ROTTERDAM SCHOOL OF MANAGEMENT

Using twitter to Explore the Effect of Social Media Buzz on Sales

Bachelor Thesis

Meeke de Jong
Tom van Berkel
Evita Bizune

3rd June, 2013

This document is written by Meeke de Jong, Tom van Berkel, and Evita Bizune, who declare that each individual takes responsibility for the full contents of the whole document. We declare that the text and the work presented in this document is original and that no sources other than mentioned in the text and its references have been used in creating it. RSM is only responsible for supervision of completion of the work but not for the contents.

Contents

Management Summary

In this thesis we have extensively explored the use of social media for predicting business outcomes. In particular, we looked at the effect of the volume of social media buzz and the size of the audience reached by buzz regarding products and its corresponding sales. By using Twitter as a social media platform and movies as a research unit, a study has been conducted in order to contribute to the existing literature about predictive analytics useful for marketing units. As described in our thesis, we have found high results that support our hypotheses, implying that our thesis is valuable for managerial implications and future research.

Introduction

The development of accurate predictive tools is important for any business across all industries (Nyce, 2007). Kucera and White (2012) have found that companies that have introduced predictive practices for their marketing campaigns have outperformed companies that have not done this. Moreover, the growth of the Internet has allowed researchers to discover and focus on new methods of conducting research regarding predictive marketing analytics instead of traditional research methods. One of the effects of the growth of the Internet is the rise of social media like Twitter. It can be concluded that Twitter can be perceived as one of the novel ways of research methods, where predictive analytics can be applied to discover popularity trends of future consumer behaviour. We believe we can develop a study that 1) benefits from the advantages of online research, 2) accounts for the disadvantages that come with several of the online tools currently used and 3) has an extensive theoretical outline based on scientific evidence.

Specifications of Our Study	
Hypothesis 1	The more there is social media buzz regarding a product in the entertainment industry, the higher on average are its sales.
Hypothesis 2	The bigger the audience that is reached by tweets about a product in the entertainment industry is, the higher on average are the sales of the product.
Complementary Variable	YouTube trailer views.
Focal Unit	Products in the entertainment industry.
Theoretical Domain	The products in the entertainment industry of all types, in all languages, in all countries and at all times.
Type of Relationship	Causal, probabilistic and positive.

Literature Review

In the past, several studies were conducted which focused particularly on predicting sales' outcomes by using social media. After analyzing and comparing the different elements of the six studies, it can be concluded that there are no remarkable historical trends noticeable, except for minor trends regarding the effect size parameter and the confidence intervals. Most of the studies conclude that the volume of tweets is a strong predictor of the sales. However, our most essential finding from the meta-analysis is that there is no consistency in the effect size of the relationship between the volume of the social media buzz itself and the sales of a product. This is due to the varying results in regression coefficients, a low overall effect size, and the fact that there are no overlapping confidence intervals. According to our additional investigation, more variables in a regression model in contrast to only one variable provide more information regarding sales' trends. This implies that by introducing multiple variables into our research, besides our main variable (volume of tweets), we will be able to assess more optimally which factors have an influence on a product's sales. These additional factors will consist of the number of total views of the official trailers on YouTube, the amount of posted blogs, and the number of theatres the movie is released in. Besides, to add additional value to the existing literature, we will introduce a complementary study in which we will take the amount of followers into consideration in relation to the product's sales. The existing literature has not yet scrutinized this variable as well as YouTube views variable.

Methods

A longitudinal research strategy will be appropriate for our given hypotheses. The tool NodeXL will be our measurement instrument, which will be used in order to collect our tweets and calculate the volume and the amount of followers. BoxOfficeMojo will provide information regarding releases dates, sales data of movies and number of theatres screening the movies. We defined a comprehensive measurement protocol to avoid possible errors and ensure validity and reliability of the study.

Results

The data were collected between 5th April and 16th May 2013, consisting of a total sample size of 30 movies. The movies were released in the United States within this timeframe. In total, we collected 161,317 tweets for 30 different movies that accumulated to 887,979,439 total followers within these timeslots. For our main hypothesis, we aimed to investigate whether the effect of the amount of followers, expressed as an average number of followers reached per hour, is higher than the effect of the volume of tweets, which is expressed as a tweet rate per hour, about a particular movie. The obtained results are presented below:

Results			
Independent variables	Pearson correlation	Confidence Interval	Sample size
Number of followers	0.975	$0.9748 \leq r \leq 0.9752$	30
Volume of tweets	0.800	$0.7982 \leq r \leq 0.8017$	30
YouTube views	0.876	$0.8749 \leq r \leq 0.8771$	30

When including the variables the following regression model was created, resulting in an R-squared value of 0.996 and an adjusted R-squared of 0.991.

$$y_{Sales} = \beta_0 + \beta_1 \times x_{youtube} + \beta_2 \times x_{followers} + \beta_3 \times x_{tweets} + \beta_4 \times x_{theatres} + \beta_5 \times x_{blog\ posts} + \varepsilon_i$$

Moreover, we have investigated the effect of number of followers and tweets when data were gathered one week prior to the release of a movie. The findings have shown a very strong relation of the number of followers (r = 0.983) and number of tweets (r = 0.998) with the sales.

Managerial Implications and Future Research

Derived from our findings, several managerial implications are applicable. For instance, a business could receive more valuable knowledge when applying our study regarding the future trend of a product, the choice of a marketing strategy, for competitor analysis, and to predict inventory. Besides that, we are unique in exploring and providing insights about the effect of the number of followers which are reached by a tweet and the corresponding sales of a product, and by including the number of YouTube views as an extra variable.

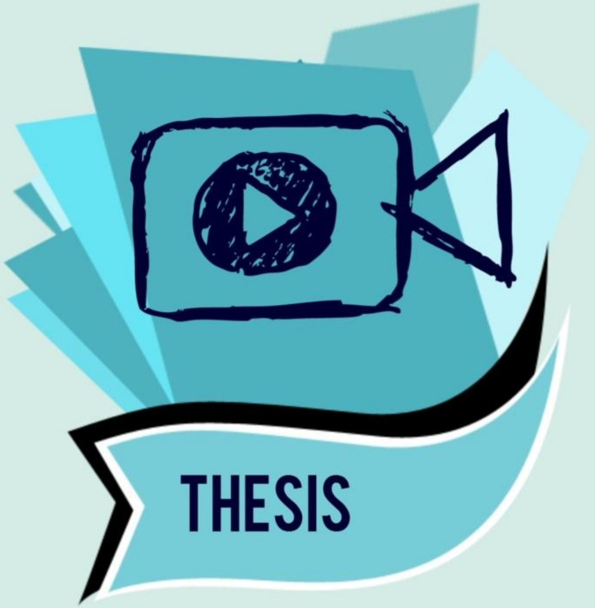
A few suggestions for future research can be made:

- Inclusion of followers and YouTube views as central variables.
- Extension of the regression model.
- Incorporating the changes in the business model of the movie industry that will certainly affect the dependent variable.
- Elaborate and expand on our study, such as a further division between different time zones, assessing differences between non-English and English speaking areas, the penetration of Twitter users per country or generalizing to other areas of interest.

- Enhancing and expand the predictive tool we created.

It can be concluded that our study contributes to the existing literature by proving that the size of the audience is a stronger predictor than the volume of social media buzz regarding the sales of products in the industry. Besides that, results show that YouTube trailer views are indeed an indicator for the success and sales of a movie. With this valuable information, predictive analytics using social media for marketing proposes can be enhanced.

On the next page, an infographic is presented with an interactive depiction of the main findings of our study.



ROTTERDAM SCHOOL OF MANAGEMENT

BACHELOR THESIS IBA

Meeke de Jong
Evita Bizune
Tom van Berkel

2013

PREDICTING SALES...

...of the Entertainment Industry by using Twitter

The aim of our study is to predict sales in the entertainment industry by using data from the social network Twitter.com. We focus on box-office movies as our research unit. Our main findings are that both the total social media buzz as well as the size of the audience that is reached by tweets can predict the sales of products in the entertainment industry. Furthermore, including more variables in the regression model improves the R square and increases its significance.

BASIC DATA

Total Sample Size:	30 MOVIES
Total Number of Tweets:	161.376
Avg Followers / Tweet:	4.190
Number of Weeks:	6
Tool Used:	NODEXL
Sales Data:	BOXOFFICEMOJO

HYPOTHESES

"The more there is social media buzz regarding a product in the entertainment industry, the higher on average are its sales."

"The bigger the audience is that is reached by tweets about a product in the entertainment industry, the higher on average are the sales of the product."

REGRESSION MODEL

$$Y_{sales} = B_0 + B_1 * X_{YouTube} + B_2 * X_{Followers} + B_3 * X_{Tweets} + B_4 * X_{Theatres} + B_5 * X_{BlogPosts} + E_i$$

EFFECT SIZE: CORRELATION

Displayed below is the correlation between our main independent variables with our dependent variable: the total revenues of the box-office movies.

0.975	0.800	0.876
Number of Followers	Number of Tweets	YouTube Views

***Abstract** - This article argues that social media buzz as well as the size of the audience reached by the buzz can be used to predict the sales of products in the entertainment industry. In particular, we examine Twitter and YouTube data to predict the revenues of box-office movies. We state that the average number of followers per tweet per hour (follower rate), the average number of tweets per hour (tweet rate) and a novel variable of the total number of YouTube views for an official trailer are all strong predictors of the movies' sales. Furthermore, we argue that including more variables such as the total number of theatres a movie is released in, improves the explanatory power of the regression model. Our paper first describes predictive analysis and how the corresponding research methods have evolved over time. Subsequently, our hypothesis is specified and the independent variables introduced. We then outline and analyze the existing literature which is ambiguous about the predictive power of the tweet rate. The results of the imported data from Twitter show a strong correlation between the tweet rate and the revenues of the box-office movies, which resolves the ambiguity around this variable. The follower rate has proven to be an even stronger predictor of the movies' sales, which highlights a focus on this novel variable for future research regarding the predictive power of Twitter. Finally, a predictive model will be established which generalizes and specifies the various variables that influences a movies' sales.*

Keywords - Predictive analytics, Twitter, social media, YouTube, NodeXL, campaign development, marketing, sales, entertainment industry, movies, followers.

Section 1. Introduction

In the first section of this paper, the importance of predictive analytics and their implications on today's business operations is discussed.

The development of accurate predictive tools is extremely important for any organization or business across all industries (Nyce, 2007). The author further argues that many business activities, such as inventory management, price setting and marketing benefit from precise forecasts. We will therefore first explain how the use of predictive analytics has evolved over time and more specifically, which drivers have facilitated an increasing interest in this field. We will analyze emerging marketing trends and relate these to using predictive tools. Subsequently, an outline of the traditional research methods that were extensively used for forecasting will be conducted. These traditional manners will be compared to novel ways which have accelerated since the growth of the Internet and offer important advantages for solving various research biases and other limitations. In particular, we will introduce the social network Twitter as medium to gather data for predictive purposes.

1.1. Predictive Analytics

Predictive analytics is a broad term that outlines a variety of statistical and analytical techniques used to develop models that can be used to make accurate forecasts (Nyce, 2007). The field of predictive analytics is not revolutionary and was first introduced in 1920 (Hair, 2007). However, this field has recently enjoyed a substantial increase in interest. More specifically, the business intelligence market has a growth rate of 9% per year, consisting of 80 billion in turnover of which 50% comes from predictive analytics (Zwilling, 2013). Two trends have facilitated a more reliable and efficient use of these forecasting tools. First of all, the availability and quality of data can be regarded as the most essential determent for developing correct estimations (Nyce, 2007). The increasing quantity of data, often referred to as Big Data, has assisted researchers to collect valuable information (Sicular, 2013). The author further states that besides the rise in volume of data, variety and velocity of the data has contributed to the growth in analytics. In particular, the worldwide data storage doubles every 12 months (Gallaugher, 2013). A second important driver is the advances in technology and software meant to retrieve and analyze these data (Nyce, 2007). Eckerson (2007) states that recently developed software that allows for graphical modelling, reduced network traffic and redundant queries and more automatic algorithms are central for making accurate predictions in a more efficient manner.

1.2. Predictive Analytics Enhance Marketing Campaigns

Marketing is a business unit that has particularly benefited from predictive analytics (Kucera and White, 2012). Kucera and White (2012) have found that companies that have introduced predictive practices have outperformed companies that have not done this for two important marketing factors. Firstly, marketers who have included predictive analytics have realized an incremental sales lift almost twice as big as the lift by marketing units that did not make use of these tools. Also, the click-through rate performed 76% better for those departments that used predictive analytics. Kucera and White (2012) explain that this is due to the fact that an accurate prediction of marketing campaigns results in a more reliable segmentation of the target markets and therefore a more optimal marketing strategy. A research by Hair (2007) has elaborated upon the various advantages of predictive modeling for marketing purposes. Hair (2007) specifically states that accurate predictions of marketing campaigns and its corresponding sales allows for a more optimal price setting, introduces distribution alternatives and determines likely responses for advertising communications. Moreover, Siegel (2008) has found that businesses that have embraced predictive models have increased their retention rates by 3% and have achieved a growth of 12% for returning customers.

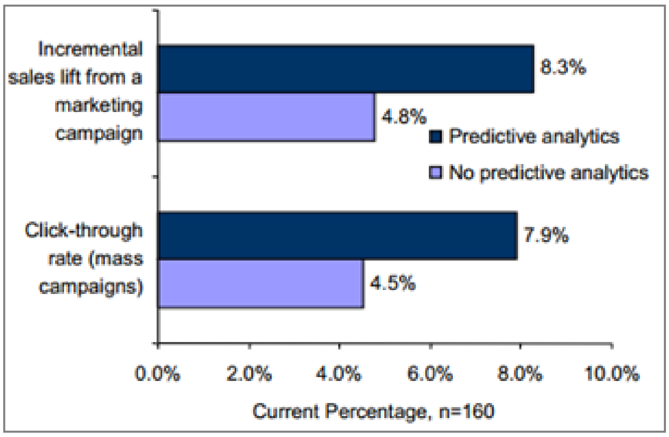


Figure 1: Sales of predictive/no predictive analytics (Kucera & White, 2012)

1.3. Traditional Predictive Marketing Research Methods

Traditionally, conventional marketing research methods have been used to make predictions about future events. Silver (2012) examines the use of expert judgements in order to make forecasts. The researcher tested whether political experts could accurately predict the outcome of presidential elections, but concluded that these professionals performed extremely poor. His explanation of these results is that experts are skilled in explaining phenomena, rather than predicting these. Mentzer (2005) has elaborated on the different methods for expert judgements and found that the most widely used principle is the consensus method, where a combination of sales people, personnel and managers agree upon one forecast. However, he states that this method has heavily suffered from political bias within the company, in which a few powerful members influence the opinions of the whole judgement group (Mentzer, 2005). Another traditional research method used for forecasting is extrapolation (Ducham, 2010). Extrapolation means that a prediction is made based on past data by using statistical models. However, Ducham (2010) argues that this method has a limited applicability, since many newly introduced products have no prior data and marketing trends are rapidly changing. Other conventional manners of making future assessments are using surveys or focus groups. Ducham (2010) has explored the most important limitations of these two practices. Most importantly, people behave differently than they say they will behave. Also, customers may have included their own assumptions without verification when their opinion is asked (Ducham, 2010). In addition, Smithson (2000) has argued that the two most important limitations of using focus groups are that this includes the tendency to comply with socially acceptable norms and the emergence of groupthink, meaning that some dominating participants might influence the thoughts of remaining part of the group, which can result in inaccurate predictions. Finally, Fowler (2002) has elaborated extensively on the use of surveys in general, and warns for the potential moderator biases, meaning that the structure of the questions is extremely error sensitive due to subjectivity of a researcher.

