARCH	Autoregressive Conditional Heteroscedasticity
GARCH	Generalized ARCH
GDELT	Global Database of Events, Language, and Tone
LAD	Least Absolute Deviations
MSE	Mean Squared Error
OLS	Ordinary Least Squares (Linear regression)
RW(T)	Random Walk (Theory)
VAR	Vector Autoregression

Dedicated to my water polo coach

Chapter 1

Introduction

1.1 Motivation

The stock market was always a point of discussion. The notion of trading drives many financiers and economist to clash over different behavioral hypothesis related to it. There is a constant massive interest in how the market behaves under different circumstances such as recessions, economic crises, wars, sanctions, and other events. At the early beginning of the 20th century, the topic of market sentiment emerged and produced a new discussion break-point between those who remained skeptical and those who found it as an essential concept for the future of markets. With the development of IT and exploring Big Data as a tool, there appeared much broader ways to investigate what exact impact on the market price prediction could a wide variety of data resources have and how those could change the attitude of investors participating in the market. It resulted in a shifting of algorithmic trading paradigms, which volume increases dramatically each year (see figure 1.1). This type of trading attracts those who seek to win on arbitrage applying a considerable amount of different techniques with a combination of numerous data sources. One of the most exciting parts of algo-trading is to "win a market," meaning to be as much precise as possible in predicting the future stock return in order to receive a potential reward if such opportunity occurs. Nowadays, traders are trying to scrutinize every possible source of information to increase the accuracy of their predictions. Among the most interesting are traders' psychologically-related data sources. It could be a combination of social networks activity, different reactions on public news, general moods, and others.

In this work, we want to validate what advantage of the world news sentiment could investor have, how distinct types of sentiment data correlate with stock return fluctuation and what are the significant benefits of including such data in the field of stock research.

1.2 Notion of Markets

1.2.1 Evolution

The market started its history hundreds of years ago from the simple agreements to engage in commercial transactions, and rights to provide these transactions. In medieval ages, the notion of future agreements took part in most of goods deliveries. Throughout the centuries, this kind of trading relations transformed into the new form when the Chicago Board of Trade was established in 1849. It was the first market for futures trading. Starting from the agricultural industry in the USA this type of trading spread all over the world. In the 20th century, largely promoted by investment banking firms, futures market reached its mature state. The subsequent

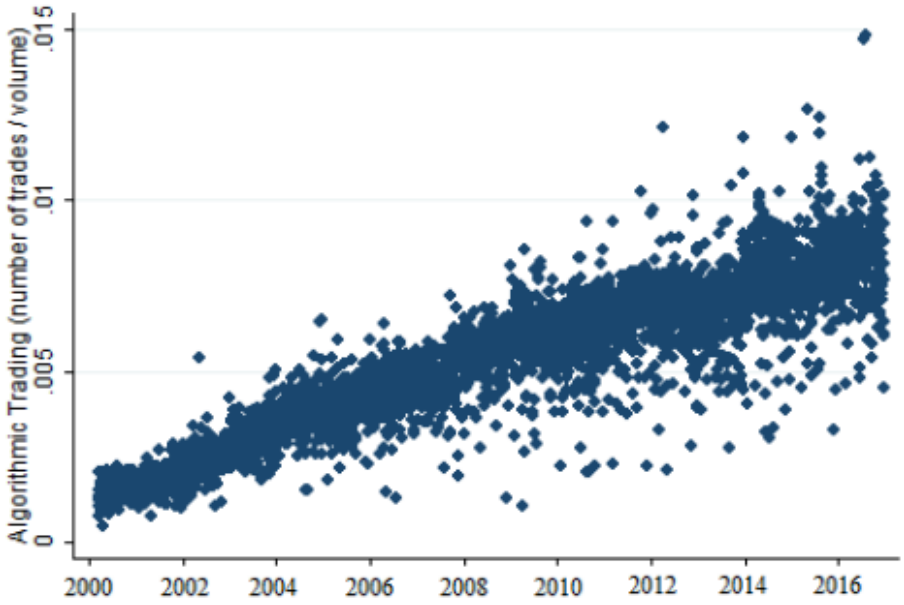

FIGURE 1.1: No. of trades over volume of algorithmic trading (Verheggen, 2017)

modification of trade came with the introduction of the Internet that provided the opportunity of electronic trading, which appeared in shifting of trading paradigms from day-specific to second-specific, from human trader to trading bot.

In the beginning of 21st century, the rapid development of data science, backed by numerous data resources, has lead to a crucial change in market perception. This point further, data has been considered as a critical strategic asset for a market industry. Processing of the world information has opened the horizons of modern trading opportunities, portfolio optimizations, and stock prediction algorithms. Simulations and optimization of trading with real-time data proved that innovative ways of the alternative data usage could provide significant results and contributed to the Malkiel and Fama, 1970 Efficient Market Hypothesis (EMH).

1.2.2 Efficient Market Hypothesis

Efficient Market Hypothesis (EMH) Malkiel and Fama, 1970 a Nobel prize winning investment theory which states that a stock share price reflects the "efficient" (root-causal) to it information and is controversial to the belief that stock price change is an unpredictable, Random Walk process. Taking into account this theory, once particular information about a specific stock occurs, it should affect its price. Broadening the essence of this theory, the more information traders gather about a specific market entity, the better fit has their prediction of its future behavior, as stocks are always traded at their fair value, with no possibility of an investor to buy an overvalued or undervalued stock. Nevertheless, not all the stocks are the same. No one could be sure that assumptions made on stocks are based on the whole available information. That is why Eugene Fama divided them into three forms: weak (only past information reflects the stock price; impossible to predict the future return), semi-strong (public information reflects the stock price; analysis could predict the return), strong (both insider and public information is incorporated into the stock price; no insiders could make profit from a stock).

That drives to the conclusion: as the full market information could never be publicly available, rational market as stated in EMH is more an utopia, then a reality. Assets still could be over-valued and under-valued due to unequal information access combined with investor actions lead by fear, confidence and risk-taking ability Baker and Wurgler, 2007.

1.2.3 Noise and Rational traders

A significant amount of time was dedicated to exploring financial market behavior and its patterns. The researches conducted in this field made the academic society branch out into two camps: those who believe that market share price could be predicted, and those who state it is the Random Walk process. Despite the conservative views, there are organizations and private investors implementing various techniques, applying numerous data sources to gain significant prediction results. Along with such stratums of trading society, there are also those who take trading as a hobby, amateurs, who seek for patterns to win the arbitrage, or rely on the third "expert" sources and act as a crowd. All these players acting on the market, regardless of their gain or loss, influence the market in a specific manner. These actions contribute to stock price volatility growth and were discussed by De Long et al., 1990 which argues that the number of financial market anomalies such as high volatility, undervaluations and overvaluations of assets, could be explained by the notion of "noise trader."

Noise trader Noise trader is a type of investors who believe that he/she has a significant information about the future price of an asset by relying on various third party informers. When a noise trader has overvalued beliefs of a specific asset, he/she tends to be riskier in order to profit more. Noise traders donate to keeping the asset price in imbalance, not allowing arbitrage to converge to fundamental value. Bullish acts of this type of trader contribute to the share price growth and vice versa. The higher percentage of noise investors versus rational investors acting on the market, the higher is the volatility level.

Rational trader The behavior of rational investor requires more in-depth investigations, mostly, for the long-term trading strategies. To bet against noise investors on the market, he/she uses different sentiment resources such as financial showings (volatility, volume, price patterns) indexes Mao, Counts, and Bollen, 2011, news, and social media sentiment. These allow him/her to assume how the noise traders could behave and how to design the investment strategy under such circumstances. Rational investor could try to guess the belief of noise investors about certain assets rather than to wait for long-term arbitrage perspective.

1.3 S&P 500

Most of the big public companies are presented on large stocks, such as NYSE (New York Stock Exchange). S&P 500 is one of NYSE indexes. It tracks the stocks of 500 large-cap companies (> 6.1 billions of market capitalization) which have 50% or more headquarters in the United States, and at least 50% of stocks are public. The total market capitalization of S&P 500 is approximately equal to 80% of the market cap of

the whole stock market. It is commonly used as a benchmark for measuring the success of portfolio management as it reports the risks and returns of major American market players.

In comparison to S&P 500, Dow Jones Industrial Average tracks only 30 companies with a combined capitalization of 25% of American stock market, whereas NASDAQ tracks a bigger amount of tech companies and privately owned.

S&P 500 index could be divided into several sectors. The largest sector is: "Information Technology" having around 25% of capitalization. In our research we will focus on the top 5 representatives of this sector by the stock weight (April 2019) (see table 1.1).

TOP 5 S&P 500 TECH AFFILIATES			
Affiliates	**Average Volume**	**Free Float**	**Market cap**
Apple Inc.	28.74m	4.68bn	986.20bn
Microsoft Corporation	24.96m	7.55bn	967.12bn
Amazon.com Inc.	4.22m	412.94m	935.83bn
Facebook Inc. Class A	21.89m	2.37bn	549.57bn
Alphabet Inc. Class C	2.33m	643.61m	808.48bn

TABLE 1.1: Stock information of top tech companies in S&P 500 (December - April) by Financial Times

Considering the trading volume and sufficient amount of shares in free float, these companies have a high interest among traders.

Chapter 2

Related Works

In this chapter, we describe the works that investigated the causality between sentiments and market behavior and how some approaches proved or refuted the different sentiment type effect or correlation on market change. There are some success and failure stories based on distinctive data sources such as market indexes, news agencies, social networks, investor surveys.

2.1 Sentiment and Market interaction

With their research Barberis, Shleifer, and Vishny, 1998, presented a model of investor sentiment. They supported the theory that stocks appearing in positive news and have high past returns are overvalued, and those associated with bad news were undervalued which proves that investors tend to underreact. Moreover, the research found an evidence that with the increasing frequency of similar news shocks the average returns become negative, which leads to overreaction. This paper centers a point of discussion on investor psychology which is a significant cause-factor for irrational events happening on the market.

In their study, Fisher and Statman, 2000 examined large (Wall Street strategists), medium (investment news writers) and small (individuals) investors. They proved three groups having different behavior on the markets. Large investors act completely different comparing to the other two groups. The sentiment of Wall Street strategists and individual investors have a negative statistically significant correlation with S&P 500 returns, and no statistically significant correlation between investment newsletters writers and S&P 500 index. However, a combination of these three groups could be used for asset allocation strategies on S&P 500 market. Another interesting discovery was that small and medium investors are affected by high S&P 500 market returns which make them bullish. Controversially to the initial belief, small investors acts had no correlations with small-cap stocks and yet appeared to have with large-cap. This study broadened the investigation of Barberis, Shleifer, and Vishny, 1998 and concluded that exploring "investor" at general is not enough, because market players favor their specific distinctive to other strategies.

Alternatively, Baker and Stein, 2004 explores the connection of liquidity and expected market returns. This theoretic evidence encourages considering market liquidity for stock returns predictions. There is a connection between trading costs along with activities as a sentiment indicator on markets because it shows a high correlation between a year turnover and share of equity in total external finance.

To contribute to famous theory of "noise" traders De Long et al., 1990 and Black, 1986 , Brown and Cliff, 2004 proved the existence of speculator (who has a bias expectation on a future asset value) and fundamentalist (who has unbiased expectation on a future asset value). They used market aggregates instead of stocks to find the effect of the sentiment on to the whole market and two surveys as investor

sentiments: *American Association of Individual Investors* (AAII) and *Investor Intelligence* (II) from which monthly and weekly *bull-bear spreads* were calculated. They created three experimental groups to explore the sentiment influence on market indexes: 1) Market performance (based the "freshest" market data), 2) Trading activity type, 3) Derivatives variables (outcomes of trading). Finally, the correlation of market sentiments and market return appeared to be strong. They used PCA and *Kalman Filter* to extract unobserved sentiment features. They showed prominent results in having a correlation between market derivatives in both weekly and monthly measures. *VAR* model for monthly data showed that sentiment could predict itself, but is inefficient for the prediction of large-cap stocks. A different situation could be observed with *VAR* model for weekly data where individual sentiment is correlated with itself and with large market derivatives. Also, it is intriguing that institutional sentiment could predict the individuals' sentiment. All in all, Brown and Cliff, 2004 concluded that with their approach they found no significant evidence that sentiment could predict market returns.

