SOCIAL MEDIA SENTIMENT ANALYSIS FOR FIRM’S REVENUE PREDICTION

Ioanna Dimadi

Subject: Information Systems
Corresponds to: 30 hp
Presented: VT 2018
Supervisor: Simone Callegari
Examiner: Görkem Paçaci

Department of Informatics and Media
Abstract
The advent of the Internet and its social media platforms have affected people’s daily life. More and more people use it as a tool in order to communicate, exchange opinions and share information with others. However, those platforms have not only been used for socializing but also for expressing people’s product preferences. This wide spread of social networking sites has enabled companies to take advantage of them as an important way of approaching their target audience. This thesis focuses on studying the influence of social media platforms on the revenue of a single organization like Nike that uses them actively. Facebook and Twitter, two widely-used social media platforms, were investigated with tweets and comments produced by consumer’s online discussions in brand’s hosted pages being gathered. This unstructured social media data were collected from 26 Nike official pages, 13 fan pages from each platform and their sentiment was analyzed. The classification of those comments had been done by using the Valence Aware Dictionary and Sentiment Reasoner (VADER), a lexicon-based approach that is implemented for social media analysis. After gathering the five-year Nike’s revenue, the degree to which these could be affected by the classified data was examined by using multiple stepwise linear regression analysis. The findings showed that the fraction of positive/total for both Facebook and Twitter explained 84.6% of the revenue’s variance. Fitting this data on the multiple regression model, Nike’s revenue could be forecast with a root mean square error around 287 billion.

Keywords:
Social Media, Facebook, Twitter, Sentiment Analysis, VADER, Linear Regression
# Table of Contents

Abstract ........................................................................................................................................ 2  
1. Introduction ........................................................................................................................... 1  
  1.1. Background .................................................................................................................... 1  
  1.2. Research Problem ......................................................................................................... 3  
  1.3. Research question and purpose ............................................................................... 4  
  1.4. Delimitations ............................................................................................................. 5  
  1.5. Outline ....................................................................................................................... 5  
2. Related work ........................................................................................................................... 6  
3. Literature Review ..................................................................................................................... 9  
  3.1. Fundamental Concepts ............................................................................................... 9  
    3.1.1. Social Media and Brand Pages ............................................................................. 9  
    3.1.2. Financial Performance ....................................................................................... 10  
    3.1.3. Sentiment Analysis ......................................................................................... 12  
  3.2. Development tools ....................................................................................................... 15  
    3.2.1. Python .............................................................................................................. 16  
    3.2.2. VADER .......................................................................................................... 16  
4. Methodology ........................................................................................................................... 18  
  4.1. Research Approach ....................................................................................................... 18  
  4.2. Data Collection ............................................................................................................ 19  
    4.2.1. Nike Inc. Revenue .......................................................................................... 19  
    4.2.2. Social media data ............................................................................................ 20  
  4.3. Sentiment Analysis process ......................................................................................... 21  
    4.3.1. Data Preprocessing .......................................................................................... 21  
    4.3.2. Sentiment Classification of the Data ............................................................... 22  
  4.4. Prediction Model .......................................................................................................... 23  
    4.4.1. Finding the Independent variables ................................................................... 23  
    4.4.2. Model Creation ............................................................................................... 24  
    4.4.3. Evaluating the model ..................................................................................... 25  
  4.5. Critical Reflection ......................................................................................................... 25  
5. Results and Discussion ............................................................................................................ 28  
  5.1. Collected Data ............................................................................................................. 29
5.2. Analyzed Data ................................................................. 30
5.3. Data Classification ......................................................... 32
5.4. Analyzing the multiple linear regression model ..................... 36
   5.4.1. Identification of the independent variables .................... 36
   5.4.2. Multiple linear regression deduction ........................... 37
      5.4.2.1. Coefficient and Multicollinearity ......................... 37
      5.4.2.2. Examining the statistical significance ................... 38
      5.4.2.3. Examining Confidence Interval for B and Coefficient ... 38
      5.4.2.4. Testing Autocorrelation ................................... 39
   5.4.3. Prediction Process .................................................. 40
      5.4.3.1. Predicted Model ............................................. 40
      5.4.3.2. Model Reliability .......................................... 40
      5.4.3.3. Evaluation .................................................. 42
6. Concluding Remarks and Future Perspective ............................ 43
   6.1. Summarizing the findings ........................................... 43
   6.2. Limitations ............................................................ 44
   6.3. Future Research ....................................................... 45
   6.4. Conclusions .......................................................... 46
7. References .................................................................. 48
List of Figures

Figure 1. Converting resources into capabilities ......................................................... 11
Figure 2. Sentiment Analysis methods ................................................................. 13
Figure 3. Steps of the thesis .............................................................. 19
Figure 4. VADER Steps ............................................................. 23
Figure 5. Total comments and tweets ............................................................. 29
Figure 6. Analyzed comments and tweets ......................................................... 31
Figure 7. Classified positive comments and tweets ........................................... 33
Figure 8. Classified negative comments and tweets ......................................... 33
Figure 9. Classified total comments and tweets ............................................... 34
Figure 10. Classified positive/negative comments and tweets ......................... 34
Figure 11. Classified positive/total comments and tweets .................................... 35
Figure 12. Normal P-P Plot of Regression Standardized Residual ................. 41
Figure 13. Standardized Residuals ......................................................... 41
Figure 14. Actual and Predicted quarterly revenue .......................................... 42
List of Tables

Table 1. Fan pages from Facebook and Twitter ........................................ 21
Table 2. Excluded variables ...................................................................... 36
Table 3. Coefficients .................................................................................. 37
Table 4. A. Coefficients, B. ANOVA ......................................................... 38
Table 5. A. Confidence Interval, B. Model Summary .................................. 39
Table 6. Autocorrelation ........................................................................... 39
List of Equations

1. Annual Revenue ................................................................. 20
2. Normalization Model .................................................................... 23
3. Multiple Regression ......................................................................... 24
4. RMSE .......................................................................................... 25
5. Predictive Model ............................................................................ 40
1. Introduction
This chapter aims to inform about the background as well as the goal of this thesis, followed by defining the research problem and the questions that help in the investigation of the under researched area. The delimitations of the current study are also discussed while an outline of the rest of the paper is provided in the last part of this chapter.

1.1. Background
The advent of social media has changed people’s daily life (Landry, 2014; Morra et al., 2018). More and more people communicate and share information, knowledge and opinions in diverse domains through social networking sites. Nowadays, most of the websites and blogs have familiarized with consumer generated media, allowing people to express their ideas and embed the content to online communities. This exponential growth of social media platforms in volume has turned people into active users that produce content able to influence the purchase decisions of their social network (Burby and Atchison, 2007). To better understand the impact on decision making, considering a case of a daily active user with a large number of followers; Sharing a positive purchase experience of a new pair of jeans, will make their social friends interested in learning more about that product (Barker, 2017). This is even more obvious if the amount of the global active users in the most well-known social media like Facebook and Twitter is taken into consideration. Currently, worldwide users of Facebook are more than 2 billion per month while for Twitter they are around 330 million (Statista, 2018; Aslam, 2018).

Therefore, the widespread demand for approaching the proper target audience has made organizations turn their interest in social media usage. Additionally, the social media marketing has enabled companies to connect with their clients more frequently, efficiently and in a low-cost manner. The online advertisement can also be beneficial for corporations in order to understand the successfulness of each product or service that they provide through monitoring and tracking of their customers preferences (Singh and Kaur, 2015). Moreover, it can even provide firms with a thorough guidance on the development of new products that match with their consumers’ needs.

Indeed, official fan pages serve as a channel for social interaction. By creating their own brand pages both on Facebook and on Twitter, companies can investigate their clients’ conversations related to their products. According to Kwon et al. (2015), the brand hosted pages on social media platforms can be characterized as open forums
where people can discuss about the brand, and in case of any problem, they can reach out to the customer support to find solutions, and also become aware of new releases. The company created pages are not limited by geographical boundaries, so international consumers can participate in the discussion and share their feedback.

However, this can also be problematic in a way because more followers eventually produce more data in social media platforms (Kwon et al., 2015). Manual data analysis is considered to be difficult and inefficient to overcome this, automated analysis processes need to be taken into consideration. Sentiment analysis is a widely applied method that processes natural language to identify meaningful data for firms (Kharde and Sonawane, 2016; Singh and Kaur, 2015). Understanding the social sentiment of the brand’s product or service is an important aspect as it can help corporations comprehend to what extent customers are satisfied with the current amenities.

This has led to a substantial number of research endeavors in the past. Those endeavors were conducted by combining customer sentiment values from social media platforms in terms of positive and negative comments about any product of the company and the financial reports, published on the company’s websites. Previous studies also have explored the possibility that a single media platform like Twitter can help in the prediction of the revenue.

Asur and Huberman (2013) have used the integration of a chatter on tweets to predict the ticket sales for 24 released movies. By gathering the tweets that were relevant to those movies and in conjunction with the use of regression modeling they created a successful predictive model. The result showed the model managed to predict the ticket sales for the first weekend with 97% accuracy outperforming the Hollywood Stock Exchange, a popular trading tool for movies. Another successful model was presented on Lassen et al. (2014) paper. In this study, the authors have developed a predicted model using the digital-world activities like tweets related to iPhones to predict the sales of each Apple’s iPhone model. The evaluation of the model showed that the predicted quarterly sales had an average error between 5% and 10% from the actual quarterly sales.

A number of authors have also considered the effects of multiple media platforms on financial outcomes. Saboo et al. (2016) have used MySpace, Facebook, Twitter and YouTube to collect data in order to find the degree to which it can influence the buy-
The authors of the music track sales study have concluded that the three social metrics that mainly influenced the music track sales were the number of users that follow an artist, the number of comments that are written for a focal brand, and the number of times a song is tested online. In a recent study, Oh et al. (2017) have investigated the extent to which different variables from social media platforms such as Twitter, Facebook and YouTube can influence the ticket sales for the new release movies. The evaluation of the developed ordinary least square (OLS) regression model showed a strong correlation between the social variables and the box office revenue. However, it came with some limitations as the movie needs to be at all three social media platforms for an accurate prediction of its revenue.

