Social media and sales: Determining the predictive power of sentiment analysis towards car sales

Author: Olivia H. Plant
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT
This paper aims at exploring the use of sentiment analysis on social media as a tool for sales forecasting in the automotive industry. Previous research on this topic has presented significant results although current literature still lacks investigation on the usefulness of this technique when it comes to more expensive items. In particular, about 500,000 social media posts and eleven car models from the Dutch market are analyzed using linear models. Furthermore, the research compares these outcomes to the predictive power of search volume by using Google Trends as an indicator. Based on variables that are assessed as strong predictors of sales a prediction model using decision tree regression is built that can potentially be used by car manufacturers as an addition to traditional forecasting methods if tested and developed further.

The results suggest that social media sentiments have little to no predictive power towards car sales. While search volume as well as the rate of attention a model receives on social media show significant results and can be incorporated into the prediction model, the sentiments itself only obtain weak correlations with car sales and clearly show a limitation to the technique of sentiment analysis. Although the findings cannot be generalized for other car models and markets, this research contributes to further understanding the field of sentiment analysis and explores its boundaries. It also presents a prediction model that allows to approximate car sales. It therefore has both practical and academic value.

Supervisors: Dr. Fons Wijnhoven
Dr. Chintan Amrit

Keywords
Sales prediction, sentiment analysis, opinion mining, search volume, social media, car sales

7th IBA Bachelor Thesis Conference, July 1st, 2016, Enschede, The Netherlands.

1. INTRODUCTION

The use of social media is continuously increasing (PewResearchCenter, 2015). Likewise, the amount of data collected on social media platforms is increasing at an exponential rate. Platforms like Twitter or Facebook are specifically designed to allow users to interact and connect while other platforms are made for content sharing (Bing, Chan, & Ou, 2014). The data which is hereby produced can be of great value for companies seeking to understand their consumers better: Social media acts as a word of mouth and allows companies to collect large-scale and up to date data that represents honest consumer opinions. (Ceron, Curini, Iacus, & Porro, 2013; Tuarob, Tucker, & Asme, 2014). Especially platforms like Twitter are interesting for companies since it allows users to post and comment statements in real time. (Bing et al., 2014) Many companies have understood this development and pay increasing attention to social media content in order to make better decisions (Liu, 2012). Research has been conducted in various industries that seeks to understand the significance of this social media data but many industries are yet to be analyzed.

Data which is that large that it requires advanced and unique data storage, management and visualization technologies is usually referred to as 'big data'(Chen, Chiang, & Storey, 2012). Big data has long been viewed as an emerging trend and was part of the Gartner hype cycle for emerging technologies for many years until 2014 (Gartner, 2014). More recent hype cycles contain various big data issues as individual technologies which emphasizes the importance and complexity of big data in modern society.

A popular method used by many companies to analyze big data created through social media or other web platforms is sentiment analysis which implies analyzing people's opinions, sentiments, evaluations, attitudes, and emotions from written language (Liu, 2012). With regards to social media this means searching social media channels for posts on a particular topic or brand and identifying the opinions which consumers express in these. The development of sentiment analysis has also been recognized by the Gartner Hype Cycle for Content Management (Gartner, 2015).

Although various research has tried to examine the usefulness of social media data, the effect and significance of the obtained information is only partly explored. Managers also lack a clear guideline that tells them how social media can be used to obtain information and what the relationship of this information and economic data is. Therefore, research has started to explore whether sentiment analysis can help to predict sales. The results were promising although it has so far been mainly focusing on predicting sales of popular items that require relatively low involvement when purchasing them. Among the researched industries are movie sales (Asur & Huberman, 2010), stock price movements (Bing et al., 2014; Nguyen, Shirai, & Velcin, 2015), books (Dijkman, Ipeirotis, Aertsen, & van Helden, 2015) or even iPhone sales (Lassen, Madsen, & Vatrapu, 2014). Nevertheless, literature still lacks information about the validity and reliability of sentiment analysis in the context of more expensive items. This research aims to make a contribution to literature in order to partly fill this gap. The explored industry will be the Dutch car industry since cars are generally seen as high involvement purchases.

The automotive industry is a very important industry in the Netherlands. It employs about 50.000 people and is seen as one of the main industries in the Netherlands. In 2013 about 300 companies were involved in this sector, most of them exported their goods to other European countries or to China since there are no major car brands produced in the Netherlands (AutomotiveNL, 2013). Sales forecasting in the automotive industry is particularly important since cars in the current system are either built-to-delivery (which means that the purchaser will have to wait for his car after he ordered it) or built-to-forecast. However, the latter one often leads to a bullwhip-effect due to uncertainty in demand and inaccurate forecasting (Suthikarnnarunai, 2008). Even if cars are built-to-delivery, accurate forecasting can still help managers to plan and allocate their resources better.

Depending on their prior knowledge about cars, consumers tend to spend a considerable amount of time searching for information about a potential vehicle. This need for external search decreases if the consumer already possesses prior knowledge of specific attributes of the models offered for purchase. However, if the consumer has more general knowledge about cars and/or purchase decisions, he is more likely to benefit from an information search and therefore tends to search for more information. (Punj & Staelin, 1983). A study by Kandaswami and Tiwar (2014) showed that a majority of customers spend more than 10 hours to identify the best vehicle for their requirements - in China this number even reached 70% while in western countries like Germany and the US a percentage of 40-50% stated to spend at least this amount of time. The study also revealed that most consumers don't consciously rely on social networking sites when searching for information about possible vehicles since social media only ranked 6th place as preferred information sources. However, the traditional word-of-mouth reference was the most important source of information to which social media can contribute significantly (Tuarob et al., 2014).

This influence of social media as word-of-mouth on consumer behavior can possibly be of practical relevance to organizations. The report at hand will investigate whether sentiments expressed on social media can also act as a predictor for sales in the Dutch car market. It will relate data obtained through opinion mining on social media to car sales in the Netherlands and find out whether there is a relationship between these. The same industry has already been researched by Voortman (2015) who analyzed a possible correlation of Google Trends data and Dutch car sales. However, he suggested that there is still a need to also research the effect of sentiment mining on car sales. The main research question that is to be analyzed will be as follows:

'What is the predictive power of sentiments expressed on social media towards sales in the Dutch car industry?'

While most research on social media has tried to identify a formula that predicts sales most accurately by including many different variables, the exact contribution of each variable to the overall predictive power is yet to be determined. The research at hand therefore intends to explore the applicability of sentiment analysis to predict high involvement purchases by first determining the usefulness of each involved variable and then building a prediction formula based on these findings. It will also compare the predictive power of sentiments to the predictive power of Google Trends.

This paper will start by providing an overview of relevant literature and theories. From this a research model is constructed that intends to explain possible correlations between social media data and sales. Subsequently, data from social media and Google Trends is analyzed with regards to sales prediction abilities and the outcomes are compared to each other. Finally, a prediction model is constructed including the variables that were assessed as most useful. The results will then be discussed in the context of academic and practical implication and limitations. Finally, the paper will end with a conclusion about this research.

2. THEORETICAL FRAMEWORK

2.1 Literature search strategy

The literature search strategy included systematic and non-systematic approaches. At first, documents that corresponded to the research question were found by searching the databases Web of Science, JStor, Google Scholar and the University of Twente library with the query *((Twitter Or Google OR social media) AND (predict*) AND (sales) AND (sentiment mining OR search volume))*. More general consumer behavior theories were found by searching specifically for the theories like the AIDA model or the consumer decision making process. Since the researched topic is quite specific, a lot of literature was taken from backwards citations found in the already analyzed articles. During the search the observation was made that much of the previously analyzed literature drew information from the same articles which was taken as a confirmation that these backward citations are academically recognized and valid.

2.2 Sentiment analysis

An often interchangeably used term for sentiment analysis is the word opinion mining (Pang & Lee, 2008). According to Serrano-Guerrero, Olivas, Romero, and Herrera-Viedma (2015), an opinion describes an either positive or negative sentiment. Furthermore, an opinion consist out of a target to which these sentiments apply (Liu, 2012). The term subjectivity implies that a person holds a personal view, feeling or belief about the topic. While objective sentences consist of facts, subjective sentences are of a personal nature. Although subjectivity does not necessarily imply a sentiment, this is often the case, for example in case of judgement or appreciation (Serrano-Guerrero et al., 2015).

As stated in the introduction, sentiment analysis refers to extracting sentiments and opinions from written text (Liu, 2012). This process requires natural language processing (NLP) which is a closely related research area that explores how computers can be used to understand and manipulate natural language text or speech (Chowdhury, 2003). Sentiment analysis has been researched at a document level, a sentence level and an aspect/entity level of which the latter one is the most fine-grained level. Both areas, sentiment analysis as well as NLP are relatively young research fields. Both have been hardly researched until about the year 2000 but have been receiving rapidly increasing attention since then (Pang & Lee, 2008).

