SCHOOL OF
ECONOMICS AND
MANAGEMENT

MSc In Finance

NEKN02

Master Thesis

Spring 2021

Sentiment building from textual data content in quarterly reports

Author:

Tarek Jalessi

Supervisors:

Anders Vilhelmsson

Johan Källstrand

Abstract

Textual analysis is increasingly becoming a reliable tool for pattern assessing and forecasting especially statistically. Implementing such techniques in the financial field is still in an infantry stage and it represents a novel research area. While previous literature tends to rely on news articles as the source of data when examining the link with financial returns, internal information such as annual or quarterly reports drew small to no interest in academic research. This study aims to investigate the relationship between textual data content in quarterly reports and the corresponding stock price movement, to end up with a high-accuracy model when it comes to predicting the return. In this paper, we apply two different methods for sentiment building, first relying on python programming language and its lexicon-based approaches, second collaborating with Parlametric to establish document-specific sentiments. While the first method, combined with modeling done through MATLAB, lacked significance even with decent accuracy levels in sign prediction, the dictionary-free framework offered higher degrees of certitude. Our results confirm the hypothesis of an existing link between the asset returns and quarterly reports which is very encouraging especially for future studies.

Key words: Textual analysis, Python Programming Language, Modeling, Return Forecast, Sentiment building.

Acknowledgements

I wish to express my gratitude and appreciation for my academic supervisor Anders Vilhelmsson for his support, guidance, and inestimable help during the whole process. I would also like to thank my co-supervisor Johan Källstrand and Parlametric for the expertise and valuable input in this work.

Table of Contents

1.Introduction

Text is increasingly used as data in economic and financial research. According to Gupta, Dengre, Kheruwala & Shah (2020) text mining techniques are evolving within the financial industry and becoming more and more reliable as a tool to predict and explain ratios and returns. Traditionally Finance-related studies and research are conducted through numerical parameters and numbers to an extent where textual data and expressions have small or no value. Numerical methods could explain well some financial indicators if the companion variables are dependent in a good way. But results generated through such methods are not perfectly reliable in a fast-paced environment where small information could easily affect the overall conclusion. The ultimate question here would be if adding textual data parameters could enhance our prediction or inference.

Financial reports statistically contain more text than numbers, though most analysts pay small or no attention to this important part when conducting studies. We could easily agree that reading through the report enhances the understanding of the numbers and gives clearer insights into the document. We can say that textual data comes with huge importance in explaining historical performance and companies' expectations.

Textual analysis concerning sentiment building, through natural language processes and machine learning, is continually and gradually evolving. Though, usage of such data in empirical studies related to finance is still in an early stage of development and trial. Giving Content to Investor Sentiment. Sentiment building could be done through several methods. As presented by Tetlock (2007) the study is relying on the Harvard dictionary, where a subset of words follow a given pattern or predefined way into sentiment building. Tracy Ke, Kelly, Xiu (2019) Followed a different method. The research creates a unique dictionary based on the frequency and usage of words according to the bag-of-words technique.

This paper aims to establish a model that could be used as a reliable path to predicting returns based on quarterly text data. Through transforming raw text and establish a frequency-based pattern, as well as exploiting historical data as both training and testing data. Working with 10Q reports extracted from the Edgar platform, and using python to apply machine learning, we make it simpler since such reports are somehow following the same patterns which facilitate the task of sentiment building. This project will incorporate 50 companies listed in S&P 500 stock market and investigate the link between terms count and frequency to first explain and then predict returns. To our knowledge sentiment prediction relying on textual data provided by companies such as quarterly reports has not been conducted yet. We aim to create our specific dictionary that is naturally done through the usage and frequency of words. This would be done following world count and normalization regarding the size of the sample to be able to contribute to this new upcoming novelty and we aim to provide a reliable model that could predict accurately the returns and build sentiment.

2.Literature Review

2.1 Qualitative & Quantitative data

An establishment of a link between reports' readability, structure and vocabulary, and financial performance was suggested through different research studies (Subramanian, G. Insley & D. Blackwell, 1993). As a continuation within the scientific scope and seeking more accuracy, a combination of textual and numerical analysis has been done by Kloptchenko, Back, Vanharanta, Eklund, Karlsson& Visa in 2004 to predict financial returns.

As the same article puts as a starting point how companies subconsciously reveal more than what they intend to through textual parts in the reports. While explaining this study we would focus on the qualitative parts since it relates the most to our research topic. This was done through prototype-matching text clustering methodology (Visa et al. 2001). The technique consists of a transformation of every text through numerical encoding and thus the establishment of a common word histogram and interpret it through a Weibull common distribution. After quantization, a density resolution is created for the most used vocabulary. The same method is then used to vectorize sentences. This representation allows comparison between companies following the Euclidian distances.

The study on a final step compares the individual results and states the obvious fact "sentence construction" impacts more the overall clustering. On a higher level, the research concludes that textual data is a bit futuristic in the prediction when compared to the quantitative data which is a fair reflection on past performances. The results prove some clustering similarities between text, vocabulary, and quality of reports between a company and its higher peers when management predicts higher returns. This study also proves that comparing reports and lexical clusters within companies could be a reliable source of analysis and prediction.

A more accurate predictive model was proposed by Jaybhay, Argiddi & Apte (2012). The study combines both classical methods such as time series, technical and fundamental analysis, and data mining.

While these methods go hand in hand toward the establishment of a clear conclusion, we are going to focus on the textual analysis part. The latter method is conducted using public financial news articles retrieved from Yahoo finance, stock watch, and Reuters, to forecast daily stock prices in the Bombay stock exchange with a training period of 4 years from 2008.

The research paper uses the nearest neighbor method as well as the artificial neural network with three layers. This trains the model to establish a link between the presence of some key phrases or sentences and the given stock quotes. The results are then presented as an approximation of the closing price for the next trading day.

Combining machine learning techniques with the “back-propagation” algorithm, with the traditional methods proved to be accurate with a 94% precision rate daily.

The previously cited studies illustrate how news textual analysis improves the overall model and enhances the prediction. This has also been proved by Khedr, Salama & Yaseen (2017). In their research paper, they combined sentiment analysis and historical numerical data intending to predict stock returns’ sign for the top three Nasdaq-listed companies.

The first step is done through the Naïve Bayes framework and machine learning techniques that after several transformations assign a given sentiment to the news article. Dividing the data sample into training and testing subsets allows for link establishment as well as stock prediction.

The model comes up with results that are aligned with the previous literature hypothesis of a strong link between news articles and analysts’ reports and the stock movements. The accuracy of the prediction is enhanced from 59% to over 89% by the combination of the two data types.

2.2 Dictionary-based analysis

When looking into textual analysis, the first tool that comes into minds is the General Inquirer[1] (GI). This program is built through a unique set of procedures into the creation of sentiment analysis (stone et al,1966). This is continuously enhanced with a combination with the Harvard IV -4, Laswell value dictionaries, and other categories. This tool allowed researchers to investigate sentiment analysis based on natural language processing.