Hence, extensive research has been carried out on either the combination of multiple firms and different social media or the relation of one organization with a single social media platform. However, very little exists in terms of how a single organization’s revenue can be affected by multiple social media platforms. Tufekci (2014) stated that the confinement of one social media platform in most studies can lead to ignorance of the ecological view on society’s diffusion. In addition, a generalized predictive model would be invalid as each and every firm’s social media activity is unique as well as their incomes. For this reason, this study investigates the relationship between the sentiments in comments and tweets found on the brand-hosted pages and the revenue of the corporation. The results of this investigation will show if a predictive model for the upcoming revenues can be developed. It is hoped that this research will contribute to the literature in terms of the comparison of the effect of multiple social networking websites on the outcome of a specific organization. At the same time, it can be used as a guide for other companies.

1.2. Research Problem
Due to the emergence of the Internet and the smart connected devices, social network platforms like Facebook and Twitter are increasing rapidly. Interacting and cultivating relationships through social media is a preferable communication tool for active users and substitutes partially the traditional form of social interactions (Morra et al., 2018). However, those platforms not only alter users’ communication habits, but also their choices as social networks can shape opinions and guide purchases’ decisions. As a result, companies need to monitor social media in order to be aware of what is written online about their products or services. By understanding their customers’ preferences, brands will be well informed about their consumers’ satisfaction.
Although social media is trending and there are many articles that are related to the impact of social media on consumer’s behavior and their purchase selections of a product. Nonetheless, few articles investigated the degree to which the analysis of the social media data can help organizations to gain useful information about their clients. Therefore, the focus of this thesis is an in depth investigation into the effect that Facebook and Twitter can have on the revenue of a single organization. Nike was chosen as it has active users on both aforementioned social media platforms.

Nike Inc. is an international company that started in 1964 as Blue Ribbon Sports and seven years later the name changed to the now known as Nike Inc. The firm is a successful athletic brand that is well-known for the stylish sports clothes and shoes while it also sells sport equipment and accessories. As the design of their products is unique and fashionable most people not only buy them for sport purposes but for leisure ones as well. The company is located in Beaverton, Oregon and has more than 70,000 employees worldwide (Kicksonfire, 2018). According to Forbes (2018), Nike Inc. is one of the world’s most valuable brands for 2017, ranked #16 with a value that reaches the 29.6 billion. An organization like Nike serves a good case for a better understanding of the impact that social media have on the company’s revenue. Performing sentiment analysis of the acquired data from Facebook and Twitter classified the comments as positive, negative and neutral is the first step of this study. Subsequently, the association between the analyzed data and the Nike’s revenue is investigated by using regression analysis. Moreover, the degree to which a powerful predictive model related to the upcoming revenues can be built is also part of this thesis.

Creating a forecasting model can be the starting point for other companies to recognize the importance of using social media as they can be benefited from in multiple ways. First of all, they can gain a deep understanding about how their customers feel for their current products. Second of all, it can provide guidance about the development of new products that will match with their customers’ preferences. It can also help them build their own predictive model for the upcoming revenues by gathering appropriate data.

1.3. Research question and purpose
The purpose of this thesis is to understand the role that social media can play in the growth of a business and the degree to which this data can influence the future revenue
of the organization. According to previous research, the field lacks in the use of sentiment analysis for multiple social media in order to create a model that forecast the revenue of a specific organization. The main questions of the particular research are:

- Can the sentiment of Facebook comments, Twitter comments and/or a combination of both have an impact on Nike’s quarterly revenue?
- Can those social media comments be used to as predictive variables?

This study shows the association between the unstructured social media data and the real-world outcomes like revenues. It can also be seen as guidance for future researches on the specific field for different types of organizations.

1.4. Delimitations

There is an important delimitation that should be addressed. This study has been conducted from a specific company's perspective, and if redone for another company presumably other results will arise depending on their domain and sources for insights. In other words, the study is based on in-depth investigation of a single organization and not in the breadth of multiple companies. For this reason, the predictive model that had been created would not be directly applicable to another type of corporation as the data that its brand hosted pages produce are not similar.

1.5. Outline

This thesis contains a total number of six chapters. As we saw in the first chapter the general information of this dissertation is given, including the research problem, the research questions as well as the delimitations. The next chapter is related to the previous work that have been done in the specific domain. The third chapter is separated into two parts. In the first part the fundamental concepts of this thesis are introduced while the second part refers to the development tools that were used. The methodology chapter details the steps of this paper in order to answer the research questions. The following chapter analyses the results from this process and discusses the information that was found during the specific procedure. The last chapter of this thesis is allied with the concluding remarks and the future perspectives. The first section of this chapter summaries the outcome of the regression analysis and answers the research questions. The two following section presents the limitations of this thesis and the future studies that can contribute to the investigation of this area. While the final section draws the conclusions of the entire paper.
2. Related work
As stated in the previous chapter, sentiment analysis is an area with a growing research interest. Researchers have relied upon the analysis of users’ generated contents created on social media platforms in order to predict real-world outcomes like revenues, stock prices, sales and so on.

For example, Lassen et al. (2014) conducted a study on predicting the sales of each Apple’s iPhone model. The authors gathered 400 million tweets including tweets, retweets, and replies between 2010 and 2013 related to iPhones. Using a social analytics tool named Topsy Pro they analyzed all those tweets in order to find their sentiments. A predictive model implemented with the use of a multiple regression analysis by combining the analyzed data with the Apple quarterly sales. The evaluation of the model showed that there is a correlation between those variables while the average error of the predicted quarterly sales and the actual quarterly sales was 5% - 10%.

Furthermore, Nann et al. (2013) investigated whether the online sources can predict the Standard & Poor’s 500 stock market and to what extent. Using online data from social media platforms like Twitter, several forums and Yahoo Finance news, the authors gathered 2,971,381 messages related to the stock market index from June 2011 to November of the same year. As this public data were not assigned to a particular stock, Nann et al. (2013) used “cash tags” to enable their categorizations. The next step was to analyze the classified data in order to find their sentiments by applying the Naïve Bayes technique. This state-of-the-art sentiment analysis is a machine learning approach that needs a relevant training data to map the remaining dataset with the corresponding sentiments. For this reason, the authors manually assigned either a positive or a negative label to a few hundred messages and then used them for training their system. After classifying all messages according to their sentiment the authors were then calculated the total amount of positive and negative messages for each month as well as the total amount of positive and negative posts per day. To find the sentiment predictor, they divided the daily related comments by the monthly sentiment. A positive result was identified as a positive movement while a negative one as a downward movement. Using this simple trading model the authors found the stock movement per day. The evaluation of the model showed that taking the sentiment analysis into consideration can increase the investment of the stocks around 0.50% per trade.
Similarly Bollen et al. (2011), found that the sentiment analysis of tweets can have a beneficial effect on the prediction of the stock market trends. In their paper, Bollen et al. (2011) collected 2.7 million tweets that were related to stocks. Those tweets were then filtered in order to remove lexical features like stopwords and punctuations that would not help in the sentiment analysis process. For the analysis of the sentiment the authors applied two different methods the Opinion Finder (OS) and the Google-Profile of Mood States (GPMOS). Using the first method they first calculated the messages polarity and then classified them as positive or negative tweets. In the second method, the mood of the collected tweets was measured and they were assigned as calm, alert, sure, vital, kind, and happy. To investigate the correlation between the analyzed data and the value of the Dow Jones Industrial Average (DJIA), the authors used two techniques including the Granger causality analysis and the Self-Organizing Fuzzy Neural Network. The former was used to examine the degree to which the analyzed data can affect the stock market while the latter to investigate the sentiment analysis method that is more suitable. The evaluation showed that the mood specification can predict the price of the stock market with 86.7% accuracy and eliminate the Mean Average Percentage Error up to 6%.

In a recent study, Debakker (2017) focused on the impact that social media platforms can have on the quarterly revenue of Fortune 500 companies. The author used both brand hosted pages and fan created pages to gather posts and comments from Facebook while tweets and replies from Twitter during the year 2014 and 2016. The collected data were then filtered to remove features like special characters, numbers, hashtags as they do not contain any sentimental value. Moreover, all words were converted to lower case in order to have symmetry. After which a sentiment analysis package named the Sentimentr was used in order to find the sentiment of the preprocessed online data (Rinker, 2017). For investigating the correlation between the online sources and the Fortune 500 quarterly revenues, the author used the Granger causality analysis. As the activity for most of the companies was higher on Twitter, Debakker (2017) found that there is a stronger relationship between the metrics of this social media platform and the quarterly revenue. However, the author noted that the insufficient amount of gathered data on Facebook could be a possible reason for the low correlation between the metrics of the aforementioned social platform and the Fortune 500 revenues. The author suggested that future research is needed for the investigation of this popular platform.
Forouzani (2016) used the sentiment of tweets to predict the Return On Assets (ROA) of two organizations: the VW and the BMW. The author collected 677.596 tweets related to BMW Company between 2007 and 2015 and 151.648 tweets for VW Company from 2012 to 2015. In the next step, two different dictionaries were used in order to analyze the sentiment of those tweets as positive or negative. The first dictionary contained words related to economics while the second one was for general purposes. After the sentiment analysis, the author counted the amount of times each word was detected. The combination of ROA and the sentiment analysis merged in the feature vectors. As for the prediction of the over performance or the underperformance of both BMW and VW, the author has used three classifications including Random Forest, Naïve Bayes and AdaBoost to train the model. The evaluation of the model showed that the Random Forest algorithm was able to predict the increase or the decrease of the ROA with an 86.17% accuracy.

Another study that used online data to predict the fluctuation of the Bitcoin price was conducted by Stenqvist and Lönnö (2017). In their paper, Twitter was chosen as a source to collect 2.27 million tweets related to Bitcoin. Using the VADER tool the sentiment of the acquired data were analyzed and classified as positive, negative and neutral. The analyzed tweets were then classified according to their interval length while the mean of those classified tweets was calculated in each group. The implemented model was based on the degree of variation in the sentiment from one time series to another in order to investigate if there is a correlation between the sentiment and the bitcoin prices. The evaluation of the model showed when the change of the sentiment is around 2.2% the predicted model can be more precise with 79% accuracy.
3. Literature Review
This chapter starts by discussing the fundamental concepts that are relevant to the thesis. The first section explains why the social brand pages become so popular and what companies can gain from them, followed by mentioning the financial performance and how it is connected to social media platforms. While the last section of the fundamental concepts refers to the different sentiment analysis methods and the way that the proper method was selected. Moreover, the development tools that were used in this thesis are also described.