Many challenges arise linked to sentiment analysis. Serrano-Guerrero et al. (2015) classify the main tasks of sentiment mining tools into five groups: The first is sentiment classification or often termed *sentiment polarity*. Common problems with identifying and classifying sentiments as positive, negative or neutral arise when the author expresses multiple opinions or when there is more than one source of opinion mentioned in the text. These multiple opinions can contradict each other or refer to different attributes of the target (Liu, 2012). The second challenge is *subjectivity classification*. This means that the tool needs to define whether a text contains factual data or expresses the subjective belief of the author. Thirdly, the tool needs to *summarize* the given opinion of the author. The fourth challenge is to *extract* the opinion from the text. Finally, a great challenge for sentiment analysis tools is *sarcasm or irony* on which a lot of research has been conducted that seeks to improve tools in this regard. Serrano-Guerrero et al. (2015) recognize that there are other minor problems with sentiment analysis and classify them into a sixth category named ‘others’.

Many of these challenges have already been dealt with to a certain extent and the accuracy of sentiment analysis tools is steadily improving. Online sentiment analyses have shown great usefulness in many fields such as when predicting elections, sales or public opinion in general.

2.3 The influence of social media on consumer behavior

By analyzing various consumer behavior theories, a model can be established on how social media plays a part in influencing consumers buying decisions.

According to Ajzen’s Theory of Planned Behavior (1991) a behavior is influenced by an intention which again is influenced by three different factors. These are the person’s attitude towards the behavior, a subjective norm and the perceived behavioral control (describing the perceived easiness of fulfilling the behavior). Applying this theory to social media and consumer buying decisions, social media can help to shape the subjective norm that consumer’s experience. If a vast amount of users posts negative comments about a car, the consumer might decide to not buy this since he feels that society does not support this decision.

The degree of influence however depends very much on the circumstances of the consumer as well as the content and source of the information he receives. Based on components of the Theory of Reasoned Action (Ajzen & Fishbein, 1980) and the Technology Acceptance Model (Davis, 1989), Erkan and Evans (2016) established the Information Acceptance Model (IACM) which states that purchase intention is also influenced by the type of information a consumer receives about the item and whether he decides to adopt this information. Information adoption depends mainly on information usefulness which is influenced by three main factors. These are the 1. Information Quality, 2. Information Credibility and 3. The Need for Information. These three factors therefore also play a crucial role when it comes to the influence of social media on purchase decisions.

The stages of the decision making process in which a consumer can be influenced through social media can be mapped through the AIDA model which describes the funnel that consumers enter when they are drawn to a product and ultimately decide to buy it. According to Lassen et al. (2014), social media can play a part in all steps of this model:

Table 1. Adapted from Lassen et al. (2014)

Attention	Reading a social media posts about a product
Interest	Searching and reading reviews e.g. on social media or via Google, comparing to other products
Desire	Forming preference based on own opinion and social influence (among others from social media)
Action	Product mention/review/recommendation on social media → serves as influence for other readers

Just like Bing et al. (2014) and Voortman (2015) stated in their research, it is expected that social media sentiments affect sales only with a delay which we will name ‘time lag’. Table 1 shows that in general the time lag between an increase in positive or negative comments and an increase/decrease in sales is variable since the consumer can be influenced by social media in any stage (early or late) of their buying process.

The definition of this time gap becomes clearer when observing the consumer buying decision process by Kotler (1994) as seen in Table 2. According to this, consumers go through five stages when buying a product. Different to the AIDA model, the initial buying decision comes from the consumer himself through recognizing his need instead of being persuaded by a product.

Table 2. Consumer decision making process (Kotler, 1994)

1.	Problem recognition
2.	Information search
3.	Evaluation of alternatives
4.	Purchase decision
5.	Post purchase behavior

Since cars are classified as high involvement purchases it can be assumed that Kotler's model is slightly more accurate with regards to the time lag. Consumers are unlikely to see a car and then instantly be persuaded to buy it. The time lag will take place between the information search phase and the purchase decision while the consumer is evaluating alternatives. Although probably not searching for information about car models on social media, the consumer can be easily influenced by social media at this stage since he is very open minded towards new information. He might then actively search for information about a model of which he read a lot of positive reviews before researching other alternatives.

2.4 Search behavior and buying intention

While mentions on social media are seen as the rate of attention that a product receives, Google Trends represents the search activity or interest that potential customers perform on the product (Voortman, 2015). Research has shown that search activity can act as a representation of buying intention and even predict consumer behavior and sales of both low and higher involvements purchases (Choi & Varian, 2012; Goel, Hofman, Lahaie, Pennock, & Watts, 2010; Yang, Pan, Evans, & Lv, 2015).

Different to social media, which sometimes even consciously intends to influence decision makers as part of marketing strategies, Google only indirectly influences the customer decision process since it requires active search behavior from the customer side. It mainly represents already present interest in a product and can therefore in some cases also act as a predictor for buying intention. However, Google can still unintentionally influence the decisions made based on the results it presents to the person searching (Epstein & Robertson, 2015). Once a person's interest is evoked he is likely to do some research on Google in order to gather information. Later on, the person will probably make a buying decision based on the information he obtained through his search. Since high involvement purchases require more research from the customer side than low involvement and routine purchases, it is possible that Google is even a better predictor of the first one, shown in research such as the one by Yang et al. (2015) who predicted Chinese tourist volume.

In the previously discussed models, search interest can be categorized under the category *Interest* in the AIDA model and the *Information Search* of Kotler's consumer decision making process. Equal to social media, a time lag between the moment of search and the actual purchase is expected.

2.5 Social media as a predictor for sales

There has been previous research that explored the relevance of social media to predict sales, such as the work of Asur and Huberman (2010) who managed to predict box office sales remarkably accurate by including many variables such as sentiments but also the frequency of tweets into their prediction formula. Their research is widely recognized and has been cited more than 1200 times according to Google scholar. Furthermore, various other prediction researches have been based on the approach by Asur and Huberman (2010).

One research based on this method is by Lassen et al. (2014) who predicted quarterly iPhone sales by analyzing the sentiments of tweets and using a seasonal weighting of tweets to calculate the given quarter's proportion of the last calendar year.

Both researches used the following definition:

p: Tweets with positive sentiment
n: Tweets with negative sentiment
o: Tweets with neutral sentiment

with Subjectivity being:

$$Subjectivity = \frac{p+n}{o}$$

and the Positivity to Negativity Ratio (PNratio) being:

$$PNRatio = \frac{p}{n}$$

It is important to note that this definition of subjectivity differs slightly from the one of Serrano-Guerrero et al. (2015) mentioned earlier. As shown in the formula, subjectivity here always implies a positive or negative sentiment and is weighted against the number of tweets without sentiment.

A similar approach will be taken in this research where the PNratio will be the main independent variable. However, p will not be defined as the *number* of positive posts but as the *percentage* of positive posts from all posts about a particular car model over the time span of one month. This also yields for n being the percentage of negative posts. This is done in order to prevent the increase of social media usage and the subsequent increase of posts over the past years to influence the outcome of the research. A weighting of posts with regards to the total number of posts available each month will therefore not be necessary. Obviously, this will only have an impact when the relationships of either positive or negative mentions with sales are analyzed. The PNratio will be the same, regardless if it is calculated with percentages or absolute number since it only describes the ratio between those.

2.5.1 Hypothesizes

Both Asur and Huberman (2010) as well as Lassen et al. (2014) stated that the PNratio had a positive influence on sales. This is explained through the work of scholars on the topic of electronic word of mouth marketing (eWOM) and its influence on consumer buying decisions. Research generally agrees that eWOM has a positive impact on purchase intention, although the strength of this impact depends on factors such as the influence of the author or the corporate image of the advertised item (Bataineh, 2015; Cheung & Thadani, 2012; See-To & Ho, 2014). The first hypothesis to be researched will therefore be:

H_1: The PNratio of social media mentions about a car model has a positive influence on sales of this model

Bataineh (2015) concluded that three eWOM factors in particular have a significant and positive impact on consumer purchase intention which are eWOM credibility, quality and quantity. This finding is also supported by Cheung and Thadani (2012). Relating to this, it is hypothesized that not only the quantity of eWOM but also the general attention an item receives can have a possible influence on purchases. Therefore, the following hypothesis is established:

H_2: The number of total mentions about a car model on social media correlates positively with the amount of car sales

Equally to positive reviews having a positive impact on sales, it is assumed that negative reviews have a negative impact. Lee, Park, and Han (2008) found that the consumer attitude towards a product becomes more unfavorable as the proportion of negative online consumer reviews increase. Since they also stated that customers tend to believe negative comments more than positive ones it is expected that this relationship is even stronger than the one between positive comments and sales and will therefore be analyzed:

H_3: The percentage of negative mentions about a car model has a negative influence on the amount of sales of this model.