As we could experience from Brown and Cliff, 2004, prediction of market derivatives could be inefficient using investor sentiments. So Qiu and Welch, 2004 explored what sentiment proxies could fit the market return predictions: Closed-end fund discount (CEFD), and their Consumer Confidence (CC) (relies on investment questions from surveys). The research found a high, statistically significant correlation between Michigan Consumer Index and USB/Gallup investor sentiment but CEFD has not, and USB/Gallup investor sentiment showed that wealthy investors are more optimistic than poor investors. Their validation over UBS/Gallup investor sentiment survey data suggest: 1) (CEFD) could not be used as a proxy for investor sentiment, but CC can, 2) CEFD does not correlate with small-cap firms, but CC does. Finally, they suggest CC over CEFD showing that relying less on investor sentiment and financial data and more on identifying proxies from certain survey questions could make a significant correlation with market returns.

In contrast to Brown and Cliff, 2004, Brown and Cliff, 2005 used surveys as investor sentiment. They helps to predict the market return for 1-3 years but have a little predictive power for the near-term return. They found that: 1) the market is overvalued when it appears in the atmosphere of optimism, 2) the sentiment is positively related to the market valuations, 3) bullish (bearish) shock to sentiment results in underperformance (overperformance) of the market for some period. By these three facts, they concluded strong support for the hypothesis that asset values are affected by investor sentiment. This finding supports the theory that irrational sentiments of investors do affect asset price level and proves the importance of surveys for market predictions argued inQiu and Welch, 2004

Along with Brown and Cliff, 2005, Prytula, 2005 attempted to estimate the noise traders' information misperception and predict their trading behavior. This study claims, that trader's habits to collect information for their trading strategies did not change through time and, likely, broadened to the bigger selection of business and financial news. The research found that Autumn is the most active for the news issuing season, regardless of the size of the companies discussed. The study proves that on regular basis noise traders do not significantly affect the stock market, and during the period of market fall noise traders' misperception becomes more volatile.

Kumar and Lee, 2006 presented the research of retail investors trading behavior and its influence on market price anomalies taking into account only individual investors as *small* as it is in Fisher and Statman, 2000. They supported the notion of noise traders and rational trader who rely on attention-based trading or investment

analysis respectively. The research found that Buy-Sell imbalance (BSI) could explain stock return co-movements and could correlate with the function of arbitrage costs. To sum up, the collective actions of the noise traders could influence the stock returns and support sentiment based theory of return co-movements presented by Barberis, Shleifer, and Wurgler, 2005 as buying (selling) of a group of stocks influence the buying (selling) of others, and also creates a tendency of buying (selling) among investors.

In addition to the above evidence, Barber and Odean, 2007 explored three types of brokerage data and Plexus Group trading data and found that individual investors reflect attention-based buying behavior. They buy when: 1) the stock experiences high volume days, 2) at the period of extremely negative and extremely positive one-day returns, 3) under news highlighting the stocks.

Research of Wang, Keswani, and Taylor, 2006 is alternative and controversial to Kumar and Lee, 2006. They studied daily and weekly volatility, returns, and different investor sentiment to find a relationship between them. They used three sentiment indicators: put-call trading volume ratio (PCV), put-call open interest relation (PCO). They found support for volatility and sentiment being influenced by market returns and no support that noise traders (their sentiment) influence either return or volatility.

Another possibility to explore sentiments was presented by Baker and Wurgler, 2007. They introduced "top-down" and "bottom-up" approaches to quantify investor sentiment. The "top-down" approach focuses on aggregate sentiment to market returns and stocks. The "bottom-up" approach centers on individual investor psychology (overconfidence, conservatism, consistency) to explain how individual investors underreact or overreact to past returns. Statistical evidence of sentiment was not strong which made it challenging to distinguish *RW* from a long-lived bubble. It is stated that stocks which are most likely to be influenced by the investor sentiment are younger, smaller, volatile, unprofitable, but "bold" stocks are less likely to have such cause-effect. Principal Component Analysis appeared to be a useful tool to detect general patterns in the number of time series omitting specific fluctuations. By applying PCA, they detected reflective general demand and reflective speculative demand as the most important source changes in mutual fund flows. They were able to identify "greed" vs. "fear" and "bullish" vs. "bearish" notion. Without in-depth specification on particular traders as in Kumar and Lee, 2006 the evidence was significant to prove the effect of sentiment on the market.

Moving to index predictions, Han, 2007 stated that investor sentiment is an important influencer of S&P 500 index and option prices. To prove that investor sentiment affects S&P 500 option prices three investor proxies were introduced: 1) Investors Intelligence's index, 2) long and short positions of large "commercial" and "non-commercial" traders, 3) valuation errors of S&P 500 index. These sentiment proxies were positively correlated with each other. They argued that S&P 500 index tends to be overvalued when news writers and large speculators are bullish about a future market return, and bull-bear active spread of large investors negatively predict the market return. Sentiment proxies appeared to have a co-movement with index "anomalies." The research finalizes that if 1) investor survey indicates more market professionals to be "bearish", 2) large speculators take bigger short positions in S&P 500 futures and 3) S&P 500 index is more depressed than fundamental, risk skewness of the monthly index return becomes significantly more negative and index *volatility smile* becomes steeper. Summing up, this study proved that the index is influenced by investor sentiment; there are some specific patterns of investors trading on the market and could be used for index movement and volatility.

With the emerging number of sentiment measurements Bandopadhyaya and Jones, 2008, wanted to identify which of the most popular: *Put-Call Ratio* (PCR) and *Volatility Index* (VIX) measures estimates better the non-economic effect on the asset prices. In their study, they also evidenced that investor sentiment could explain short-term movements in asset prices better than any other fundamental approaches. Based on RW regressions of S&P 500 index for non-economic market influence, PCR appeared to be a better measure of investor sentiment then VIX.

With their pioneer investigation of international equity markets Baker, Wurgler, and Yuan, 2012, attempted to analyze global, country-specific sentiments to find how their combination could influence stock returns. They took separate and combined indexes of Germany, Canada, France, Japan, the United Kingdom, and the USA to experiment with the stocks and *Siamese twins* share prices. They formed cross-correlational portfolios on characteristics of firm size, sales growth, total risk, book-to-market equity ratio. Investor sentiment appeared to affect the international market level and cross-sectional returns with both global and local sentiments considered important. This research showed that global markets are connected, so are the companies. So in terms of stock return predictions, one has to consider not only directly related sentiments but also external sentiments.

A Machine Learning approach to enhance existing algorithms introduced by Passalis et al., 2019 recently. They developed Temporal Logistic Neural Bag-of-Features (TLo-NBoF) 2.1 for high-frequency order book data. The idea behind is to enhance the Bag-of-Features (BoF) model with neural networks in order to improve predictions of time series.

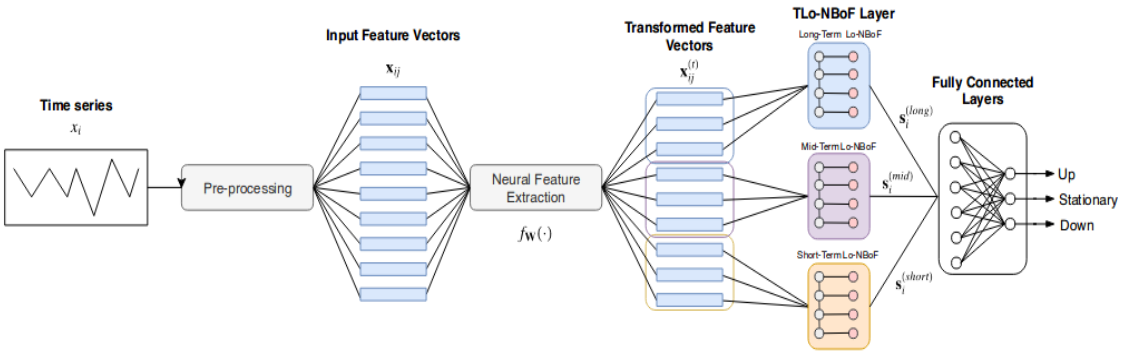

FIGURE 2.1: TLo-NBoF for Time Series forecasting (Passalis et al., 2019)

They took time series data as a separate vector for each time stamp. The feature vectors were processed by the series of neural transformation layers. Here the convolutional layers were used to extract higher-level features, to capture the relationship between previous and future vector. To convert the resulted set of transformed higher level features into a specific length representation, invariant to the length of TS input TLo-NBoF was used. The results obtained showed using BoF modification (TLo-NBoF) performed better than CNN, LSTM, GRU and other state-of-the-art methods.

This section covered important evidence of how investor and market sentiment correlate with market behavior and could be used for prediction of future market movements, and how different approaches applied on various data resources could be used to improve existing trading paradigms.

2.2 News, Twitter and Market indexes as sentiments

Moving closer to the context of the research, we want to show the existing approaches of how to benefit from the Internet as a source coverage, as well as Big Data modern sentiment could be applied for market prediction problems, and what are results of such experiments are.

2.2.1 Market indexes and Google Search sentiment

Mao, Counts, and Bollen, 2011 were the pioneers in combining various market sentiments and Google Search data to identify the investor mood and how these correlate with stocks. These were: stock log return, DSI bullish percentage, Investor Intelligence (II), Volatility (VIX), Tweet volumes of financial searches (TV-FST), Negative News Sentiment (NNS) and Google search volumes of financial search terms (GIS), Twitter Investor Sentiment (TIS) :

$$TIS_t = \frac{N_{bull}}{N_{bull} + N_{bear}}$$

Weekly analysis: GIS, based on Granger causality can be a market predictor, but surveys of investor sentiment. They evidenced that trading volume can increase direction accuracy for DJIA and VIX. The key points of this research are: 1) Granger causality analysis showed that adding GIS improves MAPE prediction error for VIX, DJIA, 2) weekly forecasting accuracy improves with the addition of GIS even if market experience high volatility. That inevitably proves that Google Search results should be considered as a market indicator.
Daily analysis: All these sentiments DSI, TIS, TV-FST, and NNS have positive statistically significant correlations and significant correlations between log returns based on Granger causality test, except DSI. The experiments showed that Twitter outperformed the survey and news indicators in terms of prediction power, and also evidenced that it outperforms GIS by volume of financial topics.

Summing up, this analysis concluded that Twitter (TIS, TV-FST), news (NNS) could predict DJIA log return outperforming surveys and news, and should be considered as an Internet source which has potential to be used as a spotter of investors' mood change and their future behavior on stocks.

2.2.2 Twitter sentiment

Zhang, Fuehres, and Gloor, 2011 offered a way to evaluate Twitter sentiment on prediction NASDAQ, Dow Jones, S&P500 indexes. Their method proposed to create the "fear" words indicating the potential flow of stock market investment decisions. They found out that people use emotional words such as "hope," "fear" or "worry" to express both positive and negative context. They created three baselines for emotions measurements: 1) number of tweets per day, 2) a number of retweets per day, 3) a number of followers per day.

The research finalized: when people express a lot of "fear," "hope" and "worry" it appears to have a causal relationship with Dow Jones Index fall. If no emotional bounce spotted - Dow rises (see figure 2.2). This pattern shows that people tend to share tweets being under negative, stressful conditions which leads to the question of how to quantify the stress and happiness power of Twitter sentiment for market predictions.

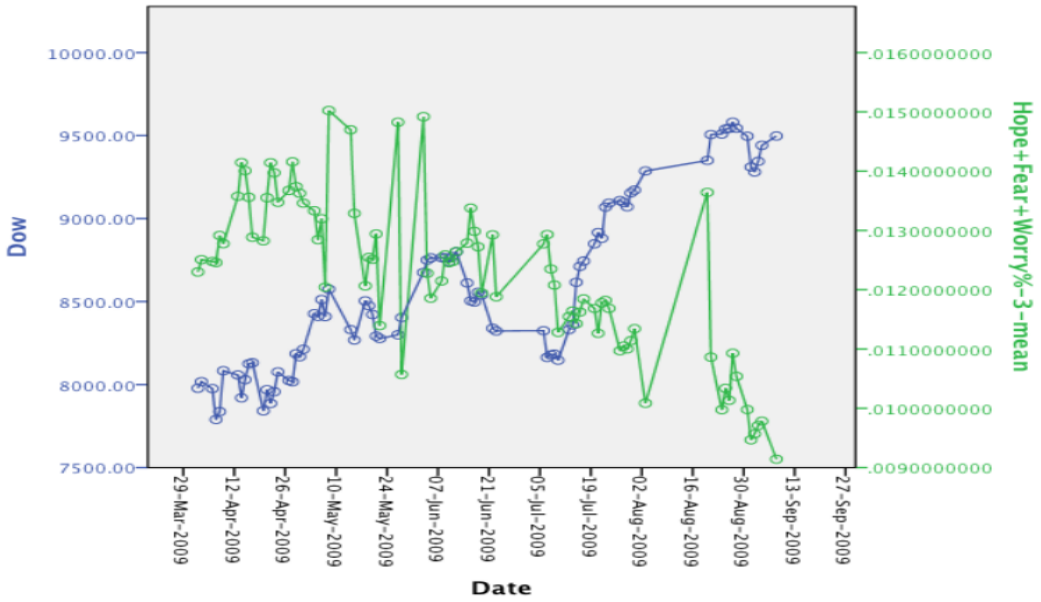

FIGURE 2.2: Correlation of hope-fear-worry and DJIA index (Zhang, Fuehres, and Gloor, 2011)

Following Zhang, Fuehres, and Gloor, 2011 Mittal and Goel, 2012 developed their mood model based on the Profile of Mood States (POMS). They dealt with missing data by implementing *concave function* because their finding argued that stock data usually follows a concave relationship if no anomalies occur. In their studies, Self Organizing Fuzzy Neural Networks (SOFNN) had higher accuracy over Linear regression in predicting actual stock values. The Granger causality showed that "calmness" and "happiness" are the best predictors for stock market movement taking 3-4 period data. An important investigation is that combining more than two different mood types (Calm, Happy, Alert, Kind) leads to overfitting. SOFNN performs the best when it contains the DJIA index and three-day Calmness and Happiness data (see figure 2.3). Finalizing the study, they argued that the filtering of text sentiment they applied contributed mainly to the accuracy of their model. Considering results obtained, selecting Twitter as a sentiment for stock predictions it is advised to filter out data that is not emotional enough and filter tweeter mood into several states using emotionality scoring.