When it comes to sentiment analysis one major publishment was done by Tetlock (2007). This analysis relied heavily on the GI analysis with the Harvard dictionary. The study is about predicting stock returns through news articles and specifically the Wall Street Journal "abreast of the market" column. Multiple transformations are done, and the sample is divided into two components to explore patterns and then test how reliable the overall analysis is. The key part is the establishment of pessimism factors and relating them using series of regressions with volumes and returns. It also accounts for the timing and specific day of the week to predict the trader's behavior. With a sample size of 16 years, this is mainly done through standardization and the use of dummy variables for each day of the week.

The Harvard IV-4 dictionary assigns predefined sentiments to each written word and could be used to establish sentiment analysis. With the huge amount of natural language, some vocabulary is under-estimated and sometimes would pass without being considered.

[1] An automated program that converts textual data into numerical ones based on frequency.

The main findings of this study are the heavy link between pessimism indicator and downward movements for the stock prices. This sign of movement would be almost irreversible or slow to reverse, which is aligned with how negative news has a higher impact on the magnitude which leads to an increase in markets' volatility.own

The overall goal of this research was the establishment of a winning portfolio since stock prices are predicted. The previously cited findings and the high volatility component would make it difficult for such strategies to take place since it requires high turnover and thus higher costs that could surpass the possible gains.

Further studies concerning the lexicon-based methodologies were conducted by Heston & Ranjan (2017). They ran a comparative work between the previously cited Harvard dictionary, Loughran & McDonald "LM" (2011), and the neural network. They based their research on 900,754 news articles from Reuters for a 7 years-period from 2003. The overall aim is to construct sentiment prediction with the three lexicons and illustrate which method is more accurate.

Several differences were observed in terms of both the magnitude and the sign of the forecasted return. These differences are due to the relevance and context especially when it comes to the LM dictionary since it is specially adapted with the financial scope. The conclusion also shows that the overall effect also depends on the occurrences and dispersion of the news.

2.3 Adaptative Framework

In contrast with the previously cited methods, Tracy Ke, Kelly, and Xiu (2019) proposed a new framework that does not rely on any dictionary under the name of "SSESTM"[2] , with a clear purpose of predicting stock returns from financial news articles.

This study was conducted with data consisting of over 10 million articles for a period of 28 years based on financial news retrieved from Dew Jones newswires and the mentioned companies. For comparative reasons, a portfolio is constructed with a simple strategy relying on the stock prediction.

The data is divided into a training sample where correlations are tested, and a validation sample where the sentiment or dictionary created would be used as a reliable model to predict asset returns.

This methodology consists of three basic steps: starting with a bag-of-words count and then isolation of the "most relevant features" relying on correlation screening. Thus, selecting or creating lists of words depending on its usage compared with the financial return sign. Then using frequency-based analysis, sentiment weights are assigned. The last step is to adapt the scores into article-based ones using an internal likelihood structure. These techniques emerge towards an establishment of an adaptative analysis depending on the terms used in each article and the results are standardized with regards to the length and context.

This way differs from the Harvard dictionary, which according to the article underestimates the overall sentiment building by not incorporating some meaningful words. This has been proven for both directions when high-frequency words could not be assigned to the

[2] Supervised sentiment extraction via screening and topic modeling.

dictionary. We end up with a model that is performing on a higher level than comparative methods such as the Raven Pack[3] dictionaries when it comes to predicting stock returns.

[3] Leading data analytics platform.

3. Hypothesis development

Previous studies have suggested a link between textual data content on media and financial returns. Strong evidence was found to prove the existence of tight links between pessimism and the decrease in financial returns. This is basically due to the rationality and the psychology of the trader's behavior (Tetlock, 2007). This is consistent with the classical financial assumption that risk aversion comes within human nature so bad news is likely to create pessimism and thus participate in a downward tendency of the financial returns.

We believe that media has a decent effect on the overall market and especially when it comes to stock returns since it consists of an outside addition in terms of news and sometimes helps cover some information that companies try to hide. This is natural that within firms, news perception is predicted, and management would try to spread it out in the best way possible to decrease the harmful effect.

The continuing development in biological science and its combination with information technology proved that sentiment could be passed subconsciously through textual data (Jaybhay, Argiddi & Apte, 2012).

Textual analysis has been increasingly used in the financial field and it has been proved to be linked to the trading attitude and thus to the returns. This research studies the link between the textual part of the quarterly reports and the financial returns of the corresponding companies. We are aiming to test whether the historical information could be a reliable way to build a relationship and to establish a given pattern to generalize these findings into a forecasting level.

While we are looking into this pattern, quarterly reports could be part of the news set surrounding each company. We believe that a wide range of variability exists in this news flow. For some industries that are continuously covered, and dynamic, such reports could be neglectable and have a minor effect concerning the trading behavior. On the other hand, for some industries, it could be the only source of information and thus its effect would be stronger and more reliable.

Our Hypothesis suggests that quarterly reports are highly reliable and could generate adequate inferences when it comes to stock returns especially for industries with low media coverage.

4. Methodology

4.1 Methodological introduction

Building sentiment analysis relying on textual data in the financial field is usually done in two macro-steps: Assigning rates and scores to each article, text news, or report then establishing a correlation link by modeling the relationship between historical data and financial return.

As mentioned in the previous literature section, there are two schools of thoughts when it comes to sentiment analysis: Pre-defined dictionaries (E.g. Tetlock, 2007; Heston & Ranjan, 2015) and Adaptative framework (E.g Tracy Ke, Kelly, and Xiu, 2019)

Our study incorporates two methods to build sentiment and try to create a reliable inference for stock returns.

The first methodology implemented during this study consists of two steps: First, assigning scores for each of the predicting variables (Positivity, Negativity, Subjectivity, and polarity) relying on the Loughran and McDonald Financial Sentiment Dictionaries (LM). Second, the establishment of a pattern relating these scores to the assets returns.

The second approach is building a company-specific dictionary based on word quantification and then establish a pattern linking highly used terms to changes in stock prices.

4.2 Data

4.2.1 Edgar

EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, is a data collection system in the United States where companies are required by law to file all reports in a standardized format. This filing system consists of different types of forms such as form10k, form10q, form8k, Forms 3,4, and 114.

For this study and for quantification reasons the “10q” forms were picked since it showcases quarterly reports presenting a more accurate way to predict financial returns rather than the classical yearly shareholders’ report. The overall structure of the files is quite similar allowing us to have a better comparative analysis between companies and their use of words.

4.2.2 Sample description:

Dealing with textual data implies accounting for language barriers and considering them. For this exact reason, our analysis is done exclusively with English-written texts. We picked the US market to test our hypothesis and conduct our research. A list of 50 companies listed in the S&P500 market was picked. The selection criteria were the performance and the variability of industries. Table 1 presents some of the firms with their corresponding sector of activity.

Table1: Sample of companies and sectors:

Symbol	Company name	Sector/Industry
ABT	Abbott Laboratories	Health care
CAN	Accenture	Information Technology
ATVI	Activision Blizzard	Communication Services
AAP	Advance Auto Parts	Consumer Discretionary
AES	AES Corp	Utilities
AFL	Aflac	Financials
APD	Air Products & Chemicals	Materials
ARE	Alexandra Real Estate Equities	Real Estate

This research study aims to implement several machine learning techniques to interpret and analyze current trends as well as trying to explain them through textual data. This was the rationale behind the test period of ten years starting from 2011 which gives us a sample of around 1500 documents.