The previous theories have started from the assumption that people who have already bought a car, place social media posts with reviews of it online which then influence other people to buy the same item. However, this is does not necessarily yield for high end priced cars. Especially luxury cars tend to receive a lot of attention with only a very small and exquisite clientele purchasing them. It is therefore possible that an increase of positive comments from 'fans' of a high end car does not lead to an increase in purchases. This leads to the following hypothesis:

H_4: The higher the price of a car, the weaker the correlation between the social media data and the sales

Lastly, the research also intends to compare the predictive power of sentiment towards sales to the one of search volume towards sales. Until now there has been no research actively comparing these two variables. However, when comparing single studies about Google Trends and Twitter as predictors for sales, Twitter usually provided higher R Square values. (Asur & Huberman, 2010; Choi & Varian, 2012). A problem with these researches is that they obtained their data by also including other variables besides sentiments which makes it difficult to judge the pure prediction power of sentiments. This again emphasizes the need for an active comparison which will be given in this research. The following hypothesis is established:

H_5: The correlation of sentiments with car sales is higher than the correlation of relative search volume with car sales.

2.6 Research model

From the above mentioned hypothesizes, the following research model is established:

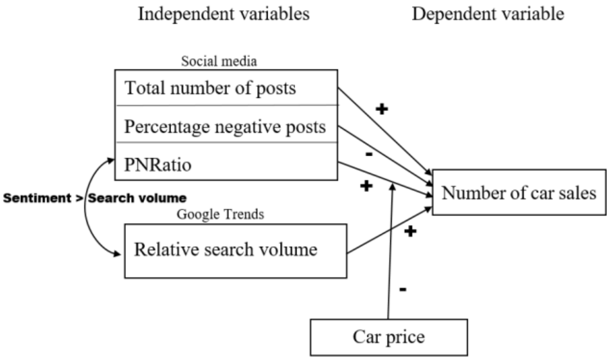


Figure 1. Research model

3. METHODOLOGY

In order to test the research hypothesizes and answer the research question sufficiently it is important to know how the group of opinion expressing people corresponds with the target group of the research (Wijnhoven & Bloemen, 2014). Obviously not all people who intend to buy a car have a social media account; however, in the Netherlands already 80% of the population with internet access also uses social media (CBS, 2015). The vast majority of car buyers is therefore assumed to also have access to social media.

The required data was gathered for a total of eleven models over a period of 52 months, ranging from January 2012 until April 2016. The cars were chosen according to the European system of car classification (Office for Official Publications of the European Communities, 1999) which divides cars based on their size and specifications. So called mini-cars are labelled as A Class while bigger cars are ranked as B Class and above. The list ends with the F Class which describes luxury cars. Extra classes include among others S (sports cars) and J (off-roaders). This research analyzes two models each from class B to E and one model each from class A, F and S which covers the most common cars as well as two luxury cars.

For each car model the following variables were collected:

1. Total number of posts about this model per month
2. Number of positive posts per month
3. Number of negative posts per month
4. Google Trends score per month
5. Number of cars sold per month

This data served to later on calculate the percentages of positive and negative comments as well as the PNratio.

The social media feeds were analyzed through the use of a student version of the tool Coosto. This tool allows to analyze social media posts placed in the Netherlands and classifies each post as either positive, neutral or negative. As sources it uses eight different social media sites, as well as various news sites, blog sites and forums. Since the research is determined to analyze social media sentiments, the news sites, blogs and forums were excluded from the search. The remaining social media sites that were searched are Twitter, Facebook, LinkedIn, YouTube, Google+, Hyves, Instagram and Pinterest. Most posts usually stem from Twitter since all data from this platform is available to the public while e.g. Facebook posts are mostly only available to read for friends of the author. In total a number of 502,681 social media posts were analyzed.

The search was conducted by entering the Dutch name of each car model into Coosto and taking down the monthly number of all comments related to this car model as well as the number of positive and negative comments. Included in the count of posts are both the original posts as well as retweets (on Twitter). This increases the validity of the measurement method since a post that is retweeted often is read by more people and could also influence more consumers in their buying decisions.

This data was entered into the statistical data program SPSS. Later on, the percentages of positive and negative comments and the PNratio were calculated. While searching, various spellings or expressions that a user could use while posting about a car model were considered. For example, the query for the car model Volkswagen Passat was: *"VW Passat" OR "Volkswagen Passat"*. Furthermore, cars with a similar name but different specifications were explicitly excluded for the search, such as the BMW 5-serie GT. Posts about this car model would also show up when searching for the BMW 5-series, however, the GT differs from the standard BMW 5-Series which means that posts about this version are unlikely to influence consumers considering to buy the standard 5-Series. The query was therefore defined as *"BMW 5-serie" –GT*. An overview of all cars, prices and search terms can be found in Appendix A.

As a comparison factor, the relative search volume per month from Google Trends was collected for the same car models using the same search terms. These sometimes had to be adjusted slightly in order to match the search language of Google Trends, for example by replacing the Boolean search term OR with the sign +. Google Trends only gives the relative search volume per month which is the query share of the searched term. The query share is calculated by dividing the query volume of the searched term by the total number of searches in the specified region and the given time frame. The month with the highest relative search volume is then normalized to 100. With regards to this research this means that the search volume of a particular car model was divided by all searches in the Netherlands between January 2012 and April 2016. After this, the month with the highest average then received the score while the scores of the other months were adjusted according to this maximum (Choi & Varian, 2012). This normalization implied the score of 100 represents a different query share (and absolute number of searches) for every model, depending on what the monthly maximum of searches was. This has to be considered when comparing the scores of different car models. The exact number of Google searches analyzed is unknown since Google did not provide for this number.

For the dependent variable, the monthly number of car sales was deducted from the documents given on the website of BOVAG, the Dutch Federation of Automotive Dealers and Garage Holders (*Bond van Automobielhandelaren en Garagehouders*)[1] which were also taken down into SPSS.

A study revealed that 60% of the buyers ('normal buyers') needed between one and six months from first thinking about buying a new car and the actual purchase while 16% needed less than a month for this decision. Only 9% needed more than a year to buy a new vehicle (Putsis & Srinivasan, 1994, 1995). The time lag will therefore be tested until a maximum of twelve months since sales more than a year after an increase in social media sentiments are unlikely to be causally related to those.

4. ANALYSIS AND RESULTS

4.1 Descriptives and linear regression analysis

After gathering the data some general analyzes were performed to map and understand the actual relationships between the variables. Since the PNratio seemed to be an appropriate predictor of sales according to the research model, a graph with the monthly mean values of the PNratios and the number of car sales was created (Figure 2).

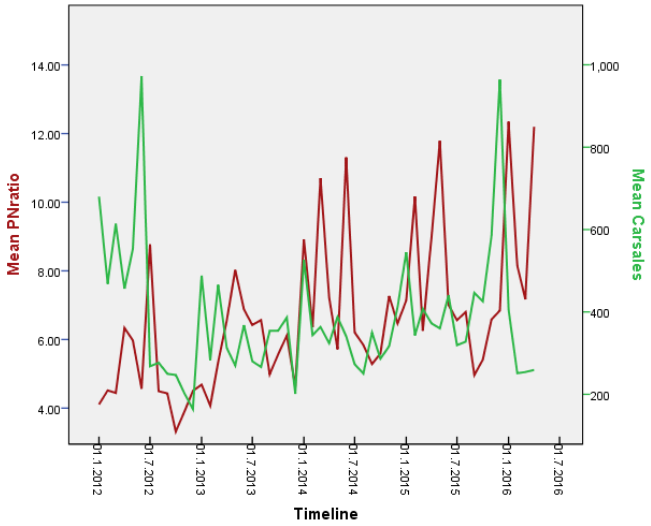


Figure 2. Graph showing the monthly means of PNatio and Sales

At first sight, the graphs seem to resemble each other slightly which could indicate a relationship between the two variables. Although spikes are amplified and shifted in some spots, a similarity is clearly visible. However, when observing more closely it becomes evident, that spikes of the PNratio seem to follow those of the car sales which is the opposite of what was hypothesized. Although this assumption would perhaps make sense, this possibility will not be tested in this research since it does not fit the research model and requires a different theoretical basis. The following sections will continue to explore possible relationships between sales and preceding social media data.