Zhang, Fuehres, and Gloor, 2011 stated that they received promising results: their greedy portfolio management algorithm applied to DJIA allowed them to win on arbitrage after 40 days of work.

Going deeper into the specific topic of Twitter sentiment, Nisar and Yeung, 2018 tried to predict stock market movement by analyzing the political context of tweets using lexicon based analytics. They decided to choose the *concave function* proposed by Mittal and Goel, 2012 as it applies to stock data. They found a causational relationship between "mood" and "close price" but not statistically significant.

Finally, they stated that political twitter content can be used to predict stock market movements and proves that political conditions make an influence on market returns.

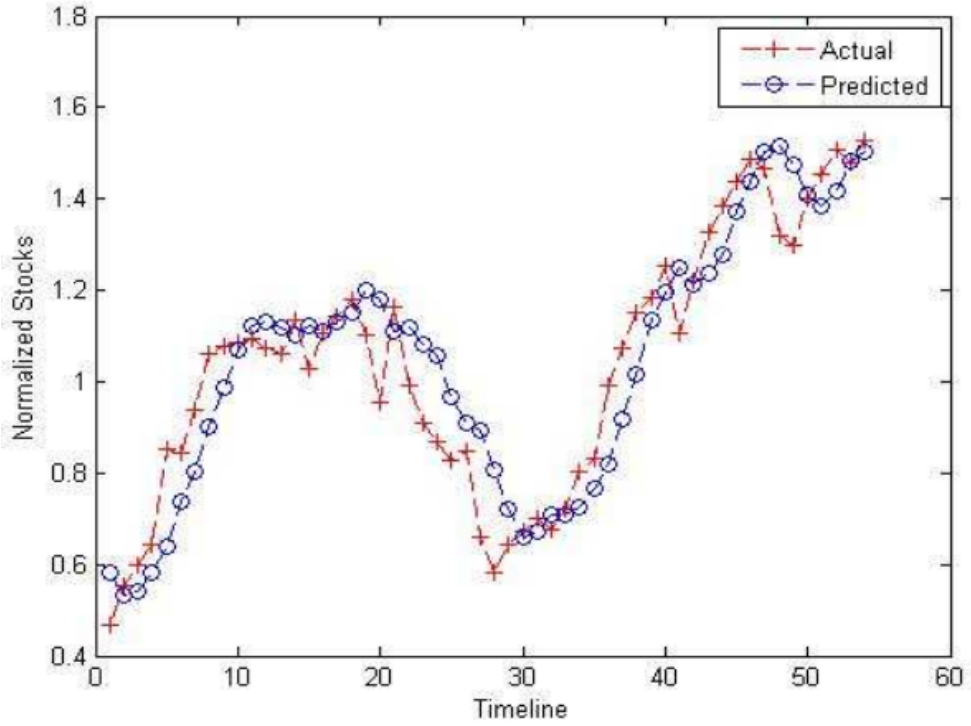

FIGURE 2.3: Predicted vs Actual Stock Values (SOFNN on Calm+Happy+DJIA for 40 days (Mittal and Goel, 2012)

2.2.3 News sentiment

Moving to news sources, Kalyani, Bharathi, and Jyothi, 2016 used Apple daily stock data and news data (titles and body) to predict the stock trend. They have split news by their positive and negative polarity and using TF-IDF algorithm defined the importance of each document, initially filtered out specific stop-words. They applied SVM, Naive Bayes, and Random Forest. Random Forest prediction accuracy for trend bounded between 88%- 89%, SVM: around 86%, Naive Bayes around 83%. By this investigation they proved that polarity based on news sentiment applies for the stock trend predictions.

Taking financial news, Atkins, Niranjan, and Gerding, 2018 proved that market volatility could be predicted better by news rather than by financial data. They used LDA to assign the news data to specific topics and using this data as features and directional data obtained from a market to train a multinomial Naive Bayes model for hourly predictions. They decided to exclude the first hour and last half an hour of trading for each day reasoning that these periods have irregular patterns that could bias the model as suggested by Schumaker and Chen, 2010. Having a minute data available for hourly predictions they split data for each hour into six equal parts and assigned weights w for them: $w_i = w_{i-1} - 0.15$, where $i = 0$ is closest to prediction period and enhanced the basic model using *Bigram model* to benefit from phrases. They applied *Chi-squared* feature reduction to select 30 most appropriate features among generated, because it had less variance then PCA. To deal with words that appear too frequently or too rarely the *Luhn cut-offs* were implemented. SVM was used as it could handle high dimensional data and position the hyperplane with a high margin between classes. To test the relationship between the news data and market data the following measures were applied: directional movement, RW movement,

technical analysis model using TA-Lib [1].

Finally, they received a strong evidence that textual news resources could predict directions of market movements even better than market price history.

It is clearly seen that both Twitter and News sentiments influence the market and could be used for market return prediction. Considering Atkins, Niranjan, and Gerding, 2018, who proved that textual news outperforms market price in terms of volatility predictions and Mao, Counts, and Bollen, 2011 who evidenced that Twitter outperforms the news sentiment as a primary source of investor mood, market investment strategists should count on such sentiment segments. Definitely, all those research outcomes strongly evidence that the influence of mass media and social network on the market at the age of Big Data are only strengthening their position so the benefits are coming with them.

2.3 Data sources

- Fisher and Statman, 2000 as a large strategists indicator data from *Merrill Lynch* was used, containing sentiment of Wall Street sell-side strategists. The survey of 15-20 investment strategists is collected monthly. For the medium investment group, the *Investor Intelligence* index was used. This survey contains results of 130 investment newsletter writers and is conducted every week. For the small investors' data from *American Association of Individual Investors* (AAII), was taken;
- Brown and Cliff, 2004 used for analysis both the monthly and weekly frequencies. Monthly data gives 406 observations on average. Weekly data consist of 596 observations on average. They used *American Association of Individual Investors* (AAII) with weekly responds in range 26 to 422 and *Investors Intelligence* with weekly bull-bear categorizations of 150 market newsletters as in Fisher and Statman, 2000 as they measure the sentiment of market players with a direct approach;
- Prytula, 2005 explored weekly data of prices and amount of news from 622 sources about 16 companies available on *Lexis-Nexis Academic Web*;
- Qiu and Welch, 2004 used monthly data from *Michigan Consumer Confidence Index* by Michigan Consumer Research Center, and UBS/GALLUP Index of Investor Optimism (survey of investors with more than $10 000 wealth) with reports on last Monday of the month;
- Kumar and Lee, 2006 focused on 62,387 individual investors over 1991 to 1996 with monthly sum portfolios values $ 2.18 billions making their study maximally centered on an individual segment of market participants;
- Barber and Odean, 2007 used three brokerage data types: small discount, large discount, full-service and stock data of professional traders presented by Plexus Group;
- Wang, Keswani, and Taylor, 2006 used daily and weekly market data at range: 1990 - 2001 and sentiment data for only weekly data: put–call trading volume ratio, put-call open interest ratio, *American Association for Individual Investors* (AAII) and *Investor Intelligence* (II);

[1] http://www. ta-lib.org/

- Baker and Wurgler, 2007 contains data of mutual funds provided by *Investment Company Institute* and sentiments: trading volume from New York Stock Exchange Turnover (NYSE), dividend premium, the closed-end fund discount, the quantity and first-day returns on IPOs and the equity share [2];
- Han, 2007 used dataset of S& P 500 provided by *Chicago Board Options Exchange* and sentiment indexes: *Investor Intelligence* and *Commodity Futures Trading Commision* (CFTC) with large traders daily position aggregated reports , Sharpe's valuation errors Sharpe, 2002 of S&P 500 index;
- Bandopadhyaya and Jones, 2008 used freely available *Put-Call Ratio* (PCR) and *Volatility Index* (VIX) using daily data from 2004 to 2006 years from *Chicago Board Options Exchange* [3];
- Baker, Wurgler, and Yuan, 2012 got the data from Datastream which covers the stocks from the largest exchange in Germany, United Kingdom, France, Japan, Canada and from NYSE, Amex, Nasdaq for the USA, data of three *Siamese twin pairs* in UK and USA;
- Passalis et al., 2019 used a large-scale high-frequency limit order book dataset of 5 Finish companies from Helsinki exchange. Data collected from 10 business days in June 2010 in a total of 4.5 million limit orders;
- Mao, Counts, and Bollen, 2011 collected Investor Intelligence and Daily Sentiment Index surveys freely available daily, Twitter posts, news from media services, Google Insights for Search (GIS) (Google Trends), Yahoo! Finance : DJIA, trading volume, Volatility Index (VIX) and gold price [4];
- Zhang, Fuehres, and Gloor, 2011 used indices of Dow Jones, S&P 500, and NASDAQ and Twitter feed on several accounts for six months in 2009, with a range of 8100-43040 tweets a day;
- Mittal and Goel, 2012 used six months of the 2009 year data of Dow Jones Industrial Average, along with 476 tweets of 17 million Twitter users applied for daily analysis and predictions;
- Nisar and Yeung, 2018 used a high-frequency Tweeter data May 4th, 2016 - May 9th, 2016 of election period in the United Kingdom. For stock calculations there was FTSE100 opening and closing values from Yahoo! Finance obtained. Due to limitations of the dataset, there was some missing data (weekends, holidays);
- Kalyani, Bharathi, and Jyothi, 2016 used Apple Inc. from 2013 to 2016 daily market data. News data was collected from news.google.com, reuters.com, finance.yahoo.com ;
- Atkins, Niranjan, and Gerding, 2018 Yahoo!Finance index data and Reuters news data of 2011-2012 was obtained;

[2] http://people.stern.nyu.edu/jwurgler/
[3] http://www.cboe.com/
[4] https://www.gold.org/goldhub/data/gold-prices

Chapter 3

Data

This chapter explains what happens in the world data availability, what news sources were explored, what was the final decision of selecting the one appeared in this work, what preprocessing steps were taken, and what is a final view of data.

3.1 Existing sentiment data sources

In our research, we are interested in high-frequency predictability research of stocks. That is why a major constraint is to find a source that could supply with frequent data. There are a lot of useful resources available which could provide with valuable information for the investigation of market behavior. These sources could be, namely, divided into two groups: Stock market & investment-oriented and General sources.

3.1.1 Stock market & investment oriented

There are sources that are directly related to the Markets. They are provided by market related companies (*Merrill Lynch*), public companies, organizations (*American Association of Individual Investors*) connected to the field of finance, web services (*Lexis-Nexis Academic*). The resources obtained are different indexes (*Investor Intelligence , Michigan Consumer Confidence Index*), 10-k reports (*Apple, Johnson & Johnson*), investors' survey studies, financial data (stock returns (*Yahoo!Finance*), market capitalization, trading activity(*Put-Call Ratio*)), and others.

3.1.2 General sources

These are the sources which could be used for market studies but are not directly related and produced for their need. These are various mass media sources, news agencies, social networks, blogs, various public entities.