1.4. Internet Provides New Research Opportunities

The percentage of the population in the United States that is active on the Internet has rapidly grown from 43.13% in 2000 to 78.24% in 2011, and is still growing (World Bank, 2013). This growth has allowed researchers to discover and focus on new methods of conducting research instead of traditional research methods. For instance, several researchers have gathered data by using online customer surveys that allowed both for cost advantages as well as reaching a larger audience (Fielding et al., 2008). In addition, online surveys offer convenience to respondents since they can respond at their own pace (Gingery, 2011). However, Gingery (2011) also argues that this method will generally lead to lower response rates and has a higher risk of sampling bias, since only certain demographic groups have a continuous access to the Internet or willing to respond online. Fielding et al. (2008) have extensively examined the use of online interviewing and argued that its main limitation is the easiness to delete or ignore the form. Also, Fielding et al. (2008) argue that this method is more prone to distraction on the respondent side which can lead to several errors in their replies. Hair (2007) warns that these new methods for predictive analytics often do not start with a hypothesis but are essentially data driven. He specifies that traditional researchers often argue that this lack of theoretical foundations can lead to finding a relationship to support preconceived notions, with no inclusion of substantive scientific evidence (Hair, 2007). As will be later explained, we believe we can develop a study that 1) benefits from the advantages of online research, 2) accounts for the disadvantages that come with several of the online tools currently used and 3) has an extensive theoretical outline based on scientific evidence.

1.5. Using Twitter to Predict Trends

Together with the raise of the Internet, came the increase in social media platforms (Smith, 2013). There are multiple types of social media platforms, but microblogs are of particular interest for research purposes. Microblogs can be defined as websites that are particularly useful for sharing time-sensitive information and opinions by using less than 200 characters (Gallaugher, 2013). By far, Twitter is the most popular microblogging service provider, hitting 400 million tweets per day (Farber, 2012). The high integration of these networks into consumers' lives and daily usage mainly for communication purposes give these platforms an interesting marketing potential (Java et al., 2007). McCormick et al. (2013) specify that Twitter has proven to be useful for research because social media data allow observing human behaviour in a real-time setting without influencing the behaviour of interest. In addition, Twitter is extremely cost effective and makes scalability possible (McCormick et al., 2013). We argue that Twitter is more effective for research purposes than other social media platforms. One reason is that the vast majority of the Twitter data are publicly accessible, while the content on other major platforms such as Facebook are behind the so-called walled garden (Gallaugher, 2013). On Twitter, only 6% of the users' accounts are private (Moore, 2009). Besides that, the platform itself is research friendly since it supports importing data through its Application Programming Interference (API). This allows researchers to rapidly import vast amounts of data (Twitter, 2012). Gupta et al. (2013) have proved that Twitter predicts the popularity trends of future consumer behaviour. They further elaborate that these findings are extremely valuable for marketing units since it will help business planners in creating more effective marketing campaigns and allocating marketing budgets more efficiently.

It can be concluded that Twitter can be perceived as one of the novel ways of research methods, where predictive analytics can be conducted to discover popularity trends of future consumer behaviour what will be of particular interest and of high value for marketing units. In our research, we will study this phenomenon by exploring social media as an example of collective intelligence. We will investigate whether a crowd gathered on one social platform has the power to predict real-world outcomes, such as the sales of a product. Our study aims to explore the degree of social media buzz regarding products and the audience reached by the social media buzz related to the average sales of products. We will elaborate on this in the next chapter.

Section 2. Specifications of the Study

In this section, the specifications of our study such as hypotheses, variables and relationships between these variables are identified. Furthermore, we discuss the contribution of our research to the already existing knowledge and estimate an expected effect size.

As described in the *Introduction* chapter, we are aiming to develop a study that benefits from the advantages of online research, accounts for the disadvantages that come with several of the online tools currently used and has an extensive theoretical outline based on scientific evidence. We will focus on exploring social media as an example of collective intelligence by using predictive analytics to investigate the power of a social platform on real-world outcomes like the sales of a product. Hereafter, we will formulate and elaborate on the hypotheses of our study.

2.1. Hypothesis 1

Firstly, we will examine the impact of the social media buzz on the sales of a product in the entertainment industry. We define social media buzz as the informal way of sharing publicly open information between people and it includes a large variety of ways of expressing such as blogs, tweets and posts on online social media platforms. We argue that an increased attention to a product on a social media platform will raise awareness and interest of other users to buy or experience the product. By tracking social media buzz about a product relative to competitive products, future performance of a product could be forecasted and possibly influenced. Hence, our first hypothesis is:

The more there is social media buzz regarding a product in the entertainment industry, the higher on average are its sales.

In our study, social media buzz will be measured by looking at the volume of tweets posted about a related product, mentioning that product in the tweet, which is a message posted by a user of Twitter. We have selected Twitter as a measurement platform because, as discussed in the *Introduction* chapter, Twitter is more effective for research purposes than other social media platforms.

This paper reports on a study which considers movies and their corresponding box-office sales as a research unit. Movies will be a suitable research subject since, according to Sadikov et al. (2009), movies have a known release date, which allows to study the ‘hype’ before the release in relation to ‘success’ after the release. Secondly, movies provide an inherent normalization compared to other domains, since the sales in n^{th} week after the release date are comparable across movies. Besides, users tend to tweet generously about movies which results in sufficient data for analysis (Sadikov et al., 2009). Finally, the sales data of movies are widely accessible and more importantly, published with merely a small time lag (BoxOfficeMojo, 2013). We believe that the findings of this study could be extended to a broader set of cases, namely products in the entertainment industry. The sales can be comparable among products in the entertainment industry since people perceive these products as being similar in their nature. This is supported by Vogel (2007) where music, movies, broadcasting and television programming are all classified as entertainment enterprises.

2.2. Hypothesis 2

An additional study that will be conducted, which will slightly differ from our first hypothesis, is a study that takes the audience that is reached into consideration. We use the total number of followers to measure the audience reached by a tweet. We believe that social media users influence followers by posting messages and this might be a valuable predictor for sales. It is widely proven that friends have a significant influence on people's thoughts and actions. For instance, Altermatt (2003) argues that friends show consistent, albeit modest similarities with regard to their self-perception and appear to be influential with regard to people's abilities. Hence, our second hypothesis is:

The bigger the audience that is reached by tweets about a product in the entertainment industry is, the higher on average are the sales of the product.

As discussed previously, Twitter as a measurement platform and movies as research unit will be suitable for our study. We believe that Twitter users influence followers by posting tweets and this might be a valuable predictor for sales. When a Twitter user tweets positively about a movie, it will raise curiosity by some of his or her followers and indirectly influence them to watch that particular movie in a theatre. It is sensible then that the more is tweeted about a movie, the more the movie is promoted, resulting in higher sales. However, an essential part of this word of mouth advertisement is the scale of the social network of each user. A user with merely ten followers that tweets positively about a movie will reach significantly less people than someone with hundreds of followers. Therefore, a tweet by someone that has many followers will have a much stronger positive effect on sales than a tweet by someone that has fewer followers. Hence, we expect that this variable could be even more reliable predictor than merely the total number of tweets about a movie.

2.3. YouTube as Complementary Variable

Besides, we will explore the effect of other variables on the sales of a product. We expect the number of YouTube official trailer views on the day of a movie release positively affecting the corresponding sales of a movie. We have chosen to include YouTube since it has never been introduced before in relation to predicting sales in the entertainment industry. The platform offers two distinct advantages. Firstly, the total number of YouTube views is publically accessible. Besides that, the platform reaches over a billion users per month, which ensures that there are data even for less popular movies (YouTube, 2013). Although the results obtained from analyzing YouTube trailer views cannot be expanded to a broader set of cases, they would nonetheless provide implications for the movie industry and related marketing campaigns.

We are aiming to clarify the relationship by investigating solely the effect of one of the above variables on the dependent variable of sales, excluding any adjacent variables. Besides that, the total effect of the several independent variables will be investigated with means of a regression in order to build a prediction model which can forecast the box-office sales of a movie for the opening weekend and the subsequent weeks.

2.4. Definition of Concepts

The aforementioned hypotheses are a guideline considering the direction of the focus of this study. In this paragraph, the focal unit, the theoretical domain and other specifications of the study's theory will be discussed.

Focal Unit

The definition of the focal unit indicates that it is a unit or entity about which the theory formulates statements (Hak, 2012). While a narrow focal unit of our research is movies, we argue that the results can be expanded to a broader set of cases, such as TV shows, music sales and musicals as mentioned before by Vogel (2007). This is supported by Vogel (2007) where music, movies, broadcasting and television programming are all classified as entertainment enterprises. Since we are aiming for a generalization of the effect, we will examine *products in the entertainment industry* as the focal unit of our study.

Theoretical Domain

Furthermore, the hypotheses are assumed to be true with regard to the entire theoretical domain, namely, the results are relevant to *the products in the entertainment industry of all types, in all languages, in all countries and at all times*. For instance, as sales of movies' and music albums' releases could be measured in a similar manner, we expect our findings to be applicable also to these instances. Besides that, Twitter is used in nearly every country in the world with the service available in more than 20 languages (Twitter, 2013).

2.5. Conceptual Model

In order to examine our hypotheses, it is essential to define the concepts under investigation. Social media buzz and the size of the audience reached are assumed to cause a change in sales of a product in the entertainment industry. Hence, these variables are considered as independent variables which affect a dependent concept. The latter corresponds to sales of a product in the entertainment industry. Conceptual models that investigate this issue are depicted in Figure 2(a) and 2(b). As these figures depict, we expect there is a causal, probabilistic and positive relationship between the independent variables and the dependent variable. Namely, the more buzz there is in social media or the bigger audience that is reached by tweets, the higher are the sales of a product in the entertainment industry. Moreover, as illustrated in Figure 2(c), we will also examine a combined effect of these independent variables and YouTube views on sales of a product. Additionally, the effect of number of theatres a move is screened in and number of blog posts as reported by IceRocket will be investigated.

Figure 2(a): Conceptual model for social media buzz

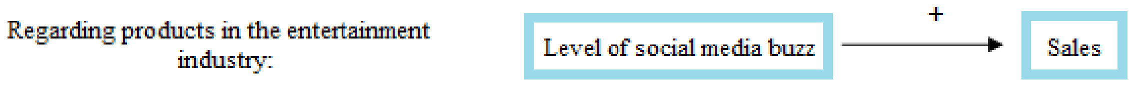

Regarding products in the entertainment industry:
Level of social media buzz
+
Sales

Figure 2(b): Conceptual model for the audience

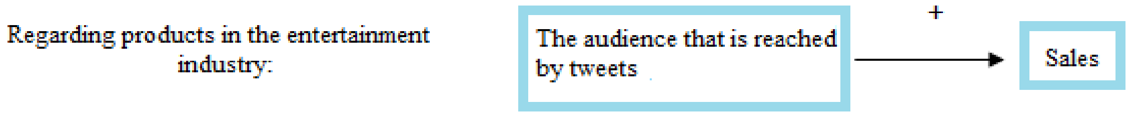
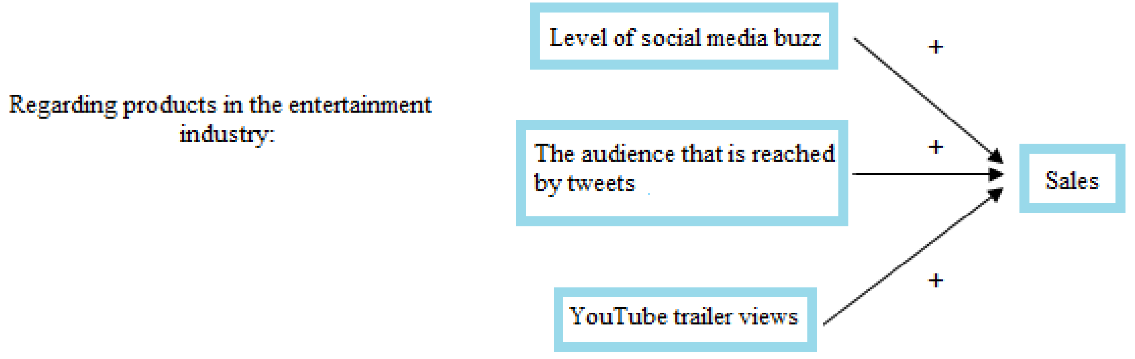

Regarding products in the entertainment industry:
The audience that is reached by tweets
+
Sales

Figure 2(c): Conceptual model for the combined effect


Level of social media buzz
+
Regarding products in the entertainment industry:
The audience that is reached by tweets
+
Sales
+
YouTube trailer views

Causal Relationship

As stated in the previous paragraph, we expect a causal, probabilistic and positive relationship between the independent variables and the dependent variable. The causal relationship is particularly applicable regarding the relationships between the size of the audience and the sales as well as the YouTube trailer views and the sales. This can be explained by the degree of visibility and therefore the possibility to influence people to visit a particular movie. For instance, when a follower of a Twitter user, who tweets about movies, reads a positive tweet about a movie of that specific Twitter user, he or she might get incentivized to watch the movie. In this case, the tweet that reaches the follower will be the cause while the follower visiting the cinema to watch the movie will be the effect. This causal relationship is similar to the YouTube trailer views. When people see a trailer of a movie on YouTube (cause) they might get enthusiastic and will go and watch the movie in the cinemas (effect). This can be explained by the fact that people usually do not watch trailers after they have already visited the particular movie, implying a causal relationship.

2.6. Contribution of Our Research

There are several reasons why our research will contribute to the existing literature and will add value for managerial implications. Liu, Y. (2001) argues that it will give a measure to help forecast box office sales. Among other things, accurate forecast allows studios and theatres to more optimally plan the screening capacity and potentially optimize exhibition contracts. In addition, we are unique in exploring and providing insights about the effect of two independent variables, namely YouTube views and the number of followers which are reached by a tweet and the corresponding sales of a product. Moreover, we will analyze shortcomings in the existing literature and give suggestions for further research.

2.6.1. Managerial Implications

- **The future trend of a product.** When it is proved that more social media buzz or a bigger audience that is reached will cause higher sales, it will be possible for companies to closely monitor the future popularity of their products. When many tweets are posted or many people are reached by related tweets, the company will realize that this product will most probably be a success. On the other hand, potential failures can be predicted beforehand. This information can be used to increase or decrease marketing budgets in an early stage, which would prevent a business from spending more money on a product that will not sell well.