Before analyzing each car model separately, the averages of each model were calculated for the variables *PNratio, total number of mentions, percentage negative mentions* and *Trends*. Each variable was mapped onto a scatterplot to examine their relationship with sales. Although some of the graphs seemed very spread out they could be interpreted as somewhat linear. Since most findings in earlier research were also based on linear models (Asur & Huberman, 2010; Goel et al., 2010), a regression analysis for the PNratio and the sales was conducted in order to analyze the first three hypothesizes. While conducting the regression analyses, the residuals histogram and the PP plot were examined to ensure that the criteria of a linear regression analysis were fulfilled. The diagrams indicated that the errors were independent from each other, were approximately normally distributed and had a constant variance. This means that the relationships between the variables were indeed approximately linear and the conditions for a linear regression fulfilled. The use of linear models would therefore not bias the outcome. The outcomes of these linear regression analyses are summarized in Table 3.

Table 3. Linear regression analyses for averages of variables with sales.

Variable	R	R^2	Significance
PNratio	.111	.012	.746
Total mentions	.804	.646	**.003***
Percentage negative comments	.253	.064	.454
Google Trends	.320	.102	.338

The regression analysis for the PNratio with sales indicates a weak, negative correlation[2] that is not significant and obtains a very low R^2 value. The negative direction of the correlation does not fit the research model which stated that an increase in positive sentiments as well as a decrease in negative sentiments will lead to an increase in sales. As indicated by the graph, the sales therefore might not be causally related to the preceding social media sentiments. However, consumer behavior is assumed to vary based on the car model a person wishes to purchase. Therefore, although a correlation of the PNratio and sales cannot be found at this general level it is possible that the ratio serves as a predictor for sales if it is analyzed under consideration of the model type and by including a time lag between social media data and sales into the model.

In order to test the second hypothesis, a regression analysis between the total number of mentions (regardless of the type of sentiment) about a car model and the number of sales was conducted. This showed a strong correlation of 0.804, significant at $p < 0.01$ level, and a quite high R^2 value of 0.606. This means that about 60.6% of all variables can be calculated through the use of this variable. The third hypothesis is tested by conducting a regression analysis for the percentage of negative comments and the number of car sales. This analysis showed a weak and nonsignificant but negative correlation which fits the expectation

[1] https://www.bovag.nl/pers/cijfers/personenauto/verkoopcijfers-personenauto-s-naar-merk-model-per

[2] The negative relationship is not shown in Table 3 since R just displays the strength of the correlation and not its direction

of the research model. Finally, an estimate of the predictive power of Google Trends was obtained by conducting a regression analysis for Google Trends and sales which also showed a not significant but positive correlation.

4.2 Inclusion of time lags into the model

After analyzing the dataset for general relationships between all data and car sales, the dataset was divided per car model into smaller sets in order to analyze each car model separately. Furthermore, a time lag was incorporated. This was done by using the cross correlation function of SPSS and checking, for which time lag (smaller than or equal to 12 months) between the independent variable and the car sales the correlation was the strongest. Hereby only time lags were considered where the independent variable precedes the dependent variable and where the correlation fits the direction of the research model (for example, strong negative correlations between the PNratio and sales were ignored since they were assumed to be not causally related to each other). Table 4 shows per car model and variable, for which time lag the strongest correlation was found. In case no number is given, a correlation fitting the research model could not be found. An asterisk marks whether this correlation was significant at $p \leq 0.5$ or not.

The found correlations varied widely among the car models. This supports the assumption that consumer behavior does indeed vary per model. As evident, the PNratio does not seem to be a useful indicator for car sales, even with the inclusion of a time lag. Only one out of eleven found correlations was significant and this correlation had a time lag of 0 which means that the social media sentiments occurred in the same month as the corresponding car sales. It can therefore not be used to predict sales. In comparison to the Google Trends values, which showed many moderate and significant correlations, the PNratio of the analyzed models does not seem to have any predictive power towards car sales. H_1 can therefore only be accepted for the model VW Golf.

Besides the relative search volume, which seems to be the strongest predictor of sales, the total number of mentions also correlates significantly with car sales in five cases. Although the relationships are not as strong as initially suggested by the average data, this makes it the second strongest predictor of sales from the analyzed variables. Furthermore, as expected, the percentage of negative mentions seems to negatively correlate with car sales; however, this is only significant in one case. The negative relationship can therefore not be seen as proven with the other ten car models.

4.3 Influence of price on correlations

In order to analyze H_4, the eleven car models were split into two price classes. The lower price class contained all cars with a starting price of less than 30 000€ (five models) while the higher price class contained the cars who started at 30 000€ or more (six models). The prices were taken from the Top Gear website[3]. The correlations found in Table 4 were then compared in an independent t-test between the two groups. Since a t-test requires the variables to be normally distributed, the Shapiro-Wilk test for normality was conducted for every group (each correlation divided into high- and low-priced cars) whereby all variables were found to be normally distributed. The results for the four independent t-tests are shown in Table 5.

Table 5. Independent t-tests for high- and low priced groups of correlations with sales.

Correlations with sales	Mean difference	Significance
PNratio	-0.0006	0.990
Total mentions	-0.017667	0.868
Negative mentions	0.0644	0.154
Google Trends	-0.1642	0.178

Since the PNratio already showed no significant connection to car sales it also was not surprising that the car price had no significant influence on this correlation either. The mean correlation for lower priced cars with sales is 0.160 and the mean correlation for higher priced cars is 0.161. This leads to a mean difference of 0.0006 which is almost negligible. Furthermore, since Levene's test showed insignificant results, equal variances between the high- and low-priced group are assumed.

Equally, when looking at the differences between high and low priced cars concerning the correlations of total mentions, the percentage of negative mentions or Google Trends with sales, no significant difference is visible. Equal to the first test, Levene's test indicated equal variances in all cases which leads to the conclusion that there is no difference between the high and low priced car samples. H_5 is therefore rejected for all analyzed cars.

4.4 Establishing a prediction formula

After testing the variables for each car model separately and finding some moderate correlations, the question remains whether a combination of independent variables can lead to a more reliable prediction of sales. Previously, the total number of mentions about a car model and the Google Trends score were identified as the strongest predictors. In the following step, two tools will be used in an attempt to build a significant and reliable prediction model.

Table 4. Optimal time lags and Pearson's correlations per car model

	PNratio x Sales		Number total mentions x Sales		Google Trends Score x Sales		Negative mentions x Sales	
Car model	***Lag***	***Correlation***	***Lag***	***Correlation***	***Lag***	***Correlation***	***Lag***	***Correlation***
Fiat Panda	11	0.027	3	**0.424***	3	**0.694***	-	-
Ford Fiesta	8	0.223	1	0.247	12	0.257	8	-0.164
Opel Corsa	3	0.164	7	0.149	2	0.197	10	-0.152
Honda Civic	9	0.136	12	0.256	1	**0.519***	9	-0.206
VW Golf	0	**0.284***	7	**0.388***	4	**0.473***	8	-0.212
VW Passat	5	0.146	4	**0.508***	5	**0.486***	2	-0.207
Ford Mondeo	9	0.132	9	**0.499***	5	**0.344***	2	**-0.296***
BMW 5 Serie	9	0.152	-	-	12	0.015	11	-0.064
Mercedes Benz E-Class	9	0.205	5	0.063	8	0.123	5	-0.118
Porsche Panamera	10	0.055	9	**0.422***	9	**0.348***	11	-0.092
Porsche 911	9	0.244	4	0.245	9	0.277	2	-0.227

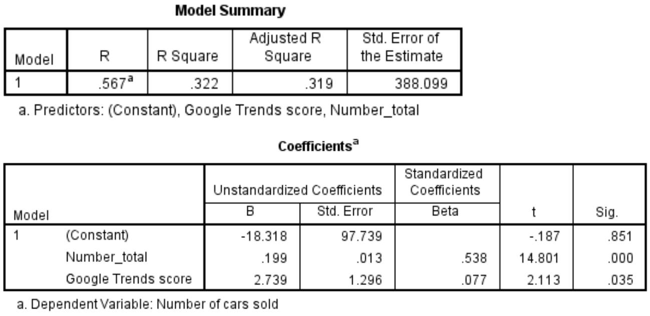
[3] http://www.topgear.nl/koopgids/nieuw/

At first, another linear regression analysis using SPSS will be executed in order to gain some general information about the variables which is not provided by the second tool. Then a decision tree regression using the M5P classifier of the data analysis tool WEKA will be conducted which also includes the different car models as a variable. Finally, if the classifier is evaluated as a successful improvement, a time lag identified as optimal for planning purposes will be implemented in order to establish the prediction formula.

Before performing any kind of parametric tests with combinations of variables, the independent variables were tested for multicollinearity. In case a multicollinearity was found this could reduce the impact of a linear regression analysis since the variable are dependent on each other. This was tested by determining the Variance Inflation Factor (VIF) for a regression analysis of the two variables that are to be included into the model. The VIF of the Google Trends score and the total number of mentions was 1.110. Since multicollinearity is only considered to be a problem in case the VIF is greater than 10 (University at Albany, 2003), both variables can be used with confidence in combinations with each other.