3.2 Issues and alternatives of data sources

The era of data, nowadays, experiences a great breakthrough in the redesigning the approaches of data usage. The need of information often leads to the illegal acts of those seeking for a "forbidden fruit." One of the recent and well-known was Cambridge Analytics data scandal. After investigation personal Facebook information of 220 millions of Americans, they designed an analytic tool to influence the US elections outcome through direct advertising. The worlds' governments responded with new regulations in data privacy on the Internet, resulting in cutting off numerous data channels, or updating the access with a smaller, more general amount of available information.

The mentioned above and other cases lead to the logical conclusion that sooner or later any people's behavioral analytics will be pushed to seek for the indirect sources of information in order not to violate the law and find alternatives to the closed information channels.

In the field of stock, the world news could be treated as the indirect resources for prediction. There is a diverse amount of news agencies sharing their content online or even providing APIs for their data usage. Such sources of information could also provide an alternative option for measuring an investors' mood, or even predict the changes on the market.

The rising value of news sources could also be seen with the appearance of numerous services gathering a tremendous amount of news agencies posting and selling the reprocessed data. In addition to this evidence Thomson-Reuters, recently, made unavailable its free dataset of news which dates back from 1896 till these days which was used by Atkins, Niranjan, and Gerding, 2018 and now it could be gathered only from the one of the news data selling companies.

3.3 Intraday data

3.3.1 News source

This research was focused on searching the alternative to Facebook, Twitter or social media in the world news segment and how it could be beneficial for stock market.

First of all, it was considered important to discover separately the news sources from where the financial news could be downloaded. They were grouped into the following: 1) from where the date, news title, and news text, could be scrapped entirely, 2) partially scrapable, where not all information could be scrapped, and 3) not scrapable, where at most one information type could be scrapped (see table 3.1).

NEWS AGENCIES		
Scarable	**Partialy scrapable**	**Not scrapable**
The New York Times	Business Insider	Wall Street Journal *(only demo content)*
CNN		The Guardian *(only demo content)*
ABC News		Bloomberg
The Economist		Financial Times *(only partial text)*
CNBC		Reuters
Google News		
La Jazeera		
Fox News		
Financial Post		
CNBC		

TABLE 3.1: News agencies web news available for scrapping

Considering difficulties faced while scrapping from multiple sources, as well as time limitations we did not manage to collect enough data. Thomson-Reuters was decided to be the most applicable news source having minute preciseness, titles and full news text. Sadly, during the news mining Thomson-Reuters shut-down their public archive access.

"News API" [1] appeared to be a useful service of news for the research, providing, link, category, description, country, and agency of certain news. But it has only one month data available for free and 500 requests per day limit. In terms of working with stock data, where the stock intraday price is available only on working days of New York Stock Exchange and from 9:30 a.m. to 4:00 p.m., one month could deliver insufficient results.

During the exploration of possible news resources GDEL T (Global Data on Event, Location, and Tone) Project [2] was found. This service unites the world's printed and web news from every corner of the world with the ability to identify location, organizations, names, sources, themes, and emotions as a processed result of the news. This service produces news updates for every 15 minute period. So, we considered the GDELT Global Knowledge Graph (GKG) V2.1 as a source of news information. The version 2.1 was optimized for scripting language usage providing each 15-minute timestamp in CSV format. We gathered data (see figure 3.1) from this source for the period of March-May 2016. Among all the information groups available we were mostly interested in several data types which we put in the table (see table 3.2).

USED DATA TYPES	
Data Type	**Description**
V2.1DATE	datetime
V2SOURCECOLLECTIONIDENTIFIER	news sources type
V2SOURCECOMMONNAME	name of the news source origin
V2DOCUMENTIDENTIFIER	direct link to the news in the source
V1ORGANIZATIONS	organizations mentioned in the news
V2.1ALLNAMES	all names mentioned in the news
V1.5TONE	percentage of positive and negative words, of emotionally polarized words and tone of the document at all
V2GCAM	result of Global Content Analysis Measures consisting of 2300 dimesions

TABLE 3.2: News data types from GDELT used in research

[1] https://newsapi.org/
[2] https://www.gdeltproject.org/

Tone We believe that "tone" from V1.5TONE should be the most important feature among others obtained for stock correlations. This feature describes an average "tone" of the news document. It is calculated as a difference between the percentage of all positive and all negative words in the document. This means that we could make assumptions based on the particular feature on whether it is related to certain company news either negatively or positively. In our assumption, this should somehow influence the traders, so the stock returns.

Polarity Besides, we believe that "polarity" which counts the percentage of emotionally polarized words in the text can, also, be useful in our studies.

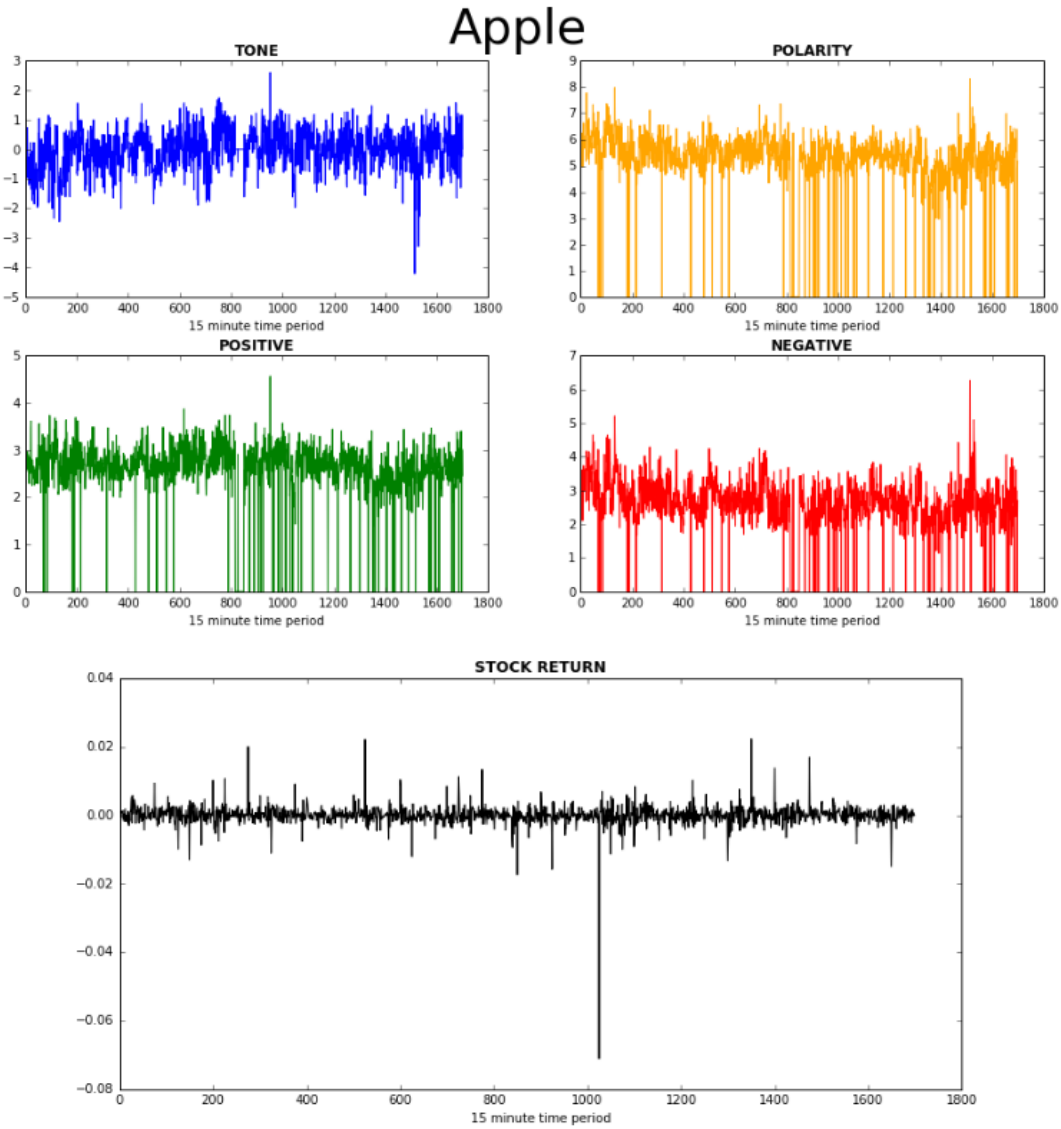

FIGURE 3.1: News sentiments and stock return data (Apple)

This work appeared to be the first one to use such news aggregation service for predictability studies of the stock returns of SP 500 companies.

3.3.2 Stock source

The intraday stock price data was provided by the Chair of Statistics, Faculty of Business and Economics at University of Augsburg which purchased the dataset from *QuantQuote* [3]. The stock data is available from 9:30 till 16:00 EST of each trading day. We used it for the period of March-May 2016.

3.4 Data preprocessing

3.4.1 Stocks

There is a need to convert stock prices into stock returns to investigate the financial gain or loss.

$$R_t = log\left(\frac{P_t}{P_{t-1}}\right),$$

where R_t - log return of the stock for a 15 minute period at time t, P_t - stock price at time t. Considering the GDELT time constraints, t corresponds to the 15-th minute timestamp of the trading day.

3.4.2 News

Search for the related news The research based on top 5 Information Technology affiliates of S&P 500 Index by performance. To search the news that relates to these affiliates a table of companies associations was created. If one of these companies is mentioned in V2DOCUMENTIDENTIFIER, V1ORGANIZATIONS or V2.1ALLNAMES than we consider the news to be related to the certain affiliate and could be used further in the study (see table 3.3).

TOP 5 S&P AFFILIATES AND THEIR ASSOCIATES	
Affiliates	**Associates**
Apple Inc.	Apple, Shazam, Emagic, Siri, Beats Electronics, NeXT Inc., Novauris, PrimeSense, Apple Pay
Microsoft Corporation	Microsoft, Windows, GitHub, LinkedIn, Skype Technologies S.A.R.L, Mojang, Yammer, Hotmail
Amazon.com Inc.	Amazon, IMDb.com, Alexa.com, Audible.com, Zappos, Annapurna Labs, Goodreads, Twitch.tv, Kiva Systems, Amazon Robotics, Whole Foods Market
Facebook Inc. Class A	Facebook, WhatsApp, Instagram
Alphabet Inc. Class C	Google Maps, AdSense, DoubleClick, YouTube, Google, Gmail, Google drive

TABLE 3.3: S&P 500 Top 5 Information Technology affiliates

This phase produced the first dataset (RAW) of all news related to S& P 500 affiliates over which we conduct our studies.

[3] https://quantquote.com/

Financial filtering Having news where company associate is mentioned could still be too noisy data, containing useless information for this research. There is a need to test the news for its financial relation. Having 2300 dimensions of news text analysis for each news, the financial filtering method was applied. We outlined the dictionaries provided by the GDELT, which suit the need for the news to be financially related: c18.59 `econ_stockmarket`, c18:60 `econ_earnings_report`, c18.61 `econ_IPO`, c18.154 `econ_monopoly`, c18.286 `econ_bubble`, c18.287 `econ_inflation`, c18.288 `econ_deflation`. If the news happens to consist words of any of such dictionaries, it is considered as financial. This method produced financially filtered dataset (FIN). Having several amounts of news for a 15 minutes, the average of all showings is taken into account as a sentiment reaction for a current timestamp.

3.4.3 News prioritization

Every source of information has unique popularity among the Internet. Some sources are often cited, and their information used to share at other platforms, some of them are not popular and even having valuable information, they do not reach and cover a significant amount of Internet consumers. As the method to create a prioritization of certain news influence, we introduce weighted mean sentiment calculation using Page Rank and Harmonics Centrality. The outcomes of these algorithms for the period of November 2018 - January 2019 were obtained from Common Crawl [4] which calculates the ranking for 90 million domains. The research of Boldi and Vigna, 2014 compared different graph algorithms, including Page Rank and Harmonic Centrality. The axioms of density and score-monotonicity were satisfied by both of them, but only Harmonic Centrality appeared to be compliant to the size axiom which results in better node importance assigning, considering a specific case of network growth. It is interesting to check which of those algorithms provide preferred ranks of the news sources so that the impact of news would be significant.

Page Rank Page Rank is an algorithm designed by Sergey Brin and Larry Page for the need of Google Search engine showing search results relative to their importance. The idea behind the algorithm is to find the probability distribution of a certain page occurrence as a result of person randomly clicking on pages arrive on it. This algorithm identifies the value of the node by importance of the neighbours directly linked to it. This could be presented as:

$$PR(x_a) = (1-d) + d\sum_{i=1}^{n} \frac{PR(x_i)}{C(x_i)},$$

where d - is a damping factor (probability of person to continue clicking the links at any step), $x_1, ...x_n$ nodes (pages) that cite the page x_a, $PR(x_i)$ - page rank of the node, $C(x_i)$ - number of edges from x_i directed to other nodes.