4.2.3 Data collection

Relying on the Python programming language, after selecting the companies list, an algorithm was created to extract the 10Q reports. This was done through Pyedgar[4]. This tool extracts the URL link of the report and then downloads it. Following this pattern, we were able to store 1830 files for our given 50 companies.

Through the Eikon Reuters Platform, a list of historical prices for each company is created. We are only interested in the closing prices and we only selected the month-to-month prices.

[4] A built-in package allowing the extraction of data from the EDGAR platform.

The time horizon of the price scope is limited to 10 years by the already cited platform. This gives us a wide range of prices allowing us to calculate the return on a simple equation as follows:

$$Return = \frac{This\ Month\ closing\ price - Last\ month\ closing\ price}{Last\ month\ closing\ price}$$

4.3 Sentiment building with dictionary:

4.3.1 Sentiment creation

Another key feature in Python was the Pysentiment2[5]. This tool is dedicated to textual analysis and it contains several dictionaries such as Harvard IV and LM dictionaries. Since we are dealing with financial context, we judged that the latter is more accurate.

The establishment of an efficient algorithm allowed for the construction of one excel file containing 50 sheets featuring the filing dates and the assigned scores for each report. Between these steps, python performs tokenization, which is a required transformation of the textual data allowing to substitute each given word with its root one. This pathway is crucial to build sentiment and construct an accurate pattern.

Tokenizing is quickly followed with score assignments relying on the occurrences and word count. Our sample is then cleaned in a way to remove outliers and companies having incomplete data sets. Our first observation once the output is created is the high variability between companies and even within the same firms, concerning the filing date. For that specific reason, we decided to predict next month’s return. This was done through the creation of new personalized columns with rounded-up dates to reach the next closing date.

[5] A built-in tool allowing for sentiment analysis of textual data.

The combination of the previously cited steps allows for a generation of structured tables as demonstrated by table 2.

Table2: Sample of Generated table:

Closing date	Return	Sign	Positivity	Negativity	Polarity	Subjectivity
31/05/2011	0.049	1	304	4978	-0.884	0.020
31/08/2011	0.100	1	610	2572	-0.616	0.006
30/11/2011	0.081	1	742	3089	-0.612	0.005
31/05/2012	-0.019	0	659	2952	-0.635	0.006
31/08/2012	-0.050	0	755	3224	-0.620	0.006
30/11/2012	-0.014	0	724	3136	-0.624	0.003
31/05/2013	0.019	1	669	3059	-0.641	0.004
31/08/2013	-0.029	0	1160	14629	-0.853	0.013
31/08/2014	0.074	1	482	2944	-0.718	0.003
30/11/2014	0.018	1	455	2842	-0.723	0.003
31/05/2015	-0.024	0	432	2628	-0.717	0.003
31/05/2015	-0.024	0	376	2856	-0.767	0.004
31/05/2015	-0.024	0	440	2273	-0.675	0.006
31/08/2015	0.005	1	524	2902	-0.694	0.005
30/11/2015	0.007	1	516	2837	-0.692	0.005

Table2 presents a sample of quarterly reports scores per category. (1) **The closing dates** (selected as the end of the next month following the publication date). (2) **The sign** (a binary variable that takes the value of one if the financial return increased compared to the previous month).

4.3.2 Regression & Modeling

As we successfully transformed our textual data into numerical values, we are basically in a common situation of several variables and one dependent variable. The simplest way to define regression is a series of statistical processes looking for a certain pattern linking one variable to other independent predictors. Since our sample contains different companies, we opted for a time-series analysis investigating the movement of the return according to the given scores.

Due to the complexity of the process and the time limitations, our prediction would be focusing on the sign of the change in terms of return. For this specific step we constructed a binary model where an increase is expressed by 1 and following the same logic, the decrease is expressed by 0.

The following formula is the mathematical representation of the equation created through the ordinary least squares (OLS) regression analysis.

$$Sign = \alpha + \beta 1t\ P + \beta 2t\ N + \beta 3t\ Pl + \beta 3t\ S + \varepsilon t$$

Where:

α: is the intercept

P: is the positivity score

N: is the negativity score

Pl: is the polarity score

S: is the subjectivity score

ε: is the Error term.

t: Closing date.

For accurate modeling of the assembled data, we ran multiple types of regressions. The rationale is to find the most accurate model that fits each company. This is mainly done through MATLAB software, where we performed four different types of regressions and for each one, we built a confusion matrix to present the correct predicted fractions.

For each model we run the regression, then relying on the fitted values we try to predict the sign of the returns. While we have the correct data, we can easily check the correctly inferred results.

The four regression models we built our estimation around are the ordinary least squares (OLS), logistic (LOGIT), linear discriminant analysis (LDA), and quadric discriminant analysis (QDA).

The LDA and QDA approaches are implemented with categorical variables. For our case, we have two classes which are 0 and 1 depending on whether an increase in the return has occurred. The LOGIT model directly extracts the probability of a given observation belonging to a particular class without making any assumptions on the distribution of the predictors' values. The obvious difference between both LDA and QDA and the logit that the probability's extraction is built in two steps: First, the sample is studied and an assumption about each class's distribution is made. Then, relying on Bayes theorem[6] the probability is calculated.

The logic is the same and both methods imply normal distribution. While the linear discriminant analysis assumes equality between variances among all classes, the quadric discriminant analysis opts for different covariances between the classes.

Since the LOGIT is limited to a binary representation and LDA and QDA work for 2 or more classes, usually the latter present a higher level of performance with categorical classification.

4.4 Adaptative framework

As we performed this study in a collaboration with PARLAMETRIC[7], this step is the main component of this partnership. For this step, we aimed to enhance the analysis of the overall research by focusing on a list englobing both the top and worst 10 performers. This limited the sample size to 612 reports.

The data set was analyzed through Parlametric's platform for narrative analytics. This software works mainly with PDF files as input, which we previously retrieved using the Python programming language. The framework allows for word quantification and

[6] Mathematical representation to calculate the conditional probability given the previous outcome.
[7] Award-winning tech company from Lund in Sweden

normalization regarding the size and structure of the reports. A list of the most occurring 100 words is created for each specific company. Since the sample presents companies from different industries and sorts the stock prices with different periods, the linear normal distribution could not be a reliable way for the analysis.

Consequently, each list would have 100 words, and the rank is used as the scoring system giving 100 for the highest occurred word and 1 for the lowest. For the next step, Mann Whitney U-test is performed on these word-frequency lists to retrieve the most significant words.

This test allows for comparison between two independent groups and investigates the differences. This statistical test would explain the differences between good and bad performers in terms of word count and checks whether that explanation is significant.

These terms would be the engine of our research. Since the formula previously implemented to quantify word-frequency is back-engineered into predicting the returns knowing the score.

4.5 Delimitations

Due to the time-factor limitation, and the complexity of the subject, we find delimitations to be unavoidable. First, we find the sample size to be relatively small and could be not representative enough. Though, the processing and computational times, and power made it harder to collect a larger sample. Second, while we are mainly interested in predicting the sign of the return, we omitted some key aspects around the variables. We did not test for the predictors' characteristics such as Heteroscedasticity and Multicollinearity. We believe the effect of such features is minor and since we are essentially trying to establish some inference pattern this would not be problematic.