The results of the linear regression analysis showed weak correlations for both variables that were significant at $p < 0.05$. The weak correlation is not surprising since it was demonstrated in section 4.2. that correlations differ widely per car model. However, it is important to know that the inclusion of the two variables into a linear model leads to significant results since the following step will be conducted with another tool that does not provide a significance measure.

Table 6. Linear regression analysis for total number of mentions and relative search volume with sales.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.567[a]	.322	.319	388.099

a. Predictors: (Constant), Google Trends score, Number_total

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-18.318	97.739		-.187	.851
	Number_total	.199	.013	.538	14.801	.000
	Google Trends score	2.739	1.296	.077	2.113	.035

a. Dependent Variable: Number of cars sold

A regression tree analysis on WEKA improves this correlation to 0.7685 by building a decision tree based on linear models that considers the various car models. Since WEKA does not give a significance or R² value for the decision tree regression the models can be further compared through the root mean squared error (referred to as 'Standard Error of the Estimate' in SPSS) which has significantly improved through the inclusion of the car model.

The equation can be found in Table 7 under 'LM num 1' (linear model number 1). The model contains some so-called 'dummy-variables where, depending on the car model, the values 1 (TRUE) or 0 (FALSE) have to be inserted as a multiplicand. For example, the equation for the sales of the Volkswagen Passat can be easily simplified to:

Sales VW Passat =   194.9348
+ 226.5298
*+ 0.0485 * Number_total*
*+ 3.5877 * Trends_score*
- 260.1935

Table 7. Formatted output of M5P classifier in WEKA without time lag.

```
=== Run information ===
Scheme:       weka.classifiers.trees.M5P -M 4.0
Relation:     ALL DATA-weka.filters.unsupervised.attribute.
              Remove-R2-weka.filters.unsupervised.attribute.
              Remove-R2-3,5-6,9-11
Instances:    572
Attributes:     4
              Car_model
              Number_total
              Trends_score
              Sales
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
M5 pruned model tree:
(using smoothed linear models)
LM1 (572/63.987%)

LM num: 1
Sales =
      194.9348 * Car_model =  Ford Mondeo, BMW 5 series,
                              Fiat Panda, VW Passat,
                              Opel Corsa, Ford Fiesta, VW Golf
    + 226.5298 * Car_model =  Fiat Panda, VW Passat,
                              Opel Corsa, Ford Fiesta, VW Golf
    +   93.0651 * Car_model = Opel Corsa, Ford Fiesta, VW Golf
    + 175.9239 * Car_model =  Ford Fiesta, VW Golf
    + 267.5853 * Car_model =  VW Golf
    +     0.0485 * Number_total
    +     3.5877 * Trends_score
    -  260.1935

Number of Rules: 1
Time taken to build model: 0.07 seconds

=== Cross-validation ===
=== Summary ===
Correlation coefficient            0.7374
Mean absolute error              163.9318
Root mean squared error          318.1139
Relative absolute error            48.2572 %
Root relative squared error        67.566 %
Total Number of Instances         572
```

The correlation of 0.7685 is moderate and can be used on 63.987% of all cases which means that about half of all car sales can be estimated through the use of this equation.

However, as stated above, when trying to find a useful prediction formula it is important to include an appropriate time lag that gives car sellers and manufacturers enough time to react to the forecast. WEKA does not have the option to include an optimum time lag for each model into one decision tree; therefore, a time lag suitable for all models needs to be defined: Production planning in the automotive industry usually starts three months before the manufacturing and can be changed at the latest one month beforehand (Suthikarnnarunai, 2008). The lead time varies from a few weeks to a few months per car producer. For this formula, the time lag will be set to four months which gives manufacturers time to plan one month ahead of production and three months for production and distribution of the cars. An optimal time lag of four or five months was also found relatively often in Table 4 and is therefore expected to bring the best results. Nevertheless, the same procedure can also be used to establish a decision tree with another time lag. The time lag was added by simply shifting the number of sales in the dataset upwards since WEKA does not have the option to incorporate a time lag by itself.

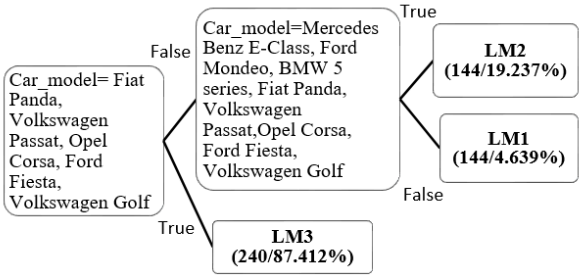
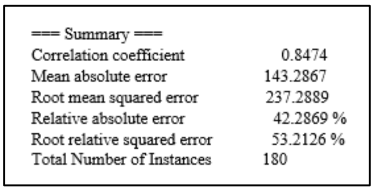

Car_model= Fiat Panda, Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
False
Car_model=Mercedes Benz E-Class, Ford Mondeo, BMW 5 series, Fiat Panda, Volkswagen Passat,Opel Corsa, Ford Fiesta, Volkswagen Golf
True
LM2 (144/19.237%)
LM1 (144/4.639%)
False
True
LM3 (240/87.412%)

Figure 3. Decision tree output M5P classifier including a four-month time lag.

The decision tree visible in Figure 3 was built on 66% of the data and tested on the remaining 34% to ensure that the model is not only applicable to this specific dataset but also to data on which it was not built. The outcome of the decision tree is one of three linear models that have to be used depending on the car model. For the complete model refer to Appendix C.

As displayed in Table 8, the results were good with a correlation of 0.8474 and an applicability of 100%[4]. Since the simple linear regression in Table 6 already brought significant results and the decision tree is also based on linear models it is assumed that this improved model is also significant. The root mean squared error of 237.3 is still quite high but is much lower than in the previous models. Practically, the model could perhaps be used as an addition in sales planning if it is further tested and improved.

Table 8. Summary of M5P classifier output for decision tree regression including a four-month time lag



```
=== Summary ===
Correlation coefficient                  0.8474
Mean absolute error                    143.2867
Root mean squared error                237.2889
Relative absolute error                 42.2869 %
Root relative squared error             53.2126 %
Total Number of Instances              180
```

5. DISCUSSION AND LIMITATIONS

5.1 Discussion

This research intended to answer the research question *'What is the predictive power of sentiments expressed on social media towards sales in the Dutch car industry?'*. The question was analyzed by searching for significant and meaningful correlations between the PNratio (quotient of positive and negative social media posts) and sales of particular car models. In order to judge whether negative sentiments had more predictive power without including the positive comments, the same procedure was conducted for the percentage of negative comments about these car models and sales. The results showed for both analyses that sentiments have little to no predictive power towards car sales. Although the directions of the relationships seemed to fit the research model (with the PNratio correlating positively and negative comments correlating negatively with sales), the found relationships were very weak. Only for one car model per variable a significant correlation was found although these were equally weak. Hypothesizes 1 and 4 where therefore rejected for 10 car models and both only accepted for one. A line chart suggested that social media sentiments might follow car sales instead of preceding it as assumed in this research. Although this would be of no practical relevance for planning purposes, this supposition would have to be analyzed in other research.

While sentiments showed very weak results, the general attention about a car model (represented by total number of mentions) and the search volume (Google Trends score) were shown to correlate much stronger with sales and in many cases significantly. It therefore seems much more likely that these variables are also causally related to sales than the sentiments, although a causal relationship was not proven in this research. Hypothesis 2 is therefore corroborated and specifically accepted for four car models. Hypothesis 5 is rejected since Google Trends showed to be a much stronger predictor for sales than the PNRatio and the percentage of negative mentions in all cases.

The results of Google Trends as a predictor for sales also match other research such as the ones of Wu and Brynjolfsson (2013), Choi and Varian (2012) and Yang et al. (2015). A combination of the two strongest predictors into a decision tree regression led to a prediction model that approximated the sales quite well. It could be used on 100% of the data in the dataset and showed a correlation of about 0.85. Although the root mean squared error was quite high, the model could possibly be used as an addition to traditional sales forecasting methods in the automotive sector if it is tested further.

Furthermore, a comparison between the higher priced and lower priced cars showed that there was no difference concerning the strength of the correlations between the analyzed variables and sales. The assumption that data of higher priced cars would show weaker correlations with sales (H_4) is therefore rejected. All conclusions per hypothesis are summarized in Table 9.

Table 9. Conclusion of hypothesizes per car model.