Harmonics Centrality Harmonics Centrality was introduced by Marchiori and Latora, 2000 and is a distance-based centrality measure in comparison to Page Rank. It depends on how far one node is placed in relation to the other nodes in graph (Internet). The more nodes are closer to it, the bigger weight (rank) it has. This could

[4] http://commoncrawl.org/

be represented as:

$$H_{(x_a)} = \sum_{i=1, x_i \neq x_a}^{n} \frac{1}{d(x_i, x_a)},$$

where x_i and x_a are nodes, d - is a distance function, $H_{(x_a)}$ - is harmonic centrality value of x_a. If there is no path between x_i and x_a, then $d(x_i, x_a) = 0$

Sentiment weighting Sentiment weighting is an approach to generate two different datasets using Page Rank and Harmonics as the coefficients of news importance based on its rank domain and produce weighted sentiment showings for each 15-minute period:

$$SS_{ts} = \sum_{i=1}^{n} \frac{Rk(x_i)}{\sum_{j=1}^{n} Rk(x_j)} x_i,$$

where SS_{ts} - is sentiment showing for 15 minute timestamp ts, $Rk(x_i)$ rank of a node (news source) x_i produced by selected algorithm, n is a number of elements (news) of 15-minute ts.

Summing up, there are Raw (RAW), Financial (FIN), Financial with Page Rank (PR), Financial with Harmonics (HR) datasets generated, which appear to be four unique types of data preprocessed from the GDELT Project to explore how they could influence the stock return predictability.

Chapter 4

Empirical research

First of all, it is important to test our time series for stationarity. For this aim, the *Dickey-Fuller test* has to be conducted. We consider a regression model:

$$\Delta y_t = \lambda + (\rho - 1)y_t + \epsilon_t,$$

where $\Delta y_t = y_t - y_{t-1}$, ρ is a coefficient, y_t - value at time t, λ - a measure of drift, ϵ_t - error term at time t. Here, $H_0 : \rho = 1$ states for unit root and $H_1 : \rho < 1$ states for possible stationarity.

All sentiment datasets and stock return appeared to have $p - value < 0.05$ as an outcome of the *Dickey-Fuller test*. Accordingly, we consider them to be stationary.

4.1 Correlations

4.1.1 Autocorrelations

To begin with, we will take a "tone" of news related to the S&P 500 affiliates and explore its autocorrelation: similarity of observed values as a function of time lag in between.

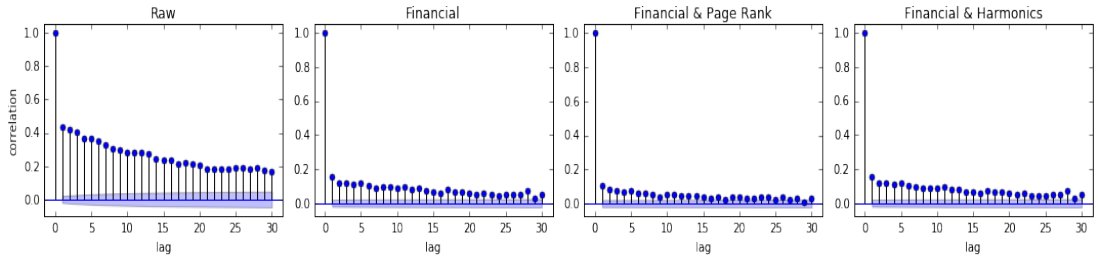

FIGURE 4.1: Autocorrelation of "tone" (Alphabet)

From the left plot (see figure 4.1), where all news about affiliate is gathered (RAW) we could see a relatively high correlation. It seems that various sources tend to repeat some news information for a particular period keeping the information in inertia. Removing non-financial news considerably decreased autocorrelation of financial (FIN) dataset. Page Rank (PR) and Harmonics (HR) have prioritized news relevance, so the plots should be even more accurate reproducers of financial news autocorrelation than FIN. Each of these plots suggests that the highest correlation exists at one lag - 15 minutes. Additionally, Raw has the highest statistical correlation (most probably due to news inertia), and Harmonics seems to have higher autocorrelation than Page Rank.

4.1.2 Cross-autocorrelations

We examined the correlation between the stock return and "tone" from *V*1.5*Tone* in terms of their time lags (see figure 4.2). Although it appeared to be statistically insignificant, one can notice that there is the poorest effect on stock return by the sentiment of Raw dataset. In any of the dataset, it seems to have no correlation at all or approximately the same showing throughout the lags. Analyzing Facebook Financial, Page Rank and Harmonics one can find that "tone" is affected by the stock return and shows a correlation on the sequence of lags (positive lags), which means that in this case, "tone" is continuously influenced by stock return. However, we cannot state that there is a significant impact of "tone" on stock returns (negative lags), considering no correlation on time lags. Amazon, in contrast to Facebook, seems to have cross-autocorrelation of both "tone" and stock return.

Summing up, filtering and prioritizing of news helped to reduce the non-related news acorrelation with stock return and proved that eliminating the news "noise" the suggested way is beneficial for "tone" predictability. In comparison to Atkins, Niranjan, and Gerding, 2018 where they found that financial news affects the stock within 20 minute period, we expect that news could correlate within, approximately 1 hour period (4 lags).

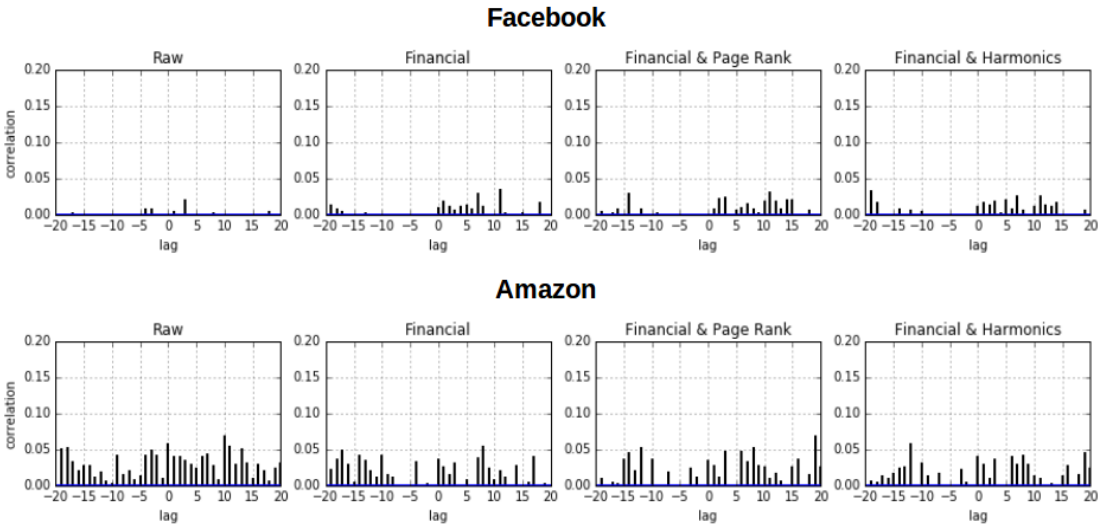

FIGURE 4.2: Cross-autocorrelation between "tone" and stock return (Facebook & Amazon)

4.2 Linear and Quantile Regression

4.2.1 Linear Regression

Linear regression is one of the simpliest approaches of predicting the quantitative response on the basis of predictor variable. It could be represented as:

$$R = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon,$$

where β_j is a parameter, x_j -the predictor, R - stock return, ϵ represents an error.

To decide which features to use in liner regression we conducted a *t-test* and obtained the *p-values* (see table 4.1). We cannot state that these features could be significant predictors of stock return.

Dataset	Tone	Polarity	Previous return	Positive	Negative
Apple					
Raw	0.0399	0.6832	0.7791	0.5504	0.2367
Financial	0.2106	0.9699	0.7791	0.4502	0.4908
Page Rank	0.3315	0.7498	0.5017	0.5477	0.3868
Harmonics	0.1927	0.3965	0.5017	0.6400	0.1748
Microsoft					
Raw	0.7634	0.0036	0.1440	0.0224	0.0042
Financial	0.7652	0.5215	0.9757	0.4743	0.7852
Page Rank	0.2697	0.1401	0.8967	0.0690	0.9967
Harmonics	0.8047	0.1706	0.0000	0.5052	0.3207
Amazon					
Raw	0.1381	0.6727	0.5705	0.2606	0.6411
Financial	0.2721	0.8046	0.5705	0.4483	0.7271
Page Rank	0.3701	0.4132	0.5089	0.8536	0.2584
Harmonics	0.3077	0.7513	0.5089	0.5300	0.3712
Facebook					
Raw	0.8772	0.0070	0.1440	0.0051	0.0186
Financial	0.8359	0.0145	0.1440	0.0156	0.0973
Page Rank	0.8559	0.3214	0.5230	0.6042	0.5340
Harmonics	0.7029	0.5791	0.5230	0.4001	0.9772
Alphabet					
Raw	0.6424	0.3804	0.0189	0.2990	0.5414
Financial	0.6447	0.5435	0.0189	0.8016	0.4884
Page Rank	0.4030	0.1710	0.0085	0.0682	0.8658
Harmonics	0.6089	0.8051	0.0085	0.4835	0.8240

TABLE 4.1: *P-value* of feature series in relation to stock return

We applied multiple linear regression of 5 sentiment features ("tone", "polarity", "previous return", "positive", "negative). Each of the datasets appeared to produce approximately the same predictions (see figure 4.3).

To compare effectiveness, we calculated a mean of stock returns of train set as a naive forecast for a test period. To compare we calculated the Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(R_i - \widehat{R}_i)^2$$

where n - number of values in the test set, R is a actual values array, $\widehat{R}$ is predicted values array. The MSE of linear regression and mean for the same period appeared to be almost the same. Harmonics showed a relatively smaller MSE than Page Rank,

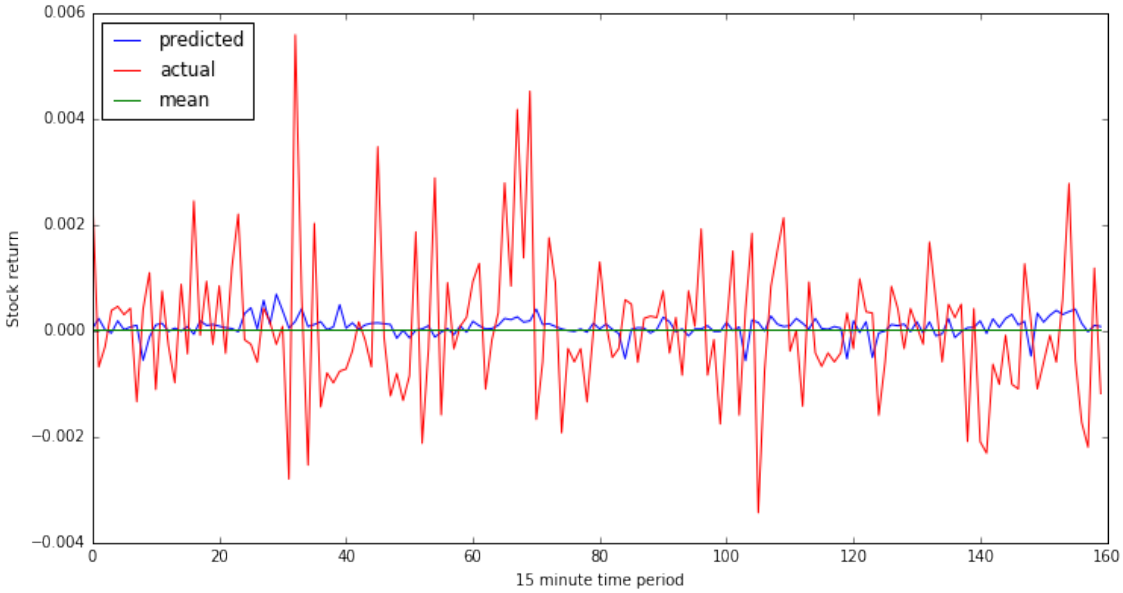

FIGURE 4.3: Linear regression based on selected features on Financial dataset (Facebook)

Raw, and Financial. Still, each of them had higher MSE than a naive forecast for the test period (see table 4.2) .

Affiliate	Raw	Financial	Page Rank	Harmonics	Mean
Apple	0.4077	0.4087	0.4198	0.4220	0.4036
Microsoft	0.2170	0.4463	0.4609	0.6437	0.2112
Amazon.com	0.2611	0.2663	0.2739	0.2698	0.2633
Facebook	0.2137	0.2154	0.2204	0.2195	0.2112
Alphabet	0.3462	0.3419	0.3531	0.3467	0.3420

TABLE 4.2: MSE of linear regression on test data and mean (naive forecast) ($*10^5$)

To crown it all, one could state that linear regression is ineffective in the case of prediction of stock returns based on sentiments from the GDELT.