The third delimitation would be around the price returns and the closing dates picked. While we certainly recognize the importance of time horizon when it comes to predicting returns and accounting for market liquidity, we were not able to generate the day-to-day prices and perhaps assess better the overall changes. This might reduce the accuracy of our predicted changes, but time pressure and the data set made it undoable.

Even though the LM dictionary is finance-specialized, we believe following an adaptative framework might be a better way to create subconscious links and thus generate more accurate results. This requires high computational power and time dedication.

We conclude this part by stating that even though, the textual analysis might be a great indicator to forecast future returns, combining it with a classical numerical analysis such as the fundamental one would create a more accurate pattern that might participate in the construction of a well-hedged portfolio.

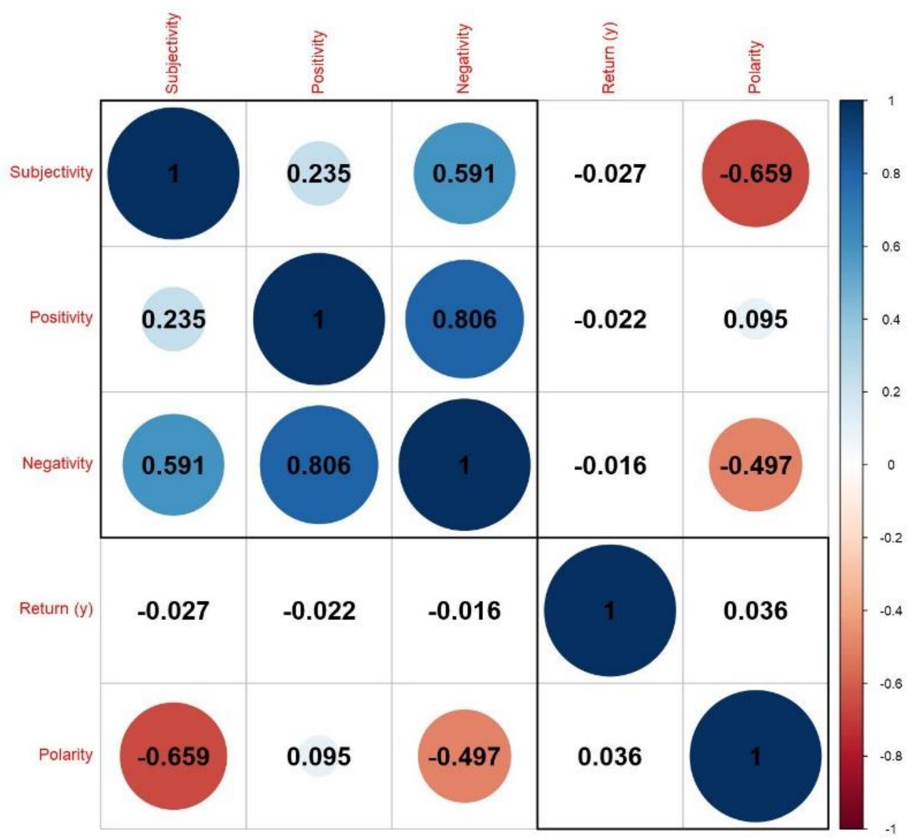
5. Empirical results

5.1 Lexicon sentiment analysis approach

5.1.1 Correlation

Chart 1 illustrates the correlation matrix for Alliant Energy (LNT) . The results prove the existence of a small correlation between return and the independent variables. This sample is somehow generalizable since all companies presented correlations between -0.2 and 0.2. This shows almost an absence of causality between the scores and the return.

Chart 1: Correlation Matrix n°1



Note: Chart1 presents correlation values for both dependent and independent variables. The fourth raw or column reports the values regarding relationship between (1) **Return** (dependent variable: financial return realized at the end of the month) and (2) **Subjectivity**, (3) **Positivity**, (4) **Negativity**, and (5) **Polarity** (Independent variables: scores generated through Python programming language and Loughran and McDonald financial dictionary.

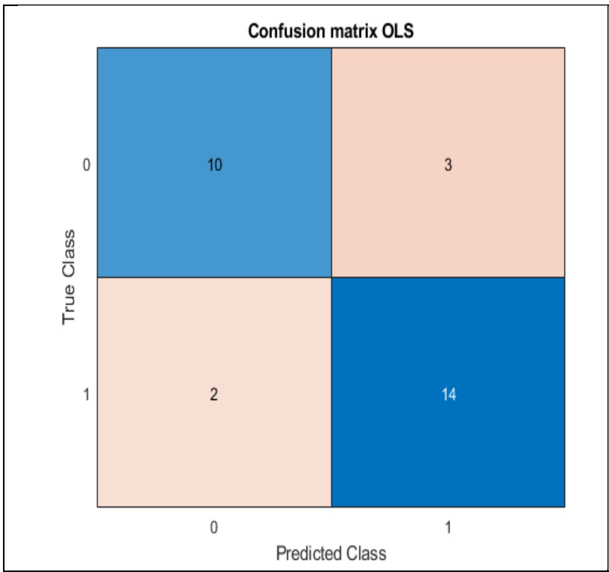
Another important finding in this chart would be the high correlation that surpasses 0.8 between positivity and negativity. This is an illustration of how management tends to compensate for negative information by adding positive words and generating more positivity into the quarterly reports. This also goes back to our delimitations about Multicollinearity.

5.1.2 Model selection

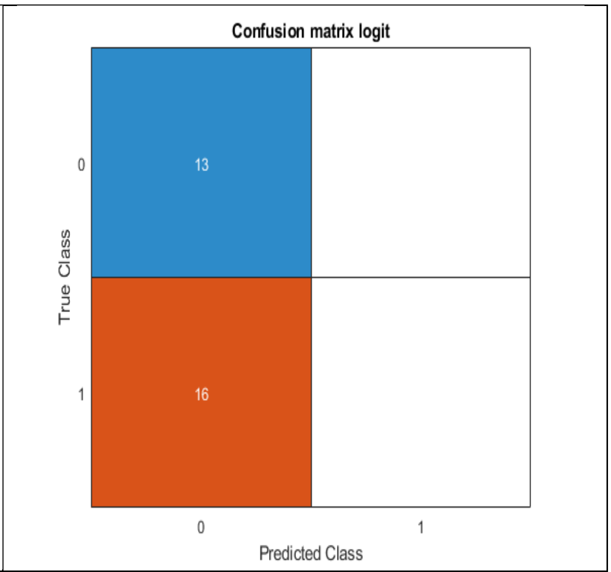
As explained in the methodology section: our results are generated through MATLAB. We heavily rely on the confusion matrix to confirm the best model that fits the data. This matrix is a performance classifier that showcases how well the prediction goes.

For illustrative purposes, we picked ABC company to interpret its results. Table 3 englobes the result of the four models and presents the efficiency in terms of forecasting.