H_1	*The PNratio of social media mentions about a car model has a positive influence on sales of this model*	**Accepted** for VW Golf
H_2	*The number of total mentions about a car model on social media correlates positively with the amount of car sales*	**Accepted** for VW Golf, VW Passat, Ford Mondeo, Porsche Panamera
H_3	*The percentage of negative mentions about a car model has a negative influence on the amount of sales of this model.*	**Accepted** for Ford Mondeo
H_4	*The higher the price of a car, the weaker the correlation between the social media data and the sales.*	**Rejected** for all analyzed correlations
H_5	*The correlation of sentiments with car sales is higher than the correlation of relative search volume with car sales.*	**Rejected**

The fact that all correlations were found to be normally distributed while testing the conditions for a t-test is interesting. Although the strengths of the correlation seem to differ per car model, it seems like all correlations per variable are centered around a specific mean value which was found more often than other values. However, this assumption would have to be tested with a bigger sample size since a sample of eleven car models is not big enough to draw general conclusions about the shape of the distribution.

[4] The sum of the percentages in Figure 3 exceeds 100 because some car models can be calculated using two formulas

While other scholars showed that sales of lower priced items such as movie tickets or even IPhones can be predicted very accurately, no evidence is found that this also yields for cars. Speculations can be made on why this is the case although they will have to be verified in other research: The results suggest that consumers do not let the opinion of other people ('subjective norm' according to Ajzen (1991)) in an online environment influence their buying choices when it comes to more expensive items. This assumption also questions the usefulness of the Theory of Planned Behavior in the context of high involvement purchases and virtual communities. Another assumption is that other decision factors simply outweigh the subjective norm when deciding for a particular car model. Car sales seem to be strongly policy driven which means that consumers buying a car might consider factors such as the tax class of a model as more important than the opinion of other people. Especially in the Netherlands, taxes between car models vary significantly due to the green policy of the Dutch government (Crisp, 2014). The data also showed that car sales seem to vary per season although this was not the same in every year: While mostly car sales dropped towards the end of the year, they increased rapidly at the end of the year 2015 (see Figure 2), presumably because a new policy starting in the next year would require buyers to pay much higher taxes (Automotiveimport, 2015).

5.1.1 *Academic implications*

The research has primarily helped to further explore the boundaries of sentiment mining. Until now, little research had been conducted on the predictive power of sentiment mining in the context of high involvement purchases. Although in two cases some influence of social media sentiments on car sales was determined, none of this seems of particular strength which clearly sets a limit to the usefulness of sentiment analysis in car sales prediction.

This research contrasts other research within the field of sentiment analysis such as Asur and Huberman (2010) and Lassen et al. (2014) to name just a few researchers who received very good results when using this technique. Furthermore, research using search engine volume to forecast high priced items such as house sales (Wu & Brynjolfsson, 2013) or tourist volume (Yang et al., 2015) equally showed significant results with high R^2 values which makes it even more surprising that the same does not yield for sentiment analysis in this context. The research therefore has found new limitations that were previously unknown and should certainly be further investigated.

5.1.2 *Practical implications*

The decision tree regression leading to specific prediction formulas for car sales can be seen as the biggest practical contribution of this research. Although it is advised to test and improve the model on more data before using it in sales forecasting, the results found in this paper look very promising. Even if the formula proves to be less successful than expected or strategists in the automotive sector decide not to use it, this research nonetheless provides insights that can be very valuable to them. It is important for sellers to know that social media sentiments do not directly represent or influence the performance of their company. A lot of positive attention will therefore not necessarily lead to high sales within a few months without the car dealer making a good effort. Equally, negative comments are no reason to expect a significant decrease in sales. Nevertheless, general or extensive negative publicity should of course still be avoided, since many practical examples have proved that negative press will eventually lead to a decrease in sales (for example in the 2015 Volkswagen scandal).

Another practical implication for companies is the use of social media as a marketing tool to enhance car sales. Since consumers do not seem to be influenced by positive remarks in their buying decisions, it should be further investigated how effective social media strategies in this sector are so they can be adjusted accordingly.

5.1.3 *Reliability and validity of data*

According to the definition of the Psychology department of the University of California (2007), reliability refers to the repeatability of findings. Since the data are based on factual data such as the number and nature of social media mentions, the Google Trends value and the number of car sales, a second study on this topic would yield the same results. Both Coosto and Google Trends are working with a fixed algorithm which means that they would give the same results again if they were presented with the same queries. The reliability of data is therefore judged as high.

Validity describes the credibility of the research (University of California, 2007). This study might be criticized by some experts for its internal validity and specifically for the use of the tool Coosto. Since the algorithm is unknown to the public, its validity can therefore not be ultimately proven. However, the tool has a high reputation and is widely used. It has an accuracy of 80% according to Team Nijhuis (2013) which is very high compared to other tools considering the results of the research from Serrano-Guerrero et al. (2015). Considering the possibility that Coosto has classified some tweets falsely as positive or negative, the internal validity of the PNRatio and other sentiment based variables can be disputed. Nevertheless, the total number of mentions as well as the Google Trends score can be classified as internally valid. The relative search volume stems from Google itself which means that the Google Trends score is based on complete data. Of course Google Trends does not include other search engine queries; however, Google has a market share of about 75% compared to other search engines (Schwartz, 2015), which means that the vast majority of online searches is covered within this score. Furthermore, the results of other search engines are not expected to differ significantly from those of Google which leads to the conclusion that Google can be seen as a valid representative of all search engines. The total number of social media mentions is drawn from the database of the tool Coosto. Although it is likely that this database does not contain all posts ever posted on social media, it can also be seen as a valid representative of all social media data.

External validity refers to the extent to which the findings of a study are generalizable beyond the research (University of California, 2007). Since sentiments showed weak correlations with sales in all cases, it seems reasonable to assume that their predictive power will be equally low when it comes to other car models on the Dutch market. However, these findings cannot be generalized outside the Dutch market since consumer behavior may vary based on culture, and other decision factors might apply due to other policies. Furthermore, the individual car models showed varying results concerning the strengths of the rate of attention and the search volume as a predictor. It is therefore not possible to deduct general assumptions about other models from the findings of this research. Equally, the prediction model is only valid for the analyzed car models and for the Dutch market. Generalization of the findings is therefore for most findings not possible.

5.2 Limitations and further research

The study at hand has provided a lot of unexpected results that open up further questions. The assumptions made in the previous section should be subject to further research since they could have important implications for the research field of sentiment mining and the automotive industry. Firstly, the question whether sentiment mining has an equally weak predictive power when applied to other car models or other high priced items should be further investigated. The car industry was chosen as representative for high involvement and expensive purchases, however, the findings cannot be generalized for other high priced items. As mentioned previously, Dutch car sales are influenced by other factors which are unique to this industry and could have mitigated the prediction ability of social media. It is therefore unknown whether other high priced items can be forecasted through social media or not.

The study has analyzed and (in some cases) found correlations between car sales and preceding social media activity. This activity can be explained by the research model which was built based on established literature. However, there is no proof resulting from this study that the research model is also the actual causal model. Based on the previous speculations it is therefore advisable for further research to focus on the nature of the relationships between the applied variables and on their directions. Since Figure 2 suggested that social media sentiments might actually follow sales, it can be useful to conduct another literature research to explore whether there is any basis for this assumption.

The relationships of the analyzed variables were found using linear models since these were determined as most appropriate and the requirements for the use of linear regression were fulfilled. Initial tests with the data set showed that using quadratic or cubic equations would not improve the results significantly and in some cases distort the direction of the data (for example by creating an upside down parabola which would mean that sales drop again if the PNratio increases beyond a certain turning point). However, since the errors with the linear regression were quite big (although normally distributed) and many correlations not very strong further research could also use nonparametric models to test the strength of the relationships.

Lastly, as mentioned before, the conclusions made are based on data for eleven car models and limited to the Dutch market. The findings are therefore also restricted to these car models and to the Netherlands. Since the car models among each other showed great differences in terms of which factor has the strongest predictive power, the findings cannot be generalized for other models and markets.

6. CONCLUSION

In contrast to what was expected, social media sentiments do not have predictive power towards Dutch car sales. The found correlations were very weak and mostly nonsignificant. These observations are academically important since they clearly set limits to the field of sentiment analysis. While sentiment analysis has proven to be of great usefulness in other fields this does not seem to be the case for car sales. On the other hand, the attention a model receives and its search interest were shown to be relatively strong predictors in most cases that could be incorporated into a potentially useful prediction model. It is uncertain whether these findings also yield for other high priced items since the Dutch car industry is possibly influenced by other factors such as tax policies that mitigate the impact of social media opinions on consumer purchase decisions. Further research should therefore focus on other high priced items as well as on other car models and markets in order to further explore and map the boundaries of opinion mining.