4.2.2 Quantile Regression

Linear regression model allows estimating the expected value of the response variable for a set of predictor variables. Thus, the prediction is focused on least squares approach omitting the possibility that specific data could act differently on the certain part (quantile). In order to have a more comprehensive picture of how a certain number of predictors influence the response variable, we apply Quantile Regression. This method splits the predictor dataset into quantiles in order to identify the influence on the response variable. It is robust to outliers in the response measurements, can be applied for flexible distribution assumptions and to fit heterogeneous data.

Linear regression seeks to update the parameters in order to minimize the ϵ^2, but Quantile regression is based on Least Absolute Deviation Equation. Here it seeks

to find such parameters that could minimize the outcome of equation taking the module instead of square as in Linear Regression:

$$OLS : arg\,min \sum_{i=1}^{n} \Big[R_i - (\beta_0 + \beta_1 x_i + ... + \beta_p x_p) \Big]^2$$

$$LAD : arg\,min \sum_{i=1}^{n} \Big| R_i - (\beta_0 + \beta_1 x_i + ... + \beta_p x_p) \Big|,$$

where $\beta_{1,..,p}$ are the parameters to be updated, x_i is the predictor, and R_i - the stock return.

The parameters of τ-th quantile can be obtained by adding a τ in LAD minimization equation:

$$\beta_0(\tau), ..., \beta_p(\tau) = arg\,min \sum_{i:R_i \geq \beta_0 + ... + \beta_p x_p} \tau \Big| R_i - (\beta_0 + ... + \beta_p x_p) \Big| + \sum_{i:R_i < \beta_0 + ... + \beta_p x_p} (1 - \tau) \Big| R_i - (\beta_0 + ... + \beta_p x_p) \Big|$$

Here we plot quantile regression parameters of "tone" and "polarity" sentiments as a function of τ (see figure 4.4). The green and red dotted lines on the plot show the upper bound and lower bound (95 % confidence intervals) of the percentiles and blue line is the actual value.

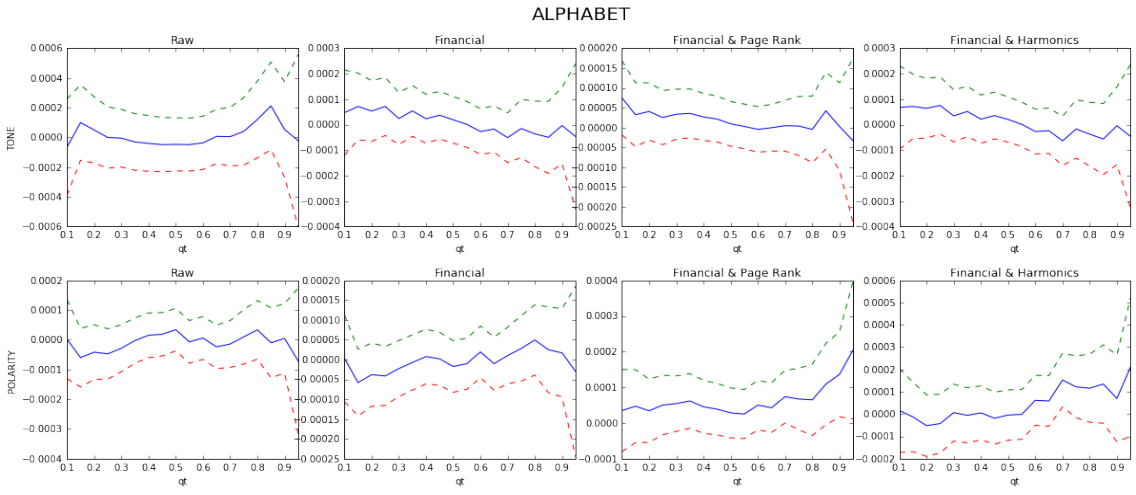

FIGURE 4.4: Quantile regression of "tone" and "polarity" in 4 datasets (Alphabet)

As it could be noticed, none of the sentiments from any datasets seem to have a significant influence on the stock return (see figure A.1). This condition remains the same for every dataset of each affiliate. So, even exploring different sentiment data parts we cannot state that there is a significant correlation that could be used to predict a stock return.

4.3 Autoregression

As we could not find any statistically significant predictor for a stock return (see table 4.1), it is desirable to explore how the stock return could predict itself. In Autoregression we will regress the previous values of stock return to predict its future value which could be expressed as:

$$R_t = \beta_0 + \beta_1 R_{t-1} + ... + \beta_p R_{t-p} + \epsilon_t,$$

where R_t is a stock return predicted from $R_{t-1}, ..., R_{t-p}$, $\beta_{(1,p)}$ are the estimated parameters, p is a number of lags to be used in prediction and error term ϵ_t. We will take a lag of 4, so our AR model will have a 4-th order. Having implemented AR(4), we try to conduct the a 4 lags one-step ahead forecast of stock return for the whole trading day which is 6 hours 30 minutes. Considering the timestamp of 15 minutes that we use in our studies, we predict 26 timestamps ahead (see figure 4.5).

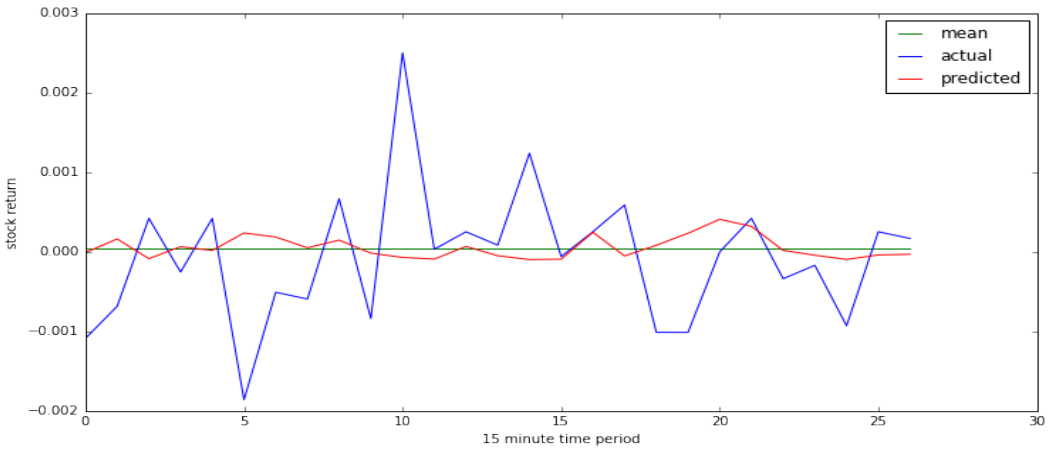

FIGURE 4.5: Autoregression of stock return (Microsoft)

Mean of the training sample was also calculated and plotted over the test period. The MSE for Apple and Alphabet is slightly lower in AR prediction than in produced mean. Others, Facebook, Amazon.com, Microsoft, appear to have the MSE of Autoregression prediction higher than in calculated mean.

4.4 Return prediction summary

We applied the *t-test* which showed that there are no common features to use as the predictors for stock return. Consequently, linear regression applied showed higher MSE then just a mean over a test set for 3/5 of observed affiliates of S&P 500. Searching for specific percentiles of features as an impact on stock returns using quantile regression did not deliver any essential results. Even Autoregression model applied on stock returns, showed slightly lower MSE than just the mean of stock returns.

In conclusion, one should state that no statistically significant information was found that could be used for predictions of stock returns. News sentiments mined from the GDELT and preprocessed by financial relation and prioritization seemed to have no visible effect, and stock returns do not have enough memory to predict themselves.

4.5 Combination of AR and Linear Regression

Taking into account the notion of noise traders De Long et al., 1990, behavior of traders using news sources as information Barberis, Shleifer, and Vishny, 1998 and Han, 2007 there should be at least some linear interdependencies between news sentiments and stock return. For that aim we implement a model which is a combination of AR model and linear regression, where returns, "tone" and "polarity" till the 8th lag are considered. In terms of prediction of stock returns it is:

$$R_t = \beta_0 + \beta_1 R_1 + ... + \beta_p R_{t-p} + \lambda_0 x_t + ... + \lambda_p x_{t-p} + \theta_0 y_t + ... + \theta_p y_{t-p} + \epsilon_t,$$

where R_t is a return predicted from $R_{t-1}, ..., R_{t-p}$, $x_{t-1}, ..., x_{t-p}$, $y_{t-1}, ..., y_{t-p}$ where R_t is stock return at time t, x_t is "tone" at time t, y_t is a "polarity" at time t and $\beta_{(1,p)}$, $\lambda_{(1,p)}$, $\theta_{(1,p)}$ are the parameters, p is a number of lags to be used in prediction and error term ϵ_t. This model exists as a part of Vector Autocorrelation Regression (*VAR*) model.

Lags of feature significance for different datasets		
Affiliate	**Feature: "tone"**	**Feature: "polarity"**
Apple	FIN: 7; HR: 7	PR: 1, 7; HR: 1
Microsoft	RAW: 3; FIN: 3; PR: 3, 6, 8	RAW: 3, 8 PR: 4
Amazon.com	PR: 6	RAW: 2
Facebook	None	RAW: 3; FIN 2, 3
Alphabet	RAW: 6; PR: 3; HR: 1, 4	RAW: 6; PR: 3,6

TABLE 4.3: Features & stock return significant interdependence on certain lags based on 4 datasets

Completing the analysis, we found some significant evidence of stock return linear interdependence with "tone" and "polarity" considering different datasets (see table 4.3). As a side effect, we found the evidence of tone being negatively correlated with the polarity of news. This can contribute to the fact that negative news is more emotional than a positive one.

4.6 APARCH

Considering some interdependence observed between the sentiment features ("tone" and "polarity") and stock return, it is advisable to investigate whether we could predict the volatility of stock returns.

For a specific case of stock returns, we decided to use Asymmetric Power Autoregressive Conditional Heteroscedasticity model (APARCH). This model is an extension to the famous ARCH model Engle, 1982 for which a Robert F. Engel received a Nobel prize in 2003.

GARCH The idea behind Generalized Autoregressive Conditional Heteroscedasticity model is that conditional volatility should depend on the p last returns and q conditional volatilities. A simple GARCH(1,1) model is:

$$
\begin{aligned}
R_t &= \mu + \epsilon_t \sigma_t \\
\epsilon_t &\sim \mathcal{N}(0,1) \\
\sigma_t^2 = \lambda_0 + \lambda_1 (R_{t-1} - \mu)^2 + \beta \sigma_{t-1}^2 &\text{ with} \\
\lambda_0, \lambda_1, \beta &> 0,
\end{aligned}
$$

where R_t is a return at time t, $\lambda_0, \lambda_1\, \beta$ are the volatility parameters, μ is a mean, σ_t is a conditional volatility at time t, ϵ_t is an error term at time t.

APARCH As APARCH is an extension to the GARCH model, it also models a conditional volatility of p last returns. But it differentiate stock losses and profits. It is really important to take this into account as there is a notion of *leverage effect* which states that volatility responds stronger on negative stock returns than positive of the same value. So APARCH is designed specifically to capture this interdependence. We designed AR(1)-APARCH(1, 1):

$$
\begin{aligned}
R_t &= \mu + \phi R_{t-1} + \epsilon_t \sigma_t \\
\epsilon_t &\sim \mathcal{N}(0,1) \\
\sigma_t^\delta = \lambda_0 + \lambda_1 \left[\left| \frac{R_{t-1} - \mu - \phi R_{t-1}}{\sigma_{t-1}} \right| - \gamma \frac{R_{t-1} - \mu - \phi R_{t-1}}{\sigma_{t-1}} \right]^\delta &+ \beta \sigma_{t-1}^\delta + \vec{x}_t^T \vec{\eta} \\
\text{with } \lambda_0, \lambda_1, \eta, \delta, \beta > 0, |\gamma| &< 1,
\end{aligned}
$$

where R_t is stock return at time t , μ - mean of stock returns, ϵ_t error term at time t , σ_t is volatility at time t, as parameter, $\vec{x}_t$ is a vector of sentiments at time t and $\lambda_0, \lambda_1, \beta_1, \phi, \gamma\, \delta$ are the parameters, $\vec{\eta}$ is a vector of parameters. In this formula γ plays a role of asymmetry (penalizing) parameter: if $\gamma > 0$ - an above average return results in volatility decline and vice versa if $\gamma < 0$ - an above average return results in volatility growth.