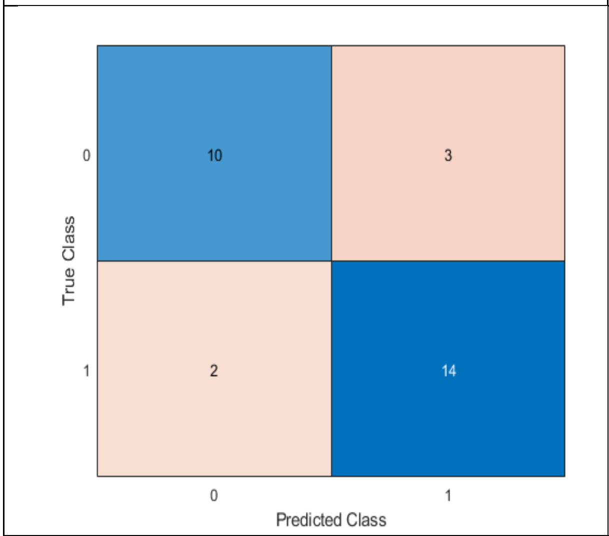
Table 3: Confusion Matrix



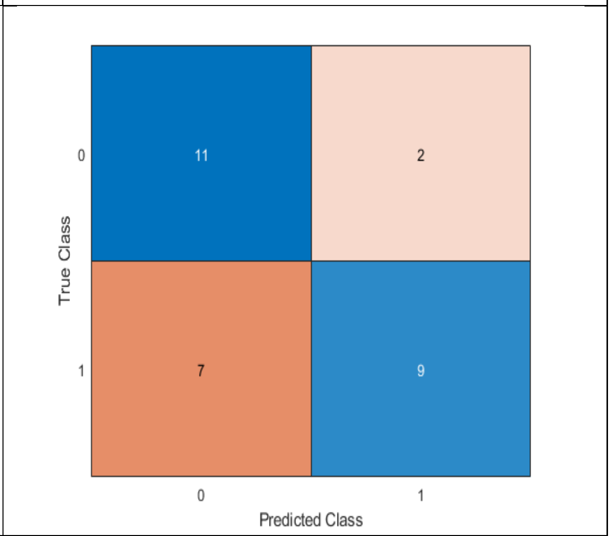
OLS



LOGIT



LDA



QDA

Note: Table3 reports (1) **the confusion matrixes** (the diagonal cases present the correct predicted fraction per model, generated through MATLAB software), (2) **OLS** (ordinary least squares regression model), (3) **LOGIT** (logistic regression), (4) **LDA** (linear discriminant analysis), and (5) **QDA** (quadratic discriminant analysis).

This specific example illustrates that the Ordinary Least Square method is the most suitable model that fits the actual data and was able to correctly predict 83% of the returns' signs.

With that being said, we must mention the wide variability between the best model when it comes to different companies. Tables 4 summarize the results of the top 10 companies in terms of return for the selected period.

Table 4: Results summary of the top 10 companies

Company	Total return	Industry	OLS	LOGIT	LDA	QDA	Significance
APA	0.54	Energy	0.59	0.48	0.59	0.66	no
ANET	0.38	Information Technology	0.80	0.35	0.80	0.85	no
ARE	0.30	Real Estate	0.65	0.42	0.65	0.65	no
ABMD	0.19	Health Care	0.83	0.45	0.83	0.69	yes
GOOGL	0.18	Communication services	0.75	0.50	0.75	0.88	no
GOOG	0.18	Communication services	0.75	0.50	0.75	0.88	no
ANTM	0.14	Health Care	0.67	0.57	0.67	0.63	no
AIG	0.12	Financials	0.60	0.53	0.63	0.60	no
ANSS	0.10	Information Technology	0.69	0.52	0.69	0.69	no
MO	0.10	Consumer staples	0.61	0.58	0.61	0.55	no

Note: Table 4 reports total return realized per company for the top 10 performers and the correct fraction of sign prediction. The significance column returns if the results are significant.

This list contains companies with positive cumulative returns for the sampling period. For each regression method, we presented the correct fraction estimated by the model and the significance level of all of them. The results confirm the efficiency of the OLS, LDA & QDA models with small variability for every one of these companies. The logit failed to reach the 50% barrier four times with this sample. This might be because this model lacks effectiveness and does not take into consideration the type of distributions for the given classes.

Such high numbers are very promising since we could predict the sign with high accuracy reaching up to 88%. While the numbers are high enough to make a correct forecast, the results are limited to the sample level since one out of the ten companies presented significance for the coefficients. This might be due to the low number of predictors or the sample as explained in the delimitations part. These same results were found when testing with the worst ten companies when it comes to cumulative returns.

Table 5: Results summary of the worst 10 companies

Company	Total return	Industry	OLS	LOGIT	LDA	QDA	Significance
AES	-0,20	Utilities	0,71	0,46	0,71	0,61	no
AAP	-0,21	Consumer Discretionary	0,67	0,43	0,63	0,47	no
AMP	-0,22	Financials	0,69	0,36	0,69	0,67	no
ABC	-0,31	Health Care	0,86	0,61	0,86	0,61	no
ALXN	-0,32	Health Care	0,43	0,53	0,43	0,63	no
ALL	-0,35	Financials	0,71	0,65	0,71	0,71	no
ATVI	-0,37	Communication services	0,67	0,43	0,67	0,67	no
AIZ	-0,47	Financials	0,68	0,68	0,64	0,68	no
AEP	-0,50	Utilities	0,71	0,40	0,71	0,58	no
ADBE	-0,50	Information Technology	0,80	0,33	0,80	0,77	no

Note: Table 5 reports total return realized per company for the worst 10 performers and the correct fraction of sign prediction. The significance column returns if the results are significant.

5.1.3 Regression

In a continuation with the presentation method, a sample of one OLS regression result is illustrated in table 6. This example is not randomly selected since Abiomed (ABMD) is the only company in our given sample that contained significance.

While all predictors have small to neglectable coefficients, subjectivity is the most important factor by far with 213.37 compared to close around zero values for the rest.

This shows how the subjectivity parameter is small for the reports and demands high numbers to quantify it.

Economically speaking a company's quarterly report would have a very small amount of subjectivity. It represents a sense of belonging and embracing of the performance.

Table 6: ABMD regression Results:

	Estimate	SE	T-stat	P-Value
(Intercept)	-2.3633	1.0889	-2.1864	0.0387
Positivity	0.0007	0.0002	2.4471	0.0221
Negativity	-0.0003	0.0001	-2.0529	0.0511
Polarity	-2.981	1.1821	-2.5219	0.0187
Subjectivity	213.37	101.03	2.1119	0.045

Number of observations: 29, error degrees of freedom: 24

Root Mean Squared Error: 0.475

R-squared: 0.246, Adjusted R-squared: 0.121

F-statistic vs. constant model: 1.96, p-value= 0.133

Note: Table 6 Presents the OLS regression results for (1) **ABMD** (ABIOMED company).

5.2 Adaptative approach

5.2.1 Significant words

As previously stated, this whole method looks into finding an established pattern that could explain the differences between both good and bad performing companies. The following table 7 presents the results generated by Mann-Whitney U-test.

Table 7: Mann-Whitney U-test between Top and worst 10 performers.