7. ACKNOWLEDGEMENTS

I would like to thank my supervisors Fons Wijnhoven and Chintan Amrit for their tremendous supervision and recommendations while writing this thesis. I would furthermore like to thank Anna Pieper and Jarno Bredenoord for their peer reviews and useful input during discussion sessions. Further thanks goes to Jaap Slootbeek who provided me with statistical advice on various occasions. Lastly, I would like to thank my parents for their enormous support throughout the whole study of International Business Administration and without whom the attainment of a bachelor’s degree would not have been possible.

8. REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179-211. doi:10.1016/0749-5978(91)90020-T

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*: Prentice-Hall.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media*. Paper presented at the International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM.

Automotiveimport. (2015). BPM 2016: De BPM gaat weer fors omhoog, dit moet je weten. Retrieved from http://www.automotiveimport.nl/bpm/bpm-2016-fors-omhoog

AutomotiveNL. (2013). De Nederlandse auto-industrie bloeit als nooit tevoren. Retrieved from http://www.automotivenl.com/nl/databases/nieuws/automotivenl/item/3292-de-nederlandse-auto-industrie-bloeit-als-nooit-tevoren

Bataineh, A. Q. (2015). The impact of perceived e-WOM on purchase intention: the mediating role of corporate image. *International Journal of Marketing Studies, 7*(1). doi:10.5539/ijms.v7n1p126

Bing, L., Chan, K. C. C., & Ou, C. (2014). Public sentiment analysis in Twitter data for prediction of a company's stock price movements. *2014 Ieee 11th International Conference on E-Business Engineering (Icebe)*, 232-239. doi:10.1109/icebe.2014.47

CBS. (2015). Gebruik sociale netwerken sterk toegenomen. Retrieved from https://www.cbs.nl/nl-nl/nieuws/2015/27/gebruik-sociale-netwerken-sterk-toegenomen

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2013). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*(2), 340-358. doi:10.1177/1461444813480466

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Cheung, C. M. K., & Thadani, D. R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems, 54*(1), 461-470. doi:http://dx.doi.org/10.1016/j.dss.2012.06.008

Choi, H., & Varian, H. A. L. (2012). Predicting the present with Google Trends. *Economic Record, 88*, 2-9. doi:10.1111/j.1475-4932.2012.00809.x

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51-89. doi:10.1002/aris.1440370103

Crisp, J. (2014). Dutch car tax regime leaves Germany far behind in curbing CO2 emissions. *EurActiv*. Retrieved from http://www.euractiv.com/section/transport/news/dutch-car-tax-regime-leaves-germany-far-behind-in-curbing-co2-emissions/

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340. doi:10.2307/249008

Dijkman, R., Ipeirotis, P., Aertsen, F., & van Helden, R. (2015). Using twitter to predict sales: a case study. *eprint arXiv:1503.04599*.

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), E4512-E4521. doi:10.1073/pnas.1419828112

Erkan, I., & Evans, C. (2016). The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. *Computers in Human Behavior, 61*, 47-55. doi:http://dx.doi.org/10.1016/j.chb.2016.03.003

Gartner. (2014). Gartner's 2014 hype cycle for emerging technologies maps the journey to digital business. Retrieved from http://www.gartner.com/newsroom/id/3114217

Gartner. (2015). Hype cycle for content management, 2015. Retrieved from https://www.gartner.com/doc/3096318/hype-cycle-content-management-

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences of the United States of America, 107*(41), 17486-17490.

Kandaswami, K., & Tiwar, A. (2014). Deloitte. Driving through the consumer's mind: Steps in the buying process. Retrieved from http://www2.deloitte.com/content/dam/Deloitte/in/Documents/manufacturing/in-mfg-dtcm-steps-in-the-buying-process-noexp.pdf

Kotler, P. J. (1994). *Marketing management : analysis, planning, implementation, and control* (8th ed.). Englewood Cliffs, N.J.: Prentice Hall.

Lassen, N. B., Madsen, R., & Vatrapu, R. (2014). Predicting iphone sales from iphone tweets. *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*. doi:10.1109/edoc.2014.20

Lee, J., Park, D. H., & Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications, 7*(3), 341-352. doi:10.1016/j.elerap.2007.05.004

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167. doi:10.2200/S00416ED1V01Y201204HLT016

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603-9611. doi:10.1016/j.eswa.2015.07.052

Office for Official Publications of the European Communities. (1999). Regulation (EEC) No 4064/89 merger procedure. Retrieved from http://ec.europa.eu/competition/mergers/cases/decisions/m1406_en.pdf

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

PewResearchCenter. (2015). Social media usage: 2005-2015. Retrieved from http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

Punj, G. N., & Staelin, R. (1983). A model of consumer information search behavior for new automobiles. *Journal of Consumer Research, 9*(4), 366-380.

Putsis, W. P., & Srinivasan, N. (1994). Buying or just browsing? The duration of purchase deliberation. *Journal of Marketing Research, 31*(3), 393-402. doi:10.2307/3152226

Putsis, W. P., & Srinivasan, N. (1995). So, how long have you been in the market? The effect of the timing of observation on purchase. *Managerial and Decision Economics, 16*(2), 95-110. doi:10.1002/mde.4090160202

Schwartz, E. (2015). Is Google's search market share actually dropping? Retrieved from http://searchengineland.com/googles-search-market-share-actually-dropping-237045

See-To, E. W. K., & Ho, K. K. W. (2014). Value co-creation and purchase intention in social network sites: The role of electronic Word-of-Mouth and trust – A theoretical analysis. *Computers in Human Behavior, 31*, 182-189. doi:http://dx.doi.org/10.1016/j.chb.2013.10.013

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Elsevier, Information Sciences 311*, 18-38. doi:10.1016/j.ins.2015.03.040

Suthikarnnarunai, N. (2008). Automotive supply chain and logistics management. *Imecs 2008: International Multiconference of Engineers and Computer Scientists, Vols I and Ii*, 1800-1806.

Team Nijhuis. (2013). Coosto presentation. Retrieved from http://www.minorinternetmarketing.nl/images/Minor_files_studenten/Social%20Media/Sheets%20presentatie%20Coosto%20Saxion%20_LR_.pdf

Tuarob, S., Tucker, C. S., & Asme. (2014). Fad or here to stay: predicting product market adoption and longevity using large scale, social media data. *Proceedings of the Asme International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2013, Vol 2b*. doi:V02bt02a012

University at Albany. (2003). PAD705 Handout: Multicollinearity. Retrieved from http://www.albany.edu/faculty/kretheme/PAD705/SupportMat/Multicollinearity.pdf

University of California. (2007). Reliability and validity. Retrieved from http://psc.dss.ucdavis.edu/sommerb/sommerdemo/intro/validity.htm

Voortman, M. (2015). Validity and reliability of web search based predictions for car sales. *University of Twente*.

Wijnhoven, F., & Bloemen, O. (2014). External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews? *Decision Support Systems, 59*, 262-273. doi:10.1016/j.dss.2013.12.005

Wu, L., & Brynjolfsson, E. (2013). The future of prediction: How Google searches foreshadow housing prices and sales. *Available at SSRN 2022293*.

Yang, X., Pan, B., Evans, J. A., & Lv, B. F. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386-397. doi:10.1016/j.tourman.2014.07.019

9. APPENDICES

A. Analyzed car models including class, price and Dutch search terms

Car model	Class	Starting price	Search term Coosto	Search term Google Trends
Fiat Panda	A	13,675€	Fiat Panda	Fiat Panda
Ford Fiesta	B	13,995€	Ford Fiesta	Ford Fiesta
Opel Corsa	B	20,950€	Opel Corsa	Opel Corsa
Honda Civic	C	21,050€	Honda Civic	Honda Civic
VW Golf	C	21,050 €	(Volkswagen Golf) OR (VW Golf)	"Volkswagen Golf" + "VW Golf" -Trends
Ford Mondeo	D	29,575	Ford Mondeo	Ford Mondeo
Volkswagen Passat	D	31,450	(Volkswagen Passat) OR (VW Passat)	"Volkswagen Passat" + "VW Passat" -Trends
BMW 5 Series	E	47,990	"BMW 5-serie" –GT	"BMW 5-serie" –GT
Mercedes E Class	E	46,800	(MB OR Mercedes Benz OR Mercedes) "E Klasse"	MB + Mercedes Benz + Mercedes "E Klasse"
Porsche Panamera	F	106,400	Porsche Panamera	Porsche Panamera
Porsche 911	S	115,000	Porsche 911	Porsche 911