- **Choice of a marketing strategy.** Mohr (2007) has discussed the increasing importance of buzz marketing which comprises word of mouth and viral marketing. In case the business learns in an early stage that a product does not reach the expected level of sales, it can adjust its marketing strategy by involving in a fierce marketing on Twitter. Since we imply that there is a causal relationship between our variables, the business could try to reach out to a large audience in order to raise awareness and potentially increase the willingness to spread the word about the product.
- **Competitor analysis.** Companies can monitor the competitors closely by gathering information from social networks such as Twitter. In this way companies can predict the popularity of the competitor's products when these sales data is not publically available or when there is a large time lag present between the actual sales and its publication.
- **Predict inventory.** If social media buzz can predict sales, companies can use this information to manage their inventories more successfully. Sales are closely related to actual demand, and many companies are forced to sell below cost price due to a demand overestimation (Simchi-Levi et al., 2009). On the other hand, companies have also run short of products because they underestimated their popularity. In case of the movie industry, tweets could be used to predict popularity of a movie and a number of theatres and sessions per day of a movie could be estimated in order to avoid losing sales.

2.7. Expected Size of Effect

The extent to which an effect matters is largely dependent on the managerial or practical implications of the effect. For this reason, it is hard to precisely state the degree of the effect size to be relevant (Hak, 2012). A cost-benefit analyses of the size of effect correlated with the sales must be performed in order to conclude if it will be worthwhile to conduct studies in which the actual effect size is measured. Assuming a practical perspective, the effect on a product's sales caused by the volume of social media buzz or the size of the audience that is reached must be at least higher than the average profit margin in the industry. On the other hand, from a statistical perspective, the results must be higher than a commonly used benchmark (for instance the correlation coefficient must be higher/lower than ±0.5) to have a considerable effect. Applying this to the above described managerial practice of competitor analysis, it can be concluded that the costs that are made to perform a competitor analysis on Twitter cannot exceed the additional sales that are caused by this effort.

Section 3. Literature Review

This section outlines the historical trends of our hypothesis, discusses the quality of the prior researches and a conducted meta-analysis. At the end of the section, complementary analyses such as vote counting and a subgroup analysis have been described.

The interest in using social media to predict outcomes was initiated around 2001 by Liu, Y., where the researchers evaluated the chatter's power to predict stock market outcomes. Thereafter, several studies were conducted which focused particularly on predicting sales' outcomes using social media, as introduced in our first chapter. This literature review aims to review the critical points of current knowledge about the predictive power of social media including substantive findings as well as theoretical and methodological contributions (Boundless, 2013). It should be noted that there has not been performed any prior research regarding the effect of audience on the sales of a product and therefore it this section is focused on the social media buzz.

3.1. Historical Trends

Because the results of a single study have none or little value outside a series of studies alike (Hak, 2012), we have performed an analysis of prior related studies in order to explore historical trends and previous contribution to the investigation of our hypothesis. Cumming (2012) mentioned that a finding in a single study can either be close or far from the actual effect in a population because of sample variation. Therefore the existence of an effect can only be evaluated after reflecting upon multiple studies, and the construction of a replication history is central to the development of a theory (Cumming, 2012). For that reason, it is necessary to conduct research among multiple studies to find a validated effect that would discuss our main research subject. Replication information is gathered by unscrambling several articles that are supporting or confronting our main hypothesis, which has resulted in a focus on articles by Liu Y. (2001), Dewan and Ramprasad (2009), Sadikov et al. (2009), Asur and Huberman (2010), Liu H. (2012), and Rui et al. (2013). The research strategies and results of the aforementioned articles are detailed in Appendix A. Another study by Bakshy et al. (2011) does not explore the effect of independent variables particularly on the sales of a product. It was included in our analysis merely to gain a more elaborative understanding of the followers' influence and their predictive power in general. In this chapter, the replication information is analyzed in order to constitute historical trends regarding our hypothesis. To conclude, there are a few small historical trends discovered from the six articles, but remarkable or significant trends are not present. The research we conducted regarding historical trends focuses on different elements of the studies such as strategy, variables, populations and results. The subsequent findings are stated below.

3.1.1. Historical Trend of the Research Strategy

The research strategies which are used in the six articles are similar, namely a causal design using four longitudinal (Liu Y. (2001), Asur and Huberman (2010), Liu H. (2012) and Rui et al. (2013)) and two time series (Dewan and Ramprasad (2009) and Sadikov et al (2009)) approaches. These two types of research strategies are both covered in the category "a sample of population elements measured repeatedly", where a time-series design observes in one instance of the focal unit how the values of variables change over time and longitudinal strategy is more focusing on the change in the value of the independent variable and a later change in the dependent variable (Hak, 2012). Both studies are perfectly applicable to our general hypothesis, where a sample of tweets is measured repeatedly over time. Although both time series designs were discovered in studies performed in 2009, it is likely that it is to mere chance that our selection of articles showed this difference only in this year. It can thus be concluded that there is no remarkable historical trend noticeable regarding the research strategy of our main hypothesis. However, a critical note should be made that the size of our research analysis is too limited to draw a valid conclusion.

3.1.2. Historical Trend of the Population, Sample Size and Observations

Scrutinizing the populations reveals again that there are none or little differences between these elements. Five out of six of the populations in the analyzed studies consist of box-office movies that are released in the time of the research period (Liu Y. (2001), Asur and Huberman (2010), Liu H. (2012), Rui et al. (2013) Sadikov et al (2009)). Dewan and Ramprasad (2009) differs in their population by researching music album releases and its sales. There is no historical trend that can be explored in the type of population used for this type of studies. If the sample sizes of the studies are compared and recalculated to account for different research periods, four out of six studies use an approximately equal sample size of around 50 box-office movies per five months. In contrast, Dewan and Ramprasad (2009) used a sample size of 962 music albums in a five month period. In addition, Sadikov et al (2009) uses a slightly larger sample size of 82 movies. It is interesting to see that both of these studies used the time series research strategy. However, there is no apparent rationale why a time series strategy would allow for a larger sample size in a five month period. Although, time series studies are usually longer which allows for a larger total sample size than longitudinal studies, all sample sizes are recalculated which allows for a consistent comparison. It is likely that Dewan and Ramprasad (2009) were able to use a larger sample size merely due to a more frequent release of music albums than of box-office movies. Besides that, the sample size in the study of Sadikov et al. (2009) is not substantially different and cannot be related to a trend. Again, a critical note should be made that the volume of the analyzed researches could be too narrow in order to conclude upon a trend in the population, sample size or the observations.

3.1.3. Historical Trend of the Variables

All main variables used in the studies can be categorized as type ratio, which implies that the difference between two values is meaningful (Graphpad, 2013). The ratio type contains the most information of all possible types and allows for sophisticated statistical measurements (Calkins, 2005). The independent variables of all the six studies can be generalized as “online social media buzz”, which includes inter alia tweets, blog posts and Yahoo! Movie posts. In addition, the dependent variables revenue of the box-office movies and the music album sales are ratio types. Since the characteristics of the variables are the same for all six studies, it can be concluded that there is no noticeable trend between time and elements of the main variables.

3.1.4. Historical Trend of the Effect Size Parameters

Three different effect size parameters are used to describe the relationship between the main variables in the six studies. Sadikov et al. (2009) uses a Pearson’s correlation, which enables for determining a relationship between a variable X and Y. Rui et al. (2013) uses a GMM estimator as an effect size parameter. In addition, four articles have applied a regression model to their studies (Liu Y. (2001), Dewan and Ramprasad (2009), Asur and Huberman (2010) and Liu H. (2012)). While some of the articles disclose beta values of their regressions, which is considered a sufficient effect size parameter, all of the studies disclose R-squared values. Although it cannot be considered as a valid effect size measurement, the R-squared assesses the level of explained variability within a regression model. We expect that there is an increased focus on revealing the most optimal approach to predict a dependent variable by using a regression model, rather than quantifying the degree to which the variables are related using correlation (Graphad, 2013). Thus, we argue that there is a historical trend noticeable, in which R-squared is increasing in popularity as opposed to using Pearson’s correlation.

3.1.5. Historical Trend of the Confidence Intervals

The confidence intervals reveal a minor historical trend. Firstly, the confidence intervals of all studies are remarkably narrow, due to the enormous amount of observations. However, the study by Liu (2001) has only 13,000 observations, while other studies have millions of observations (Sadikov et al. (2009), Dewan and Ramprasad (2009), Asur and Huberman (2010), Liu H. (2012) and Rui et al. (2013)). We argue that over time more observations have been included in the studies due to several reasons. Firstly, this trend can be explained by the fact that over the years, the amount of Internet users has grown from half a billion in 2001 to more than 2.5 billion in 2013 (Internetworldstats, 2013). Moreover, an increasing amount of the online time is spent on social media platforms (Habeshian, (2013). On the other hand, technological developments in the measurement tools and the introduction of APIs have significantly supported researchers to collect large amounts of data (Kashyap, 2010). Because of these reasons, the confidence intervals have become smaller over the years.

3.1.6. Conclusion

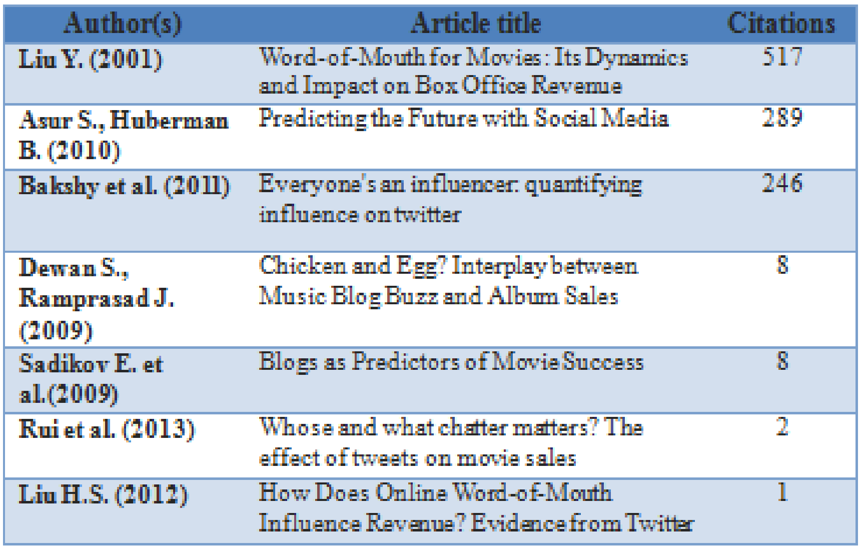
After analyzing and comparing the above elements of the six studies, it can be concluded that there are no remarkable historical trends noticeable, except for minor historical trends regarding the effect size parameter and the confidence intervals. After evaluating the replication history, a legitimate explanation can be found for this exceptional finding. As explained, the popularity of the Internet and social media platform in particular has increased and improvement in measurement tools allows for gathering more data. However, two critical notes should be made regarding the historical trends. Firstly, the research subject is an extremely new topic which makes it virtually impossible to go back into far history. Also, the limited amount of six studies makes it more rigid to draw conclusions regarding historical trends.

3.2. Evaluation of the Quality

In this chapter a more elaborative evaluation will be made regarding the quality of the aforementioned studies which are relevant for our hypothesis.

3.2.1. Citation Analysis

According to the citation analysis in Table 1, there are several articles which have been cited a remarkable number of times while others have received little attention among researchers. The article by Liu Y. (2001) has been cited 517 times, indicating a high interest regarding their research. The researches by Asur and Huberman (2010) and Bakshy et al. (2011) have also been cited 289 and 246 times respectively and could be classified as core publications in the field. The fact that the articles by Liu H.S. (2012) and Rui et al. (2013) have such an insignificant number of citations might be due to the recent publishing of these papers. Although the papers by Dewan and Ramprasad (2009) and Sadikov et al. (2009) have only been cited 8 times, after assessing the quality of these studies, we have decided to include these articles for a more extensive current standpoint about the predictive power of social media.



Author(s)	Article title	Citations
Liu Y. (2001)	Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue	517
Asur S., Huberman B. (2010)	Predicting the Future with Social Media	289
Bakshy et al. (2011)	Everyone's an influencer: quantifying influence on twitter	246
Dewan S., Ramprasad J. (2009)	Chicken and Egg? Interplay between Music Blog Buzz and Album Sales	8
Sadikov E. et al.(2009)	Blogs as Predictors of Movie Success	8
Rui et al. (2013)	Whose and what chatter matters? The effect of tweets on movie sales	2
Liu H.S. (2012)	How Does Online Word-of-Mouth Influence Revenue? Evidence from Twitter	1

Table 1: Citation analysis

3.2.2. Validity of Research Strategies

Each study itself was critically examined, where all research strategies were reflected upon their research methods. Studies that did not explain the way of conducting research were excluded from our final selection of six articles.

Comparing the performance of music album sales or movies box-office revenues with regard to their social media buzz enables us to observe results over a continuous period of time. In the analyzed articles, the advantage of these conditions was exhausted and a longitudinal or time series research strategy was used in order to track the trends. It is true that the most valid results for this type of research could be indeed obtained through focusing on repeatedly investigating many individual cases over a substantial period of time. Although a time series research strategy is considered as a good practice according to the "new" methodology, we believe that a longitudinal study is stronger and more informative due to its higher internal validity (Hak, 2012). Besides, a time series strategy is characterized by its low number of cases measured with a high number of occasions, while a longitudinal strategy focuses on a high number of cases and a low number of observations (StackExchange, 2013). Since we are limited in our time period and there are considerable movie releases per week, a longitudinal research strategy is more suitable than a time series strategy regarding this specific research subject.

3.2.3. Populations

The ease of obtaining data for products' sales within the entertainment industry explains why the aforementioned populations are a reasonable choice for the research purposes. Moreover, movies and music albums have a specific date of release which enables for a longitudinal study and investigation of the effect of the social media buzz on the products' sales depending on the customer (non-)experience with a product. Hence, due to a possibility of determination of several essential characteristics of the data, the strategy of collecting the data within the entertainment industry should be considered as a good practice.

A critical note should be made that five out of six studies examine box-office revenues, and therefore all focus on the same sub-industry. Together with the study about music sales, these sub-industries cover a significant part of the entertainment industry. However, certain domains that fall within this industry such as television series are not considered. Yet, a general conclusion is sought for the whole entertainment industry by using a limited set of articles. This should be taken into consideration throughout the research process and when drawing conclusions.

3.2.4. Results

Four out of six studies focus on the influence of social media buzz on sales. The study by Liu (2001) argued that the volume of word of mouth has a significant influence in predicting the box-office sales of a movie. Moreover, using a conceptual framework, the author suggests that word of mouth generated in week *t*, along with critical reviews and numbers of screening, is an effective predictor of sales for the next week (*t+1*). Sadikov et al. (2009) found that the amount of references in blog posts has a significant effect on the box-office sales for the subsequent weeks after the release of a movie. Remarkably, Sadikov (2009) argues that pre-release data have no significant relationship to the first week of sales. The findings of these two studies are therefore contradictory and require further research. However, a critical note can be made comparing the analyses conducted by both researchers. The contradictory results might be caused by the fact that Liu, Y. (2001) uses a natural log for its word of mouth messages, while Sadikov et al. (2009) removed movies that received less than 1000 blog references. Both methods can impact the results.