3.1. Fundamental Concepts

3.1.1. Social Media and Brand Pages
The increasing prevalence of the Internet in conjunction with the need of people to exchange information online has driven the creation of web 2.0 and social media platforms. According to Kaplan and Haenlein (2010), there is a difference between the terms social media and web 2.0 that most people ignore. Web 2.0 is a technology that enables users to generate content and use it easily even if they are non-experts. However, social media are an electronic word of mouth platforms that allow people with common interest to interact and spread information by taking advantage of the benefits that web 2.0 makes available (Njeri, 2014; Barnes, 2014).

Those platforms empower customers to participate actively and create meaningful content by expressing their opinion in various ways (Debakker, 2017). Previous research showed that user reviews influenced customers’ buying preferences, especially when a negative comment was written related to a product of interest (Saboo, Kumar, and Ramani, 2016; Borah and Tellis, 2015).

However, those platforms not only are used by customers to express their opinion and exchange information but also by marketers to interact with a huge amount of people (Li and Shiu, 2012). Knowing that social media can affect the brand recognition, companies turn their interest to create brand pages that enable communication with their customers. This interaction can be beneficial in many ways as firms not only lower their costs and accelerate their marketing process but increase their personal earnings as well. Moreover, customers’ recommendations and enhancements related to current products can lead to a better understanding of their needs. (Zaglia, 2013; Debakker, 2017).
Social platforms like Facebook and Twitter are a combination of blog, social networking sites and texting (Kwon et al., 2015; Miller, 2008) that enable companies to gain insights from their consumers in different ways. This becomes increasingly evident if we take account of the vast amount of daily messages that were sent by users on Facebook and Twitter: 734 million and 500 million respectively (Zephoria, 2018; Aslam, 2018). Apart from messages, users have also the ability to subscribe to official brand pages by using functions such as like and follow (Kwon et al., 2015). Moreover, Facebook’s share button and Twitter’s retweet button are two functions that allow users to spread data related to the company among their social network (Webtegrity, 2018). This huge amount of unstructured user-generated data is also known as social media big data (Debakker, 2017).

According to Kwon et al. (2015), motivation is the key that leads customers to follow the brand and produce relevant data. For some people the ‘social-interaction seeking’ is their primary impetus as they can exchange opinions with other costumers that are part of the firm community or express their thoughts to the brand manager. Others are more ‘information seeking’ motivators as they focus on understanding how the brand works or finding recent information about a brand which is relevant to discounts and competitions. Still, for others the ‘brand usage and likability’ is a motivation as brand advocates are more likely to follow the firm and influence their social network in favor of the company.

Overall, those incentives are closely related to firms in terms of how strong their relationships are with their customers. Those companies that fail to build good connections will face consumer disapproval as they will feel excluded from the brand’s community (Kwon et al., 2015).

### 3.1.2. Financial Performance

According to Paniagua and Sapena (2014), business performance is an important measurement that shows whether firms convert data into meaningful actions that will give them a competitive edge over their competitors (Paniagua and Sapena, 2014; Barney, 1991; Day, 1994). Business performance consists of three types: the corporate social, the financial, and the operational (Paniagua and Sapena, 2014), but for the purpose of this thesis only financial performance is investigated. The term financial performance refers to the extent to which a company can use the assets effectively in order to increase their revenues (Investopedia, 2018). This type of performance is
a multifaceted construct that can be calculated by the level of sales, growth, profitability, and share price (Paniagua and Sapena, 2014; Investopedia, 2018). As shown in figure 1, online conversations, sharing information, and presence are resources that converted into financial performance capabilities through the consumer’s preferences channel and the social marketing channel.

![Diagram](image)

**Figure 1. Converting resources into capabilities (Paniagua and Sapena, 2014)**

The customers’ revealed preferences channel, as the name indicates, is a way for people to express their opinion or preferences in social media platforms for the products that they use. According to O’Connell and O’Sullivan (2016), non-financial measurements are essential for the enhancement of the firm’s financial performance. Customer preference is one of the measurements that can help the organization to identify their consumers’ satisfaction. As customers become active users of social media, especially Facebook and Twitter, companies are able to use the unstructured user-generated data wisely to forecast the demands in the market in correlation to their current products (Paniagua and Sapena, 2014).

Companies can also use the social marketing channel to promote services or products in social media to garner additional revenue. Social media marketing has a different approach from the regular media as the marketing process is not a direct advertisement process rather than a collective one; firms aim to collaborate with their customers as it is a two-way communication in order to facilitate their brand advocacy. Although social media marketing has a different tactic compared to regular media, they still have the same goal to approach their customers and heighten their brand reputation (Weinberg and Pehlivan, 2011; Paniagua and Sapena, 2014). However, social media marketing seems to outweigh the traditional media when it comes to the effect
that it has on the financial performance of the business (Hulland, Wade, and Antia, 2007; Paniagua and Sapena, 2014). As a traditional media like television delivers a controlled marketing message from the company while in the case of social media, the marketing messages are delivered directly from the satisfied customers (Weinberg and Pehlivan, 2011; Paniagua and Sapena, 2014).

Customers’ revealed preferences as well as social marketing channels play an important role in the financial performance due to the fact that they can impact the level of the outcome. Therefore, there is an elaborate body of investigation done on the relationship between social media and business outcome. For instance, Rui, Liu, and Whinston (2013) found that the Word of Mouth on Twitter has a huge effect on the revenue of Hollywood movies especially when it comes from more active users. Moreover, Ding et al. (2017), Oh et al. (2017), and Asur and Huberman (2010) also discuss the importance of social media in the prediction of box-office revenues. On the other hand, Debakker (2017) conducted a study in which she investigated the effect that social media have on multiple companies’ sales regardless of their field. Another research used different variables like the volume of likes and users that were found on Facebook fan pages in conjunction with volume of queries gathered by Google Trends in order to investigate their effect on the revenue (Boldt et al., 2016). In addition, Hyunyoung and Hal (2012) also used Google Trends data to create a seasonal model that was able to predict the sales of Ford Company. Liu et al. (2015) investigating the effects that tweets have in the forecast of the stock co movement for two industries. Forouzani (2016) also used tweets in order to create a model that would foresee the return on asset for two automobile companies the BMW and the VW. Last but not least tweets were also used as a metric from Lassen et al. (2014) study so as to foretell the future sales of each Apple’s iPhone model. Saboo et al. (2016) used multiple social media like MySpace, Facebook, Twitter and YouTube to examine how the buying process of 36 music artists works and predict their upcoming sales. However, no previous study has focused on the impact that Twitter and Facebook consumers’ opinion can have in the prediction of the revenue for a single organization.

3.1.3. Sentiment Analysis
The arrival of social media created a new area of investigation: sentiment analysis. As an increasing number of online pages permits user comments, analyzing the content is useful in order to understand the sentiment classification of the comment. Sentiment analysis or opinion mining is aptly named; it is a process of detecting opinions, and sentiments that have been found in the user-generated content through entities
like products and services (Duan et al., 2015; Liu, 2010). Those one-sided information is derived from the way of user thinking expressed in text (Rajni and Rajdeep, 2015). As shown in figure 2, two methods of sentiment analysis are widely performed: (i) the machine learning based, and (ii) the lexicon-based (Profitaddaweb, 2017).

![Figure 2. Sentiment Analysis methods (Medhat et al., 2014)](image)

The machine learning, as the name indicates, is a method that enables machines to learn and get improved through experience by using input examples. There are two types of machine learning: the supervised and the unsupervised learning. In the first case, data labeling is a manual process and the model uses different classifiers like Naïve Bayes and Maximum entropy for mapping the output with the right sentiment. While in the second case, the unlabeled input data are defined by using different patterns in the data set (Dhaoui et al., 2017). Meaning to say that the input data are divided into clusters according to which the model makes the prediction of the output.

On the other hand, the lexicon-based is a method that uses sentiment oriented words and phrases to calculate the outcome. There are three types of lexicon-based analysis: one manual approach, and two automated approaches, the dictionary-based, and the corpus-based. Manual analysis is not broadly used as it is time-consuming and the process should be operated by the researcher (Bing, 2012). In the case of the automated approaches, the dictionary-based method is a generalized dictionary that contains words with different level of sentiment classification. Each word has a list of
other phrases with the same and the opposite meaning (Dhaoui et al., 2017; Bing, 2012). While the corpus-based approach is a specific dictionary that is related to a particular field of interest. This dictionary is developed by using statistical and semantic techniques in order to find words that are relevant to the specific domain (Bing, 2012).

The type of the data that will be analyzed play an important role in the selection of the appropriate sentiment analysis approach as it can affect the outcome. Johanson and Lilja (2016) made a comparison between three different machine learning methods, including Naïve Bayes, SVM, decision tree and one lexicon based tool named AFINN. Using three different datasets one from IMDB reviews and two twitter datasets, they investigated the performance of those methods in classifying the sentiment of the aforementioned datasets. Based on the small number of datasets that were examined, the authors found that the lexical based analysis is the most suitable approach for short texts like social media content compared to the machine learning methods.

This thesis uses a lexicon-based method, the Valence Aware Dictionary and Sentiment Reasoner (VADER), as it is implemented for analyzing the user-generated content from social media platforms. A common way to represent the text in lexicon-based method is known as Bag of Words where each sentence is treated as a bag of independent polarized words (Kolchyna et al. 2015). Otherwise stated the sentences were looked as words, each of them is marked as positive, negative or neutral. After which the total mark of those unigrams defines the sentiment score of the whole sentence (Debakker, 2017; Liu, 2010).

Moreover, in a lexicon-based approach the researcher is free to determine which words are more important for his investigation. As a result, a number of studies used different metrics in order to calculate the sentiment score of a sentence. For instance, Benamara et al. (2007) took into consideration only the adverbs and the adjectives for computing the sentiment of the sentence, while Pozzi et al. (2013) and Hogenboom et al. (2013) investigated the case where emoticons were important for finding the sentence’s sentiment score. Fersini et al. (2016) appear to be positively related to the above investigations as they used the previous studies’ metrics in conjunction with the lengthening words. The result showed that this combination created a powerful sentiment outcome as the words were classified more accurately.
Sentiment analysis is really useful when it comes to social media platforms as other large volume of unstructured data can be analyzed more efficiently (Rajni and Rajdeep, 2015). By analyzing the content and identifying the sentiment score for posts and tweets, companies can gain insights related to their products. These data are extremely difficult to be processed manually and sentiment analysis seems to be the most suitable solution (Rajni and Rajdeep, 2015).