A	B	A	B	A	B
1	66	34	1	1	1
3	19	1	1	1	9
1	1	1	1	29	63
49	51	1	1	1	58
1	1	70	1	1	1
1	26	1	1	1	66
1	6	47	31	70	14
33	41	1	1	1	63
1	1	1	1	38	27
1	16	70	1	1	46
shares		Billion		Credit	

A	B	A	B	A	B
1	1	1	1	55	50
43	24	1	41	21	54
1	1	1	71	31	46
10	1	1	50	57	55
30	8	26	51	43	61
47	1	62	20	57	80
38	1	45	43	26	59
1	1	1	72	56	75
32	48	42	6	50	20
30	1	26	47	43	66
Certain		Due		Statements	

Note: Table 7 reports (1) **Mann-Whitney U-test** (Statistical and mathematical tool that compares differences between two sets of observations), (2) **Shares**, **Billion**, **Credit**, **Certain**, **Due**, and **Statements** (Most used terms that are significant at 10%).

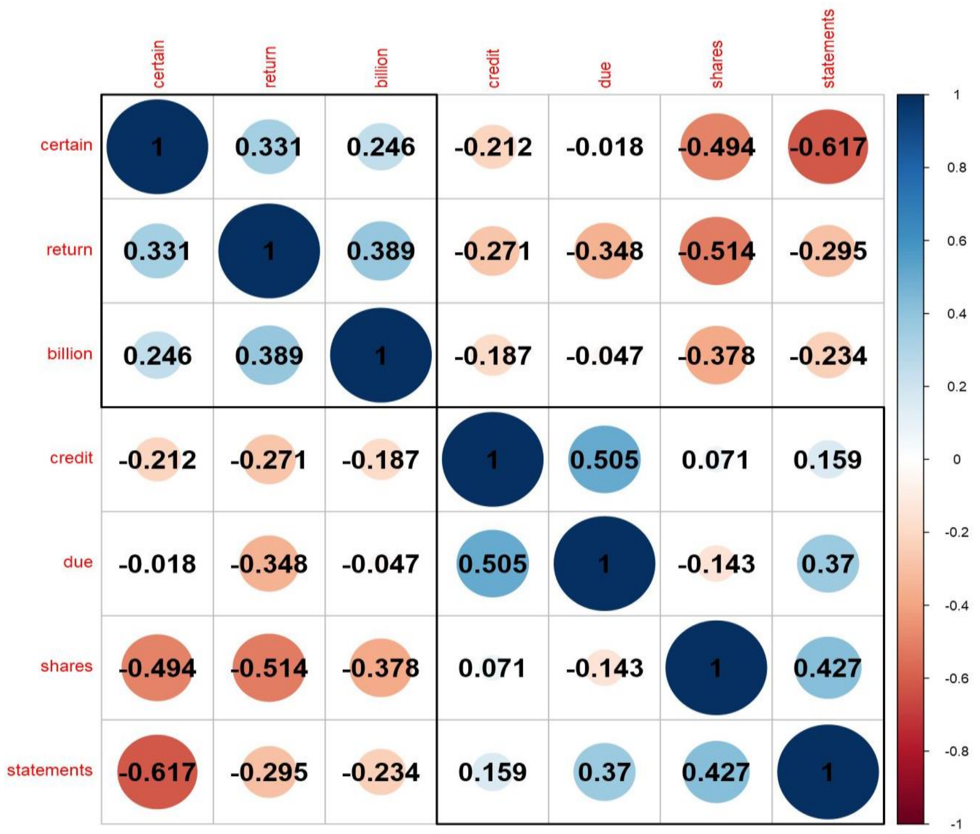
All these results have P-values smaller than 10%. This list of words is the most significant and has the highest impact on the dependent variable (financial return).

5.2.2 Correlation

Chart 2 presents the correlation between the highest 6 words and the corresponding company's return. This shows very interesting findings since we could clearly distinguish that the words "billion" and "certain" have positive correlations with numbers around 0.35 while the words “shares”, “credit”, “due”, and “statements” have negative correlation numbers and therefore the most occurrences they have the lower the stock return would be.

When mixing the six words we end up with a correlation of 0.7 which is very encouraging and mirrors a strong link between the words and the financial returns.

Chart 2: Correlation matrix n°2

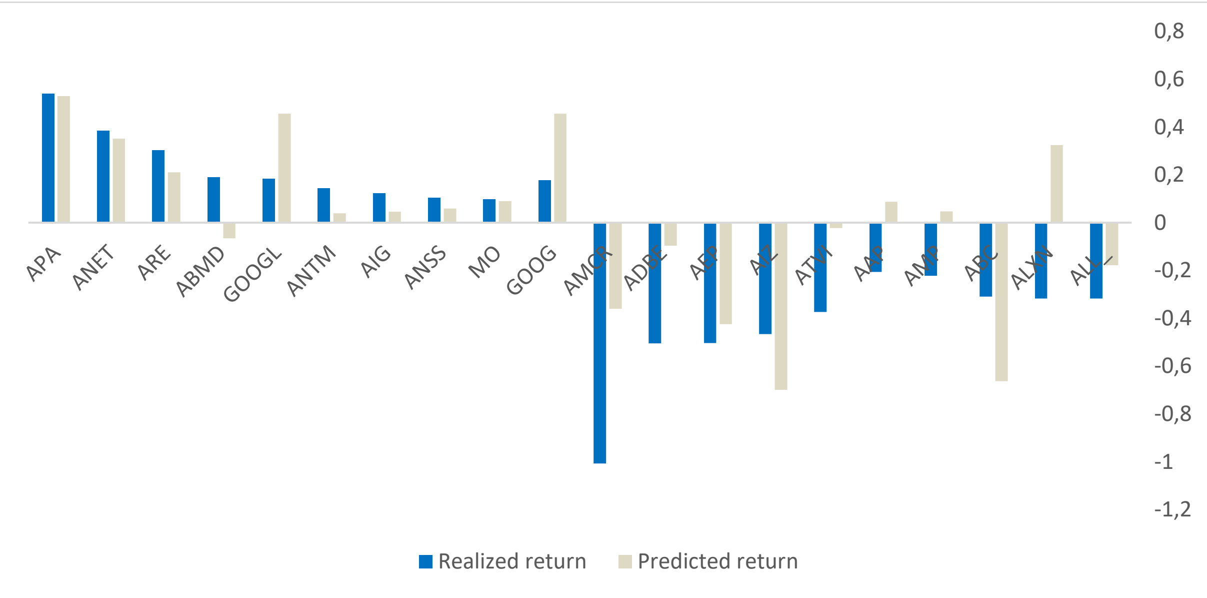


Note: Chart2 presents correlation values for both dependent and independent variables. The second raw or column reports the values regarding relationship between (1) **Return** (dependent variable: financial return realized at the end of the month) and (2) **Certain**, (3) **Billion**, (4) **Credit**, (5) **Due**, (6) **Shares**, and (7) **Statements** (Independent variables: Most significant words used in quarterly reports).

5.2.3 Prediction

The regression for this model relied on constants for each of the words to reach the highest total prediction. This perfect fit iteration is used for illustrative reasons and to look for the maximum fit between the terms and financial return. After scaling back the words a list of predicted values of the stock price change was established. Chart 3 showcases the differences between the total outcome generated and the estimated one.

Chart 3: Actual and predicted total returns



Note: Chart3 is a histogram presentation of the predicted returns compared to the realized ones.