Moreover, in contrast to the majority of the prior researches, Liu, H. (2012) states that the volume of tweets is not a valuable factor due to the insignificant results, which is supported by a study performed by Chevalier and Mayzlin (2006), who investigated the volume of reviews on Amazon.com. This could be explained by the fact that both studies included the volume of tweets in the regression analysis, without examining the direct relationship solely between the volume of tweets and its sales. After conducting this comparison, it can be concluded that there is still an ambiguity what exactly is affecting the consumers' willingness and actual purchasing or using a service and whether they are influenced by the tweets of other users.

We included two articles in our literature review that have a focus on followers. Rui et al. (2013) concluded, besides the effect of number of tweets on the sales of a movie, a more contributing finding regarding the influence of users. Classifying them in two different groups of more influential and less influential users, the researchers found the users with more followers posted more influential tweets. Although this is an essential finding, it can be criticized for an uncertain definition of the influence. Namely, it is arguable whether solely the division in “influential” and “less influential” regarding the number of followers of a particular user can be perceived as a measurement of the impact of a tweet. The researchers divide the groups between less and more than 400 followers, which seems an arbitrary division. Besides, this article does not include a predictive analysis, for instance relating followers to the sales of a product. We will investigate the relationship between the total number of all followers and the sales of a product.

The article by Bakshy et al. (2011) further discusses the relationship between the number of followers and the influence a user possesses. Since it is very difficult to observe, the researchers included all users irrespective of their possible influence score in their sample. They found that the users with the largest number of followers tend to post the most tweets which comprise far-going cascades of reposting and retweeting. Bakshy et al. (2011) do not discuss how these cascades and a number of followers that are reached by a tweet can be related to predicting the sales of a product. However, in case these findings can be expanded to a larger set of tweets, we can argue that it is not the volume of tweets, but rather the number of followers which are reached through these tweets will be a better predictor of the sales of a product.

A critical note that can be made is that not all articles reveal the results of the analyses they conducted. For instance, out of six articles only three stated the coefficients of regression instead of merely the R-squared value. It is therefore not possible to make valid conclusions regarding the obtained results and analytical methods of all studies.

It can be concluded that there are contradictory results between the different elements of the studies. This analysis gave us valuable information for our research. Firstly, we will attempt to makes a conclusion whether tweets do have an influence on sales, and more importantly if followers have a bigger influence than the amount of tweets. Besides testing for pre and post release correlation between the tweets and its sales, we are aiming to test this hypothesis by assessing the predictive power of the tweets in the first week after the release and the corresponding sales of the second week after the release.

3.3. Meta-Analysis

Further, we will make a conclusion regarding the current empirical evidence of our hypothesis by conducting a meta-analysis. Two essential presumptions of a meta-analysis are that the research strategy and the effect size for a series of studies investigated should be the same.

The complete set of research papers that we have analyzed, applied a causal design research strategy by using mainly a longitudinal view. Generally, this strategy explores the change of the value of the independent variable and a later change of the value of the dependent variable in the population during a certain time period (Hak, 2012). The correlation or the beta values of the regression models can both be used in order to measure the effect size of a change in the values of the variables. Taking into account these two presumptions, five out of six articles were found to be suitable for further investigation because of the use of regression or correlation coefficients. The research by Rui et al. (2013) was disregarded due to the usage of a non-standardized GMM estimator as an effect size parameter. Three articles (Liu Y. (2001), Dewan and Ramprasad (2009) and Liu H. (2012)) have disclosed standardized regression coefficients which will be compared within the meta-analysis framework and will serve as the main point of reference for the further research. We were unable to perform a meta-analysis for investigating a relationship between the number of followers and sales since there are no prior studies with a similar hypothesis available for this purpose. Moreover, to support our conclusions, a meta-analysis comprising five studies with disclosed Pearson correlation coefficients will be performed.

3.3.1. Forest Plots

A forest plot is a confidence interval picture that reflects the results from a series of studies and a meta-analysis of those studies (Cumming, 2012). The plot depicts an obtained effect size of the volume of the social media buzz related to the sales of a product, including the confidence intervals for each study. We have scattered two forest plots where the initial one displays results of standardized regression coefficients and the second supports our findings.

Standardized Regression Coefficients

We have drawn a forest plot for a set of studies which used standardized regression coefficients for measuring the effect size of the volume of social media buzz (Figure 3). We

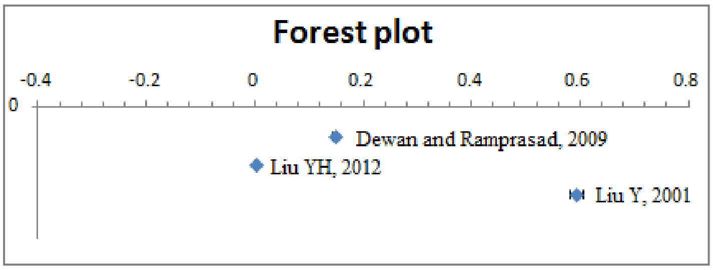


Figure 3: Forest plot of standardized regression coefficient

identified two essential trends displayed in the forest plot. Firstly, the results of the researches (depicted as diamonds) vary significantly across the studies with a range from 0.0000373 to 0.5920000. These varying results imply that a finding within one study, even if it is significant, can be considerably different from the actual effect in the population and therefore cannot be taken as a valid result of a research. However, it is essential to note that a high variability in the results could be also due to the heterogeneity between studies. For instance, Dewan and Ramprasad (2009) use blog posts for measuring social media buzz, while Liu H. (2012) uses tweets for the same purpose. Secondly, due to the substantial sample size, the confidence intervals are so small that they are not seen on the forest plot. Since the confidence intervals do not overlap at any point, the congruence between the studies is low. Because of this inability to combine two similar and independent results, it is impossible to make the confidence interval shorter than a confidence interval for a single result. The analyzed articles used regression coefficients with a beta value which measures how strong each independent variable influences the dependent variable. Since in the findings the overall beta is fairly low, the impact of the independent variable on the criterion variable is low. The forest plot of the beta's cannot be used to support the hypothesis that the social media buzz can predict sales. However, this does not necessarily mean that there is no significant relationship, but that the analyzed articles do not agree on the effect size of this relationship. Thus, an additional meta-analysis was conducted focusing on the correlation to receive more information about this exceptional finding (Figure 4).

Coefficients of Correlation

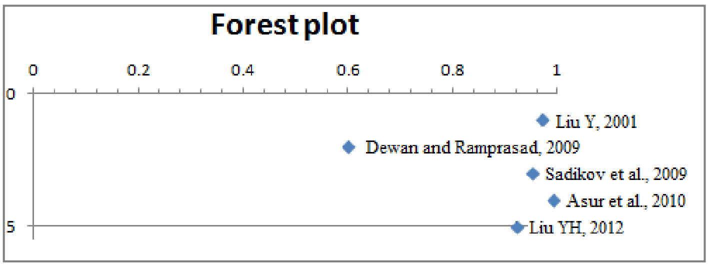


Figure 4: Forest plot of coefficients of correlation

As we cannot yet conclude that the social media buzz is a significant predictive power of sales, we are willing to investigate what would the combined effect of several variables on the dependent variable be. For this purpose, we scattered another forest plot using the correlations of the five articles (Liu Y. (2001), Dewan and Ramprasad (2009), Sadikov et al. (2009), Asur et al. (2010) and Liu, H. (2012)). Since the regression models of the studies have included more than one independent variable and do not report these values individually, we cannot use the obtained R-squared values to support our main hypothesis which focuses merely on the volume of the social media buzz. The findings serve as a proof that more variables combined can be a sufficient predictor of the sales of a product. In four of the articles, we measured the coefficient of correlation by taking a root of the reported R-squared values. We should critically note that this is not the best practice. However, this was done solely for the purposes of showing a general trend of higher predictability when using more than one independent variable.

In comparison with the coefficients of regression, the results do not significantly vary across the studies with one exception (Dewan and Ramprasad, 2009). Due to the immense sample size, the confidence intervals are extremely narrow and reveal no overlap between the studies. However, the results of the study are so similar to each other that repeating the same research will most probably lead to the same results. The overall effect size of 0.9437 is very high, which implies that a model with several independent variables has a higher predictive power of sales than merely the volume of the social media buzz.

3.3.2. Conclusion of Empirical Evidence of the Hypothesis

Referring to the prior analysis, several conclusions can be drawn regarding our hypothesis. Most of the studies conclude that the volume of tweets is a strong predictor of the sales. However, our most essential finding from the meta-analysis is that there is no consistency in the effect size of the relationship between the volume of the social media buzz itself and the sales of a product which is due to the varying results, a low overall effect size, and the fact that there are no overlapping confidence intervals. According to our additional investigation, more variables in a regression model in contrast to only one variable provide more information regarding sales' trends. Moreover, the sample sizes of the studies investigating the effect of the social media buzz on sales are extremely large which results in the fact that that the confidence intervals of the regression coefficients do not overlap. Although small confidence intervals are reliable indicators of the obtained results, it is harder to identify a pattern of an approximate overall effect across the studies. Furthermore, it is essential to note that a high variability in the results could be also due to the heterogeneity between studies. There are several possible explanations to the varying results across the studies. Firstly, the studies we have examined have used different social media platforms and it is an interesting issue for future researches. Namely, it could be of interest to investigate whether the predictive power of a product's sales varies depending on the social media platform examined. Another possible reason for the different results could be that social media buzz about different entertainment products, for instance movies and music albums, differently affects the prediction of the sales or is actually not related and can be attributed to other factors. Moreover, the overall predictive power of the regression models in Figure 4 differs due to using varying independent variables. It is difficult to assess the predictive power of social media buzz alone and results should be viewed from the perspective of the entire model instead.

3.4. Additional Analyses

In order to excel we conducted further analyses to include supplementary elements regarding the studies which support our main hypothesis.

3.4.1. Vote Counting

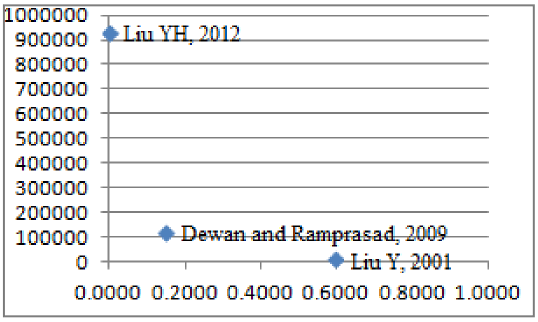
Vote counting can be used for a larger set of researches where effect sizes differ across the studies (Cooper and Hedges, 1994). We have taken five studies into account when performing the vote counting and tested them for the presence of statistically significant positive results. Each study's results were compared to the benchmark of the 95% significance level (α = 0.05) and were analyzed whether an obtained result falls within this interval. Four of the studies fall within the category that illustrates a statistically significant positive result with a p-value ranging from <0.001 (Liu Y., 2001) to 0.01 (Dewan and Ramprasad, 2009). However, the study by Liu H. (2012) showed that there is no significant relationship between the social media buzz and the sales of a product (p-value of 0.459). With a plurality of the studies falling in the category which states that there is a significant relationship between the social buzz and the sales of a product, this category can be considered as a winning one. Hence, according to the vote counting, there are reasonable grounds to assume that there is a true relationship between the social media buzz and the sales.

These results illustrate that the relationship between the independent variable and the dependent variable is present, as was also shown by the initial meta-analysis. Although this method enables for the analysis of a larger set of data, it is essential to note its main drawback. Namely, if only the vote counting is taken into consideration, it can be concluded that the social media buzz is a predictor of the sales of a product. However, a more substantial analysis of the effect size should be performed in order to reveal more details regarding this relationship. Although this relationship is present, the effect size is minor as showed on the aforementioned forest plot.

It can be concluded that the overall research community agrees that a higher value of prediction can be obtained by using more variables in a model instead of using only the social media buzz. However, the effect size of the analyzed articles varies, which leaves us with an interesting issue for further investigation.

3.4.2. Analysis of Potential Publication Bias

A publication bias is a tendency to report different positive significant results from negative or insignificant findings (Song et al., 2010). A funnel tool is frequently used in order to explore the presence of this bias. Usually, a funnel plot assesses whether the size of a standard error corresponds to the size of effect. Due to the limitation of undisclosed standard errors, we have used a sample size instead as advised by Peters et al. (2006). Cumming (2012) also describes a relationship between the standard error and sample size. Theoretically, studies with a large sample size, which is the case for our meta-analysis, should be plotted closer to the upper end of the funnel plot. Studies with a smaller sample size tend to have a considerable standard error and would be dispersed at the bottom of the plot (Cumming, 2012). We have scattered two funnel plots, however, due to the small number of studies, it is difficult to identify whether a potential publication bias is present.



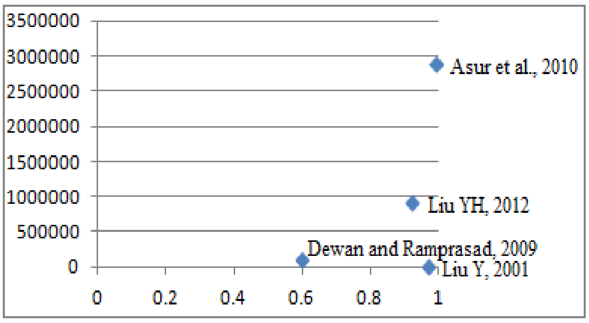


Figure 5: Funnel plot of coefficients of correlation

Figure 6: Funnel plot of coefficients of correlation

Figure 5 is a depiction of the three initially selected articles, where the horizontal axis displays values of a regression coefficient and the vertical axis reveals a sample size N. Although these results should not serve as a firm ground for further research due to the small number of articles, it clearly shows a trend that the greater the sample size is, the lesser is the observed effect of the social media buzz on the sales of a product. Figure 6 displays the relationship between the correlation coefficient and a sample size N. This funnel plot surprisingly illustrates that irrespective of the sample size, the obtained correlation coefficient, as long as it is significant, does not vary across studies. The results should be interpreted with caution though since these coefficients include more than the variable of volume of social media buzz.

Although the funnel plot did not allow determining the presence of a potential publication bias, it depicted several trends. Firstly, a greater sample size illustrates a smaller effect of the social media buzz on the sales of a product. Secondly, regarding the correlation coefficient, the sample size does not affect the results of a study with a presumption that they are significant.

3.4.3. Subgroup Analysis

Subgroups enable for the assessment of the simplest form of the moderator analysis (Cumming, 2012). By sub-grouping studies for those which used movies and music albums as a sample for the research, we have found that for the initial meta-analysis there is no difference in the results with regards to the chosen sample (Figure 7). However, the correlation coefficient meta-analysis has shown that the effect of social media buzz is considerable more evidential in the studies which used movies as their sample size (Figure 8).