According to Rui, Liu, and Whinston (2013), the sentiment score of tweets have an impact on the revenue of box office as people are influenced by the perception of their close friends. In a different field, Ringsquandl and Petković (2013) found that the sentiment analysis is a reliable method for politics as it was able to label the political sentiment on Twitter more accurately. Moreover, the research showed that there is a strong correlation between tweets and politics as most of those tweets reflect on the offline political background (Ringsquandl and Petković, 2013). Another research discussed the use of the sentiment analysis for different social media variables to predict the revenues of the Fortune 500 companies (Debakker, 2017). Oh et al. (2017) had also investigated the impact that the sentiment of different Facebook variables such as likes, and posts can have on the sales of Hollywood Movies. In addition, sentiment analysis was used from Lassen et al. (2014) to analyze tweets relevant to iPhone in order to predict its model sales. Similarly, Forouzani (2016) also analyzed the sentiment of tweets referring to BMW and VW companies to foretell their quarterly return on asset or ROA.

Much of the available literature has shown that the sentiment analysis of various social media variables has a positive effect on the real world outcome (Rui et al., 2013; Oh et al., 2017; Debakker, 2017; Forouzani, 2016). Surprisingly, there have been no detailed investigation of the impact that the sentiment of consumers’ comments, found on Twitter and Facebook brand pages, can have on the revenue of a particular organization. This study focuses on analyzing the user-generated data that were found on Nike official pages both on Facebook and Twitter in order to investigate the extent to which their sentiment affects the firm’s income.

3.2. Development tools
This section introduces the tools and the packages used for collecting and analyzing the data from Facebook and Twitter.
3.2.1. Python

Python is a programming language that supports several paradigms such as object oriented, functional, and so on. Furthermore, its simple syntax facilitates readability and comprehension. It has been characterized a general purpose language as it can be applied in several domains. Moreover, Python due to its simplicity and its supply of high-level tools is widely used for extracting data from online sources such as forums, social media platforms and so on (Felix, 2016). For this reason, it was selected to be the main programming language for gathering relevant data from both Facebook and Twitter brand pages. To implement a web scraper that would be able to acquire comments and tweets from the aforementioned platforms a Python library called Beautiful Soup was used. This library enables programmers to collect data from HTML files and is far less time consuming than the previous processes (Richardson, 2017). In addition, Python packages like Aiohttp, Asyncio, Argparse, Urllib, Pandas were used to make easier the extraction and the storage of the social media data.

3.2.2. VADER

The acronym VADER stands for Valence Aware Dictionary and Sentiment Reasoner and is an open source tool that identifies the sentiment of text by combining the lexicon-based approach with the rule-based classifiers (Hutto and Gilbert, 2014). Due to the aforementioned combination, VADER takes into account the order of the words as well as their emotion’s intensity instead of treating the text as a bag of independent polarized words. This means that the tool measures the sentiment strength of each word in the text and according to the total sentiment score maps it as positive, negative or neutral comment. VADER was implemented in order to facilitate the analysis of social media content producing more accurate outcome (Hutto and Gilbert, 2014). The tool is based on three goals:

“1) The development and validation of a gold standard sentiment lexicon that is sensitive both the polarity and the intensity of sentiments expressed in social media microblogs; 2) The identification and subsequent experimental evaluation of generalizable rules regarding conventional uses of grammatical and syntactical aspects of text for assessing sentiment intensity; and 3) Comparing the performance of a parsimonious lexicon and rule-based model against other established and/or typical sentiment analysis baselines.” (Hutto and Gilbert, 2014)

To achieve the first goal, Hutto and Gilbert (2014) have used three sentiment lexicons including LIWC (Pennebaker and Francis, 1996), ANEW (Bradley and Lang, 1999) and GI (Stone, 1966) to gather the appropriate lexical features. As the lexicon that the
authors wanted to implement was relevant to social media data, they also collect features like emoticons, slangs and abbreviations. The result of this process was 9000 lexical features. Hutto and Gilbert (2014) have used a wisdom of the crowd method to validate the sentiment valence of each feature that they gathered in order to select 7500 features that would be incorporated into their lexicon.

As for the second goal, Hutto and Gilbert (2014) have acquired 800 tweets and have examined their sentiment intensity. Using a qualitative analysis, they investigated the attributes that influence the sentiment strength of the tweets. Those tweets with similar contextual elements were grouped together, resulting in five rules that can be generally applied.

Combining the lexical features with the generalized rules, the authors implemented the VADER tool. According to Hutto and Gilbert (2014), the specific method was tested with eleven tools including lexicon, and machine learning classifiers and the outcome showed that VADER outweighed the other procedures. This “gold standard lexicon” was also validated in relation to the human sentiment analysis and the correlation coefficient depicted that they had around the same performance (Hutto and Gilbert, 2014).
4. Methodology
This section details the process followed to answer the research questions. Starting by explaining the research approach that best fit in this study and the procedure that was used to achieve it. Moreover, the way that the data were gathered from the different online sources as well as the method that was used to analyze the social media data are described in detail. The next section refers to the approach that was performed in order to find the most suitable variables for the creation of the prediction model and the technique that was used for evaluating the model. In the last part of this chapter a critical reflection of the selected methodology is presented by pointing out the weaknesses of this thesis.

4.1. Research Approach
This section explains the research paradigm as well as the approach that was chosen in order to answer the research questions (section 1.3). The thesis is aligned with the positivist paradigm. A cause and effect process where a researcher investigates the results of the experiment in order to make conclusions by accepting or rejecting the null hypothesis (Oates, 2006). In this case, the main objective is to reject the following null hypothesis:

- The sentiment of Social media data like Facebook comments and Twitter tweets have not an association to Nike’s quarterly revenue.

Since investigating the association between the classified social media data and the revenue is the main part of this thesis, a quantitative data analysis was the most suitable approach to be used. Analyzing the data and observing the results that have been produced through statistical analysis is a way to understand if the gathered data can provide meaningful information (Dudovskiy, 2018). The steps that were taken to collect this data from the different sources is described in the next section (see section 4.2). Once the data were gathered their combination was used in order to develop a predictive model that would assist to determine the revenue projection of Nike Company. The development of this model as well as its evaluation is elaborated in section
4.4. However, a brief illustration of the thesis stages is showed in the figure below (see figure 3).

![Figure 3. Steps of the thesis](image)

### 4.2. Data Collection

As stated in the previous section, the first step in this study was to collect the relevant data from three online sources. Two social media platforms were used such as Facebook and Twitter and the official website of Nike Company. As for the former, comments and tweets were retrieved from 26 Nike hosted pages that exist on the aforementioned platforms; while the latter was used for gathering the quarterly Nike Inc. revenue due to the fact that it contains more accurate metrics regarding the company’s outcome. The data collection process is explained thoroughly in the next two subsections (see section 4.2.1 and section 4.2.2).

#### 4.2.1. Nike Inc. Revenue

The primary benefit of considering a large-scale entity such as Nike for this research is the sheer volume of user-generated content available on the relevant social media platforms. As a recognizable brand with highly visible products can create a huge amount of user-generated content on social media platforms compared to the unknown ones. According to Forbes (2018), Nike Inc. is one of the world’s most valuable brands for 2017, ranked #16 with a value that reaches the 29.6 billion. For the analysis, I selected the total Nike Inc. revenue instead of the total Nike brand revenue as the subsidiaries companies including Converse and Jordan were also investigated. The quarterly Nike Inc. revenue for the last five years (2014 - 2018) was gathered from their official website in order to avoid any inaccuracy.

Finding the annual revenue of an organization the sum of the quarterly earnings of four consecutive quarters need to be calculated. For most companies the calculation of the annual revenue is done by adding the four quarters of the year as the first three
months period stands for the first quarter of the year and so on. However, Nike has a different approach when it comes to the computation of the yearly revenue. For instance, the annual revenue for 2014 would be calculated as shown below:

\[
\text{Annual Revenue}_{2014} = Q3_{2013} + Q4_{2013} + Q1_{2014} + Q2_{2014}
\]  

(1)

4.2.2. Social media data

Social media platforms allow both users and companies to create a new fan page by using any name of their interest including those of well-known brands. As a result, a large amount of Nike fan pages from different geographic locations exist on those platforms. However, only the certified social media pages were used as those pages can be found on both social media platforms. This would help to make a comparison between the two platforms and inspect if either one of them or their combination affect the revenue the most.

Investigating those pages, I discovered that the corporation uses each brand-created page to refer to a specific product category. Although the term brand-hosted page does not exist on Twitter an organization can create an account for each product category that works as an official fan page. In addition, social media platforms have each own indicator for helping companies to know the level of communication for their pages. For instance, Facebook uses the like function; while Twitter has followers as a sign of users’ activity.

The main reason that the brand-created pages were selected is due to the fact that they can be tracked easier by the brand in order to gain useful insights. An additional reason is that those pages are public and visible to any social media user which gives the opportunity to people with shared interest to express their opinion about company’s services even if there are not supporters of the firm. Furthermore, the accessibility of those pages on both platforms enables the gathering of the appropriate data.

In this thesis, only the Nike-hosted pages with the identical appellation on Facebook and Twitter were examined. Finding 16 fan pages that are common in both social networking websites I decided to investigate the sentiment of their content. However, three of those pages were excluded as their content was not in English. Eventually, the appropriate data were gathered from 13 Nike-created pages from each platform (see table 1). The dataset includes only comments and tweets while the actual posts
were not taken into consideration as they mainly produced by the company and they do not contain any valuable information. The data collected between June 2013 and February 2018 using a web scraper that was developed with Python (see section 3.2.1).