This graph sets a very clear view of how close this analysis predicts the financial returns. Surely, the model varies a bit with 75% of underestimating the total gain or loss, but this number is very encouraging and most of all the results are significant.

6. Discussion

6.1 Analysis & discussion

This study is built around the hypothesis of an existing relationship between the 10 Q, quarterly reports, and the financial returns. Our motivation is to construct a decent model that allows sign forecasting based on textual analysis. We went further with our hypothesis and predicted that the importance of such files varies between industries and thus we should have different values and accuracy levels between the selected sectors.

Following a coherent logic, we expected the utility sector to have the best numbers when it comes to forecasting the financial returns and that is due to the lack of media coverage. This means that we also expect low predictability for technological and financial companies.

As shown in the previously explained table 5, there is no remarkable differences between the sectors, and the models correctly predict the sign with a 70% accuracy. Adding to that, the high p-values generated to each model, we must state that not only the given sample do not contain enough evidence to support the hypothesis, but its results are also not generalizable on a larger scale.

This lack of significance might be due to the reduced sample size since 50 companies might not be highly representative. In addition to that the next month's closing price date, while it is convenient, might miss out on other events happening during a whole month. As explained before this was selected due to the time pressure.

If we focus on the second method implemented, we can clearly state that the adaptative approach offers more accuracy in terms of predicting the sign since 80% of the change's directions were correctly estimated. This analysis goes an extra step and aims to predict the financial returns. The findings were close enough to realized ones.

When comparing both methods we cannot correctly assess the differences and perhaps the list of words generated could be underestimated by the LM lexicon since the latter did not show a significant level when only predicting the sign.

Within the financial atmosphere and when digging for findings around the asset prices, we need to emphasize the behavioral aspect that represents the engine of the whole prediction methods. As we cannot omit the human factor, this whole study was based on the assumption of a complete market and that the financial stock prices reflect well enough the published information. We cannot say that this assumption was respected since different results were realized.

We are in favor of Parlametric's frame and their adaptative system that allowed for a better understanding and thus higher accuracy in terms of prediction as well as presenting highly significant predictors.

6.2 Future research

The data-mining field is continuously improving and enhancing. The first question that we tried to answer was how well we could predict the financial returns based on textual analysis. While the study was done on a small sample with only 10 years as a prediction timeframe, we were able to generate decent models and predict the sign of the returns but with no significance.

This leads us to think that as long as the textual analysis is interesting and inference could be made upon it, the selected source of information might not be highly impactful on the traders' actions. This might be due to the level of complication and choice of words within the selected reports.

This urges our minds to look into other sources of textual data that could impact the financial market and thus would help create a highly profitable portfolio. This new era is built around technology and social media. The latter is highly used daily and contains a high amount of natural language data.

Our next research would study the textual data content on social media platforms and look into a relationship or a causality effect on the financial returns. This would be a highly impactful step towards the construction of strong recommendations with a high degree of accuracy and thus create a smart trading strategy that allows for profitable portfolio construction.

7. Conclusion

While there was no clear distinction between the models in terms of data-fitting, the ordinary least square regression was highly accurate and effective. This Illustrates that a linear representation between a binary dependent variable (Sign) and the four predictors (Positivity, Negativity, Subjectivity, and Polarity), exists and could be used to infer the stock movement. This combination converges toward decent results and how we could link the returns to the presented criteria. This showcases a relationship of causality between the textual data and the financial returns.

As we cannot forget that we are looking into a human reflection upon these results and the trading behavior tends to follow some trends and the resulted reactions differ when it comes to the perception and interpretation of the published reports.

Most of the previous literature relied on the media news texts and showed a strong link and market reflection of the articles' sentiment (pessimism, negativity, positivity…). Unfortunately, relying on the LM dictionary concluded that there was no level of significance for the predictors for the results to move from data fitting into sing inferring with confidence. Therefore, creating our adaptative dictionary and accounting for the frequently used words enhanced the prediction level and established a clear pattern for the financial reports. These findings are encouraging since with more time-horizon, and thus a bigger sample we could reach and establish high-performing portfolios.

References

C.Tetlock, P., 2007. Giving Content to Investor Sentiment:. *The journal of Finance*, 62(3), pp. 1139-1168.

Gupta, A., Dengre, V., Kheruwala, H. A. & Shah, M., 2020. Comprehensive review of text-mining applications in finance. *Financial innovation,* 6(1), pp. 1-25.

Heston, S. L. & Heston, N. R. S., 2017. News versus Sentiment: Predicting Stock. *Financial Analysts Journal,* 73(3), pp. 67-83.

Jaybhay, K. M., Argiddi, R. V. & S.S.Apte, 2012. Stock Market Prediction Model by Combining Numeric and News Textual Mining. *International Journal of Computer Applications,* Volume 57(19), pp. 16-22.

Khedr, A. E., S.E.Salama & Yaseen, N., 2017. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *I.J. Intelligent Systems and Applications,* Volume 9(7), pp. 22-30.

Kloptchenko, A. et al., 2004. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance and Management*, 12(1), pp. 29-41.

Stone, P. J., Dunphy, D. C. & Smith, M. S., 1966. *The general inquirer: A computer approach to content analysis.*. s.l.:MIT Press.

Subramanian, R., Insley, R. G. & Blackwell, R. D., 1993. Performance and Readability: A Comparison of Annual Reports of Profitable and Unprofitable Corporations. *The Journal of Business Communication,* 30(1), pp. 49-61.

Visa, A. J. et al., (2001). *Data mining of text as a tool in authorship attribution.* SPIE Proceedings 2001.

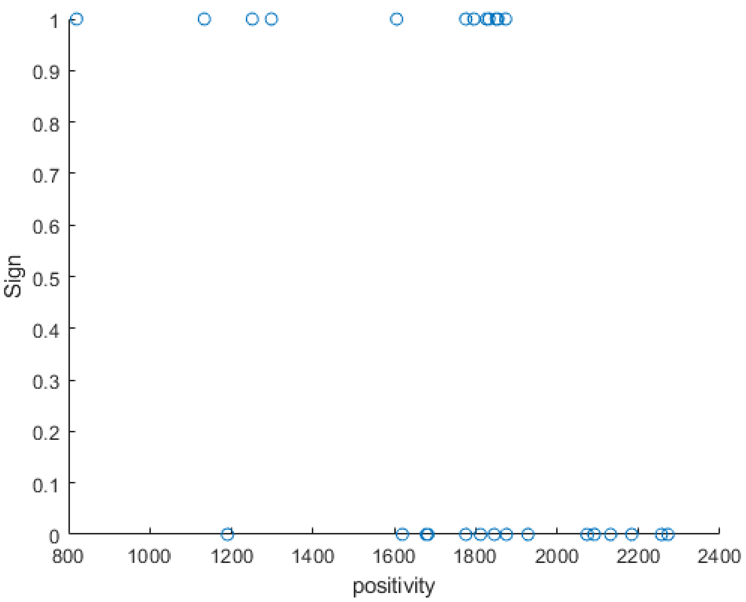
Zheng, T. K, Brian, K, Dacheng, X (2019), *Predicting Returns with Text Data,* Working paper, no 2019-69, University of Chicago, Becker Friedman Institute for Economics.