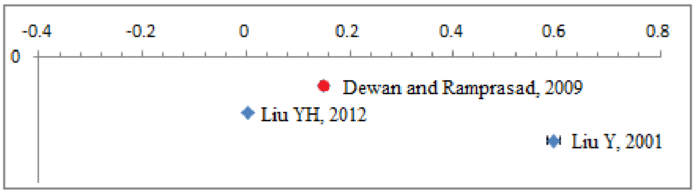


Figure 7: Subgroups analysis (movies vs. music albums) of standardized regression coefficients(X axis – beta, Y axis – number of studies)

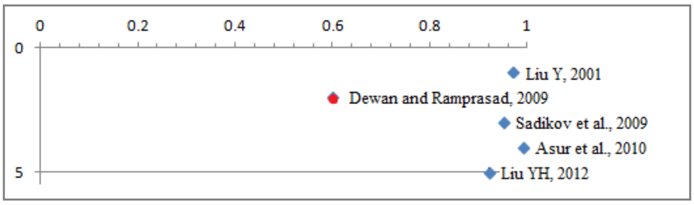


Figure 8: Subgroups analysis (movies vs. music albums) of coefficients of correlation (X axis – correlation coefficient, Y axis – number of studies)

Moreover, the studies could be also classified by those which have used microblog Twitter for their studies and those that used other online media such as blogs. The results of this subgroup analysis are depicted in Figure 9 and Figure 10.

The initial meta-analysis illustrates tweets as a source of data having considerably more predictive power on the sales of a product. Although this is not a representative picture of the reality due to the limited number of studies analyzed, the difference in the predictive power of posts and tweets is substantial, according to the Figure 9. The supporting meta-analysis (Figure 10) does not display any major trends in the subgroup analysis.

After performing a subgroup analysis, it can be concluded that it might be possible that the results obtained from a movies' sample would have higher predictive power with regards to the box-office revenues and should be interpreted carefully when applied to predict sales of other goods. Moreover, Twitter appears to have a higher predictive power compared to other social media.

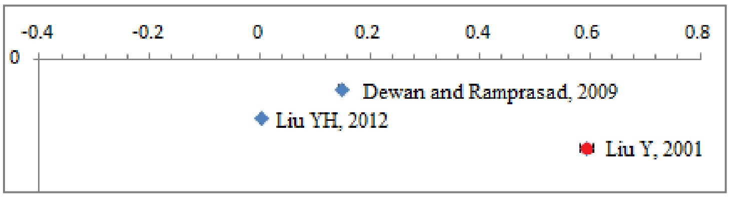


Figure 9: Subgroups analysis (tweets vs. other online posts) of standardized regression coefficients (X axis – beta, Y axis – number of studies)

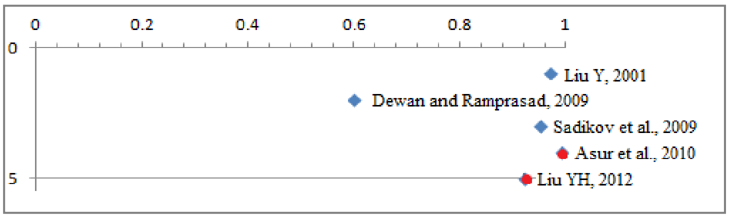


Figure 10: Subgroups analysis (tweets vs. other online posts) of coefficients of correlation (X axis – correlation coefficient, Y axis – number of studies)

3.4.4. Meta-Regression

This meta-analytical feature is used for identifying continuous moderators within the set of studies. Once identified, they could be the underlying cause of the heterogeneity between the observed effect sizes (Cumming, 2012). While there is a majority's agreement that the sales of a product can be predicted by the amount of social media buzz in combination with other variables, for instance valence, it might be possible that this prediction is caused by other variables which are beyond the scope of the research. Due to the ever increasing access to internet and mobile online usage the number of online users is growing tremendously over time, we believe that the predictive power of the social media buzz on the sales could be affected by the continuous moderator in the form of the number of online users. Besides, another moderator with a potential misrepresentation could be the difference between the number of registered users and the actual users. These factors should be taken into account when planning further research.

3.4.5. Different Meta-Analyses

Although in all of our studies there was more than one independent variable, we were mainly focusing on analyzing the relationship between the extent of the social media buzz and the sales of a corresponding product. However, we are aware that amount of followers could be of significant contribution to predict sales by using social media. Nevertheless, due to the impossibility to compare the varying effect size parameters of different studies, we were unable to perform a meta-analysis for the followers of Twitter users. Instead, we have performed a meta-analysis which accounts for the overall predictability of a model. The results of this meta-analysis have been described above in the forest plot of the correlation coefficients.

3.5. Overall Conclusion

Based on the previous analyses in the literature review referring to the prior studies, several conclusions can be drawn regarding our first hypothesis. Our most essential finding is that it is ambiguous whether the volume of the social media buzz is a valid predictor of the sales of a product or not. This is due to the varying results and a low overall effect size of the regression coefficients. Since there is currently no consensus about this issue, we will contribute by giving a definite answer. We will use the volume of social media buzz as a central variable in our correlation analysis as well as our regression model. By further examining this variable, we will be able to resolve the current ambiguity around this variable. Secondly, to add additional value to the existing literature, we will introduce a complementary study in which we will take the amount of followers into consideration in relation to the product's sales. The existing literature has not yet scrutinized this variable. However, we believe that the amount of followers will give valuable insights for predicting sales.

Moreover, according to our additional investigation, more variables in a regression model in contrast to only one variable provide more information regarding the sales' trends. This implies that by introducing multiple variables into our research, besides our first independent variable (volume of tweets), we will be able to assess more optimally which factors have an influence on a product's sales. These additional factors will consist of the number of total views of the official trailers on YouTube, the amount of posted blogs and the number of theatres the movie is released in. The number of YouTube views has never been considered as a variable in the existing literature. We provide a valuable extension on the existing studies by introducing this novel variable which is relevant for the literature on box-office movies.

Finally, based on our previous analyses, a longitudinal strategy will be preferred due to its higher internal validity and superior applicability to relatively shorter research periods. The ideal subset of domain will be movies with a regression and correlation coefficient as effect size parameter.

Section 4. Methods

In this section, we will elaborate on the methods and techniques that will be used while conducting our study. Moreover, an extensive data matrix will be introduced that will outline the research cases.

4.1. Longitudinal Research Strategy

For our research we used a longitudinal research, since this strategy fits our study most optimally. A longitudinal study suggests correlation because the change in value of the dependent variable is measured related to the change in value of the independent variable (Hak, 2012). This is suitable since our hypothesis investigates the existing correlation between social media buzz about a product in the entertainment industry and the size of the audience in relation to its sales. Besides that, the internal validity of this research design regarding a general claim is high due to the chronology of the change in the values of the independent and dependent variables (Hak, 2012). Using a longitudinal strategy, data are collected for a limited number of points in time while looking at multiple cases. Our hypothesis requires us to gather tweets about several movies for a relatively short period of two months. It can thus be concluded that a longitudinal research design is the most optimal choice for our research. Moreover, our research strategy is causal. We tracked relevant tweets one week prior to the release (cause) and the corresponding sales for the opening week (consequence). Secondly, we tracked tweets from the date of release (cause) and corresponding sales for the second week after the release (consequence). This will enable us to make more valid conclusions regarding the predictive power of the volume of social media buzz and the amount of followers.

4.2. Selection of Cases

Total Population

A population is a set of instances of the focal unit. It is, by definition, a subset of the theoretical domain (Cumming, 2012). As described in the *Introduction*, the focal unit of our hypothesis is the products in the entertainment industry and its corresponding population is the box-office movies released in the US. Movies are a suitable subject of research as mentioned before in the paper. According to Sadikov et al. (2009), movies have a known release date, which allows to study the hype' before the release in relation to 'success' after the release. Secondly, movies provide an inherent normalization compared to other domains, since the sales in n^{th} week after the release date are comparable across movies. Besides, users tend to tweet a considerable amount about movies which results in sufficient data for analysis (Sadikov et al., 2009). Finally, the sales data of movies are widely accessible and more importantly, published with merely a small time lag (BoxOfficeMojo, 2013). This provides us the opportunity to gather all the sales data within our limited time span.

Data Matrix

To establish the data matrix and the relevant cases, information from Box Office Mojo movie releases has been analyzed. Box Office Mojo is a widely recognized independent online medium which is used to analyze and describe movies and to collect reviews. We will elaborate on Box Office Mojo in the M*easurement Instrument* paragraph of this chapter.

One of our data matrixes, as depicted in Table 2, consists of six release dates and has no data available prior to the release. The tweets of these movies were collected for a time period of one week after the release date. The correlation between these data and the sales of the week can be statistically derived. However, it is more interesting to relate the collected data to the week of sales following the week the data are measured, to ensure that predictions can be made. Therefore, the sales of the movie for week two (*t + 1)* has been included. For all these cases, the change in number of total followers of the first week has been compared with the change in sales in the first and second week after release.

Cases 5th of April*	**Change in *independent variables* (5 – 11 April)**		**Changes in *sales***	
***Movies with ambiguous names such as “24” or “Shelter” have been excluded from analysis**	**Tweet rate**	**Follower rate**	**5 – 11 April**	**12 – 18 April**
Evil Dead	2,546	4,086,041	$32,041,782	$12,303,168
Jurassic Park	1,050	3,273,672	$23,109,165	$11,361,925
The Company You Keep	57	193,182	$172,610	$381,487
Simon Killer	7	7,595	$7,089	$7,962
Upstream Color	30	311,548	$38,731	$103,405
The Brass Teapot	7	46,835	$5,222	$1,775
Free Angela and All the Political Prisoners	12	12,629	$69,306	--
Eddie: The Sleepwalking Cannibal	4	46,479	$1,294	$338

Cases 12th of April	Change in *independent variables* (12 April – 18 April)		Change in *sales*	
	Tweet rate	Follower rate	12 – 18 April	19 – 25 April
Scary Movie 5	1,590	2,201,180	$16,647,592	$7,389,344
The Angels' Share	56	170,156	$25,738	$46,620

Cases 19th of April	Change in *independent variables* (19 April – 25 April)		Change in *sales*	
	Tweet rate	Follower rate	19 – 25 April	26 April – 2 May
Oblivion	4,578	11,871,168	$47,287,500	$22,884,415
The Lords of Salem	260	2,696,962	$855,293	$264,036
Filly Brown	186	999,401	$1,745,091	$689,333

Cases 26th of April	Change in *independent variables* (19 April – 25 April)		Change in *sales*	
	Tweet rate	Follower rate	26 April – 2 May	3 – 9 May
Tai Chi Hero	7	15,492	$19,474	$4,986
Arthur Newman	32	242,141	$177,225	$19,764
The Big Wedding	180	585,315	$10,334,565	$5,453,126
Kings Faith	2	12,334	$31,988	$13,306
Midnight's Children	14	45,469	$14,266	$33,030
Pain and Gain	1,176	4,947,349	$26,319,075	$10,289,125
An Oversimplification of Her Beauty	44	302,876	$13,196	$6,551
Reluctant Fundamentalist	55	326,240	$43,754	$126,704

Cases 3rd of May	Change in *independent variables* (26 April – 2 May)		Change in *sales*	
	Tweet rate	Follower rate	3 – 9 May	10 – 16 May
Iron Man 3	4,276	72,759,825	$212,421,084	$89,470,779
The Iceman	40	214,686	$127,529	$171,209

Kiss of the Damned	3	4,947,349	$4,618	$6,515
What Maisie Knew	10	214,686	$31,152	$38,753
Scatter my Ashes at Bergdorf's	3	5,173	$32,357	$73,748
Cases 10th of May	**Change in *independent variables* (3 – 9 May)**		**Change in *sales***	
	Tweet rate	**Follower rate**	**10 – 16 May**	**17 – 23 May**
The Great Gatsby	2,518	9,223,705	$66,743,604	Data were not obtained due to time limitations.
Peeples	44	715,049	$5,707,777	
Venus and Serena	7	41,310	$14,971	
Sightseers	2	2,721	$10,936	

Table 2: Data matrix with 6 batches

As a complementary research, we used two release dates to obtain the total tweets and the total number of followers one week prior to the official release dates. These data, presented in Table 3, can be correlated to the sales of the first week after the release date. In this way, a predictive measurement tool can be developed since this allows us to perform a pre and post statistical test. The corresponding data are displayed in the table below.

Cases 3rd of May	**Change in *independent variables* (26 April – 2 May)**		**Change in *sales***
	Tweet rate	**Follower rate**	**3 – 9 May**
What Maisie Knew	2	6,449	$31,152
Kiss of the Damned	2	5,730	$4,618
The Iceman	186	2,556,107	$127,529
Scatter my Ashes at Bergdorf's	3	8,619	$32,357

Cases 10th of May	**Change in *independent variables* (3 – 9 May)**		**Change in *sales***
	Tweet rate	**Follower rate**	**10 – 16 May**
The Great Gatsby	4,695	16,846,746	$66,743,604
Peeples	168	3,622,102	$5,707,777
Venus and Serena	13	31,750	$14,971
Sightseers	24	29,426	$10,936

Table 3: Data matrix with data one week prior to release of a movie

4.3. Measurement Instruments

In order to collect the data for our study, several measurement instruments have been used that reveal information about the dependent and independent variables.

NodeXL

The exact procedure for the identification of and for extracting relevant evidence is as follows. The tool NodeXL was our measurement instrument, which was used in order to collect the tweets and calculate the volume and amount of followers. NodeXL is a free, open-source template for Microsoft Excel that makes it easy to explore network graphs. With NodeXL, you can enter a network edge list in a worksheet and analyze data graphs, all in the familiar environment of the Excel window (NodeXL, 2013). This tool aims at making an analysis and visualization of network data easier by combining the common analysis functions with the familiar spreadsheet paradigm for data handling (Smith et al., 2009). NodeXL is regarded as a valid and reliable tool because it is possible to virtually include all possible data that Twitter exposes (such as the time, the amount of followers, the number of users they follow themselves, etc.). According to Hansen et al. (2011) NodeXL is a reliable tool for gathering data about concepts. As argued by Smith et al. (2009) NodeXL is created to be "a tool that avoids the use of a programming language for the simplest forms of data manipulation and visualization, to open network analysis to a wider population of users, and to simplify the analysis of social media networks." However, since Twitter has put an import restriction of 18,000 tweets per hour, it is impossible to gather all tweets. After taking a sample of movies, it was observed that popular movies such as Evil Dead reached 18,000 tweets within 30 minutes. Therefore the tracking took place according to frequent time intervals per day. Moreover, tweets can simply be filtered by date and can be eliminated if there is an overlap, meaning that certain tweets have been measured twice. Finally, a set of

30 random Twitter users have been reviewed that appeared in our data and have been compared to all the results with real-time data from Twitter to ensure the validity of these data. We have not found any differences in the NodeXL data and the data on Twitter.