B. SPSS Syntax[5]

Graph of mean PNratio and mean car sales over whole time period

```
* Chart Builder.

GGRAPH

  /GRAPHDATASET NAME="graphdataset" VARIABLES=Year_month
MEAN(PNratio)[name="MEAN_PNratio"]

    MEAN(Sales)[name="MEAN_Sales"] MISSING=LISTWISE REPORTMISSING=NO

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

  SOURCE: s=userSource(id("graphdataset"))

  DATA: Year_month=col(source(s), name("Year_month"), unit.category())

  DATA: MEAN_PNratio=col(source(s), name("MEAN_PNratio"))

  DATA: MEAN_Sales=col(source(s), name("MEAN_Sales"))

  GUIDE: axis(dim(1), label("Timeline"))

  GUIDE: axis(scale(y1), label("Mean PNratio"), color(color."3E58AC"))

  GUIDE: axis(scale(y2), label("Mean Carsales"), color(color."2EB848"), opposite())

  SCALE: y1 = linear(dim(2), include(0))

  SCALE: y2 = linear(dim(2), include(0))

  ELEMENT: line(position(Year_month*MEAN_PNratio), missing.wings(), color.interior(color."3E58AC"),

    scale(y1))

  ELEMENT: line(position(Year_month*MEAN_Sales), missing.wings(), color.interior(color."2EB848"),

    scale(y2))

END GPL.
```

[5] For processes requiring multiple repetitions of the same steps with different variables only one example is given

Linear regression analysis of average number of total mentions with average sales (averages calculated with MS Excel)

```
DATASET ACTIVATE DataSet1.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Sales_average
  /METHOD=ENTER Total_mentions_average.
```

Cross correlation for VW Passat: total number of mentions with sales

```
CCF
  /VARIABLES=Number_total Sales
  /NOLOG  /MXCROSS 12.
```

Shapiro-Wilk Test for correlations of Google Trends with sales (low priced car group)

```
EXAMINE VARIABLES=Corr_Trends_lowpriced
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Independent t-test between correlations PNratio with sales of cars <30,000€ and ≥30,000€

```
DATASET ACTIVATE DataSet5.
T-TEST GROUPS=Price(30000)
  /MISSING=ANALYSIS
  /VARIABLES=Corr_PNratio
  /CRITERIA=CI(.95).
```

Linear regression of total mentions and Google Trends including Variance Inflation Factor

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Sales
  /METHOD=ENTER Number_total Trends_score.
```

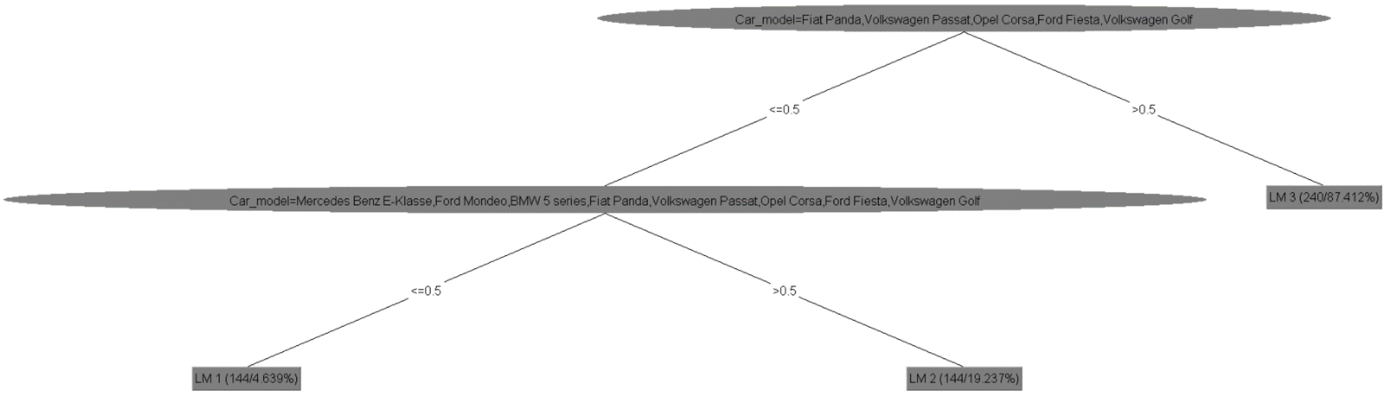
C. M5P classifier decision tree regression for sales prediction including a four-month time lag

Classifier output

```
=== Run information ===
Scheme:     weka.classifiers.trees.M5P -M 4.0
Relation:    ALL DATA-weka.filters.unsupervised.attribute.Remove-R2
Instances:   529
Attributes:  4
             Car_model
             Number_total
             Trends_score
             Sales
Test mode:   split 66.0% train, remainder test

=== Classifier model (full training set) ===
M5 pruned model tree:
(using smoothed linear models)

Car_model=Fiat Panda,Volkswagen Passat,Opel Corsa,Ford Fiesta,Volkswagen Golf <= 0.5 :
|   Car_model=Mercedes Benz E-Klasse,Ford Mondeo,BMW 5 series,Fiat Panda,Volkswagen Passat,Opel Corsa,Ford Fiesta,
                 Volkswagen Golf <= 0.5 : LM1 (144/4.639%)
|   Car_model=Mercedes Benz E-Klasse,Ford Mondeo,BMW 5 series,Fiat Panda,Volkswagen Passat,Opel Corsa,Ford Fiesta,
                  Volkswagen Golf >  0.5 : LM2 (144/19.237%)
Car_model=Fiat Panda,Volkswagen Passat,Opel Corsa,Ford Fiesta,Volkswagen Golf >  0.5 : LM3 (240/87.412%)

LM num: 1
Sales =
          22.0196 * Car_model            =Honda Civic, Mercedes Benz E-Klasse, Ford Mondeo, BMW 5 series, Fiat Panda,
                                          Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 2.3903 * Car_model           =Mercedes Benz E-Klasse, Ford Mondeo, BMW 5 series, Fiat Panda,
                                            Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 21.7974 * Car_model          =Ford Mondeo, BMW 5 series, Fiat Panda, Volkswagen Passat, Opel Corsa,
                                           Ford Fiesta, Volkswagen Golf
          + 13.22 * Car_model            =Fiat Panda, Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          - 5.6026 * Car_model           =Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 4.497 * Car_model            =Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 8.4623 * Car_model           =Ford Fiesta, Volkswagen Golf
          + 12.9356 * Car_model          =Volkswagen Golf
          + 0.012 * Number_total
          + 0.4681 * Trends_score
          - 22.379

LM num: 2
Sales =
          2.2157 * Car_model             =Honda Civic, Mercedes Benz E-Klasse, Ford Mondeo, BMW 5 series, Fiat Panda,
                                          Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 2.3903 * Car_model           =Mercedes Benz E-Klasse, Ford Mondeo, BMW 5 series, Fiat Panda,
                                           Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 143.7119 * Car_model        =Ford Mondeo, BMW 5 series, Fiat Panda, Volkswagen Passat, Opel Corsa,
                                          Ford Fiesta, Volkswagen Golf
          + 13.22 * Car_model            =Fiat Panda, Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          - 5.6026 * Car_model           =Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 4.497 * Car_model            =Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 8.4623 * Car_model           =Ford Fiesta, Volkswagen Golf
          + 12.9356 * Car_model          =Volkswagen Golf
          - 0.153 * Number_total
          + 1.5994 * Trends_score
          - 30.5101

LM num: 3
Sales =
          13.1475 * Car_model            =Ford Mondeo, BMW 5 series, Fiat Panda, Volkswagen Passat, Opel Corsa,
                                          Ford Fiesta, Volkswagen Golf
          + 15.7084 * Car_model          =Fiat Panda, Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          - 311.5853 * Car_model         =Volkswagen Passat, Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 5.3435 * Car_model           =Opel Corsa, Ford Fiesta, Volkswagen Golf
          + 10.0553 * Car_model          =Ford Fiesta, Volkswagen Golf
          + 475.613 * Car_model          =Volkswagen Golf
          + 0.0035 * Number_total
          + 21.2021 * Trends_score
          - 854.9641

Number of Rules : 3
```

```
Time taken to build model: 0.09 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

=== Summary ===

Correlation coefficient                  0.8474
Mean absolute error                    143.2867
Root mean squared error                237.2889
Relative absolute error                 42.2869 %
Root relative squared error             53.2126 %
Total Number of Instances              180
```

Original decision tree (equal to Figure 3)


Car_model=Fiat Panda,Volkswagen Passat,Opel Corsa,Ford Fiesta,Volkswagen Golf
<=0.5
>0.5
Car_model=Mercedes Benz E-Klasse,Ford Mondeo,BMW 5 series,Fiat Panda,Volkswagen Passat,Opel Corsa,Ford Fiesta,Volkswagen Golf
LM 3 (240/87.412%)
<=0.5
>0.5
LM 1 (144/4.639%)
LM 2 (144/19.237%)