We conducted APARCH for stock returns of every selected affiliate of S&P 500 top 5 tech with 4 extra feature sets which were applied as $\vec{x}_t$ in APARCH, and without any. Those four feature sets were: "tone", "polarity", "tone" and "polarity", absolute value of "tone". We compared them via Akaike Information Criterion (AIC) which evaluates a statistical model over information loss. The lower information loss, the higher is the quality of the model. In the table (see table 4.4), we want to show some results that occurred in our investigation and additionally compare them to GARCH, which could be transformed from APARCH by specifying two parameters: $\gamma = 0$ and $\delta = 2$. Comparing with GARCH we review whether it is important to count on *leverage effect*. As we can notice from the M1 comparison, using statistically significant "tones" for APARCH in financially filtered (FIN) Microsoft dataset resulted in better AIC score. In some cases, as in M2 where we tested "polarity" in APARCH, even the feature being statistically insignificant it brought better AIC score then either APARCH or GARCH with no extra features. But it does not always hold. While testing both "tone" and "polarity" in Apple raw dataset, they appeared to be better than a simple APARCH but information loss was higher than in GARCH.

AIC performance results												
Parameters	**M1.1**	**M1.2**	**M1.3**	**M2.1**	**M2.2**	**M2.3**	**A1.1**	**A1.2**	**A1.3**	**AM1.1**	**AM1.2**	**AM1.3**
AIC	-9.530	-9.425	-9.504	-9.034	-8.7	-9.026	-8.889	-8.771	-8.894	-8.965	-8.833	-8.909
μ	0.000	-0.000	0.000	0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ϕ	0.209	0.197	0.142	-0.002	0.483	-0.074	0.030	0.061	0.059	0.024	0.034	0.007
λ_0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
λ_1	0.001	0.057	0.011	0.066	0.041	0.069	0.069	0.062	0.018	0.010	0.039	0.009
(β_1	0.999	0.906	0.982	0.937	0.000	0.913	0.958	0.909	0.974	0.986	0.92	0.986
γ	0.044	0.052		0.038	1		0.044	0.082		0.968	0.147	
δ	1.678	2.594		1.294	4		1.09	2.455		1.304	2.819	
$x1 : \vec{x}^T$	0.000 (0.0*)			0.000 (0.96)			0.000 (0.974)			0.000 (0.0*)		
$x2 : \vec{x}^T$							0.000 (0.968)					

TABLE 4.4: M1.1 = APARCH with "tone", M1.2 = APARCH, M1.3 = GARCH on Microsoft Harmonics (HR) dataset; M2.1 = APARCH with "polarity", M2.2 = APARCH, M2.3 = GARCH on Microsoft financial (FIN) dataset; A1.1 = APARCH with "tone" and "polarity", A1.2 = APARCH, A1.3 = GARCH on Apple raw (RAW) dataset; AM1.1 = APARCH with "tone" , AM1.2 = APARCH, AM1.3 = GARCH on Amazon.com Page Rank (PR) dataset

It seems that APARCH with no extra sentiment features, even capturing the *leverage effect* could not benefit from it that strong to outperform GARCH. However, supplying APARCH with extra features enhances the model statistical robustness that allows to exceed GARCH performance (the results for each 5 tech S&P affiliate in these tables: A.1, A.2, A.3, A.4, A.5).

Despite performance measures, we tried to capture differences between the efficiency of data preprocessing. First of all, surprisingly, news financial filtering (FIN) appeared to be inefficient. The statistical models did not improve at all. Nonetheless, for all cases, but Alphabet, Page Rank (PR) and Harmonics (HR) prioritization produced the datasets that APARCH model mostly benefited from, and for each of S&P 500 affiliates, resulted in better AIC score. We found no visible differences between Page Rank and Harmonics prioritization. It could be presumed that in terms of stocks the most influential sources do not differ much in both ranking algorithms, resulting in approximately the same news importance weighting.

4.7 Return volatility summary

Finalizing, we prooved that in all cases using APARCH with news sentiments will lead to less information loss, so it makes the statistical model stronger (see figure4.6). We found no significant evidence that could prove that selecting only financially related news could contribute to the robustness of the prediction models, but news prioritization, based on financial news, enhances the statistical model and enables to make use of *leverage effect*. Finally, we consider that volatility produced via including

preprocessed news sentiment in our research should be preferred as it appeared to be relatively accurate, so the trader could enhance his/her investment portfolio management, calculating the volatility in the proposed way.

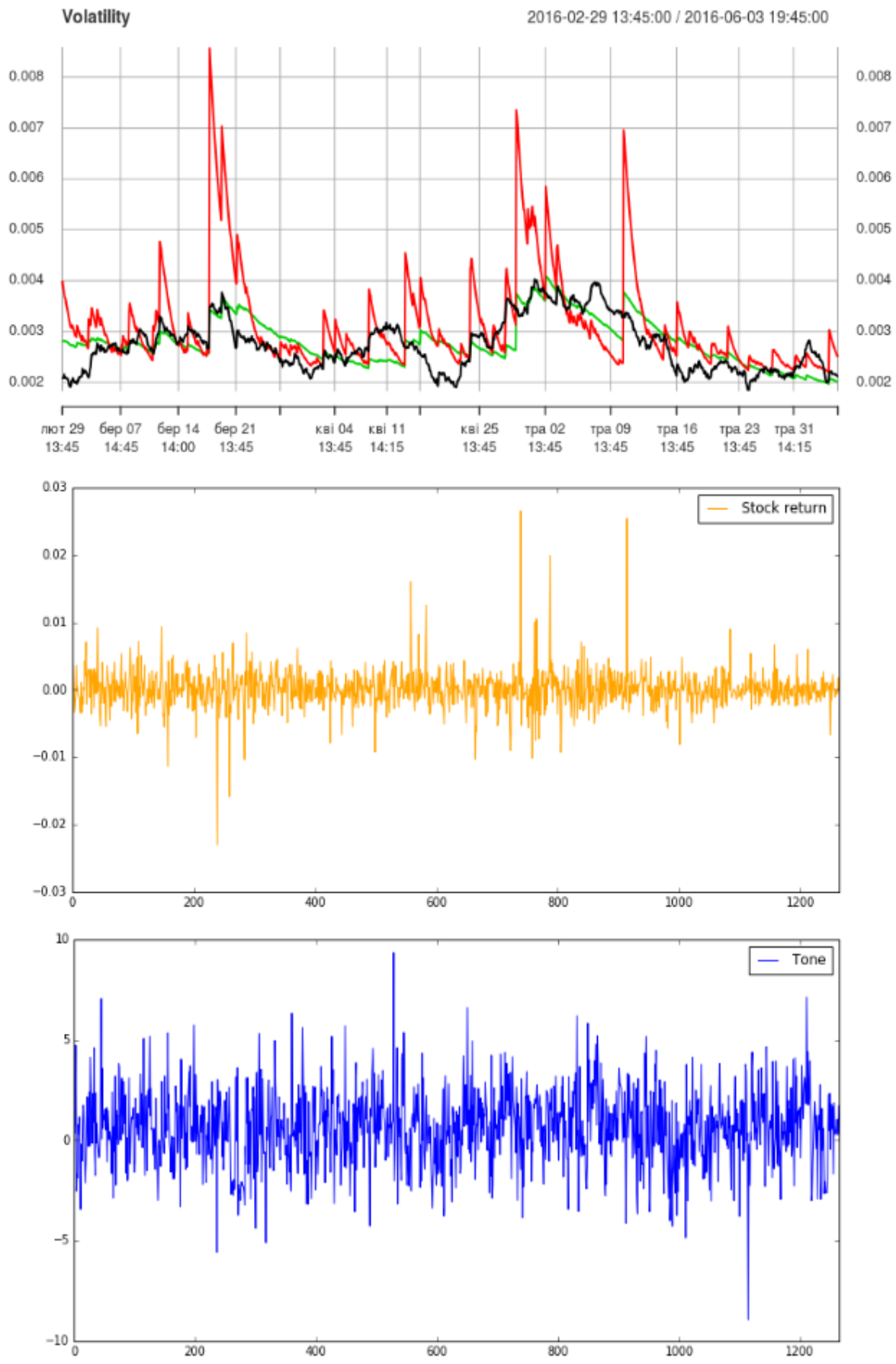

FIGURE 4.6: Comparison of models (Amazon: PR): APARCH with "tone": black, APARCH: red, GARCH: green

Chapter 5

Conclusion

In this work, we introduced the world news as an alternative resource of data for stock predictions. We aimed to explore how the trader can benefit from the massive data that includes almost 190, 000 news sources around the globe.

Our research contributions:

- We are the first to introduce a way to make use of the GDELT (Global Database of Events, Language, and Tone) Project for stock return predictability and produce four data preprocessing algorithms in order to increase efficiency of news for stock predictions.
- We proved the efficiency of news as a significant factor for APARCH models of risk to react on *leverage effect* in the behavior of stocks.
- We showed that world news is not efficient for stock return predictions, but is a great addition to stock volatility forecasting. The stock trader can benefit from volatility produced by APARCH model with news sentiments to be more confident in calculation of *Value at Risk* measure, and *Sharpe ratio* as a part of his/her investment strategy.

Appendix A

Glossary and Attachements

A.1 Glossary

- **Arbitrage** a simultaneous purchase and sell of an asset in order to profit from price imbalance
- **Average Volume** the money-equivalent amount of shares traded
- **Book-to-equity-ratio** a ratio used to find the company's value by comparing book value to the market value
- **Big Data** a large volume of different data (structured and unstructured) that has potential to be used for solving problems appearing in different fields through advanced analytics or machine learning
- **Bigram** a combination of two adjusent elements
- **Black-Sholes option pricing model** states that price of of the asset changes continuously through the option's expiration date
- **Bubble** an unusual temporal behavior that is characterized by specific features state under peculiar circumstances. This event appears unpredictively, and continues for specific timespace and dissapears
- **Bullish and Bearish** a trading behavior characterized by willing to invest (bullish) or willing to sell (bearish)
- **Buy-Sell Imbalance (BSI)** an extreme change in buying or selling amount mostly affected by good or bad news shocks
- **Chi-squared feature reduction** an algorithm used to compare features with a certain values and exclude those which are independent and leave only important ones
- **Free Float** the money equivalent of shares that could be publicly traded without restrictions
- **Granger Causality** a statistical concept of causality between signals. If signal A "Granger causes" a signal B that means that A has some information that could be used to predict B
- **Kalman Filter** or Linear Quadratic Estimation is an algorithm processing multiple series of measurments with statistical noise, and produces estimates of variables that are more accurate than, those based on a single measurment

- **Leverage Effect** a part of stock behaviour, in which volatility responds stronger on negative stock returns than positive of the same value
- **Linear Regression** a model which uses certain data as explanatory for a dependent data to qualify how well such dependent data could be predicted by explanatory one
- **Logistic Regression** a model which uses certain data as explanatory for a dependent classification data to qualify how well such dependent data could be classified by explanatory one
- **Luhn's Cut-off** used to deal with the overuse and underuse of words in documents so the remaining words become significant
- **Market cap**: the value of the company
- **Naive Bayes Classifier** a probabilistic model used for classification considering independent features probabilities
- **Portfolio** as a grouping of financial assets (here stocks), cash equivalents; individually managed group of assets
- **Principal Component Analysis** a dimension reduction algorithm used to reduce the large set of variables to leave those which carry the most information
- **Put-Call Ratio** used to understand the market mood (sentiment) buy comparing the number of put and call actions
- **Random Forest** an algorithm used for classification and regression that constructs multiple decision trees and outputs the mode of classes or mean of regressions' prediction
- **Random Walk (Theory), RW(T)** sequence has the same distribution over observation and the past information cannot be used to predict the future movement
- **Sharpe Ratio** a measure of compensation of returns for investment risk
- **Support Vector Machines** a discriminative classifier that outputs an optimal hyperplane which categorizes a new examples of features from the previously labeled dataset
- **Shock** an unpredicted or unexpected event that affects the economy
- **"Siamese twin"** the two companies operating in different countries as a single business, having separate stock showings and legal identities. Theoretically, their stock prices should move in lockstep
- **T-test** used to determine significant difference of sample means of datasets
- **Value at Risk (VaR)** a measure of risk of loss for stock investments
- **Volatility smile** a volatility parameter which appears at pricing options