<table>
<thead>
<tr>
<th>Nike-created pages</th>
<th>Facebook pages likes</th>
<th>Twitter pages followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nike</td>
<td>30,534,372</td>
<td>7,429,622</td>
</tr>
<tr>
<td>Nike Football</td>
<td>44,814,500</td>
<td>3,402,872</td>
</tr>
<tr>
<td>Nike Baseball</td>
<td>975,948</td>
<td>199,267</td>
</tr>
<tr>
<td>Nike Court</td>
<td>1,255,578</td>
<td>334,466</td>
</tr>
<tr>
<td>Nike Golf</td>
<td>1,955,345</td>
<td>634,991</td>
</tr>
<tr>
<td>Nike BMX</td>
<td>2,447,890</td>
<td>1,652</td>
</tr>
<tr>
<td>NikeiD</td>
<td>2,508,741</td>
<td>376,618</td>
</tr>
<tr>
<td>Nike Skateboarding</td>
<td>9,836,955</td>
<td>415,622</td>
</tr>
<tr>
<td>Nike+ Run Club</td>
<td>17,076,432</td>
<td>899,685</td>
</tr>
<tr>
<td>Nike Basketball</td>
<td>8,557,923</td>
<td>2,355,269</td>
</tr>
<tr>
<td>Nike Soccer</td>
<td>176,088</td>
<td>2,263,327</td>
</tr>
<tr>
<td>Jordan</td>
<td>9,468,632</td>
<td>3,605,456</td>
</tr>
<tr>
<td>Converse</td>
<td>44,256,475</td>
<td>1,086,848</td>
</tr>
</tbody>
</table>

Table 1. Fan pages from Facebook and Twitter

4.3. Sentiment Analysis process
By using the web scraper written in Python that was discussed in the previous chapter (see section 3.2.1), 550,000 comments from Nike-hosted pages on Facebook and 113,000 tweets on Twitter were gathered. The next step is to assign the sentiment to the text. The two stages of the sentiment analysis process are mentioned in this section. In order to reduce the noise and improve the accuracy in the text categorization the data is first pre-processed (see section 4.3.1) and then classified by using the VADER tool (see section 4.3.2).

4.3.1. Data Preprocessing
The pre-processing of the acquired data occurred by removing features that were not useful and would lead to computational complexity; texts such as usernames, Twitter hashtags, and hyperlinks were omitted as they do not contain any sentiment value. Moreover, the process of part of speech tag (POS) was used in order to classify the words according to the part that they belong to such as noun, adjectives and so on. The extraction of the most frequent distributions of the text played an important role in the pre-processing as each word carries a different sentiment score. Every sentence also contains some connecting words that appear frequently, and they do not add any polarity to each sentence. Those stopwords were also removed during the pre-processing as they did not contribute to the classification of the text.
4.3.2. Sentiment Classification of the Data

After pre-processing, the text was split into words which were ready for the actual classification. As stated in the previous chapter, VADER was selected for the sentiment analysis as it seemed to be the most suitable for social media content (see section 3.2.2). The main steps of VADER analysis are illustrated in figure 4. The comment or tweet was split into a bag of words for the lexicon checker in order to join the word with the right sentiment score. As social media texts includes informal expressions, VADER examines if there are an acronym such as “YOLO” or any slang word like “BTW” in order to measure it in the total score. As emoticons are also part of the social media communication, VADER takes them into consideration. Each word has a different sentiment score as the strength of those phrases varies in a scale between -4 and 4.

After the lexicon checking, the words penetrated into rule-based checker to examine the intensity of their emotions (see figure 4). In this part words are validated as n-grams instead of bag of words which will help to generate a valid polarity score of the sentence. The punctuation is the first heuristic as the sentiment that a sentence conveys can differentiate with the use of special marks. For instance, a sentence with more than one exclamation mark imparts a higher sentiment compared to the sentences with no punctuation. Another heuristic is the capitalization where the capitalized words are ranked are important for the intensity of the overall sentiment score of each sentence. The degree modifier or booster is also a contextual element that needs to be examined. The farness of a modifier from the polarized words of each comment or tweet can modify the total polarity of the text. The next heuristic is the “but” polarity which is a conjunction that connects sentences with different sentiment. The second part of the clause has a greater strength compared to the first and will affect the polarity of the comment or tweet (Hutto and Gilbert, 2014). Negation is the last heuristic where a tri-gram or the combination of three neighboring words is examined as it can shift the sentiment score from the opposed direction.

The final step of the sentiment analysis is to calculate the total sentiment score for each comment or tweet. VADER takes account of every sentiment bearing word that existed in the sentence which was between -4 and 4 and produced the overall sentiment score. However, the sentiment score of the text got a value between -1 and 1. Hutto and Gilbert (2014) introduced a normalization model to map the sentiment score in this scale.
As shown in the equation 2, VADER sums the sentiment scores of all the words in the sentence and set the value to an x variable, then adds its square to a stable variable a with the value 15. This happens due to the fact that the higher the x value is, the closer to -1 or 1 the variable y will be. In this thesis three sentiment categories exist: values less or equal to -0.1 were characterized as negative, while for values more or equal to 0.1 were positive and values in between were referred as neutral sentiment.

\[
y = \frac{x}{\sqrt{x^2 + a}}
\]  

(2)

Figure 4. VADER Steps

4.4. Prediction Model
Comparing the revenue with the independent variables is the next step. This section refers to the process that was used in order to find the independent variables that explain the revenue at a higher level. Moreover, it describes the analysis that was used for the creation and the evaluation of the model.

4.4.1. Finding the Independent variables
Having more than one independent variables I decided to use a method that is known as multiple regression. This method is mainly used in businesses, in order to make the right decision. However, it can be used for other cases like the identification of the
most important variables for the outcome (Hair et al., 2014). A multiple regression
has four approaches including: the enter, the forward, the backward, and the stepwise.

In this thesis a stepwise multiple linear regression, which is a mixture of forward and
backward methods, was first used to identify the independent variables. (Kerr, Hall
and Kozub, 2002). Stepwise regression is performed by creating a basic model that
contains only the dependent variable. The next step is to search in the list of the in-
dependent variables in order to find a predictor that correlates with the constant in the
highest degree. After adding the first independent variable the model searches for the
second predictor that in conjunction with the first will explain at a high percentage
the outcome (Field, 2013). As it continues with this process by adding one by one the
independent variables, the model examines its statistical significance by taking into
consideration the p-value. If the p-value is higher than 0.05 the approach will use a
removal test where the predictors are eliminated one by one until it reaches in a level
that the predictors can explain in a higher degree the selected constant. In other words,
the stepwise method tries to create the most suitable model that will describe thor-
oughly the dependent variable. Regarding this paper, all independent variables for
both Facebook and Twitter and in conjunction with the revenue were added and the
stepwise multiple linear regression was performed for the selection of the most suit-
able variables.

4.4.2. Model Creation
After finding the independent variables, the enter multiple linear regression was used
for the creation of the predictive model. It is a standard technique where the inde-
pendent variables are entered, all simultaneously to find if there is a strong relation-
ship with the dependent variable (Kerr, Hall and Kozub, 2002). By assuming causality
between the predictors and the dependent variable, this method determines the
equation model that can predict the constant in the most accurate way. The multiple
regression equation is in the form:

\[ Y = b_0 + b_1 \times x_1 + b_2 \times x_2 + \ldots + b_n \times x_n \]  \hspace{1cm} (3)

In equation 3, Y illustrates the outcome or the predicted variable, \( b_0 \) is the b value of
the constant variable while the rest of the \( b \) variables represent the unstandardized
coefficients b-values for each of the predictors that participate in the multiple regres-
sion model. The b-value shows the extent to which the predicted variable has changed
for every 1-unit of change in the independent variables. This means that the unstand-
ardized coefficients b-values show the impact that each independent variable has in

24
the constant if the other predictors remain stable. Those $b$-values are also used in order to eliminate the squared residuals that exist between the real values and the predicted ones. In this case, the model consists of one constant (the revenue) and a number of $b$ variables. Each of these variables represents the predictors that were previously found to explain the dependent variable the most (see section 4.4.1).

4.4.3. Evaluating the model

After creating the regression equation, it is time to calculate the quarterly predicted revenues to see how accurate the model is. As there is no flawless predicted model some predicted values will be higher than the actual ones while others will be lower (Field, 2013). The Root Mean Standard Error (RMSE) equation 4 below calculates how far the data is from the regression line that was created.

$$RMSE = \frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$ (4)

In the equation $y_j$ represents the actual revenue for each quarter, $\hat{y}_j$ is the predicted revenue for the same quarter while the $n$ is the sample size that I used. The RMSE is important in order to verify that the new model can forecast the actual revenue by showing the total errors.

4.5. Critical Reflection

In chapter 4 the methods that best fitted with the investigated problem were introduced. However, the selected methods can have some implications in both the quality of the data and the outcome of the study. For this reason, they are discussed in the following section.

Quality of Social Media Data

As stated in section 4.2.2 only comments and tweets written in English were collected. The main reason for this selection is due to the fact that most of the sentiment analysis tools support the specific language. However, this choice can have some consequences when it comes to the quality of the acquired data. For instance, comments and tweets that were found on the brand-hosted pages in a different language were excluded as the sentiment could not be calculated accurately. This means that some strong sentimental comments might not be taken into account as they were written in a non-English language. Nevertheless, the Nike hosted pages that were selected for this thesis are the universal brand pages and therefore comments and tweets are mainly written in the English language.
Quantity of Social Media Data
Due to this study has not been conducted in cooperation with Nike Company, the gathered data might not represent the total amount of comments and tweets that exist on the brand-hosted pages. Even though, a huge amount of unstructured social media data was acquired from those public pages we should keep in mind that the quality of the outcome might be influenced.

Time series
The time series of the data might also have an effect on the outcome of this study. The social media data that were acquired and analyzed are between June 2013 and February 2018. In this thesis, the specific time series were chosen as the 26 Nike hosted pages were active. While some pages have been created before this period of time we wanted to make sure that the data would be acquired from all the selected brand-hosted pages.

Selection of two social media platforms
As Facebook and Twitter are two well-known social media platforms, they were selected to be examined in this paper. Taking into consideration that other social media platforms are used by users in order to express their opinion, the result might be affected from the specific selection. This means that the investigation of Facebook and Twitter does not totally predict the revenue owing to other variables can have an impact on the financial performance of Nike Company. Therefore, the examination of other social networking websites is important in future research.