YouTube Views

Besides, we used the online video platform YouTube to include the number of views for the official movie trailer. We have chosen to include YouTube because it has never been introduced before in relation to predicting sales of products in the entertainment industry. The platform offers two distinct advantages. Firstly, the total number of YouTube views is publically accessible. Besides that, the platform reaches over a billion users per month, which ensures that there are data even for less popular movies (YouTube, 2013). However, only the accumulated number of views is presented, but it does not reveal how this accumulation occurred per week. Several trailers have a graph representing this trend, but the majority has disabled the public to view these statistics. The YouTube API reports that only owners are allowed to access the data regarding the views per day or week (YouTube, 2013). Another critical note that should be made considering YouTube is the technique in which the total number of views is counted. A study by Haddad (2010) has critically assessed the use of YouTube and social media in researches and judged the method these YouTube views are measured. He states that the total number of views of a video is far from exact science, since a considerable part of the trailers are embedded on external websites and are viewed there, which are not consistently counted by YouTube. More problematically, a view is only counted when the video has been watched in full (Haddad, 2010). In addition, a research by video hosting platform Wistia (2011) reveals that naturally, shorter videos will be more likely to be fully watched than longer videos. This study found that 78% fully watches a video that has a length between thirty seconds and one minute, while only 51% completely views a video between the two and three minutes (Wistia, 2011). For our study, this implies that shorter movie trailers would have relatively more views than longer ones. A possible solution is to take this error into account by assigning weights to the views based on the length of the movie trailers. However, more research has to be performed around this topic to offer a reliable calculation for the weights. For this reason, we have not adjusted the total number of YouTube views. Although the total number of views on YouTube is not completely reliable, the platform has a more sophisticated protocol for counting views than alternatives do. We introduce the number of YouTube views since it is a novel variable, but cautions should be made about its validity.

Box Office Mojo

The movies that were released, their corresponding theatres, and our dependent variable (sales data), were collected from Box Office Mojo, which is an online movie publication service that exposes the revenues and other data of virtually all movies. Box Office Mojo is the number one box-office destination internationally and is quoted regularly in renowned media channels such as the Wall Street Journal, Bloomberg and Forbes (BoxOfficeMojo, 2013). The sales data are exposed on a daily basis and provides real time information about the performance of the opening weekend of the movies. The total revenues of the movies can be easily split into weeks or days in order to make the measurement consistent.

IceRocket

Also, in order to introduce the variable of total blog mentions, the tool IceRocket was used. IceRocket is an online database that collects all the publically available blog data and displays the information in one database. It allows us to measure the total blog mentions of the different movies without having to accumulate the different blogs. According to Crunchbase (2013), Ice Rocket has a special focus on displaying the most up to the second results.

4.4. Measurement Protocol

The measurement protocol is a set of detailed instructions for identifying, selecting, and accessing evidence and for generating a valid and reliable score for each of the variables specified at the outset of the study (Hak, 2012). It lists the precise definition of the variable, the precise procedure for identification of and for extracting relevant evidence and the exact specifications of procedures for recording the evidence (as data) and for generating scores from them (Hak, 2012).

For all the cases that we have listed in the previous section, real-time raw data were imported from Twitter. It is extremely critical to define a comprehensive measurement protocol as this will avoid possible errors. All our researchers have downloaded NodeXL and authorized the program to ensure that the maximum of import data can be obtained. We have ensured that the same timeslots are measured for all movies, such that a valid comparison can be made. The data that overlap with previous days have been eliminated to avoid double measurements. We have critically assessed the number of tweets per movie for the previous day, and have ensured that our measurement frequency takes the changing popularity of the movies into consideration.

A measurement of an attribute (or a variable attribute) is valid if variations in the attribute causally produce variation in the measurement outcomes (Hak, 2012). To make sure that all tweets that are measured refer to a movie, we took several samples of different time slots and different movies to ensure that the data correspond to movie tweets and do not include noise. This was a labour intensive practice, but it assured that there were no tweets included that did not address the movie. Moreover, to ensure that the time period of our research is valid, we converted our times into American time since the tweets are displayed in Central European Time (CET), but the movies are released in the United States. In this way, tweets were measured from the exact release and several hours earlier. Lastly, we encourage other researchers to duplicate our studies. In this way, the practices in our measurement protocol can be reviewed, which would further increase the validity of our research. The reliability of the measurement protocol and the specification of the procedures for recording the evidence are provided below.

4.4.1. Reliability

Moreover, we ensured that our data were reliable. Reliability is the precision of the scores obtained by a measurement (Hak, 2012). We saved all the raw data on a hard drive and in the cloud, which makes the data accessible for all researchers at any time. After collecting all the tweets, the tweets were accumulated for every day to arrive at the number of tweets and the corresponding number of followers. Luckily, NodeXL reveals the number of followers in relation to a particular tweet. Therefore, the average number of followers per movie per hour, the follower rate, could be calculated easily. Since Twitter profiles are mostly open, we took a sample of users from our data set to verify if the number of followers is consistent with the number as displayed on Twitter.com. Again, this increased the reliability of our data set.

Furthermore, we have defined other practices to ensure that measurement instrument NodeXL is a reliable source to conduct our study with. First of all, when one movie has an extraordinary large or small set of tweets compared to other movies, which cannot logically be explained by its sales, we would further assess the reliability of NodeXL. Secondly, when there are sudden spikes in certain days for one movie that are completely inconsistent with the trend, we would extensively explore the potential causes to assure that this is not due to a deficiency in our measurement program. Thirdly, we measured the same movies multiple times a day to check if NodeXL provided the exact same tweets and results. Finally, we have included a complementary manual assessment. We have strategically picked a batch of tweets and reviewed these ourselves to ensure there are no obvious deficiencies in our data set. In the Appendix B, we have also discussed possible missing scores for our research.

4.5. Correlation as Main Effect Size Parameter

The correlation describes the degree of relationship between two variables. It explains if the relationship is positive or negative and the strength of the relationship (i.e. the change in the sales of a product when one more tweet is posted or one more follower is reached (Choudhury, 2009)). A correlation has mainly an informative value, because the higher the correlation between X and Y, the more precisely the value of Y can be predicted in a case if we know the value of X in that case (Hak, 2012). Thus the advantage of the correlation method is that predictions can be made since a strong correlation between two variables ensures that one variable can be predicted based on the other variable. The correlation can be seen as a standardized measure, if standard deviations of both values are known. By using this effect size parameter we can conclude if the number of tweets and more importantly, the total number of followers, do have an impact on sales.

Besides correlation, we conducted a regression analysis taking into account the several independent variables. A regression analysis is a statistical approach to forecasting change in a dependent variable (sales revenue) on the basis of change in more independent variables (volume of tweets, amount of followers, number of views, number of theatres, and blog mentions) (Business Dictionary, 2013). More specifically, a regression analysis supports to understand how the value of the dependent variable changes when any of the independent variables is varied, while other independent variables are held constant (Freedman, 2005).

Section 5. Results

In this section, the obtained results are described. Furthermore, a predictive model for future research is proposed.

Chapter 5.1: Correlation

Chapter 5.2: Complementary Research - YouTube

Chapter 5.3: Additional Variables

Chapter 5.4: Regression

Chapter 5.5: Confidence Intervals

Chapter 5.6: Predictive Analytics

In order to obtain our results, we performed several analytics' tests by using the SPSS tool. The data were collected between the 5th April and 16th May 2013, consisting of a total sample size of 30 movies. The movies were released in the United States within this time frame. In total we collected 161,317 tweets for 30 different movies that accumulated to 887,979,439 followers within the research period. The different analyses and related results will be elaborated upon in the following chapter.

5.1. Correlation

The correlation describes the degree of relationship between two variables. For our main hypothesis, we aimed to investigate whether the effect of the amount of followers, expressed as an average number of followers reached per hour, is higher than the effect of the volume of tweets, which is expressed as a tweet rate per hour, about a particular movie. The obtained results for our main hypothesis are presented in Table 4 and the descriptives are attached in the Appendix C.

Independent variables	Pearson correlation	Confidence Interval	Sample Size
Number of followers	0.975	$0.9748 \leq r \leq 0.9752$	30
Volume of tweets	0.800	$0.7982 \leq r \leq 0.8017$	30

Table 4: Estimated effect size values Week t – Sales t

Our initial investigation regarding the correlation of the rate of followers for each of the 30 movies for the research period and the sales for each movie showed a very strong positive correlation, with a Pearson coefficient of 0.975. Taking into account the sample size, a valid conclusion can be made that there is a strong relationship between the number of followers and movie sales when they are measured for week *t*.

Moreover, we explored the relationship between the number of tweets posted for a particular movie and the corresponding box office revenues. The total absolute results for all 6 batches showed a positive relationship, with a coefficient value of 0.800. While the nature of the relationship is considered strong, it is lower than the results presented by the number of followers. When comparing the relationship of the two independent variables and the sales, it can be concluded that the number of followers is an important factor for predicting sales.

5.2. Complementary Research - YouTube

Next to our main study, we performed an analysis investigating a predictive effect of the number of views for the official YouTube trailer on the movie's sales. The findings revealed a surprising positive relationship between the two variables with a coefficient of 0.876, as presented in Table 5. Although YouTube views do not have a predictive power as high as the number of followers, this is still a significant finding which could be investigated in further studies by improving measurement of this variable.

Independent variables	Pearson correlation	Confidence Interval	Sample size
YouTube views	0.876	0.8749 ≤ r ≤ 0.8771	30

Table 5: Estimated effect size value of YouTube views

5.3. Additional Variables

Furthermore, we analyzed the relationship between the number of theatres a movie has been screened in and the dependent variable. The correlation value of 0.687 was obtained, implying a medium positive correlation. Since a strong relationship between the number of theatres and sales of a movie was expected, the results are surprising. This could be explained by the fact that some of the movies were released in a limited number of theatres, restraining all the willing customers to watch a movie in the cinema and thus boost sales of a movie. Moreover, we also examined the total amount of blog posts revealed by IceRocket, a database of all public available blogs on the internet. The results show a high correlation of 0.965. However, it can be concluded that using Twitter as a tool and particularly measuring the size of audience is a better predictor than blogs posts. The obtained results are presented in Table 6.

Independent variables	Pearson correlation	Confidence Interval	Sample size
Theatres	0.687	0.6809 ≤ r ≤ 0.6930	30
IceRocket blog posts	0.966	0.9656 ≤ r ≤ 0.9664	30

Table 6: Estimated effect size values of theatres and blog posts

5.4. Regression

From the previously discussed independent variables, a regression model can be developed to explain the dependent variable corresponding “sales” of the movies. A linear regression is an approach of modelling the relationship between a scalar dependent variable and more explanatory variables, where R-squared accounts for the proportion of variability in the data (Cumming, 2012). The R-squared can be adjusted for the number of terms in a model. When including the variables the following regression model can be made:

$$y_{Sales} = \beta_0 + \beta_1 \times x_{youtube} + \beta_2 \times x_{followers} + \beta_3 \times x_{tweets} + \beta_4 \times x_{theatres} + \beta_5 \times x_{blog\ posts} + \varepsilon_i$$

$$i = 1, \dots \dots., n.$$

	R-squared	Adjusted R-squared
Regression Model	0.996	0.991

Table 7: Regression analysis

As presented in Table 7, the outcome, an adjusted R-squared of 0.991, can be interpreted as a good result. If the R-squared were 1.0, the given value of one term will perfectly predict the value of another term. Since the results are both very close to 1, it implies that our regression model is an extremely strong predictor for the sales of a movie. The obtained regression coefficient for the number of tweets is 0.267 and for the number of followers it is 1.354. The significantly higher beta for the latter is an indicator that the number of followers has five times more impact on the sales of a product than the number of tweets.

5.5. Confidence Intervals

A confidence interval states the estimated range of values which is likely to include a population parameter, being calculated from a given set of sample data. Our sample data consist of a total of 161,317 observations, tweets, which have reached in total 887,979,439 followers. The results in Table 8 show that the confidence intervals are remarkably narrow which is due to the enormous amount of observations. These findings signify a high validity of the results.

Confidence Interval	Correlation	N	LL	UL
Followers	0.975	887,979,439	0.9747	0.9753
Tweets	0.800	161,317	0.7978	0.8022

Table 8: Confidence Intervals

5.6. Predictive Analytics

We have also looked at the effect of independent variables collected in week *t* on the box-office revenues in week *t+1*. Due to the popularity of a movie on Twitter in the first week after release, people will be influenced to watch the movie in the second week. For these results, we analyzed data of 5 movie batches dating from 5th April to 9th May comprising 25 movies. As presented in Table 9, there is a strong positive relationship between the volume of tweets and the sales, with a correlation coefficient value of 0.796. The number of followers has a coefficient value of 0.987, signifying a strong positive relationship between the two variables. The obtained results depict an interesting trend. Namely, the same number of tweets has a slightly weaker relationship with the sales' figures when predicting box-office revenues for a consequent week after measurement. Whereas the same number of followers is a better predictor of the sales of a consequent week than the week the data have been collected.

Independent variables	Pearson correlation	Confidence Interval	Sample size
Followers	0.987	$0.9869 \leq r \leq 0.9871$	25
Tweets	0.796	$0.7942 \leq r \leq 0.7978$	25

Table 9: Estimated effect size values

Moreover, we have investigated the effect of number of followers and tweets when data are gathered one week prior to the release of a movie. In this way we can more optimally look at the effect of tweets on movie sales, because tweets of visitors that have already seen the movie are excluded. Data from two batches with eight movies were used in order to explore this effect. The findings have shown a very strong relation of the number of followers and number of tweets with the sales. The correlation coefficient values are 0.983 and 0.998

respectively and are presented in Table 10. These results further support our research and validate the predictive power of Twitter.

Independent variables	Pearson correlation	Confidence Interval	Sample size
Followers	0.983	$0.9828 \leq r \leq 0.9832$	8
Tweets	0.998	$0.998 \leq r \leq 0.998$	8

Table 10: Estimated effect size values

Although in this instance the number of tweets has a slightly higher result than the followers, the findings from the effect on week *t* and *t+1*, where the results were higher for the number of followers, imply that overall the number of followers has a better predictive power. In order to exploit the predictive power of the aforementioned variables, we have proposed a predictive model for the sales forecast, presented in the *Discussion* section.

Section 6. Discussion

In this chapter, elements of our study will be critically examined in order to find possible discussion notes and contributions for future research.

6.1. Criticism of Our Study

It is essential to highlight the possible errors of our study, which is valuable information for future researchers who are interested in expanding on our findings. Firstly, due to technical limitations, our measurement tools did not capture all possible data for all movies. We have accounted this issue by filtering out different time slots to ensure that all the data across all the movies are consistent. We evaluated the most popular movies and compared the data to see which timeslots are overlapping. The tweets that did not have an overlap were eliminated from our data set. In this way, we ensured the consistency of our data. However, in the ideal situation, all tweets of all different movies would have been included in the analysis. By developing a program that uses the Twitter streaming API, it is possible to computerize this process and eliminate manual importing (Twitter, 2012). Secondly, the total number of YouTube views of the official trailer has been included as a variable and has a strong correlation with the box-office revenues. However, as explained before, another critical note that should be made considering YouTube is the technique in which the total number of views is counted. A study by Haddad (2010) has critically assessed the use of YouTube and social media in researches and judged the method these YouTube views are measured. He states that the total number of views of a video is far from exact science, since a considerable part of the trailers are embedded on external websites and are viewed there, which are not consistently counted by YouTube. More problematically, a view is only counted when the video has been watched in full (Haddad, 2010). In addition, a research by video hosting platform Wistia (2011) reveals that naturally, shorter videos will be more likely to be fully watched than longer videos. Although YouTube has several limitations, it is currently the best alternative.