A.2 Quantile regression for Apple

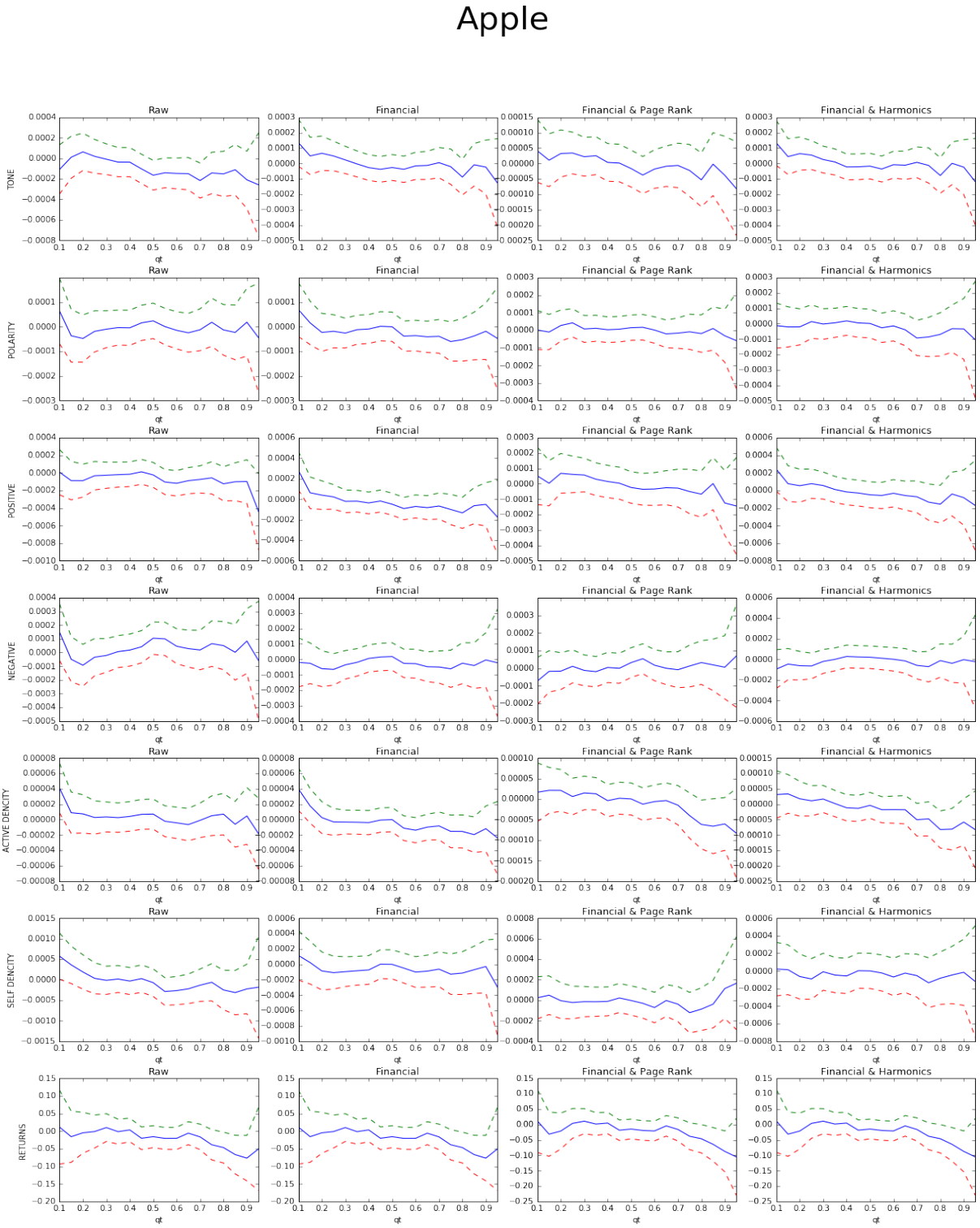

FIGURE A.1: Quantile regression of sentiments from 4 datasets (Apple)

A.3 APARCH model execution results

T - "tone" as an extra feature; P - "polarity" as an extra feature; TP - "tone" and "polarity" as the extra features; AT - "absolute value of "tone" as an extra feature; N - no extra features;

Apple	**RAW: execution results**					**FIN: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-8.888	-8.885	-8.889	-8.888	-8.771	-8.888	-8.885	-8.889	-8.888	-8.771
x1 p-value	0.974	0.966	0.974	0.976		0.951	0.967	0.974	0.978	
x2 p-value			0.968					0.969		
Apple	**PR: execution results**					**HR: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-8.954	-8.953	-8.890	-8.892	-8.299	-8.897	-8.953	-8.891	-8.894	-8.299
x1 p-value	0.973	0.977	0.974	0.977		0.973	0.977	0.974	0.975	
x2 p-value			0.969					0.969		

TABLE A.1: APARCH execution results for Apple. x1 p-value and x2 p-value describe statistical significance of sentiment features

Microsoft	**RAW: execution results**					**FIN: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	9.442	-9.444	-9.445	-9.445	-9.427	-9.033	-9.034	-8.910	-9.033	-8.700
x1 p-value	0.000	0.898	0.000	0.831		0.969	0.966	0.000	0.972	
x2 p-value			0.826					1.000		
Microsoft	**PR: execution results**					**HR: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-8.965	-8.935	-8.934	-8.936	-8.833	-9.530	-9.529	-9.522	9.527	-9.425
x1 p-value	0.000	0.999	0.475	0.000		0.000	0.990	0.000	0.962	
x2 p-value			1.000					0.999		

TABLE A.2: APARCH execution results for Microsoft. x1 p-value and x2 p-value describe statistical significance of sentiment features

Amazon	**RAW: execution results**					**FIN: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-8.702	-8.702	-8.700	-8.702	-8.716	-8.711	-8.702	-8.700	-8.702	-8.716
x1 p-value	0.000	0.898	0.000	0.831		0.003	0.999	0.999	1.000	
x2 p-value			0.826					0.999		
Amazon	**PR: execution results**					**HR: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-8.965	-8.935	-8.934	-8.936	-8.833	-8.960	-8.935	-8.934	-8.937	-8.833
x1 p-value	0.000	0.999	0.475	0.000		0.000	0.990	0.001	0.000	
x2 p-value			1.000					1.000		

TABLE A.3: APARCH execution results for Amazon.com x1 p-value and x2 p-value describe statistical significance of sentiment features

Facebook	**RAW: execution results**					**FIN: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-9.442	-9.444	-9.442	-9.447	-9.427	-9.443	-9.443	-9.445	-9.447	9.427
x1 p-value	1.000	0.901	0.962	0.777		0.987	0.891	0.924	0.794	
x2 p-value			0.876					0.999		
Facebook	**PR: execution results**					**HR: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-9.296	-9.389	-9.296	-9.482	-9.393	-9.476	-9.485	-9.482	-9.481	-9.473
x1 p-value	0.990	0.001	1.000	0.000		0.986	0.933	0.962	0.680	
x2 p-value			0.000					0.913		

TABLE A.4: APARCH execution results for Facebook x1 p-value and x2 p-value describe statistical significance of sentiment features

Alphabet	**RAW: execution results**					**FIN: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	-9.495	-9.494	-9.491	-9.496	-9.446	-9.492	-9.495	-9.492	-9.498	-9.446
x1 p-value	0.000	0.977	0.000	0.868		0.000	0.981	0.000	0.924	
x2 p-value			0.970					0.967		
Alphabet	**PR: execution results**					**HR: execution results**				
Parameters	**T**	**P**	**TP**	**AT**	**N**	**T**	**P**	**TP**	**AT**	**N**
AIC	9.487	-9.488	-9.487	-9.491	-9.442	-9.487	-9.489	-9.488	-9.495	-9.442
x1 p-value	0.000	0.900	0.000	0.943		0.000	0.904	0.000	0.931	
x2 p-value			0.957					0.953		

TABLE A.5: APARCH execution results for Alphabet x1 p-value and x2 p-value describe statistical significance of sentiment features

Bibliography

Atkins, Adam, Mahesan Niranjan, and Enrico Gerding (2018). "Financial news predicts stock market volatility better than close price". In: *The Journal of Finance and Data Science* 4.2, pp. 120–137.

Baker, Malcolm and Jeremy C Stein (2004). "Market liquidity as a sentiment indicator". In: *Journal of Financial Markets* 7.3, pp. 271–299.

Baker, Malcolm and Jeffrey Wurgler (2007). "Investor sentiment in the stock market". In: *Journal of economic perspectives* 21.2, pp. 129–152.

Baker, Malcolm, Jeffrey Wurgler, and Yu Yuan (2012). "Global, local, and contagious investor sentiment". In: *Journal of financial economics* 104.2, pp. 272–287.

Bandopadhyaya, Arindam and Anne Leah Jones (2008). "Measures of investor sentiment: A comparative analysis put-call ratio vs. volatility index". In: *Journal of Business & Economics Research* 6.8, pp. 27–34.

Barber, Brad M and Terrance Odean (2007). "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors". In: *The review of financial studies* 21.2, pp. 785–818.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998). "A model of investor sentiment". In: *Journal of financial economics* 49.3, pp. 307–343.

Barberis, Nicholas, Andrei Shleifer, and Jeffrey Wurgler (2005). "Comovement". In: *Journal of financial economics* 75.2, pp. 283–317.

Black, Fischer (1986). "Noise". In: *The journal of finance* 41.3, pp. 528–543.

Boldi, Paolo and Sebastiano Vigna (2014). "Axioms for centrality". In: *Internet Mathematics* 10.3-4, pp. 222–262.

Brown, Gregory W and Michael T Cliff (2004). "Investor sentiment and the near-term stock market". In: *Journal of empirical finance* 11.1, pp. 1–27.

— (2005). "Investor sentiment and asset valuation". In: *The Journal of Business* 78.2, pp. 405–440.

De Long, J Bradford et al. (1990). "Noise trader risk in financial markets". In: *Journal of political Economy* 98.4, pp. 703–738.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the Econometric Society*, pp. 987–1007.

Fisher, Kenneth L and Meir Statman (2000). "Investor sentiment and stock returns". In: *Financial Analysts Journal* 56.2, pp. 16–23.

Han, Bing (2007). "Investor sentiment and option prices". In: *The Review of Financial Studies* 21.1, pp. 387–414.

Kalyani, Joshi, Prof Bharathi, Prof Jyothi, et al. (2016). "Stock trend prediction using news sentiment analysis". In: *arXiv preprint arXiv:1607.01958*.

Kumar, Alok and Charles MC Lee (2006). "Retail investor sentiment and return comovements". In: *The Journal of Finance* 61.5, pp. 2451–2486.

Malkiel, Burton G and Eugene F Fama (1970). "Efficient capital markets: A review of theory and empirical work". In: *The journal of Finance* 25.2, pp. 383–417.

Mao, Huina, Scott Counts, and Johan Bollen (2011). "Predicting financial markets: Comparing survey, news, twitter and search engine data". In: *arXiv preprint arXiv:1112.1051*.

Marchiori, Massimo and Vito Latora (2000). "Harmony in the small-world". In: *Physica A: Statistical Mechanics and its Applications* 285.3-4, pp. 539–546.

Mittal, Anshul and Arpit Goel (2012). "Stock prediction using twitter sentiment analysis". In: *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)* 15.

Nisar, Tahir M and Man Yeung (2018). "Twitter as a tool for forecasting stock market movements: A short-window event study". In: *The Journal of Finance and Data Science* 4.2, pp. 101–119.

Passalis, Nikolaos et al. (2019). "Temporal Logistic Neural Bag-of-Features for Financial Time series Forecasting leveraging Limit Order Book Data". In: *arXiv preprint arXiv:1901.08280*.

Prytula, Yaroslav (2005). "The impact of news of the stock market: a noise trader approach". In: *Academia* 17.

Qiu, Lily and Ivo Welch (2004). *Investor sentiment measures*. Tech. rep. National Bureau of Economic Research.

Schumaker, Robert P and Hsinchun Chen (2010). "A discrete stock price prediction engine based on financial news". In: *Computer* 43.1, pp. 51–56.

Sharpe, Steven A (2002). "Reexamining stock valuation and inflation: The implications of analysts' earnings forecasts". In: *Review of Economics and Statistics* 84.4, pp. 632–648.

Verheggen, Rick (2017). "The rise of Algorithmic Trading and its effects on Return Dispersion and Market Predictability". In: *Master Thesis: Tilburg University* 52.

Wang, Yaw-Huei, Aneel Keswani, and Stephen J Taylor (2006). "The relationships between sentiment, returns and volatility". In: *International Journal of Forecasting* 22.1, pp. 109–123.

Zhang, Xue, Hauke Fuehres, and Peter A Gloor (2011). "Predicting stock market indicators through twitter "I hope it is not as bad as I fear"". In: *Procedia-Social and Behavioral Sciences* 26, pp. 55–62.