Selection of the sentiment analysis method
Another reason that might affect the quality of the data is the selection of the sentiment analysis tool. Even though, VADER seemed to be the most suitable tool for analyzing the social media content there are also some implications. Every sentiment analysis tool has a specific way of classifying the text into different categories. In this case, comments and tweets that were below -0.1 characterized as negative while those that were above 0.1 classified as positive. Performing an approach with a different sentiment range can also have an impact on the amount of positive, negative, and neutral comments that exist. As a result the distribution of the data can be affected from this choice resulting in a different outcome.
Selection of the independent variables

In order to find the independent variables that can explain the revenue, the stepwise multiple regression was used. However, this method does not work deductively and sometimes creates unreasonable correlations. As a result, each and every combination of the independent variables was also examined using the enter multiple regression. This investigation helped to identify the independent variables that were the most appropriate to be used for the analysis. Nonetheless, this is not the same as saying that the enter regression method guarantees that the results are not affected by the selection of those variables.
5. Results and Discussion
This chapter details the results from the previously gathered data. As we discussed in the methodology section (see section 4.2.2) 26 Nike official pages (13 Nike-created pages from each social media platform) were used in order to gather comments and tweets between June 2013 and February 2018. In the first part of this chapter a line graph depicts the quarterly amount of the assembled tweets and comments. This section refers to the changes of the total comments and tweets compared to the quarterly revenue. After depicting the collected data, the comments and tweets were analyzed through VADER in order to find their sentiment (see section 4.3). The result of this analysis is illustrated in the following section where the opinionated data are presented in relation to the quarterly revenue. In the next section, the analyzed data are used for the creation of ten independent variables, five from each social media platform, including positive, negative, total, positive/total, and positive/negative comments. These variables were then classified into one of the three revenue categories (low, medium, high) and their figures are used for the better understanding the range of the data and the standard deviation in each class. After explaining the data and the differences between the two social media platforms, a stepwise regression approach was applied in order to find the independent variables that are related more with the constant. However, as it was described in the critical reflection (see section 4.5) the independent variables were also re-examined in the enter regression method to verify their correlation with the revenue. After selecting the most suitable predictors the built in F-test of the multiple regression was operated to investigate the significance of those variables. If the p-value of its variable is higher than 0.05 the null hypothesis \( (H_0) \) that was introduced in section 4.1 is accepted. This means that there is no relation between the independent variable and the dependent and therefore the coefficient of each variable is zero.

\[
H_0: \beta_j = 0
\]

However, if the p-value of each variable is less than 0.05 then the alternative hypothesis is verified and the coefficient of each variable is different than zero.

\[
H_1: \beta_j \neq 0
\]

Afterwards, it is time to create the predictive model. Using the regression equation that was described in the previous chapter (see section 4.4.2), the model was created. In order to verify its validity and to trust the results of the model, the residual plots
were also checked. In the last part of this chapter, an evaluation of the model which test its predictive accuracy is presented.

5.1. Collected Data
Using the process that was described in the methodology chapter the data were gathered from three online sources (see section 4.2). The figures below (see figure 5A and figure 5B) illustrate the collected comments and tweets in quarters along with the revenue.

![Figure 5A](image1.png)

**Figure 5A.** The orange line shows the quarterly Facebook total comments, and the blue line depicts the quarterly Twitter total tweets, B. The blue line shows the quarterly Nike Inc. revenue.

It is visible from figure 5A that Facebook’s total comments had a declining tendency over time. They started with 45,100 comments and decreased during the following 3 quarters. Apart from a spike in Q1-2015 around 88,000, comments continued falling steadily, reaching their lowest point in the third quarter of the same year with 11,723. In contrast, Twitter’s total tweets (see figure 5A) showed an increasing inclination having their greatest rise during mid-2014 with 16,000 tweets. Despite of a sharp
decline in the first two quarters of 2015, tweets increased steadily during the whole period of investigation.

Except for the comments and tweets, Nike’s quarterly revenue was also gathered (see section 4.2.1). The figure 5B indicates that there is a general increase over the examined period. Beginning with 6,971 dollars in million in Q1-2014 and reaching the 8,984 dollars in million in the third quarter of 2018. As it can be seen from the figure 5B the first quarter of each year was profitable for the organization due to the steady increase of its revenue. While in Q1-2017 Nike’s revenue reached its highest point with 9,067 dollars in million. However, revenues seem to decline in the second quarter of each year and in Q2-2014 Nike’s revenues were at a much slower pace than the rest. The third and the fourth quarter can be described as the transitional period due to the fact that the revenue had a slight increase.

Interesting to note is that the quarterly comments and the revenue seem to have a similar tendency in almost the whole period. While the quarterly tweets follow the same pattern only in some cases. For instance, in the second quarter of 2014 that the revenue had a decreasing inclination the amount of comments and tweets also declined (see figure 5A and figure 5B). However, a different scenario can be seen in Q1-2015 where both the revenue and comments had a slight increase while tweets went down.

5.2. Analyzed Data

After assembling the data and removing the noise, the sentiment score was computed (see section 4.3). The total amount of assembled comments was 550,000 including: 89,260 positive comments, 35,740 negative comments and 425,000 neutral comments. While for tweets the total amount was 113,000 containing: 52,110 positive tweets, 11,890 negative tweets, and 49,000 neutral tweets. Those polarized comments and tweets were then distributed in quarters in order to investigate their changes over time (see figure 6).
Facebook negative comments (see figure 6A) went up and down the whole period that was examined. Starting by 2,868 negative comments and fell slightly over the next two quarters. After which they started increasing until it reached its highest point with 4,060 negative comments in the first quarter of 2015. The following two quarters comments decreased and in the last quarter of the same year they fell to a low of 628. The remaining quarters saw a steady increase in the negative comments. Similarly, Twitter negative tweets (see figure 6B) decreased gradually for the first three quarters after which they increased dramatically reaching a peak of 1,430 in the fourth quarter of the same year. In the following three years negative tweets saw a downward trend until Q2 of 2018 where they reached their lowest point with 127 tweets.
Facebook positive comments (see figure 6A) had a consistent increase over time apart from the first quarter of 2015 where they reached a peak of 11,329 positive comments. While, Twitter positive tweets saw a significant rise during the first year, followed by a substantial decrease that continued until mid-2018 were positive tweets fell to a low of only 547. However, the third quarter of the same year saw a dramatic growth in the amount of positive tweets as they almost doubled.

Interesting to note that the quarterly polarized comments and tweets graphs (see figure 6A and figure 6B) show that users’ activity on social media varied. However, it is visible that in some quarters the polarized comments and tweets had a similar inclination. For instance, between Q4-2014 and Q1-2015 comments and tweets had an increasing tendency. This is possibly congruent with the fact that Nike’s huge campaign on Football World Cup 2014 which led to rise firm’s sales not only to the football equipment but also to the other amenities that the company offers (BBC, 2014; Fidelman, 2014).

Furthermore, the fourth quarter of 2016 is another period where both comments and tweets had an increasing trend. During this period, Nike spent some money in order to promote its products in Rio Summer Olympics 2016. As we can see from figure 5B, there is a rise in the last quarter of 2016 where Nike revenue had an 8% increase. This suggests that the advertisement of a company not only increased its popularity on social media platforms, but also had a positive impact on the firm’s revenue (Kell, 2016).

An opposite case can be seen during the next year where there is a declination, both on social media data and on Nike’s revenue. This is due to the fact that the year 2017 was a competitive annual period for Nike as firms with similar facilities in the field of basketball affected the overall economic performance of the company (Forbes, 2017). During the aforementioned year, users’ activity on the brand-hosted pages also reduced which caused a decline on the amount of comments and tweets on social media platforms.

5.3. **Data Classification**

In this section the revenues were sorted in three categories: (i) low where Nikes’ quarterly income was between 6,500 and 7,500 dollars in million, (ii) medium where the range of the revenue is 7,501-8,500 dollars in million, and (iii) high where Nikes’ earnings are from 8,501 to 9,500 dollars in million. After which five independent
variables related to comments and tweets were selected and distributed in those classes. Classifying those variables in three categories can help to investigate their changes in relation to the revenue. Moreover, it shows the standard deviation of those variables in each category and the degree to which there are similarities between them.

Facebook positive comments (see figure 7 left) increased in relation to the revenue. Among the positive comments those in the low category had the lowest comments ($M=4,481$, $SD=1,617$) while in the high category the comments had increased ($M=4,894$, $SD=1,878$). However, the opposite had happened with positive tweets (see figure 7 right). The positive tweets in the lowest category had a mean of 4,559 ($SD=2.108$) and fell significantly to a mean of 1,200 ($SD=601$).

For Facebook negative comments (see figure 8 left) it can be seen that they have a similar distribution as the positive ones. The mean of negative comments in the low category was 1,600 comments with a standard deviation of 761 after which they increased until the high class where the comment reached a mean amount of 2,323 ($SD=870$). Even though the negative comments increased accordingly to the revenue,
the augmentation was not as sharp as in the case of positive comments. On the other hand, negative tweets (see figure 8 right), had a rapid reduction from a mean of 908 with a standard deviation of 345 in the low category they dropped to a mean of 291 (SD=130) in the high category.

Facebook total comments, were at the lowest point in the first category (M= 26,959, SD=11,662), whilst Twitter total comments reached its highest point (M=9,006, SD=4,064). In the next category Facebook total comments increased having a mean of 30,814 with a standard deviation of 25,916 (see figure 9 left). In contrast, Twitter total mean comments were lower compared to the low category as the mean was 5,719 with a standard deviation 0.41 (see figure 9 right). In the high category, Facebook total comments were slightly less than the medium category with a mean of 28,605 (SD=6,704), whereas the mean of Twitter total tweets was almost half of its medium category.

**Figure 9.** The blue bar shows the mean of Low total comments, the orange bar indicates the mean of Medium total comments, and the grey bar depicts the mean of High total comments.

**Figure 10.** The blue bar shows the mean of Low positive divided negative comments, the orange bar indicates the mean of Medium positive divided negative comments, and the grey bar depicts the mean of High positive divided negative comments.
Both social media started with a higher fraction of positive/negative in the low category having a mean of 2.99 (SD=0.65) for Facebook and a mean of 4.95 (SD=0.64) for Twitter (see figure 10). After which comments and tweets declined steadily in the medium category which led to a slight decrease to their standard deviation as well. In the high category the mean for the Facebook fraction was 2.13 with a standard deviation of 0.26 while for the Twitter fraction the mean was 4.08 with a standard deviation of 0.43.