Language Bias

An extremely important limitation of all the studies from the literature review is the language that is used by Twitter users compared to the keywords that are used for measurements. A research by Bryden et al. (2013) has found that online communication on Twitter can be classified in different communities, and that a frequency of words used within these networks varies significantly. Therefore, for each movie, it is hard to predict in which way Twitter users will express themselves about the different movies. Also, Ingram (2011) has argued that the 140 character limit of Twitter is one of the most important differentiating features of the social network. However, this same restriction will most probably have an influence on the language style that is used to formulate movies with longer titles. Additional research should investigate this issue further.

Related to this is the problem that arises because of variation in expressions. This can be due to the 140 character limit, but can be explained by several additional reasons. For example, Wauters (2010) argues that merely 50% of the tweets are in English, leaving us with an enormous amount of data that is not analyzed. Also, the data of the English tweets will be prone to bias due to differences in the way of using English around the globe. For instance, with *Upstream Color* it is relatively straightforward to expect that some Twitter users will tweet about Upstream Colour, due to the different way of spelling between British and American English (British Library, 2013). An issue that relates specifically to movies is the way sequences of movies are expressed. Movies such as *Scary Movie 5* are more prone to biases, since it is hard to assess whether Twitter users will indeed include the number 5 in there expressions and when they do not, if the tweet is still referring to Scary Movie 5 or perhaps one of the other movies in the sequence. A possible solution is to start with a brainstorm session and include all the possible keywords in the measurement tool. Multiple tools such as Google's Keyword Tool are offered online and suggest alternatives and synonyms of different keywords (Kim, 2013). However, this practice will lead to an enormous amount of redundant data, which can be consequently reduced by including negative keywords (Kim, 2013).

How Online Profiles Differ

A final limitation of our and other studies relying on Twitter is that Twitter users may have a different profile and opinion than the society at large. For this reason, in some cases Twitter users will not be a valid representation for conducting research. For instance, there is an overrepresentation of the age group 19-28 using Twitter (Smith & Brenner, 2012). However, we have found that the age distribution on Twitter is reflected in the movie visit trend where there is a similar overrepresentation of younger people visiting movies in the theatres (Fetto, 2010). Still, the age dispersion is stronger favoured towards young people on Twitter than in the statistics of the movie visits. By using Twitter as a platform, there is a risk of underestimating the sales of movies that are more popular among elderly who are less likely to use Twitter or express their opinions on this platform (Smith & Brenner, 2012). Similarly, Mitchell and Hitlin (2013) state that different groups choose to share their opinions on different events. More research should be conducted to analyze which demographic profiles are more likely to share their views on products in the entertainment industry on social media.

6.2. Contribution to the Effect Size

The previous studies have explored the effect size of the number of tweets on the sales of a product; however, have not come to a unanimous conclusion regarding it. The presented results of regression coefficients vary between 0.000 and 0.592, leaving much room for the variation. The findings of our study fall in the middle of the previous studies' results with a beta of 0.267 and are incorporated in the meta-analysis as presented in Figure 11. Since all of the studies, including this study, show significant results, it is difficult to argue which effect size would be the most accurate. Especially since different studies have taken into account different independent variables which affect the effect size of a regression coefficient. However, we believe that our study is a reliable source since it falls in the middle of the prior researches, signifying an optimal effect size of all the significant studies investigated. Nevertheless, in case several studies such as these would be performed in the future, it would be more likely to accurately estimate the effect size of the number of tweets on the sales of a product.

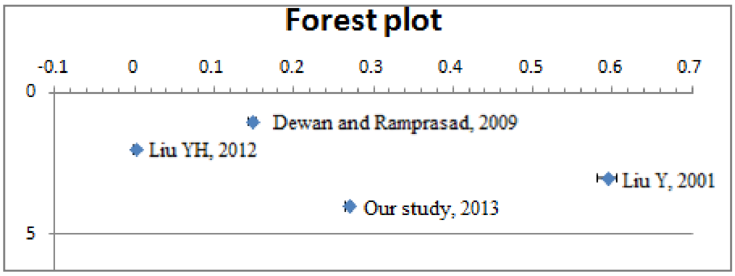


Figure 11: Forest plot of standardized regression coefficient, including our study

With regard to correlation coefficient, the previous researches showed that there is a high relation between the volume of social media buzz and the corresponding sales with an overall effect size of 0.94 across the studies. Since some of the results used for meta-analysis were obtained by reading a correlation coefficient from a study and others by taking a root of R-squared, a contribution of our study to the understanding of the effect size can be twofold. In case the correlation coefficient of the number of tweets is taken, the effect size in our study, i.e. correlation value of 0.80, is weaker compared to the vast majority of prior studies. However, an adjusted R-squared value of 0.991 would bring the result of our study close to the previous findings and hence support the conclusion that a higher set of independent variables, including number of tweets, is a better predictor of the sales of a product than an individual variable. Our study's relation to the previous studies with regard to coefficient of correlation (value of 0.80) is depicted in Figure 12.

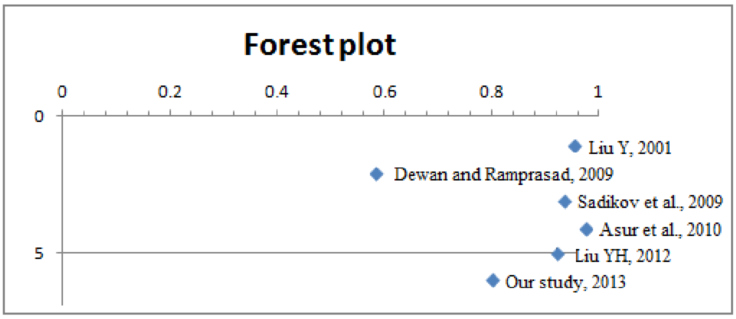


Figure 12: Forest plot of correlation coefficient, including our study

6.3. Future Research

Inclusion of Followers and YouTube as Central Variables

We strongly encourage future researchers to include the number of followers as a central variable regarding studies about Twitter. Including the number of followers in our research was a unique practice that has resulted in extremely strong outcomes, which were even stronger than assessing the total number of tweets alone. In addition, the number of YouTube views prior to the release date was proven to be highly correlated to the box-office sales. However, it should be critically noted that the number of YouTube views is only valuable for the prediction of box-office movies and cannot be generalized to other fields of research. The total number of followers on the other hand, is an extremely interesting variable since this can be extended to all different fields of research in relation to Twitter. However, it should be highlighted that there is only a rationale for including the number of followers when influencing and popularity are at hand. For instance, the number of followers would be a valuable attribute when tweets are used to predict the outcome of a talent show, since one Twitter user has an influence on its follower's opinions.

Extension of the Regression Model

Moreover, there is space for an elaboration of the regression analysis within the entertainment industry's perspective. Although we introduced two unique variables, additional variables such as the popularity of the actors or the number of weeks a movie is in the theatre can be included. Similarly, there are several other variables that might influence the sales of the movies but are not yet included in our research. Variables such as the season the movie is released in, the number of competing "strong" movies and the general economy can have an influence on the sales of the movies (Sadikov et. al, 2009).

Changing Business Model

Additional research should dive deeper into the changing business model of the movie industry that will certainly affect the dependent variable. Cameron and Bazelon (2013) argue that the impact of digitization on the movie industry is significant where an increasing number of TV's is connected to the Internet and online streaming services emerge. Moreover, Cameron and Bazelon (2013) propose three business models such as online subscription rentals, online video rentals and advertising based services to move away from the traditional revenue model that focuses on theatre sales. These new income platforms should be included in the dependent variable by future researchers. In addition, Chan (2006) argues that the box-office revenue trends vary across countries, which are correlated to different downloading behaviours. The effect of these trends on the dependent variable should be assessed, since there has been found a statistically significant evidence that these non-paid consumptions of movies replaces paid consumption in the theatres (Rob and Waldfogel, 2004).

Additional Suggestions

In addition, there are several areas that are extremely interesting for expansion on our research. For example, a further division between different time zones can be made to see if trends in non English speaking areas are similar to those in English speaking countries. Also, a study by Comscore (2010) reveals that the penetration of Twitter users per country is different, which will most probably lead to different results after comparing the total number of tweets or followers from one country to the other. Moreover, our findings can be generalized to other areas of interest, such as album or book sales. However, in comparison with box-office movies, other product groups might face the following limitations (Sadikov et. al, 2009). Firstly, there is not always a straightforward release date which might lead to problems when pre-release tweets should be measured. Secondly, the sales distribution might not be comparable across these other product domains. Thirdly, these products might not generate enough tweets to make reliable and valid predictions. Fourthly, it might be virtually impossible to measure sales of some of the other product domains which are less publicly accessible than movie sales.

6.3.1. Predictive Formula

As discussed in the *Introduction* chapter, predictive analytics can be used to forecast future success of a product and facilitate making sound marketing decisions (Kucera and White, 2012). In the previous chapters, we have proved that information gathered on the Internet can contribute to a business intelligence system by transforming raw data into valuable information for forecasting purposes. Hence, taking into account the strong correlations and high explanatory power of the regression models, it would be useful to exploit the obtained knowledge in order to quantifiably predict the future outcomes of a product. Particularly, based on the high scores of the data collected in week *t* and sales for week *t+1*, as described in the *Results* section, we suggest to establish a predictive model which would be able to forecast future sales of a product. There have been prior attempts to produce a predictive sales tool. For instance, in their study, Asur and Huberman (2010) have suggested a regression model for predicting sales with three independent variables, including sentiment of tweets. However, the process of classifying tweets based on valence is time-consuming and not necessarily reliable. According to Barbosa and Feng (2010), due to the shortness of messages, there is a considerable possibility of errors in assigning sentiment to tweets. Looking at the obtained results of our study, we argue that future outcomes can alternatively be predicted by looking at variables such as tweet rate, follower rate, YouTube views and number of theatres a movie is released in. The regression model of the relationship between these independent variables and the sales a week after data collection showed an R-squared value of 0.996 and an adjusted R-squared value of 0.991, meaning that the relation can be completely explained by these variables. This implies that given that the data can be obtained, the sales of a movie for the opening week can be predicted a week prior to release. It should be outlined that this model is only relevant for predicting the sales of movies and not for all products in the entertainment industry, since YouTube views and number of theatres specifically apply to box-office movies. The possible overall formula for a model would be the following:

$$y_{Sales} = \beta_1 \times x_{youtube} + \beta_2 \times x_{followers} + \beta_3 \times x_{tweets} + \beta_4 \times x_{theatres} + \varepsilon_i$$

where *β*s denote the proposed coefficients of the regression model and *X*s are the data obtained for each of the variables. Furthermore, a critical note should be made that this model employs unstandardized coefficients which are a better reflection of the causal relation within a function (Jaccard et al., 1990). According to Smith (2013), the number of Twitter users is continuously growing and this could have an impact on the model. Namely, an increase in the number of users would imply that more people would possibly tweet about a product leading to less importance of a single tweet or a follower and hence undermine accuracy of the model. This should be taken into consideration when building a predictive tool.

In conclusion, this model could serve as a base for future researches which would be aiming to develop a predictive sales tool. However, it could be further elaborated upon and possibly other variables should be included in the model in order to reach a function with an even higher explanatory power of the relationship between variables.

Bibliography

Altermatt, E. & Pomerantz, E. (2003) *“The development of competence-related and motivational beliefs: An investigation of similarity andinfluence among friends”,* Journal of Educational Psychology.

Anonymous (2013), Available: http://www.crunchbase.com/company/icerocket, last accessed: 13th of May 2013.

Anonymous (2013), Available: http://www.boxofficemojo.com/about/?ref=ft, last accessed: 13th of May 2013.

Anonymous (2013), Available: http://www.businessdictionary.com/definition/regression-analysis-RA.html#ixzz2TGD9paaS, last accessed: 13th of May 2013.

Anonymous (2013) Available:http://stats.stackexchange.com/questions/7110/difference-between-longitudinal-design-and-time-series, Last accessed: 13th May 2013.

Asur, S. Huberman, B.. (2010). Predicting the Future With Social Media.*Social Computing Lab HP Labs Palo Alto, California*. 1 (1), 1-8.

Barbosa, L. and Feng, J. (2010) Robust Sentiment Detection on Twitter from Biased and Noisy Data. *Coling 2010: Poster Volume*, p.36-44. Available at: http://www.aclweb.org/anthology-new/C/C10/C10-2005.pdf [Accessed: 25 May 2013]

Boundless (2013) Available: https://www.boundless.com/writing/academic-writing/practice-writing-in-social-sciences/creating-your-literature-review/, last accessed: 19th May 2013.

British Library. (2013). *Minority Ethnic English*. Available: http://www.bl.uk/learning/langlit/sounds/case-studies/minority-ethnic/. Last accessed 14th May 2013.

Bryden, J. Funk, S. Jansen, V. (2013). Word usage mirrors community structure in the online social network Twitter. *licensee Springer*. 2 (3), 1-6.

Calkins, K. (2005). *Definitions, Uses, Data Types, and Levels of Measurement*. Available: http://www.andrews.edu/~calkins/math/edrm611/edrm01.htm. Last accessed 13th May 2013.

Cameron, L. Bazelon, C. (2013). The Impact of Digitization on Business Models in Copyright-Driven Industries: A Review of the Economic Issues.*The Brattle Group, Inc*. 1 (1), 1-51.

Chan, J. (2006). The Impact of Unpaid Movie Downloading on Box Office Sales. *Wharton Research Scholars Journal*. 5 (1), 1-60.

Choudhury, A. (2009) Available: http://zencaroline.blogspot.nl/2011/03/correlation-coefficient-r.html, last accessed: 28th Feb 2013

Comscore. (2010). *Indonesia, Brazil and Venezuela Lead Global Surge in Twitter Usage.* Available:

http://www.comscore.com/Insights/Press_Releases/2010/8/Indonesia_Brazil_and_Venezuela_Lead_Global_Surge_in_Twitter_Usage. Last accessed 14th May 2013.

Cooper, H. and Hedges, L. (1994) The Handbook of Research Synthesis. New York: Russel Sage Foundation, p.193-195

Cumming, G. (2012) *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, Taylor et Francis Group: 8-51.

Dewan, S. & Ramprasad, J. (2009) *"Chicken and Egg? Interplay between Music Blog Buzz and Album Sales"*, Pacific Asian Conference on Information Systems.

Ducham, P. (2010). *Measuring Market Opportunities.* Available: http://answers.mheducation.com/marketing/marketing-strategy/measuring-market-opportunities. Last accessed 14th May 2013.