Appendix

Table A1: The full list of the selected companies:

AAL	AAP	AAPL	ABBV	ABC	ABMD	ABT	CAN	ADBE	ADI
ADM	ARR	ARP	ARS	AFL	AIG	AIZ	AJG	AKAM	ALB
ALGN	ALK	ALL	ALXN	AMAT	AMCR	AMD	AME	AMGN	AMP
AMT	AMZN	ANET	ANSS	ANTM	AON	AOS	APA	APD	APH
APTV	ARE	ATVI	AWK	AXP	GOOG	GOOGL	LTN	MMM	MO

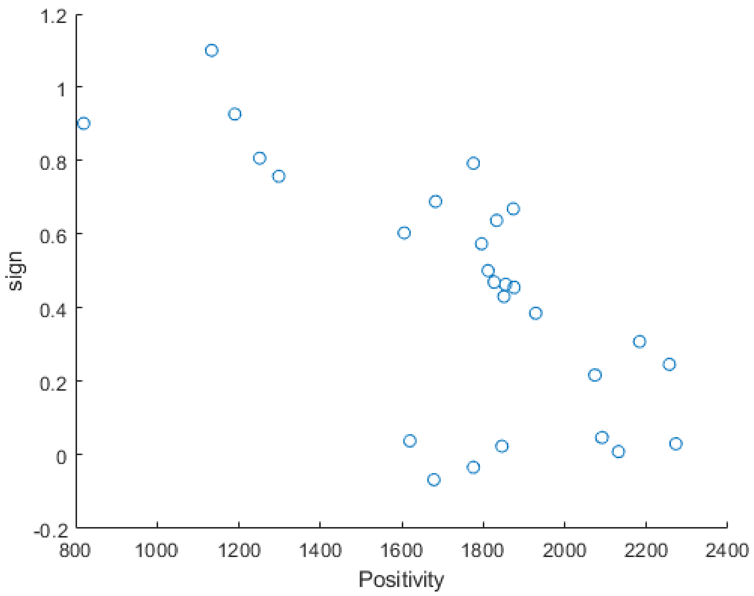
Graph A1: Graphic representation of the sign as function of positivity (Company: APTV)



Note: This is a graphic presentation of (1) the **Sign** (Dependent variable that takes the value of one if the financial return increases from month to month) in terms of (2) **Positivity** (Independent variable, the assigned score generated via Python Programming Language relying on LM dictionary).

Graph A2: Graphic representation of the predicted sign as function of positivity

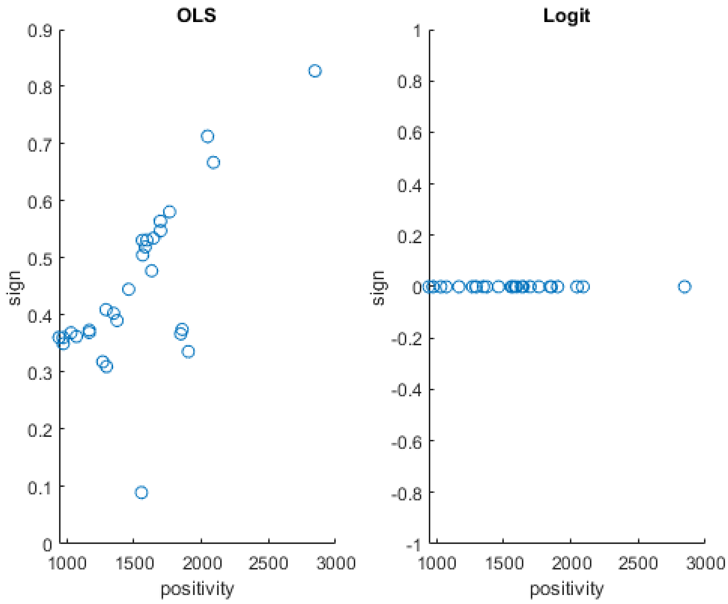
(Company: APTV)



Note: Graph A2 reports the prediction of (1) the **Sign** (Dependent variable that takes the value of one if the financial return increases from month to month) based on (2) **Positivity** (Independent variable, the fitted value generated through OLS regression).

Graph A3: A comparative illustration of the Sign prediction between OLS & Logit

(Company: AMGN)



Note: Graph A3 presents a comparative prediction of (1) the **Sign** (Dependent variable that takes the value of one if the financial return increases from month to month) based on (2) **Positivity** (Independent variable, fitted value), between (3) **OLS**, and (4) **LOGIT** models (Ordinary Least Squares and Logistic regressions).

Table A2: Most used 15 word by company for the top 5 performers

APA		ANET		ARE		ABMD		GOOGL	
to	100	our	100	to	100	to	100	to	100
for	99	to	99	our	99	for	99	our	99
s	98	or	98	for	98	our	98	as	98
on	97	we	97	as	97	impella	97	for	97
company	96	for	96	we	96	or	96	we	96
apache	95	may	95	on	95	s	95	other	95
million	94	on	94	real	94	company	94	on	94
as	93	that	93	estate	93	million	93	revenues	93
cash	92	be	92	or	92	is	92	ended	92
quarter	91	products	91	interest	91	ended	91	months	91
gas	90	could	90	income	90	as	90	are	90
other	89	as	89	million	89	months	89	google	89
net	88	with	88	from	88	on	88	net	88
oil	87	are	87	that	87	are	87	or	87

Table A3: Most used 15 word by company for the worst 5 performers

AAP		AMP		ABC		ALXN		ALL_	
to	100	to	100	to	100	to	100	to	100
our	99	for	99	million	99	our	99	for	99
for	98	company	98	company	98	or	98	company	98
as	97	million	97	our	97	we	97	as	97
net	96	on	96	as	96	for	96	on	96
or	95	net	95	for	95	with	95	or	95
ended	94	or	94	other	94	may	94	is	94
on	93	as	93	s	93	as	93	are	93
we	92	financial	92	on	92	on	92	other	92
by	91	our	91	from	91	be	91	that	91
cash	90	is	90	or	90	are	90	months	90
financial	89	value	89	ended	89	other	89	income	89
sales	88	by	88	by	88	that	88	with	88
income	87	are	87	with	87	soliris	87	million	87

Table A4: Predicted and realized

Company	Realized outcome	Predicted outcome	Reports
APA	0.539	0.529	28
ANET	0.384	0.352	21
ARE	0.302	0.210	27
ABMD	0.189	-0.066	31
GOOGL	0.184	0.456	17
ANTM	0.142	0.040	31
AIG	0.123	0.045	30
ANSS	0.103	0.060	30
MO	0.098	0.090	47
GOOG	0.177	0.456	17
AMCR	-1.006	-0.361	7
ADBE	-0.504	-0.097	30
AEP	-0.504	-0.424	45
AIZ	-0.466	-0.698	30
ATVI	-0.373	-0.023	30
AAP	-0.207	0.087	30
AMP	-0.223	0.047	39
ABC	-0.309	-0.662	30
ALXN	-0.317	0.325	31
ALL_	-0.317	-0.178	61

Note: Table A4 reports the (1) **realized** (financial return retrieved from Eikon-Reuters Platform), and (2) **forecasted return** (Quantified prediction based on word count through (3) **Parlametric platform** (Award-winning Swedish tech company's platform)).