The fraction of positive/total for both social media platforms (see figure 11) had a conflict resolution. More precisely, the mean of the Facebook fraction was 0.17 (SD=0.02) in the low category, while for Twitter the mean of the fraction was 0.50 with a standard deviation of 0.02. The medium category is slight lower in both social networks with a bit lower standard deviation. As for the high category, the Facebook fraction grew slowly, however the mean was higher than the low category (M=0.18, SD=0.09). In contrast, Twitter continued decreasing gently reaching a mean of 0.38 with a standard deviation of 0.03.

All in all, this section has shown that the mean of comments is in proportion to the revenue in almost all cases. However, the mean of positive/negative comments is in inversely proportional to the revenue. This outcome might be explained by the fact that the third category (high) contains a higher mean of both positive and negative comments which leads to a smaller fraction compared to the first category (low). The mean of tweets, on the other hand, is in an inverse proportion to the revenue in all cases. This means that when the revenue increases the amount of tweets decrease regardless of their classification. It is difficult to explain this result, but it might be related to the limited social network activity on Twitter’s Nike hosted page over time.
As the data differ between Facebook and Twitter I calculated the lower bound (mean-1SD) and the upper bound (mean+1SD) of the error bar for each group to investigate if the groups. Calculating those bounds is important as it shows the variability of the data on each group. As we can see there is no statistical significance differences between the three categories for both Facebook and Twitter as the distributed data overlaps. This indicates that the gathered data in the different categories (low, medium, high) have indeed some similarities.

5.4. Analyzing the multiple linear regression model

This section is divided into three parts. In the first part, the values of the ten independent variables, that were previously classified into one of the three categories (see section 5.3), are examined in order to identify the degree to which they can explain the revenue. Those variables that have a strong correlation with the revenue are then used for the Nike’s revenue prediction. While the last part refers to the creation and the evaluation of the forecasted model. The result of this assessment shows the extent to which an accurate prediction can be done with the use of the specific independent variables.

5.4.1. Identification of the independent variables

In order to implement the stepwise regression all the independent variables from both social media platforms were added to find the model that best explained the constant. After sequentially including independent variables and removing irrelevant ones, the fraction of the positive/total for both Facebook and Twitter seemed to be the most suitable. The table below shows the excluded variable and their significance compared to the dependent variable. As we can see their p-value (significance) is higher than 0.05 which indicates that there is no strong correlation between those variables and the revenue and therefore they excluded.

<table>
<thead>
<tr>
<th>Model</th>
<th>Beta</th>
<th>In</th>
<th>T</th>
<th>Sig.</th>
<th>Partial Correlation</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB positive</td>
<td>.200</td>
<td>2,203</td>
<td>.044</td>
<td>.494</td>
<td>.936</td>
<td></td>
</tr>
<tr>
<td>FB negative</td>
<td>.220</td>
<td>2,517</td>
<td>.024</td>
<td>.545</td>
<td>.946</td>
<td></td>
</tr>
<tr>
<td>FB neutral</td>
<td>.204</td>
<td>2,059</td>
<td>.057</td>
<td>.469</td>
<td>.815</td>
<td></td>
</tr>
<tr>
<td>FB total</td>
<td>.200</td>
<td>2,123</td>
<td>.051</td>
<td>.481</td>
<td>.889</td>
<td></td>
</tr>
<tr>
<td>TW positive</td>
<td>.177</td>
<td>1,114</td>
<td>.283</td>
<td>.276</td>
<td>.378</td>
<td></td>
</tr>
<tr>
<td>TW negative</td>
<td>.152</td>
<td>.998</td>
<td>.334</td>
<td>.250</td>
<td>.413</td>
<td></td>
</tr>
<tr>
<td>TW neutral</td>
<td>.100</td>
<td>.801</td>
<td>.435</td>
<td>.203</td>
<td>.638</td>
<td></td>
</tr>
<tr>
<td>TW total</td>
<td>.139</td>
<td>.977</td>
<td>.344</td>
<td>.245</td>
<td>.474</td>
<td></td>
</tr>
<tr>
<td>FB positive/negative</td>
<td>-.157</td>
<td>-1,360</td>
<td>.194</td>
<td>-.331</td>
<td>.683</td>
<td></td>
</tr>
<tr>
<td>TW positive/negative</td>
<td>.030</td>
<td>.237</td>
<td>.816</td>
<td>.061</td>
<td>.628</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Excluded variables
5.4.2. Multiple linear regression deduction

After deciding the independent variables the multiple linear regression analysis was performed in SPSS. The specific model helped to identify the degree to which a set of factors can predict the dependent variable. In this model the dependent variable was the revenue while the two independent variables were the positive/total comments and tweets respectively. The intention was to examine whether the null hypothesis was accepted or rejected (see section 4.1).

5.4.2.1. Coefficient and Multicollinearity

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>12,945</td>
<td>,629</td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>4,442</td>
<td>1,456</td>
<td>.304</td>
</tr>
<tr>
<td>Twitter</td>
<td>-12,983</td>
<td>1,403</td>
<td>-.921</td>
</tr>
</tbody>
</table>

*Table 3. Coefficients*

To implement multiple linear regression analysis the first step is to examine the degree to which each independent variable contributes to the prediction of the constant variable. This was done by checking the b unstandardized coefficients. As we can see the b0=12.945 while b1 = 4.442 and b2=-12.983. This means that for every one unit of change for the value of Facebook positive/total variable the revenue value increased about 4.44 if the other predictor (Twitter positive/total) was held constant. While for every one unit of change in the value of Twitter positive/total variable the Revenue value decreased about 12.9 if the other predictor (Facebook positive/total) was held constant.

However, the b2 seems not to be in the right direction as negative signs usually indicate a multicollinearity problem. In other words, I had to examine if the two independent variables were intercorrelated (Investopedia, 2018). Investigating the collinearity statistics (last two columns) can show if there is a misleading in our results. As shown in table 4, the tolerance for both predictors was higher than 0.20. As for the variance inflation factor (VIF), Field (2013) stated that the value needs to be lower than 3 for each variable and their mean close to 1. In our case both independent variable had VIF below 3 and their average was 1.029 which is not much higher than 1. As a result, the multicollinearity problem cannot be accepted.
5.4.2.2. Examining the statistical significance

A

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>T</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>12,945</td>
<td>-</td>
<td>20,585</td>
<td>0.000</td>
</tr>
<tr>
<td>Facebook</td>
<td>4,442</td>
<td>1,456</td>
<td>3,050</td>
<td>0.008</td>
</tr>
<tr>
<td>Twitter</td>
<td>-12,983</td>
<td>1,403</td>
<td>-9,253</td>
<td>0.000</td>
</tr>
</tbody>
</table>

After rejecting the multicollinearity problem, it is time to examine if the two independent variables are statistically significant for the constant variable. The Coefficients table (table 4A) implies that both Facebook positive/total variable and Twitter positive/total variable had a statistically significant impact on the outcome variable.

To illustrate that we need to check their p-value and their absolute t-value. As we can see their p-value =0.008, p-value=0.000 respectively (less than 0.05) while their absolute t-value was more than 1.96. This means that the null hypothesis can be rejected as b0 ≠ 0 , b1 ≠ 0, b2 ≠ 0.

Moreover, the ANOVA table (see table 4B) shows if the specific model is significant. The model significance was F(2,16)= 43.954 with p-value=0.000 (less than 0.05) which means that the percentage for F to be that high in a null hypothesis is less than 0.1% and therefore the model was statistically significant.

5.4.2.3. Examining Confidence Interval for B and Coefficient

A

<table>
<thead>
<tr>
<th>Model</th>
<th>95.0% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>(Constant)</td>
<td>11,612</td>
</tr>
<tr>
<td>Facebook</td>
<td>1,355</td>
</tr>
<tr>
<td>Twitter</td>
<td>-15,957</td>
</tr>
</tbody>
</table>

Table 4. A. Coefficients, B. ANOVA
The Confidence interval for the unstandardized coefficient is also really important to calculate as it implies the actual value of the variables. Taking into consideration the standardized coefficient we can examine the degree to which each independent variable impacts on the constant variable. A predictor with a higher value has a greater impact on the dependent variable. In our case, we are 95% confident that both social media platforms had a significant effect on revenue as absolute t-value was higher than 1.96. Moreover, checking the range of values I discovered that Facebook can take values between 1.355 (lowest value) and 7.530 (highest value) while Twitter can take values from -15.957 (lowest value) to -10.009 (highest value) (see table 5A). As for the constant value, it can be between 11.612 and 14.278 (see table 5A). This means that we are 95% confident that if the b coefficient of Facebook and Twitter are between the aforementioned ranges then the values are valid.

In addition, the model summary table shows the proportion of the variance in the dependent variable that can be explained by the predictor variable (see table 5B). In multiple regression, the adjusted R square is taken into consideration only if there is a huge difference in the value from the R square value. In our case, the two values were close enough and for this reason we can use the R square value instead. This value implies that 84.6% of the variance in revenue can be explained by the fraction of positive/total for both comments and tweets. Even though, other factors might affect the variation in the revenue as there is a 15.4% that remains unexplained the model was still adequate to be used for the revenue’s prediction.

5.4.2.4. Testing Autocorrelation

Having a data that is in time series it is also important to examine the autocorrelation. This means that the previous data might influence the future data and that the errors are related to one another (Field, 2013). Using the Autocorrelation table (see table 6) in relation to the ANOVA table (see table 4B) we can find if there is autocorrelation between the specific time series and its lagged version. The Durbin-Watson table for

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.920</td>
<td>.846</td>
<td>.827</td>
<td>.313052479</td>
<td>1.890</td>
</tr>
</tbody>
</table>

Table 6. Autocorrelation
5% of significance was used for F(2,16) in order to find that the dL=0.982 and the du=1.539 (Zaiontz, 2018). In general the DW value should not be between 0 and dL. In our case, the DW was equal to 1.890 which is not between 0 and 0.982 (Zaiontz, 2018). As a result, we can conclude that there is no autocorrelation and the multiple linear regression model that was created is infallible and valid.

5.4.3. Prediction Process

This section refers to the forecasted model and its evaluation by using the process that was described in the methodology chapter (see section 4.4.2 and section 4.4.3).