Eckerson, W. (2007). PREDICTIVE ANALYTICS Extending the Value of Your Data Warehousing Investment. *TDWI BEST PRACTICES REPORT* . 1 (1), 1-34.

Farber, D. (2012). *Twitter hits 400 million tweets per day, mostly mobile.* Available: http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/. Last accessed 31th Jan 2013.

Fetto, J. (2010). *2010 American Movie-Goer Consumer Report*. Available: http://www.experian.com/blogs/marketing-forward/2010/02/20/2010-american-movie-goer-consumer-report/. Last accessed 2th May 2013.

Fielding, N. Lee, R. Blank, G. (2008). The Internet Survey. In: The SAGE Handbook of Online Research Methods. London: Sage Publication, Inc. 175-254.

Fielding, N. Lee, R. Blank, G. . (2008). Internet-Based Interviewing. In:The SAGE Handbook of Online Research Methods. London: Sage Publication, Inc. 271-287.
Fowler, F (2002). Survey Research Methods. 3rd ed. London: Sage Publication, Inc. 1-181.

Freedman, D. (2005) “Statistical Models: Theory and Practice”, Cambridge University Press.

Gallaugher, J (2013). *Information Systems: A Manager's Guide to Harnessing Technology, v. 1.4*. 4th ed. London: Flat World Knowledge. 1-199.

Gingery, T. (2011). Advantages and Disadvantages of Online Surveys.Available: http://survey.cvent.com/blog/market-research-design-tips-2/advantages-and-disadvantages-of-online-surveys. Last accessed 15th May 2013.

Gonzalez-Bailon, Sandra, Wang, Ning, Rivero, Alejandro, Borge-Holthoefer, Javier and Moreno, Yamir, Assessing the Bias in Communication Networks Sampled from Twitter (December 4, 2012). Available at SSRN: http://ssrn.com/abstract=2185134 or http://dx.doi.org/10.2139/ssrn.2185134

Gupta, M. Gao, J. Zhai, C. Han, J. (2012). Predicting Future Popularity Trend of Events in Microblogging Platforms. *ASIST*. 1 (1), 1-10.

GraphPad. (2013). *What is the difference between ordinal, interval and ratio variables? Why should I care?*. Available: http://www.graphpad.com/support/faqid/1089/. Last accessed 13th May 2013.

Habeshian, V. (2013). *SSocial Takes Up 27% of Time Spent Online.*Available: http://www.marketingprofs.com/charts/2013/10582/social-takes-up-27-of-time-spent-online. Last accessed 13th May 2013.

Haddad, F. (2010). ‘An Undiscovered Archive? Online Video Sharing, Alternative Narratives and the Documentation of History.’. Available: http://www.academia.edu/743142/Using_YouTube_and_Social_Media_in_Research. Last accessed 8th May 2013.

Hair, J. (2007) *“Knowledge creation in marketing: the role of predictive analytics”,* European Business Review, pp. 305-315.

Hak, T. (2012) *“Course Book: Research Training & Bachelor Thesis Course 2012-2013”,* Rotterdam School of Management, Erasmus University

Hansen, D., Schneiderman, B. & Smith, M. (2011) “Analyzing Social Media Networks with NodeXL: Insights from a Connected World”, Morgan Kaufman.

Ingram, M. (2011). *Why changing Twitter’s 140-character limit is a dumb idea.* Available: http://gigaom.com/2011/07/21/why-changing-twitters-140-character-limit-is-a-dumb-idea/. Last accessed 14th May 2013.

Internetworldstats. (2013). INTERNET GROWTH STATISTICS. Available: http://www.internetworldstats.com/emarketing.htm. Last accessed 13th May 2013.

Jaccard, J., et al. (1990) The Detection and Interpretation of Interaction Effects between Continuous Variables in Multiple Regression. *Multivariate Behavioral Research*, 25 (4), p.467-478. Available at: http://public.kenan-flagler.unc.edu/faculty/edwardsj/jaccardetal1990.pdf.

Java, A., Song, X., Finin, T. & Tseng. B. (2007) *"Why We Twitter: Understanding Microblogging Usage and Communities",* Available: http://dl.acm.org/citation.cfm?id=1348556, Last accessed 13-05-2013.

Kashyap, V. (2010). What Is An API & What Are They Good For? [Technology Explained]. Available: http://www.makeuseof.com/tag/api-good-technology-explained/. Last accessed 13th May 2013.

Kim, L. (2013). *Keyword Suggestion Tool - Yahoo! & Google Suggestion Tool: Drive More Traffic Today!*. Available: http://www.wordstream.com/keyword-suggestion-tool. Last accessed 14th May 2013.

Kucera, T. White, D. (2012). Predictive Analytics for Sales and Marketing. Aberdeen Group. 1 (1), 1-6.

Liu, Y. (2001) "*Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue*", Journal of Marketing, pp. 1-49.

Liu, Y.H. (2012) "*How Does Online Word-of-Mouth Influence Revenue? Evidence from Twitter",* Department of Economics, Suffolk University.

McCormick, T. Lee, H. Cesare, N. Shojaie, A. (2013). Using Twitter for Demographic and Social Science Research: Tools for Data Collection.*Center for Statistics and the Social Sciences*. 1 (1), 1-31.

Mitchell, A. Hitlin, P. (2013). *Twitter Reaction to Events Often at Odds with Overall Public Opinion.* Available: http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/. Last accessed 14th May 2013.

Mentzer. (2005). *Qualitative Sales Forecasting.* Available: http://www.sagepub.com/upm-data/4914_Mentzer_Chapter_5_Qualitative_Sales_Forecasting.pdf. Last accessed 14th May 2013.

Moore, R. (2009). *Twitter Data Analysis: An Investor's Perspective.* Available: http://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/. Last accessed 16th Jan 2013.

Nyce, C. (2007). Predictive Analytics White Paper. *American Institute for CPCU*. 1 (1), 1-15.

Rob, R. Waldfogel, J. (2004). Piracy on the High C's: Music Downloading, Sales

Displacement, and Social Welfare in a Sample of College Students. University of Pennsylvania and NBER. 1 (1), 1-48.

Rui, H. Liu, Y. Whinston, A. (2013) *"Whose and what chatter matters? The effect of tweets on movie sales"*, Decision Support Systems, available online: 6th January 2013.

Sadikov, E., Parameswaran, A. & Venetis, P. (2009) *"Blogs as Predictors of Movie Success*" Proceedings of the Third International ICWSM Conference.

Sicular, S. (2013). *Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s.* Available: http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/. Last accessed 14th May 2013.

Siegel, E. (2008). *Predictive Analytics Applied: Marketing and Web.*Available: http://www.predictionimpact.com/predictive-analytics-docs/PredictionImpactAnalyticsPresentation.pdf. Last accessed 14th May 2013.

Silver, N. (2012). The Problems With Forecasting and How to Get Better at It. Available: http://fivethirtyeight.blogs.nytimes.com/2012/06/25/the-problems-with-forecasting-and-how-to-improve/. Last accessed 14th May 2013

Simchi-Levi, D., Kaminsky, P. and Simchi-Levi, E. (2009), Designing and managing the supply chain, 3rd ed., McGraw-Hill Irwin

Smith, A. Brenner, J. (2012). Twitter Use 2012. PewResearchCenter. 1 (1), 1-12.

Twitter. (2012). The Streaming APIs. Available: https://dev.twitter.com/docs/streaming-apis. Last accessed 13th May 2013.

Smith, A & Brenner, J. (2012). Twitter Use 2012. *PewResearchCenter*. 1 (1), 1-12.

Smith, C. (2013) *Twitter approaching 500m users, growth far outpacing Facebook*. [online] Available at: http://www.techradar.com/news/internet/web/twitter-approaching-500m-users-growth-far-outpacing-facebook-1128483 [Accessed: 23 May 2013].

Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E., Barash, V., Dunne, C.,

Capone, T., Perer, A. & Gleave, E. (2009) "Analyzing (Social Media) Networks with

NodeXL", Microsoft Research-Cambridge, Microsoft Research-Redmond.

Smith, C. (2013). (May 2013) How Many People Use the Top Social Media, Apps & Services?. Available: http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/. Last accessed 15th May 2013.

Smithson, J. (2000). Using and analysing focus groups: limitations and possibilities. International Journal of Social Research Melthodology. 3 (1), 103-119.

Twitter. (2012). *The Streaming APIs*. Available: https://dev.twitter.com/docs/streaming-apis. Last accessed 13th May 2013.

Twitter. (2013). *The fastest, simplest way to stay close to everything you care about*. Available: https://twitter.com/about. Last accessed 13th May 2013.

Vogel, H. (2007) Entertainment Industry Economies, Cambridge University Press, 7th edition.

Wauters, R. (2010). Only 50% Of Twitter Messages Are In English, Study Says. Available: http://techcrunch.com/2010/02/24/twitter-languages/. Last accessed 14th May 2013.

Wistia. (2011). 4 Ways To Keep Viewers Engaged In An Online Video.Available: http://wistia.com/blog/4-ways-to-keep-viewers-engaged-in-an-online-video/. Last accessed 8th May 2013.

World Bank. (2013). Internet users as percentage of population. Available: https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&met_y=it_net_user_p2&idim=country:USA&dl=en&hl=en&q=number%20of%20internet%20users#!ctype=l&strail=false&bcs=d&nselm=h&met_y=it_net_user_p2&. Last accessed 15-05-2013.

YouTube. (2013). *Statistics*. Available: http://www.youtube.com/yt/press/statistics.html. Last accessed 8th May 2013.

Zwilling, M. (2013). *Predictive Analytics is a Goldmine for Startups.*Available: http://www.forbes.com/sites/martinzwilling/2013/03/11/predictive-analytics-is-a-goldmine-for-startups/. Last accessed 14th May 2013.

Appendices

A. Replication table

	Research Strategy			Independent variable*			Dependent variable*			Results		
Author (-s)	Research Strategy	Sample size	Total observations	Definition	Range	Data collection method	Definition	Range	Data collection method	Effect Size Measure	Observed Effect Size	Confidence Interval (95%)
Liu, Y. (2001)	Longitudinal study	40 box-office movies	12136	Yahoo! Movies posts	≥ 0 Yahoo! movie posts	Message Board on http://movies.yahoo.com/	Box-office revenues of a movie	Mean: $67,005,348 Standard Deviation: $56,681,186	*Variety* magazine	Regression coefficient	β = 0.5920	0.5803 ≤ β ≤ 0.6034
Dewan, S., Ramprasad, J. (2009)	Time series study	2694 music albums	120596	Blog posts	≥ 0 blog posts	Google BlogSearch	Music Album sales	≥ $0	Nielsen SoundScan	Regression coefficient	β =0.1470	0.1415 ≤ β ≤ 0.1525
Sadikov, E, Parameswaran, A., Venetis P.(2009)	Time series study	197 box-office movies	300000000 (appx. number of blog posts in 2008)	Blog posts	≥ 0 blog posts	Spinn3r.com	Box-office revenues of a movie	≥ $0	Rottentomatoes.com	Pearson's correlation	r = 0.847	0,847 ≤ r ≤ 0,847

Author (-s)	Research Strategy	Sample size	Total observations	Definition	Range	Data collection method	Definition	Range	Data collection method	Effect Size Measure	Observed Effect Size	Confidence Interval (95%)
Asur, S., Huberman, B.(2010)	Longitudinal study	24 box-office movies	2890000	Tweets	≥ 0 tweets	Twitter Search Api program	Box-office revenues of a movie	$263,941 - $142M	Boxoffice mojo.com	R squared	$R^2 = 0.973$	$0.97257 \leq R^2 \leq 0.97343$
Liu, Y. H. (2012)	Longitudinal study	57 box-office movies	929858	Tweets	≥ 0 tweets	twittercritics.com	Box-office revenues of a movie	$1,183,354 - $366,007,900	Boxoffice mojo.com	Regression coefficient	$\beta = 0.000$	$-0.0020 \leq \beta \leq 0.0021$
Rui, H. Liu, Y. Whinston, A.(2013)	Longitudinal study	63 box-office movies	4166623	Tweets	≥ 0 tweets	Twitter Search API program	Box-office revenues of a movie	≥ $0	Boxoffice mojo.com	GMM estimator	GMM = $76,348.75	$58,098.06 ≤ GMM ≤ $94,599.44

* Indepedent variables and dependent variables in all studies are of ratio type.

B. Missing Scores

In order to arrive at a valid conclusion, it is essential to assess the missing scores from different perspectives regarding our studies.

First of all, the movies *42* and *Trance* are excluded from our study. The reason for this is that these movies have ambiguous names. After running a test import on NodeXL, we found that the majority of the tweets were not about the movie but about other events. However, it should be clearly stressed that the exclusion of these movies does not make our conclusions less reliable. Excluding these movies because the names are ambiguous on Twitter can be seen as a random exclusion. The fact that the movie has an ambiguous name regarding the NodeXL importing does not have any correlation to the final box office revenues.

Besides that, the importing practice itself has let to missing data for two different reasons. As discussed before, due to the popularity of certain movies and the limit of importing 1500 tweets per run, it was impossible to capture all the tweets for some movies. Although we have made sure to include different time slots for the cases consistently, the ideal practice would have been to include all the tweets that were posted on Twitter. Besides that, it is virtually impossible to grasp all the information just because Twitter users do not express themselves in perfect language. We included the most relevant key words in our research and expect that we imported the vast majority of tweets. However, there will be some tweets missing due to spelling mistakes, slang or popular language. It should again be stressed that this missing error should be quite similar for each of our cases and therefore should not have a strong effect on our results.

A worst case is the situation in which the missing information, if known and added to the matrix, would cause a maximum change in the observed effect. Because the missing information is unknown, the best practice is not to estimate the error but to prevent it. We have done everything within our technical capabilities to limit the missing information. Although it was inevitable to exclude the two ambiguous movies from our data set, we can perfectly argue why this did not lead to a considerable nonresponse bias. Once again, the exclusion of these two movies is completely random.

A research by the PewResearchCenter explored Twitter usage and concluded that the demographics of Twitter users are different than those of society (Smith and Brenner, 2012). Twitter is used by mainly young adults and less by the more aged citizens. This fact can have an influence on the studies, because this group might not be a fair representative for our studies. However, the American movie-goer consumer report of 2010 found that the demographics of customers visiting the cinema is almost equally scattered (Fetto, 2010). Although the imbalance is slightly stronger on Twitter as a medium than for cinema visits, this fact strongly flattens the bias.

Another research found that that the search API over-represents the more central users (Gonzalez-Bailon et al., 2012). However, if this overrepresentation is consistent across all our movies, this should not have a considerable impact on our study. It is a technical process to limit this bias, however, using the stream API could lead to more accurate results.

C. Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Tweet Rate Week t	30	.0350877192982456	76.3026634382566600	10.442635367568588	20.864541341595597
Follower Rate	30	2721.4737	72759824.5714	3846464.036739	13316403.3607972
YouTube Views	30	2118	17309358	1459797.40	3622158.520
Total Posts	30	5	36526	3004.47	7218.037
Theatres	30	1	4253	987.57	1509.749
Valid N (listwise)	30				