5.4.3.1. Predicted Model

In the previous section the unstandardized coefficients were calculated and their values were proved to be different than zero. Moreover, it was proved that the null hypothesis can be rejected as the model was statistically significant. This means that the two independent variables can be used to create a model that can predict the dependent variable. By using the regression equation (see section 4.4.2) that was mentioned in the methodology section and by substituting the b variables with the values that was found in the coefficient table (see table 3) the following predictive model was created.

\[
Revenue = 12.945 + 4.442 \times Facebook - 12.983 \times Twitter
\]  (5)

5.4.3.2. Model Reliability

After creating the model by using the aforementioned equation (see equation 5) it is time to examine its accuracy. As the linearity between the exploratory variables and the constant does not always exist, it is important to test the normality of the residuals. The term residuals refers to the faults that exist in the prediction and their calculation can be done by subtracting the predicted value from the actual value. Figure 12 shows that the residuals follow a diagonal line which points out that they are normally distributed. It is visible from the plot that, except for a deviation between 0.3 and 0.4 the data points are close to the regression line.
Apart from normality it is also important to examine the homoscedasticity. The term homoscedasticity describes the degree to which the residuals are equally distributed. This means that there is no systematic process to make the residuals be either greater than zero or less than zero but they are scattered. As we can see from figure 13 the predicted values against the standardized residuals conform the homoscedasticity as they do not follow a clear pattern and they are fairly divided into the two parts of the Y axis. As a result, this bivariate plot indicates that there is a linear connection between the independent variables and the outcome.
5.4.3.3. Evaluation
After proving the normality and linearity of the model that verified its validity it is time to evaluate the results of the regression equation by comparing the actual values with the predicted ones.

![Graph showing actual and predicted quarterly revenue](image)

*Figure 14. The blue line shows the actual quarterly revenue, the orange line depicts the predicted quarterly revenue.*

According to the graph as the time passes the prediction seems to be more accurate. In the first quarter of 2014 the predicted value was higher than the actual value. Despite some fluctuation between mid-2014 and Q1-2015, in the second quarter of 2015 the predicted value overlapped with the actual value. Similarly, in the fourth quarter of the same year the two values also crossed each other. The following period had some inaccurate predictions until the fourth quarter of 2016. Even though for the next one and half year the forecasting value did not overlap the real one, the prediction was closer compare to 2014. Therefore, the forecasted revenue overlapped in some cases to the real revenue.

However, it is important to calculate the root mean standard error (RMSE) in order to measure the accuracy of the predicted model. RMSE shows the noise or the points that are away from the fitted regression line. Using the equation that was described in the methodology section (section 4.4.3) the value of the RMSE was calculated around 287, 279 in million. This means that the typical point is equal to 287, 279 in million away from the line when the actual revenue is between 6 billion and 9.5 billion.
6. Concluding Remarks and Future Perspective

This chapter offers concluding remarks and future perspectives on the relationship between the two social media platforms and the Nike’s revenue. The summary section refers to the aims of the thesis, the research questions and the results that were found from the regression analysis. The next section discusses the limitations of this study and recommends possible future researches that could be done in this area. While the last section details the concluding observations of this paper.

6.1. Summarizing the findings

Investigating the relationship between the data produced by multiple social media and the revenue of a single firm was an under-researched area. The aim of this thesis was to contribute to this gap by examining if the sentiment of user-generated content is related to Nike’s revenues.

The first question which this paper tried to address, was:

- *Can the sentiment of Facebook comments, Twitter comments and/or a combination of both have an impact on Nike’s quarterly revenue?*

To achieve it, three online sources were used, including two social media platforms such as Facebook and Twitter, and one webpage the Nike’s official website. After classifying the social media data according to their sentiment, and finding the independent variables that best explain the constant, a multiple regression analysis was performed. The analysis showed that the mixed model was able to explain 84.6% of the revenue's variance. Moreover, by taking into consideration the findings of the Coefficient table which indicates that the b values are different from zero, the below alternative hypothesis can therefore be accepted:

*The sentiment of Social media data like Facebook comments and Twitter tweets have an association to Nike’s quarterly revenue.*

The second question was:

- *Can those social media comments be used to as predictive variables?*

In order to answer this question the accuracy of the model was first examined. We saw in the Results and Discussion chapter (see section 5.4.3.2) that the residuals were
normally distributed as the data was close to the regression line. Moreover, it was noted that a linearity existed between the independent variables and the dependent variable as the data was scattered without having a specific pattern. The multiple regression model was then evaluated to investigate how well it could perform. The results found that Nike’s revenue could be forecast as the value of the RMSE was low. Therefore one can conclude that the regression model was valid and the fraction of positive/total comments can be used for the prediction of Nike’s financial outcome.

6.2. Limitations
Using the fraction positive/total comments a regression model was created. Even though the study has reached its goals, there is a 15.4% of the revenue’s variance that the selected independent variables could not explain. A number of important limitations need to be considered as there are different possible reasons for not finding a model that can completely explain the dependent variable.

- **The exclusion of non-English brand pages**
  This study was concentrated on comments and tweets written in English. Therefore, fan pages that contained content written in a different language were excluded. For instance, Nike Sportswear and Nike Women pages are two popular brand pages with 15 and 5 million likes on Facebook respectively, however, they were not included as the written comments are in Swedish. In a similar manner, Hurley with around 4.5 million likes on Facebook was left out as the content is in French.

- **The use of common brand pages**
  Furthermore, I decided to collect data only from official brand pages that existed on both social media platforms. This limitation means that I took into consideration Nike’s pages that are available for both Facebook and Twitter in order to investigate if one of the examined platforms overweighed the other. For this reason, an active Facebook brand created page, i.e. the Nike Air Force 1, with almost three million likes, was excluded as it did not exist on Twitter.

- **Not gathering data from Greater China**
  Another issue is related to the predictive model that I created for the revenues. It seems to be able to predict the revenue quite accurately, though the sample that I used, was limited. According to Beulah (2017), during 2017 the sales in Greater China region including China, Hong Kong, Macau, and Taiwan accounted for 13%
of Nike’s global revenue. However, Facebook and Twitter are two of the social media platforms that are banned in China mainland from 2009. As a result, data were not collected from respective countries.

6.3. Future Research

**Improving the dataset**
Due to the limited time, I decided to investigate the impact on Nike’s global revenue by doing the sentiment analysis of customers’ opinions, written on Facebook and Twitter Nike’s pages. A follow-up research would also contribute to investigate other possible variables that might have an association with Nike’s revenue.

- **Investigating different social media platforms**
A giant firm like Nike uses several social media platforms such as YouTube, Pinterest and Google+ to connect with its clients. Moreover, as we discussed in the previous section (see section 6.2) Greater China highly contributes to Nike’s sales growth. However, most of the aforementioned social media platforms are banned in the region. Therefore, it is recommended that further research could investigate Nike’s hosted pages from other social media platforms that are available in Greater China such as Weibo, Ren Ren and QQ (Helixa, 2013).

- **Investigating different metrics**
I decided to focus only on two variables such as comments and tweets that gathered from the Nike hosted pages on both platforms. Future research can also take into consideration other volumes that might influence revenues such as the amount of likes, the volume of shares, and the volume of retweets. Moreover, the most influencing people of those pages could also be part of the investigation.

- **Investigating fan pages**
Moreover, further research would be of great help in investigating the sentiment of customers’ opinion on other Nike pages, and how it relates to Nike’s global revenue. For instance, fan pages that are created by the supporters of the brand, could also be considered. As it might be easier for people to express their opinions on a page that are not monitored by the actual brand. Taking into account those comments might lead to a stronger relationship between the social media platforms and the company’s financial outcome.
Changing the sentiment analysis technique
For this study a lexicon and rule based sentiment analysis named VADER was used as it was implemented for social media analysis. Even though, it was a suitable model for the sentiment analysis, I would recommend that future studies can focus on machine learning technique to examine how well they will perform in labelling text on social media. A machine learning method might also be beneficial in the investigation of comments written in different languages as the researcher can train the data which enables the process compared to techniques like VADER that are implemented only to support a single language like English. It would be interesting to see how the result might vary if other languages are taken into consideration.

Improving the predicted model
For the implementation of the predictive model, I used linear regression and the results showed that there is a significant relation between the comments and tweets from the subsidiaries fan pages and Nike’s quarterly revenue. I have considered quarterly revenues in this research paper. A different approach can be deliberated in future studies. Seasonal regression might be a suitable method for conceptualizing revenue’s predictive model as the data is classified into quarters.

6.4. Conclusions
This thesis had two main goals: to investigate if there is a connection between the comment’s sentiment of different social media platforms and the revenue of a single organization, and whether or not it is feasible to create a model that could predict the financial performance of the specific company with the use of those comments.

The financial prediction based on the people’s opinion by using a single social media was not a new field of investigation. However, most of those studies focused only on Twitter as a successful social network for this prediction. Similarly, this paper proved that Twitter has a role to play when it comes to generating revenues. A significant difference compared to earlier studies, is that the findings of this investigation showed a higher association between the use of a mixed model and the firm’s revenue. Thus, unlike previous research works that focused on positive and negative comments this thesis also took natural comments into account. By calculating the fraction of positive/total, only the positive comments out of total were taken into account. This shows how much consumers have been satisfied with Nike Company. A multiple regression model was implemented to convert comments and tweets into revenue’s prediction to
show the importance of considering the social media data, produced by multiple platforms.

However, this study is not without limitations. One source of weakness is the number of people that exist on those platforms. As we discussed in a previous section of this chapter (see section 6.2) the selected platforms are not available in all countries. Moreover, this dissertation might have some kind of self-selection bias as not all the population that purchase products from Nike Company have an account on social media platforms. In addition, those users that are active on Facebook and Twitter do not always share their opinion online. As a result, the predictive model that was created in this dissertation cannot be totally reliable for the prediction of the universal revenues due to the limited sample that was examined.

Another implication of the current paper is related to the short time series. The predictive model was created by taking into consideration the data that were acquired between 2014 and 2018. This means that the prediction can be accurate to some extent only if the tendency of the data remains the same. In other words, we saw that the tweets have a different inclination compared to the Nike’s revenue. However, if the pattern changes, the predictive model will not be able to forecast the revenue.

All in all, the impact that Nike-created pages have on its revenue, can be seen as a starting point for other companies. The methodology that has been performed in this paper is applicable to any active organization on social media platforms. Hence, the proposed procedure is not only be limited to investigating companies like Nike but also corporations in different fields. Nevertheless, the suggested model lacks generalization as every firm has a different level of interaction with its social networking websites. Thus, these findings can only be used as a guidance for future research.
7. References


